GEOProcessing 2012

The Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services

ISBN: 978-1-61208-178-6

January 30- February 4, 2012

Valencia, Spain

GEOProcessing 2012 Editors

Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische

Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany

Bernd Resch, Massachusetts Institute of Technology - Cambridge, USA

GEOProcessing 2012

Forward

The fourth edition of The International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2012), held in Valencia, Spain, on January 30th – February 4th, 2012, addressed the aspects of managing geographical information and web services.

The goal of the GEOProcessing 2012 conference was to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of advances in geographic information systems and the new applications related to them using the Web Services. Such systems can be used for assessment, modeling and prognosis of emergencies

GEOProcessing 2012 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects from fundamentals to more specialized topics such as 2D & 3D information visualization, web services and geospatial systems, geoinformation processing, and spatial data infrastructure.

We take this opportunity to thank all the members of the GEOProcessing 2012 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the GEOProcessing 2012. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the GEOProcessing 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that GEOProcessing 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in geographic information research.

We also hope the attendees enjoyed the beautiful surroundings of Valencia, Spain.

GEOProcessing 2012 Chairs

Monica De Martino, Consiglio Nazionale delle Ricerche - Genova, Italy
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität

Münster / North-German Supercomputing Alliance (HLRN), Germany
Bernd Resch, Massachusetts Institute of Technology - Cambridge, USA

GEOProcessing 2012

Committee

GEOProcessing 2012 Advisory Chairs

Monica De Martino, Consiglio Nazionale delle Ricerche - Genova, Italy
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance (HLRN), Germany
Bernd Resch, Massachusetts Institute of Technology - Cambridge, USA

GEOProcessing 2012 Technical Program Committee

Riccardo Albertoni, Consiglio Nazionale delle Ricerche - Genova, Italy
Thierry Badard, Université Laval - Québec, Canada
Petko Bakalov, Environmental Systems Research Institute, USA
Fabian Barbato, Universidad de la Republcia - Montevideo, Uruguay
Thomas Barkowsky, University of Bremen, Germany
Reinaldo Bezerra Braga, Joseph Fourier University - Grenoble 1, France
Budhendra L. Bhaduri, Oak Ridge National Laboratory, USA
Ling Bian, University at Buffalo, USA
Sandro Bimonte, CEMAGREF, France
Giuseppe Borruso, University of Trieste, Italy
Boyan Brodaric, Geological Survey of Canada, Canada
Mete Celik, Erciyes University, Turkey
Xin Chen, NAVTEQ Corporation - Chicago, USA
Chi-Yin Chow, City University of Hong Kong, Hong Kong
Christophe Claramunt Naval Academy Research Institute, France
Eliseo Clementini, University of L'Aquila, Italy
Alfredo Cuzzocrea, ICAR-CNR & University of Calabria, Italy
Chenyun Dai, Purdue University, USA
Monica De Martino, Consiglio Nazionale delle Ricerche (CNR) - Genova, Italy
Anselmo C. de Paiva, Universidade Federal do Maranhão, Brazil
Cláudio de Souza Baptista, University of Campina Grande, Brazil
Antonios Deligiannakis, Technical University of Crete, Greece
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada
Javier Estornell Cremades, Universidad Politecnica de Valencia, Spain
Stewart Fotheringham, University of St Andrews in Scotland, UK
W. Randolph Franklin, Rensselaer Polytechnic Institute - Troy NY, USA
Sébastien Gambs, Université de Rennes 1 - INRIA, France
Betsy George, Oracle America Inc., USA
Diego Gonzalez Aguilera, University of Salamanca - Avila, Spain
Björn Gottfried, University of Bremen, Germany
Enguerran Grandchamp, Université des Antilles et de la Guyane, Guadeloupe
Malgorzata Hanzl, Technical University of Lodz, Poland

Herman Haverkort, Eindhoven University of Technology, The Netherlands

Cory Henson, Wright State University / Kno.e.sis Center, USA

Gerard B.M. Heuvelink, Wageningen University and Research Centre, The Netherlands

Erik Hoel, Environmental Systems Research Institute, USA

Zhou Huang, Peking University - Beijing, China

Cengizhan İpbüker, Istanbul Technical University, Turkey

Vana Kalogeraki, Athens University of Economics and Business, Greece

Ibrahim Kamel, University of Sharjah UAE / Concordia University, Canada

Baris M. Kazar, Oracle Corp., USA

Rosa Lasaponara, CNR, Italy

Robert Laurini, INSA de Lyon - Villeurbanne, France

Fabio Luiz Leite Junior, UEPB - State University of Paraíba, Brazil

Ki-Joune Li, Pusan National University, South Korea

Qing Liu, CSIRO, Australia

Xuan Liu, IBM T.J. Watson Research Center - Yorktown Heights, USA

Victor Lobo, Portuguese Naval Academy / New University of Lisbon, Portugal

Qifeng Lu, MacroSys LLC. - Arlington, USA

Jesus Marti Gavila, Universidad Politecnica de Valencia, Spain

Hervé Martin, Université Joseph Fourier - Grenoble, France

Stephan Mäs, Technische Universität Dresden, Germany

Nicola Masini, Institute for Archaeological and Monumental Heritage / University of Basilicata - Potenza, Italy

Mark McKenney, Southern Illinois University Edwardsville, USA

Tomas Mildorf, University of West Bohemia - Pilsen, Czech Republic

Ralf Möller, Hamburg University of Technology, Germany

Beniamino Murgante, University of Basilicata, Italy

Shawn D. Newsam, University of California - Merced, USA

Daniel Orellana Vintimilla, Charles Darwin Foundation - Galápagos, Ecuador

Matthew Perry, Oracle Corp., USA

Dieter Pfoser, Institute for the Management of Information Systems, Greece

Maurizio Pollino, ENEA - Italian National Agency for New Technologies - Rome, Italy

Alenka Poplin, HafenCity University Hamburg, Germany

David Prosperi, Florida Atlantic University, USA

Sigrid Reiter, University of Liège, Belgium

Matthias Renz, Ludwig-Maximilians Universität München, Germany

Bernd Resch, Massachusetts Institute of Technology/Senseable City Lab - Cambridge, USA

Kai-Florian Richter, University of Melbourne, Australia

Claus-Peter Rückemann, Westfälische Wilhelms-Universität Münster/Leibniz Universität Hannover/ North-German Supercomputing Alliance, Germany

Mauro Salvemini, EUROGI, Italy

Markus Schneider, University of Florida, USA

Spiros Skiadopoulos, University of Peloponnese - Tripoli, Hellas

Frank Steinicke, Institut für Mensch-Computer-Medien & Institut für Informatik - Würzburg, Germany

Juergen Symanzik, Utah State University - Logan, USA

Ali Tahir, University College Dublin, Ireland

Naohisa Takahashi, Nagoya Institute of Technology, Japan

Ergin Tari, Istanbul Technical University, Turkey

Copyright Information

Table of Contents

Which Service Interfaces fit the Model Web? 1
Sven Shade, Nicole Ostlander, Carlos Granell, Michael Shulz, Daniel McInerney, Gregoire Dubois, Lorenzino Vaccari, Michele Chinosi, Laura Diaz, Lucy Bastin, and Richard Jones

One Click Focusing: An SQL-based Fast Loop Road Extraction Method for Mobile Map Services 7
Daisuke Yamamoto, Hiroki Itoh, and Naohisa Takahashi

Schema Transformation as a Tool for Data Reuse in Web Service Environment 17
Lassi Lehto

Bus Coming: A Service for Tracking Buses in Rural Areas based on Passenger Locations 23
Arthur Thompson and Wayne Goodridge

Live Cities and Urban Services - A Multi-dimensional Stress Field between Technology, Innovation and Society 28
Bernd Resch, Alexander Zipf, Philipp Breuss-Schneeweis, Euro Beinat, and Marc Boher

Comparison of Stacking Methods Regarding Processing and Computing of Geoscientific Depth Data 35
Claus-Peter Ruckemann

Automatic Classification of Points-of-Interest for Land-use Analysis 41
Filipe Rodrigues, Ana O. Alves, Francisco C. Pereira, Shan Jiang, and Joshep Ferreira

Algorithm Based Methodology for the Creation of a Regularly Updated Global Online Map Derived From Volunteered Geographic Information 50
Marcus Goetz, Johannes Lauer, and Michael Auer

Integrated Geoprocessing for Generation of Affected Assets and Rights Reports for High Voltage Electrical Infrastructures 59
Federico-Vladimir Gutierrez-Corea, Miguel-Angel Manso-Callejo, Francisco-Javier Moreno-Marimbaldo, and Emerson Castaneda-Sanabria

Towards Identifying the Best New Connection in a Spatial Network: Optimising the Performance of Hole Discovery 66
Femke Reitsma, Tony Dale, and William Pearse

Development of an Acoustic Transceiver for Positioning Systems in Underwater Neutrino Telescopes 70
Giuseppina Larosa, Miguel Ardid, Carlos David Llorens, Manuel Bou Cabo, Juan Antonio Martinez Mora, and Silvia Adrian Martinez

Data Cleansing and Selective Query Processing in Sensor Networks 76
Maria Drougka and Theodore Tsiligiridis

e-Report Generator Supporting Communications and Fieldwork: A Practical Case of Electrical Network Expansion Projects 82
Emerson Castaneda-Sanabria, Miguel-Angel Manso-Callejo, Francisco-Javier Moreno-Marimbaldo, and Federico-Vladimir Gutierrez-Corea

Level Sets and Voronoi based Feature Extraction from any Imagery 89
Ojaswa Sharma, Francois Anton, and Darka Mioc

Spatial Characterization of Extreme Precipitation in Madeira Island Using Geostatistical Procedures and a 3D SOM 98
Jorge Gorricha, Ana Costa, and Victor Lobo

Fast and Accurate Visibility Computation in a 3D Urban Environment 105
Oren Gal and Yerach Doytsher

Combining Territorial Data With Thermal Simulations to Improve Energy Management of Suburban Areas 111
Anne-Francoise Marique, Maud Petel, Asma Hamdi, and Sigrid Reiter

Vehicle Position Determination — Using Markers and Speed Reports 117
Bruce Beyeler and David C. Pheanis

A GIS-Based Approach for Representation of Air Pollution- A Case Study: Tabriz City 122
Davood Parvinnezhad Hokmabadi, Arash Rahmanizadeh, and Shahla Taghizadeh Sufiani

Study on the Composition of the Residential Environment and Environmental Cognition in Collective Housing 127
Hirotomo Ohuchi, Setsuko Ouchi, Katsuhito Chiba, and Yuta Takano

Automated Extraction and Geographical Structuring of Flickr Tags 134
Omair Chaudhry and William Mackaness

From Point Clouds to 3D City Models: The Case Study of Villalba (Madrid) 140
Juan Mancera-Taboada, Pablo Rodriguez-Gonzalvez, Diego Gonzalez-Aguilera, Benjamin Arias-Perez, David Hernandez-Lopez, and Beatriz Felipe-Garcia

A Novel Health Monitoring System using Patient Trajectory Analysis: Challenges and Opportunities 147
Shayma Alkobaisi, Wan D. Bae, Sada Narayanappa, and Cheng C. Liu

Open source GIS Tools to Map Earthquake Damage Scenarios and to Support Emergency 152
Maurizio Pollino, Antonio Bruno Della Rocca, Grazia Fattoruso, Luigi La Porta, Sergio Lo Curzio, Agnese Arolchi, Valentina James, and Carmine Pascale

Incorporating the Analyses of Urban Form into the Geocomputational Modelling The Morphological Approach 158
Malgorzata Hanzl

Using GIS for Impact Analysis from Industries Installation 164
Camila Thomasi, Gerson Nunes, Marcio Jugueiro, and Diana Adamatti

Distributed System Architectures, Standardization, and Web-Service Solutions in Precision Agriculture 171
Katja Polojarvi, Mika Luimula, Pertti Verronen, Mika Pahkasalo, Markku Koistinen, and Jouni Tervonen

Footprint-Based 3D Generalization of Building Groups for Virtual City Visualization 177
Shuang He, Guillaume Moreau, and Jean-Yves Martin

A Web-based Environment for Analysis and Visualization of Spatio-temporal Data provided by OGC Services 183
Maxwell Oliveira, Claudio Baptista, and Ana Falcao

A Heuristic-based Approach for Merging Layers Information in a GIS 190
Wilfried Segretier, Martine Collard, and Enguerran Grandchamp

Computationally Feasible Query Answering over Spatio-thematic Ontologies 198
Ozgur Lutfu Ozcep and Ralf Moller

O*-W: An Efficient O* algorithm to Process Optimal Sequence Traversal Query in a Complete Directed Graph 204
Qifeng Lu and Kathleen Hancock

An Architecture for Geographic Information Systems on the Web - WebGIS 209
Mariano Pascual, Eluzai Alves, Henrique Roig, Tati de Almeida, George Sand de Franca, and Maristela Holanda

D-WCPS: A Framework for Service Based Distributed Processing of Coverages 215
Michael Owonibi and Peter Baumann

Modeling and Querying Mobile Location Sensor Data 222
Iulian Sandu Popa and Karine Zeitouni

Land-use Mapping of Valencia City Area from Aerial Images and LiDAR Data 232
Txomin Hermosilla, Luis Angel Ruiz, Jorge Abel Recio, and Jose Balsa-Barreiro

Generation and Validation of Workflows for On-Demand Mapping 238
Nicholas Gould and Omair Chaudhry

A new methodology for the collection and handling of geological data 244
Rodrigo Cipullo and Henrique Roig

Comparison Between Generalization Processes at Large and Small Scales 251
Jacqueleen Joubran Abu Daoud and Izabela Karsznia

Which Service Interfaces fit the Model Web?

Sven Schade, Nicole Ostländer, Carlos Granell,
Michael Schulz, Daniel McInerney, Gregoire Dubois,
Lorenzino Vaccari, Michele Chinosi
Institute for Environment and Sustainability
European Commission – Joint Research Centre
Ispra, Italy
e-mail: {sven.schade, nicole.ostlaender, carlos.granell,
michael.schulz, daniel.mc-inerney, gregoire.dubois,
lorenzino.vaccari, michele.chinosi}@jrc.ec.europa.eu

Laura Díaz
Institute of New Imaging Technologies
University Jaume I
Castellón, Spain
e-mail: laura.diaz@uji.es

Lucy Bastin, Richard Jones
Aston University,
Birmingham, UK
e-mail: {l.bastin, jonesrm1}@aston.ac.uk

***Abstract* - The Model Web has been proposed as a concept for integrating scientific models in an interoperable and collaborative manner. However, four years after the initial idea was formulated, there is still no stable long term solution. Multiple authors propose Web Service based approaches to model publication and chaining, but current implementations are highly case specific and lack flexibility. This paper discusses the Web Service interfaces, which are required for supporting integrated environmental modeling in a sustainable manner. We explore ways to expose environmental models and their components using Web Service interfaces. Our discussions present work in progress for establishing the Web Services technological grounds for simplifying information publication and exchange within the Model Web. As a main outcome, this contribution identifies challenges in respect to the required geo-processing and relates them to currently available Web Service standards.**

Keywords - Model Web; Web Services; SOA; WPS; WSDL

I. INTRODUCTION AND PROBLEM STATEMENT

Historically, the Web played a role only in environmental data transport, but is now proposed as the foundation of the 'Model Web', - a distributed, multidisciplinary network of interoperating infrastructures of data and models communicating with each other using Web services [1]. As a central concept, complex environmental models and the data required to execute them should be modularized into self-containing geo-processing units [2]. These modules, as well as their compositions, should be made available as services on the Web [3]. The benefits of a Model Web are clear; models, such as similarity calculation between ecosystems or predictions of forest change, and associated data can be more easily accessed, reused and chained for multiple purposes, increasing the repeatability of research and allowing end-users to address more complex issues than when models are used in isolation. Difficult operations can also, to a certain extent, be made available to end-users who do not have the expertise required to solve an indispensable step in a compound modeling process [4]. The gain in flexibility when linking models will be directly proportional to the granularity of the services provided. Basic and generic Model Web components are more likely to be shared than sophisticated ones that are less likely to meet end-users' requirements.

Nonetheless, scientists are currently not using the Model Web to discover, re-use or chain models to the extent that was envisaged. There are significant drawbacks to the full implementation of an information system offering increased access to resources. In this paper, we analyze these drawbacks by outlining some of the central requirements and discussing the lessons learned during our previous work in numerous research projects. In doing so, we provide the foundation for standards-based implementations of the Model Web.

For this work, we assume that models are available and discoverable, so we focus on issues relating to geo-processing for the Model Web, such as ways of approaching model exposure and chaining challenges. Furthermore, we focus on the exposure of models, i.e., sets of algorithms to be used over the Web, and exclude pure offerings of geo-processing tools (such as Sextante [5] or GRASS [6]). Challenges such as the use of ontologies for achieving high-level semantic interoperability, as for example discussed in [7] and [8], are out of the scope of the presented work.

Following the concept of Service Oriented Architecture (SOA) [9], we use standards-based Web service interfaces, namely the Open Geospatial Consortium (OGC) Web Processing Services (WPS) and the Web Service Description Language (WSDL) of the World Wide Web Consortium (W3C) as a starting point for our discussions. We decided to focus on these two technologies, because of their popularity in the scientific community and many research efforts have attempted to use them in a variety of ways in service composition settings.

The following Section II presents relevant background and pointers to related work. In Section III, we outline the central requirements of services for the Model Web and reflect on the specific roles of WPS and WSDL interfaces. We conclude with a brief summary and our future roadmap in Section IV.

ISBN: 978-1-61208-178-6

II. Background and Related Work

The requirement to chain scientific models has led to a wide variety of coupling approaches and frameworks of varying specificity and interoperability [10]. These include standards such as Common Component Architecture (CCA) [11] and Open Modeling Interface (OpenMI) [12], but also orchestration tools such as the Invisible Modelling Environment (TIME) [13] and Taverna [14]. Although some of these solutions have been successfully applied to specific modeling settings, we focus our work on the use of the Web service interfaces, starting from WPS and WSDL, as the necessary standards-based enablers for a sustainable Model Web.

A. Web Processing Service

Initiated in 2004/2005, the idea of the WPS [15] was to provide a generic interface for the publication of any kind of operation or model, since at that point in time, none of the OGC attempts to create specific geo-processing services had been successful. The WPS specifies *GetCapabilities*, *DescribeProcess* and *Execute* operations (Figure 1). The response to the *GetCapabilities* request contains generic information about the WPS and details how to launch the other two request types. The second operation (*DescribeProcess*) identifies all available processes, which might be executed on the concrete WPS and defines all model inputs as parameters in the generic *Execute* request. These can be described as simple types, such as integers, Boolean or String, and Complex types, which have their own schema. There is no restriction on how to define inputs, which results in a user having to define input and output types. Finally, the *Execute* operation triggers a specific model run.

Notably, the WPS was not developed specifically for chaining purposes, but it was envisioned to allow generic client applications to read the *DescribeProcess* document, and, on the fly, to present a corresponding input form to a user. Thus it requires human-to-machine communication and interpretation of the presented form by the user.

The WPS standard has been welcomed at the time as a means by which scientific models may be published and linked, consuming as input parameters data from other standard OGC services, such as the Web Feature Service (WFS) [16] and Sensor Observation Service (SOS) [17]. Projects, such as UncertWeb [18] focus on handling and propagating uncertainty in Web-based models [19]. Several WPSs have been used in modeling chains as proofs of concept and these include: INTAMAP WPS for the automatic interpolation of measured point data [4], the eHabitat ecosystems and habitat similarity modeling WPS [20], and WPS to monitor forest change in the context of the European Forest Fire Information System (EFFIS) [21]. However, describing and profiling models as Web-based processes, and making them available as stable nodes for the scientific community is still a demanding task, with few examples of best practice [22].

B. Web Service Description Language

WSDL [23] is a W3C developed standard for describing Web services. As a mature technology, it has a large community and broad range of software support. It is an XML-based language to describe the functional properties of a service such as its method signatures, input and output messages, details of the transport protocol used (endpoint, SOAP envelope, etc.). As a description language, a WSDL file contains all of the operations or methods offered by a given service. However, WSDL does not specify how client applications access to WSDL files because it is not a communication interaction protocol like WPS. This means that each service provider may offer proprietary rules to access WSDL files.

Figure 2 reveals the differences in the level of granularity between WSDL and WPS. While a WPS provides well-defined interaction operations and each WPS process is a resource by its own, i.e., it has a unique endpoint (see also Figure 1), a WDSL file acts as a public endpoint to access all methods contained in a service. Every single WSDL method remains hidden behind this endpoint. In contrast to the case of the WPS, dedicated tools can automatically generate clients from a WSDL description file to invoke a certain service's method (machine-to-machine communication).

Any service described using WSDL can be only orchestrated by WSDL-compliant workflow software and standards. This means that WSDL is coupled to specific workflow languages, which can be a limitation in certain modeling settings. On the contrary, WPS services can be effectively composed by themselves (e.g., service cascading) because the communication protocol is made explicit. A number of wrapper solutions exist [24], where WSDL documents are created for WPS processes. These contain either abstract message descriptions, or concrete schemas for each process.

III. WPS and WSDL in the Model Web

For the context of geo-processing, we identified five Model Web challenges (Figure 3). We excluded challenges that are related to wider topics, such as model and service discovery, as well as technical issues, such as network fragility or general trust in model results. In this section, we discuss the possible roles of WPS and WSDL service interfaces as well as

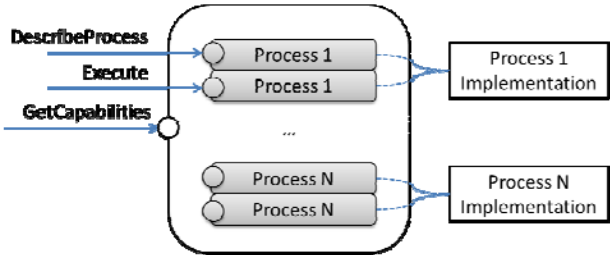


Figure 1. WPS description level.

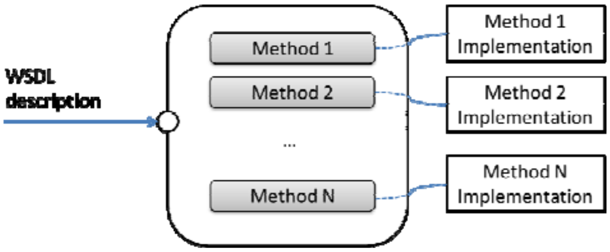


Figure 2. WSDL description level.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-178-6

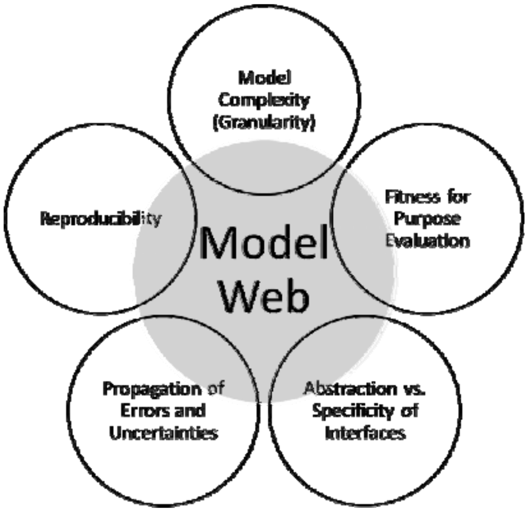

Model Complexity (Granularity)
Fitness for Purpose Evaluation
Abstraction vs. Specificity of Interfaces
Propagation of Errors and Uncertainties
Reproducibility
Model Web

Figure 3. Overview of identified Model Web challenges, which are related to geo-processing.

their combination in relation to each of these challenges. All sub-sections have the same structure: presentations of the challenge are followed by reflections on the use of WPS and WSDL in the given context. Examples are included where appropriate.

A. Model Complexity

Models can be wrapped to be exposed with standard interfaces (WPS or WSDL) at different levels of abstraction. Exposing models necessitates finding the right level of service granularity, i.e., the amount of exposed functionality [25]. Coarse-grained services encapsulating a whole model within a single interface reduce the number of service requests from the client; however they might be difficult to reuse in new scenarios. Examples of the coarse-grained approach are the EFFIS WPS, where a forest change model is encapsulated as a service; and eHabitat, which provides access to a similarity calculation for ecosystems and habitats. More generic or finer-grained services normally require less complicated input and output data, and they are more easily reused in new chains, although multiple calls are needed to run more complex models [2]. The INTAMAP WPS, while offering access to complex interpolation algorithms, also fits essentially into this category because of its clear and modular purpose: automated spatial interpolation of point data on a requested region, with or without a consideration of input uncertainties.

A careful consideration of the appropriate granularity level for services could have a positive impact on component reusability and performance [2]. Wrapped models ease the publishing and execution of entire models, but limit re-use, since successful sharing will require some adaptation of inputs and outputs for the new context. In contrast, highly distributed model compositions pose new challenges such as an increase in traffic and network latency as potentially huge quantities of data are exchanged between model components. Above all, the fragmentation of a model over distributed nodes increases the likelihood of breakable nodes and reduces the overview on the processing chain unless each transaction is documented and consistent error handling is implemented.

The well defined relationship between the WPS specification and other geospatial data services, such as WFS and SOS, has encouraged its use within the geospatial community [26]. However, though multiple authors note the need to increase interoperability with mainstream approaches to SOA, such as WSDL, several design decisions make this difficult to implement [27]. While services interfaces that are described with WSDL expose separate execution endpoints to execute an individual process, depending on the kind of binding selected (e.g., SOAP, HTTP Get/Post), WPSs offer a common end point mechanism to expose a service that holds several processes and which delivers each process description as response to a particular *DescribeProcess* request. In this sense a WPS endpoint, both at service- and process-level, serves as a unique identifier. We re-visit this issue in the Section III.C.

B. Fitness for Purpose Evaluation

Assuming that a model has been wrapped as a service and discovered, the evaluation of whether that model is fit for a given purpose might not be answered by simply consuming the process description or the corresponding metadata. It might be a matter of interpreting various model runs with varying input parameters, or of running a sensitivity analysis for the potential user's inputs. The encoding of inputs and outputs can also be an issue requiring some investigation and testing, e.g., if the desired input or output data model is not supported. Additionally, a misconfiguration of the model could lead to wrong results and some of the input parameters could be conditional, i.e., several model runs might be required before the final results will provide the user with the desired response. The more complex and unique a model is, the more model runs might be required before a user considers the result as final and is able to assess whether the model is fit for the intended purpose.

Unlike WSDL, the WPS standard has been built assuming human intervention. Thus, the formulation and execution of processing requests controlled by the user is foreseen, and the interface is tailored to human-to-machine communication. Thus, the focus in WPS developments has been at least partly on the development of specific clients through which they can be accessed and generic clients to allow the immediate visualization of results returned by a WPS.

However, conditional inputs and outputs cannot be expressed through any of the interfaces; they can only be described in the process documentation, which might be misunderstood. This might result in various invalid test runs, which could have been avoided if the *Execute* request had been validated against the conditions foreseen by the service provider.

C. Abstraction and Specificity

As model complexity increases, more translations and validations are needed for the various input and output data schemas. The huge variety of available data encodings to define inputs and outputs in modeling scenarios and their inaccuracy in defining parameter types are becoming a burden in delivering the Model Web vision.

WPS profiles can help to describe interface specifics in a more detailed and reusable way. For example, it is possible to refer to a Geography Markup Language (GML) application schema [28] in order to specify that only certain types of spatial features should be accepted, or to exploit the validation rules of a schema (such as the XML implementation of the Uncertainty Markup Language (UncertML) [29]) to specify that an input grid, which contains probabilities can be expected to contain continuous values between 0 and 1. These additional data models and dictionaries can be extremely useful in clarifying whether a discovered service is suitable for a user's data, but they must be used rigorously and consistently – an extra pressure on users and clients and an entry barrier. It is also impossible to automate all the necessary validation simply through profiling, when more complex scientific data exchange formats such as netCDF [30] are required. Developments within the UncertWeb project try to support this validation of netCDF datasets, e.g., by extending the existing netCDF Climate and Forecast (CF) metadata convention to encode and identify variables using UncertML references.

Wrapping approaches for WPS use a generic WSDL document to describe any WPS instance at once or on a per process basis [24]. In the WPS specification, a *DescribeProcess* request reveals additional process details, such as required inputs and formats. Due to the generic nature of WSDL, not all the information of a WPS and its processes can be adequately represented. This extra layer of complexity and lack of precision leads to a drastic reduction in the benefits of using WSDL. Graphical workflow composition and code generating tools require the user to know the information provided by the *DescribeProcess* operation, and how to subsequently build an *Execute* request document. In practice, this means that a user has to examine the WPS responses and understand the required parameter types and formats, before they can actually benefit from WSDL's widespread support in chaining environments and orchestration engines in order to enable automation of a process chain.

When defining schemas for each process, which appears to be the solution for the described above problems of a generic WPS, the benefits of WPS appear to be negated, since the request and response messages defined do not validate against WPS schemas. Therefore, the additional overheads of implementing the WPS specification become unnecessary, if not argued by any of the other challenges. In this case, a simpler solution would be to implement the processes using a SOAP/WSDL framework [9]. Such frameworks are able to automatically convert code into a usable Web service.

It is vital to reach the right balance between specific and generic/abstract interfaces and data specifications so as to increase usability and subsequently model sharing. Complex spatio-temporal data and models require careful description and validation, which is beyond the current capacity of the generic interfaces available.

D. Propagation of Errors and Uncertainties

When diverse data sources and processes are composed within a chain whose ultimate outputs will be used for decision-making, a need arises for properly-documented propagation of inherent or introduced errors and uncertainties.

Error and uncertainty propagation are necessary ingredients in assessing the effects of data and model uncertainty on the reliability of the outputs of a model chain. Uncertainty must be properly quantified and communicated to decision makers – for example, by supplying error estimates, quantiles or examples of equally likely alternative scenarios as outputs. In a Model Web context, the language and formats used to do this must be standardized and interoperable. There are several examples of WPS, which use the UncertML approach to characterize the uncertainty on their inputs and outputs. The INTAMAP WPS accepts an Observation and Measurement (O&M) [31] request document containing point observations, which may have associated measurement uncertainties. Depending on the nature of that uncertainty, an appropriate algorithm is selected, and a document returned, which contains interpolated grids of predicted means and variances. Some cross-validation is performed, but this is internal to the service rather than directly accessible to the user as a model-evaluation service. The uncertainty-enabled version of eHabitat, currently under development at the Joint Research Centre (JRC) of the European Commission, samples (or simulates) datasets from inputs with known or inferred uncertainty, and produces summary statistics such as exceedance probabilities and example realizations (again encoded using UncertML), on top of the usual mean predicted habitat suitability map. Again, the simulation (which could effectively be seen as a form of sensitivity analysis) is embedded inside the service and not exposed separately.

An alternative approach, which more clearly illustrates service chaining, is to use a model that in itself does not handle uncertainty. Instead, requests on that model are executed multiple times, using perturbed parameters and inputs, which are sampled from statistical estimates of the uncertainty on those inputs. This allows fairly straightforward propagation of uncertainty on inputs, model parameters and initial conditions, and with some adaptation might even help to estimate and propagate the uncertainty within the model itself. The approach was successfully demonstrated for an air quality assessment WPS by Gerharz and others [32] but raises interesting questions about the pressures of increased network traffic, especially with large and multi-dimensional datasets.

E. Reproducibility

Complex chains of diverse models make it difficult to ensure the reproducibility of model runs, and raise particular curation challenges when the recorded implementations become outdated. Given that most integrated modeling approaches use Monte Carlo simulation to incorporate and assess the impact of uncertainties [33]; this also raises the issue of ensuring the reproducibility of model runs and simulations with a random element. For instance, someone else should be able to reproduce the results of a published model, even though a component might have changed in the meantime, or some manipulation to the input data has been performed.

The eHabitat WPS is a typical example where, depending on the actual values of an input, the model algorithm might use assumptions or omit values and currently, because of the encapsulated nature of these decision rules, there is no way of recording or propagating the branches that occurred for a particular run. While lineage and provenance information [34] provide a partial solution, full reproducibility would also require some form of workflow curation and versioning.

The most obvious means of storing and documenting workflows are orchestration tools, such as Taverna, Kepler [35], or Vistrails [36]. While these tools are designed to produce workflows that can be run and shared, they are far more frequently used to describe the logic, parameters and components of a sequence of processing steps. Even in this limited role, workflow tools are extremely useful as a step towards reproducibility. For this reason, it is very relevant in this context that models described using WSDL documents are far more immediately interoperable with and easier to chain using these tools. On the other hand, this straightforward interoperability is partly because of an assumed simplicity in model inputs/outputs. For example, Kepler has little capacity for declaring complex types and ensuring a correct mapping between them, which caused difficulties in an experimental attempt to expose it as a WPS [37].

IV. CONCLUSIONS AND FUTURE WORK

This paper presented a condensed view on our currently ongoing work that investigates suitable service interfaces for the Model Web. Table 1 (below) provides a direct comparison between the WPS and the WSDL approach. While OGC's WPS requires human-to-machine and machine-to-human communication in order to fully exploit the automated capabilities, W3C's WSDL addresses purely machine-to-machine interactions. Both separate and combined approaches have their role in addressing central geoprocessing tasks in the Model Web. However, it is clear that neither of the approaches for generating WSDL wrappers for WPS-based services/processes is an adequate solution for supporting interoperability outside the OGC community.

Due to the variety of issues and approaches, the Model Web is likely to evolve towards a set of ecosystems of components of different granularities that will evolve independently, largely because of the many chasms (e.g., scientific disciplines, independent networks of developers and projects) between the different communities (see also [38]). This will certainly require in depth investigations on the relation between the Model Web and the Geospatial Semantic Web [39]. Starting points are for example provided in [7] and [8].

Besides further elaborations on service interfaces, our future work will particularly address the impact of harmonized data models for environmental information, as currently being developed in the context of the Infrastructure for Spatial Information in Europe (INSPIRE) [40]. We will base our investigations on the aforementioned eHabitat and EFFIS case studies, following a holistic approach.

Lastly, it should be noted that the interface issues stressed in this paper represent only one research field in geoprocessing for the Model Web. Assuming that generic models will become available as modules, better means for orchestration and chaining will be required. We doubt that solutions from the business sector, such as Business Process Modelling Language (BPEL) [41] or Business Process Modelling Notation (BPMN) [42], will suit the arising needs, but this is a different story.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 248488 (UncertWeb project), n° 226487 (EuroGEOSS project) and n° 284898 (ENVIROFI project). The authors would also like to thank the anonymous reviewers, which have helped to improve the overall value of this paper.

TABLE I. SUMMARY OF WPS AND WSDL COMPARISON.

WPS	***WSDL***
Niche-market	Mass-market
Several endpoints (resource-based approach)	Single endpoint (service-based approach)
Functional description of process/method signatures + other descriptive fields	Functional description of method signatures
Support for profiling	No support for profiling
No need of third-party languages (cascading composition) to enable service composition	Need third-party languages (BPEL, etc.) to enable service composition
Human-to-machine interaction	Machine-to-machine interaction
Support for (mimic) WSDL description	No support for WPS description

REFERENCES

[1] G. Geller and W. Turner, "The model web: a concept for ecological forecasting", Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International, pp. 2469 – 2472.

[2] C. Granell, L. Díaz, and M. Gould, "Service-oriented applications for environmental models: Reusable geospatial services", Environmental Modelling and Software, 25(2), 2010, pp. 182-198, ISSN: 1364-8152.

[3] D. Roman, S. Schade, A. J. Berre, N. Rune Bodsberg and J. Langlois, "Model as a service (MaaS)", AGILE Workshop - Grid Technologies for Geospatial Applications, Hannover, Germany, 2009.

[4] E. Pebesma, D. Cornford, G. Dubois,G. Heuveling, D. Hristopoulos, J. Pilz, U. Stöhlkerg, G. Morinh, and J. Skøien, "INTAMAP: the design and implementation of an interoperable automated interpolation Web service", Computers & Geosciences, 37(3), 2011, pp. 343-352.

[5] Sextante, official web page, http://sextante.forge.osor.eu/ (last access: November 18, 2011).

[6] Geographic Resources Analysis Support System (GRASS), official web page, http://grass.fbk.eu/ (last access: November 18, 2011).

[7] L. Vaccari, P. Shvaiko, J. Pane, P. Besana, and M. Marchese, "An evaluation of ontology matching in geo-service applications", GeoInformatica, 2011, DOI: 10.1007/s10707-011-0125-8.

[8] D. Fitzner, "Formalizing cross-parameter conditions for geoprocessing service chain validation, International Journal of Applied Geospatial Research 2 (1), 2011, pp. 18-35.

[9] G. Alonso, F. Casati, K. Harumi, and V. Machiraju, "Web services: concepts, architectures and applications", Springer, Heidelberg, 2004.

[10] H. Jagers, "Linking data, models and tools: an overview", proceedings of iEMSs, 2010.

[11] Common Component Architecture (CCA), official web page, http://www.cca-forum.org/ (last access: November 18, 2011).

[12] Open Modeling Interface (OpenMI), official web page, http://www.openmi.org/reloaded/ (last access: November 18, 2011).

[13] Invisible Modelling Environment (TIME), official web page, http://www.toolkit.net.au/Tools/TIME/ (last access: November 18, 2011).

[14] Taverna, official web page, http://www.taverna.org.uk/ (last access: November 18, 2011).

[15] Open Geospatial Consortium (OGC), “OGC web processing service (WPS) version 1.0.0”, OGC Standard Document, 2007.

[16] Open Geospatial Consortium (OGC), “ OpenGIS web feature service (WFS) implementation specification – version 1.1.0” OGC Standard Document, 2004.

[17] Open Geospatial Consortium (OGC), “OpenGIS Sensor Observation Service (SOS) implementation specification”, OGC Standard Document, 2007.

[18] UncertWeb project, official web page, http://www.uncertweb.org/ (last access: November 18, 2011).

[19] D. Cornford, R. Jones, L. Bastin, M. Williams, E. Pebesma, and S. Nativi, “UncertWeb: chaining web services accounting for uncertainty”, Geophysical Research Abstracts, Vol. 12, EGU 2010, p. 9052.

[20] G. Dubois, J. Skøien, S. Peedell, J. De Jesus, G. Geller, and A. Hartley, “eHabitat: a contribution to the model Web for habitat assessments and ecological forecasting”, 34th International Symposium on Remote Sensing of Environment, Sydney, Australia, 2011.

[21] European Forest Fire Information System (EFFIS), official web page, http://effis.jrc.ec.europa.eu/ (last access: November 18, 2011).

[22] F. Lopez-Pellicer, W. Rentería-Agualimpia, R. Béjar, P. Muro-Medrano, F. Zarazaga-Soria, “Availability of the OGC geoprocessing standard: March 2011 reality check”, Computers & Geosciences, 2011, DOI: doi:10.1016/j.cageo.2011.10.023.

[23] World Wide Web Consortium (W3C), “Web services description language (WSDL) version 2.0 part 1: core language”, W3C Recommendation, 2007.

[24] Open Geospatial Consortium (OGC), “OWS 5 SOAP/WSDL common engineering report”, OGC Discussion Paper, 2008.

[25] R. Haesen, M. Snoeck, W. Lemahieu, and S. Poelmans, “On the definition of service granularity and its architectural impacts”, International Conference on Advanced Information Systems Engineering (CAiSE’08). LNCS, vol. 5078. Springer, 2008, pp. 375–389.

[26] P. Maué, C. Stasch, G. Athanasopoulos, and L. Gerharz, “Geospatial standards for web-enabled environmental models”, Internal Journal for Spatial Data Infrastructures Research (IJSDIR), Vol.6, 2011.

[27] M. Gone and S. Schade, “Towards semantic composition of geospatial web services - using WSMO in comparison to BPEL”, International Journal of Spatial Data Infrastructures Research (IJSDIR), Vol.3, 2008.

[28] Open Geospatial Consortium (OGC), “OpenGIS geography markup language (GML) encoding standard - Version 3.2.1”, OGC Standard Document, 2007.

[29] Open Geospatial Consortium (OGC), “Uncertainty markup language (UncertML)”, OGC Discussion Paper, 2008.

[30] Open Geospatial Consortium (OGC), “OGC network common data form (NetCDF) core encoding standard version 1.0”, Candidate OpenGIS® Encoding Standard , 2011.

[31] Open Geospatial Consortium (OGC), “Observations and measurements – XML implementation version 2.0”, OGC Standard Document, 2010.

[32] L. Gerharz, B. Proß, C. Stasch, and E. Pebesma, “A web-based uncertainty-enabled Information system for urban air quality assessment”, Geophysical Research Abstracts, Vol. 13, EGU2011-5554, 2011.

[33] L. Bastin, D. Cornford, J Richard, G. Heuvelink, E. Pebesma, C. Stasch, S. Nativi, P. Mazetti, and M. Williams , “Managing uncertainty in integrated environmental modelling frameworks”, submitted to Environmental Modelling and Software, 2011.

[34] R. Devillers and R. Jeansoulin, “Fundamentals of spatial data quality, ISTE, London, UK, 2006.

[35] Kepler, official web page, https://kepler-project.org/ (last access: November 18, 2011).

[36] Vistrails, official web page, http://www.vistrails.org/ (last access: November 18, 2011).

[37] Pratt, A., et. al (2010). Exposing the Kepler Scientific Workflow System as an OGC Web Processing Service. iEMSs 2010, Ottawa, Canada.

[38] S. Schade, P. Mazzetti, Z. Sabeur, D. Havlik, T. Uslander, A. Berre, and L. Mon, “Towards a multi-style service-oriented architecture for earth observations”, EGU 2011, Vienna, Austria, 2011.

[39] M. Egenhofer, “Toward the semantic geospatial web. Proceeding GIS”, 10th ACM international symposium on Advances in geographic information systems, 2002.

[40] European Parliament and Council, “Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)”, Official Journal on the European Parliament and of the Council, 2007.

[41] Organization for the Advancement of Structured Information Standards (OASIS) “Web services business process execution language version 2.0”, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.pdf, 2007 (last access: November 18, 2011).

[42] Object Management Group (OMG), “Business process model and notation (BPMN), version 2.0”, http://www.omg.org/spec/BPMN/2.0/, 2011 (last access: November 18, 2011).

One Click Focusing: An SQL-based Fast Loop Road Extraction Method for Mobile Map Services

Daisuke Yamamoto
Nagoya Institute of Technology
Gokiso-cho, Showa-ku,
Nagoya, Aichi, Japan
daisuke@nitech.ac.jp

Hiroki Itoh
Nagoya Institute of Technology
Gokiso-cho, Showa-ku,
Nagoya, Aichi, Japan
hito@moss.elcom.nitech.ac.jp

Naohisa Takahashi
Nagoya Institute of Technology
Gokiso-cho, Showa-ku,
Nagoya, Aichi, Japan
naohisa@nitech.ac.jp

***Abstract*—This paper proposes a method for the fast loop roads extraction for mobile Web map services and applications. Since the existing loop road extraction method has drawbacks such as those related to the processing speed and interactivity, it has been difficult to apply the method to real-time applications such as mobile Web map services directly. Therefore, this paper proposes a fast extraction method that involves acquiring information on all loop roads with high efficiency in advance and storing the information in a database and querying those with SQL statements. The proposed method is 51.0 times faster than the previous method, and for expanded loop road extraction, it is 16.4-25.3 times faster than the previous method. Further, when used to build a loop road database, the proposed method, which involves the use of a tabulation method, is 3.86 times faster than the conventional method. We have developed the Web API function in order to acquire loop roads easily from other Web services. For the application of the proposed method, we have developed the One Click Focusing function that can modify the size, position, and scale of the focus automatically in fisheye-view maps.**

***Keywords*-fisheye views, focus+glue+context,Web map service**

I. Introduction

Advanced Web map services, such as Google Maps, have become available in recent years. Smart phones equipped with GPS sensors, such as iPhones, have also become increasingly popular, which means that anyone can access mobile Web map services. Web map services allow users to determine their current position, but they also enable the sharing of content mapped with positions in many applications, such as location-based SNS (Social Network Services) [1] and pedestrian navigation systems [2].

Many applications are based on existing Web map services, such as location mapping systems based on points (latitude and longitude) and navigation systems based on lines (road networks), but few services are based on polygons, such as areas and city blocks.

However, it is important that mobile web map services include surfaces, such as areas and city blocks, as well as points and lines. For example, let us consider a situation where a user wants to find certain areas using mobile maps. A user must adjust the scale and position of the map to visualize the whole area in the display, after scrolling through the map. However, this operation is slightly too complex in a mobile environment. Therefore, there is a user need for target area adjustments that use simple one-click or one-touch operations on mobile maps. Some map databases are based on polygon data for areas and structures, such as the National Land Numerical Information download service [3] and U.S. Census Bureau Tiger/Line [4], so we aimed to develop a method for automatically adjusting maps using these polygons. In general, humans create polygons, but polygon data is not available for all areas and structures. Therefore, we must automatically extract these polygons from road databases.

Our study was focused on a loop road. A loop road is the smallest road network, which surrounds a target point on a road network map, as shown in Figure 1. The area surrounded by a loop road is a city block. Many areas consist of some city blocks, including shopping centers, universities and parks, so a city block is an important unit. Thus, if we can handle city blocks easily on Web map services, more area-based Web map applications will be made available.

The purpose of this study is to develop a fundamental system to enable Web map services that can handle city blocks easily. We propose a fast extraction method for loop roads and expanded loop roads. Moreover, we developed a “ One Click Focusing ” function on this method.

The following requirements had to be satisfied to implement the proposed system.

Requirement 1 The processing speed must be fast and stable, in order to allow a serviceable response with Web map services.

Requirement 2 Loop roads must be calculated from existing road networks without human costs.

Requirement 3 A Web API must be defined in order to implement the proposed system easily from other Web services.

Requirement 4 Applications based on the proposed system must be available.

Thus, the proposed system possesses the following characteristics, in order to satisfy the requirements.

ISBN: 978-1-61208-178-6

Characteristic 1 Loop roads extraction can be rapid and stable by querying to a loop road database with SQL statements. Since SQL is one of the fastest technologies that involve searching massive data, we believe that our approach is the best solution for extracting loop roads. (Corresponding to Requirement 1)

Characteristic 2 The loop road database can be built automatically from existing road databases efficiently. (Corresponding to Requirement 2)

Characteristic 3 The Web API can communicate with clients and other Web servers. (Corresponding to Requirement 3)

Characteristic 4 We developed a One Click Focusing function for fisheye view maps as an application of the proposed system. (Corresponding to Requirement 4)

Our method will provide a fundamental technological contribution to innovative and intelligent Web map services.

This paper has eight sections. Section II summarizes existing loop road extraction methods. Section III describes a fast method for extraction. Section IV presents the structure of the proposed system. Section V contains the experimental results. Section VI describes an application based on the proposed system. Section VII contains related work. Section VIII concludes this paper.

II. Loop Road Extraction Method

We previously proposed a loop road algorithm [5] for extracting loop roads and expanded loop roads. This section summarizes existing methods and provides definitions of a loop road and an expanded loop road. These methods can extract loop roads and expanded loop roads with high accuracy, but the processing speeds of these methods are too slow for Web map service applications. Therefore, we propose fast extraction methods in the next section.

A loop road is the smallest loop road network, which surrounds a target position, as shown in Figure 1. A city block is the area surrounded by a loop road. A level 1 expanded loop road is the loop road network that surrounds a loop road. A level N expanded loop road is the road network that surrounds a level N-1 expanded loop road.

A loop road is important for the following reasons. Some areas, including parks, universities, and shopping centers, generally consist of city blocks, so we can treat a city block as a semantic unit that divides an area. Thus, a loop road treats road networks as areas. Some parks and universities contain several city blocks, so we need to address neighboring city blocks as well as single city blocks. We can process neighboring city blocks simultaneously by extracting expanded loop roads.

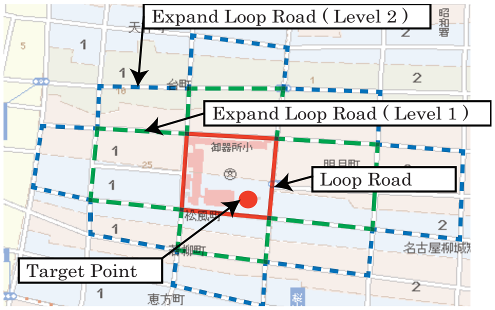
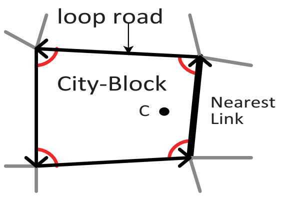


Figure 1. Sample of a loop road and expanded loop roads. The area surrounded by a loop road is a city block.



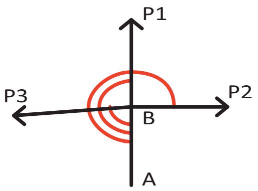


Figure 2. Example of a loop road. Figure 3. Selection of a loop road.

A. Method for Loop Road Extraction

We propose a loop road algorithm that extracts a loop road surrounding point C. We initially remove orphan links that do not form a loop road, such as straight roads.

First, we find the nearest link from point C then we follow the link in a counterclockwise direction. When the link connects to the tail of the first link, the path connecting these links is a loop road. However, when the link comes to a dead end, we follow a neighboring link using a depth-first search algorithm, as shown in Figure 2.

Algorithm 1 finds the loop road algorithm in the counterclockwise direction. Here, *start.tail* and *start.head* are the front and tail nodes, respectively, of the nearest link from point C. *setVisitedLink(node1, node2)* is a function that sets a *visited flag* on the link that connects node1 and node2. *node.UnvisitedLeftChild()* is a function that returns the unvisited and leftmost links connected to the *node*. Figure 3 shows that the function returns node P3 when node B is the current node and no links (P1, P2, P3) are not as visited flags. Likewise, *node.UnvisitedRightChild()* is a function that returns the unvisited and rightmost links connected to the current node.

B. Method for Extracting Expanded Loop Roads

The expanded loop road algorithm expands to city blocks surrounding a loop road in a radial direction, block-by-block.

The expanded loop road algorithm proceeds as follows. For each link in the level N expanded loop road, apply the loop road algorithm in the clockwise direction, as shown in Figure 4. This finds the level N+1 expanded loop road and expands to city blocks in a radial direction. However, we

Algorithm 1 *LoopRoad* algorithm (counterclockwise)

```
Require: $Link : Start$
$stack := \textbf{new}Stack()$
$setVisitedLink(Start.tail, Start.head)$
$stack.push(Start.head)$
while not $stack.empty()$ do
    $node := stack.top()$
    if $node = Start.tail$ then
        return $stack$
    else
        $child := node.UnvisitedLeftChild()$
        if $child = null$ then
            $stack.pop()$
        else
            $setVisitedLink(node, child)$
            $stack.push(child)$
        end if
    end if
end while
```

must apply the loop road algorithm in a counterclockwise direction for some links when expanding to city blocks in a complex road network. Thus, we must apply the loop road algorithm in both clockwise and counterclockwise directions for each link. The algorithm is provided below.

Step 1 Initialize List A.

Step 2 Store links for an initial loop road (Figure 4-a) in List A.

Step 3 For each link stored in List A, apply the loop road algorithm in both clockwise and counterclockwise directions, as shown in Figure 4-b.

Step 4 Remove links stored in List A from links generated in Step3.

Step 5 Apply the loop road algorithm in the counterclockwise direction using the links generated in Step4, as shown in Figure 4-c. This provides an expanded loop road.

Step 6 Repeat from Step 2 if the city blocks need to be further expanded, as shown in Figure 4-d and 4-e.

List A is the list structure storing the loop road. Finally, we find the radius and center point of the Focus connected to the loop road at each level.

III. Fast Extraction Method

This section describes fast extraction methods for loop roads and expanded loop roads.

The basic principle of the proposed method is as follows. First, we build a loop road database to store all loop roads, which are calculated in advance to improve efficiency. The loop road database contains a loop road table with the road links of loop roads and a neighboring table with relationships between roads and loop roads. Thus we can rapidly find any loop road by querying the loop road database without searching road networks, because searching huge networks would be too computationally intensive. We can also find expanded loop roads by combining neighboring loop roads, which also eliminates the need to search road networks.

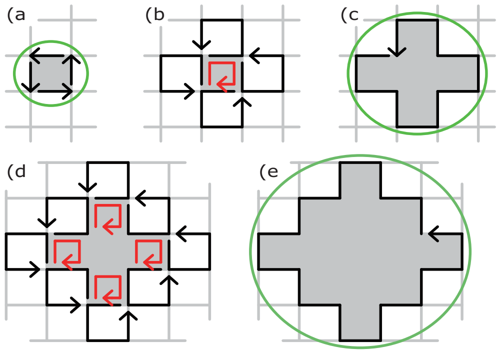


Figure 4. Expanded loop road algorithm. Gray lines, black arrows, red arrows, and the green circle indicate roads, links, previous links, and the Focus area, respectively.

A. Definitions of databases

First, we provide details of the road database and the loop road database.

The road database contained a road table with road links between intersections for navigation. The columns of the road table contained road link IDs, coordinates of road start points and end points, and other road link IDs connected to this link. The road database was converted from the “Navigation Road Map 2007” published by Yahoo Japan.

A loop road database includes the LoopRoadTbl table and the NeighborTbl table. The LoopRoadTbl table contains road links for loop roads. The columns of the LoopRoadTbl table contain loop road ID(id), lists of road link IDs and coordinates connected with a loop road (roadList, coordList), and a rectangular region containing a loop road (north, south, west, east), as shown in Table I. The roadList column is a text field containing a list of road link IDs in Comma Separated Value (CSV) format. For example, the CSV format of a loop road with road IDs of 100, 101, and 102 is ”100,101,102”. The coordList column is also a text field containing a list of coordinates for the loop road.

The NeighborTbl table contains relationships between a road and loop roads containing this road. Figure 5 shows that a road connects with two loop roads: a left loop road and a right loop road. The columns of the NeighborTbl table contain road link ID (id), left-hand side loop road ID (leftId), and right-hand side loop road ID (rightId), as shown in Table II. We can find loop road IDs connected with any roads by searching the neighboring table.

B. Loop Road Database Construction

Here we describe an automatic method for building a loop road database from an existing road database to satisfy requirement 2 in Section I.

Table I
DEFINITION OF THE LOOP ROAD TABLE: LOOPROADTBL

column	type	description
id	integer	ID of the loop road
north	real	northernmost latitude of loop road
south	real	southernmost latitude of loop road
west	real	westernmost longitude of loop road
east	real	easternmost longitude of loop road
roadList	text	list of road IDs
coordList	text	list of coordinates

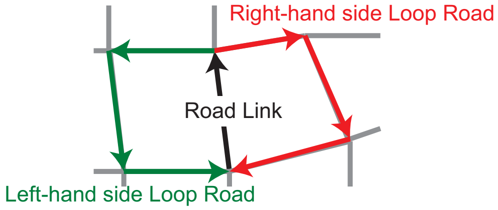


Figure 5. Road connected to two neighboring loop roads (left-hand side and right-hand side).

A very high number of road links exist in Japan, so it is difficult to generate a loop road database in a short period of time. We needed to generate a loop road database as quickly as possible, so we tested the following three methods.

1) Grid Division Method: First, we discuss a simple method of dividing the map into a grid made of vertical and horizontal lines at a constant interval before applying the loop road algorithm described in Section II to all vertices of the grid.

The algorithm for this method is as follows. First, we divide a map into a grid with a distance interval n. The following steps are then applied to each vertex P of this grid.

Step 1 Apply the loop road algorithm described in Section II to point P.
Step 2 Check whether the same loop road is stored in the loop road database.
Step 3 Insert the loop road to the loop road database if the loop road does not exist in the database.

This method can extract loop roads containing any vertices in the grid, but this method cannot extract loop roads that contain no vertices in the grid. A loop road with multiple vertices in the grid must be calculated redundantly with the same number of vertices included in the loop road. We may extract all loop roads if the distance n is sufficiently small, such as 1 meter, but the extraction time is too great since there are a lot of vertices. In contrast, we can extract loop roads in a short time if distance n is sufficiently high, such as 10 km, but we might not extract many loop roads because small loop roads might include no vertices. The density of roads varies in different areas, which makes it difficult to determine the best distance n for solving this dilemma.

2) Road Link Method: We next discuss the road link method, which extracts loop roads by searching from each road instead of vertices on a grid. This method applies clockwise and counterclockwise loop road algorithms based on Algorithm 1 to all road links and inserts these loop roads into the loop road database.

Table II
DEFINITION OF THE NEIGHBORING TABLE: NEIGHBORTBL

column	type	description
id	integer	ID of the road
leftId	integer	left-hand side loop road ID
rightId	integer	right-hand side loop road ID

The algorithm is as follows. We apply the following steps to each road link L.

Step 1 Apply the clockwise loop road algorithm based on Algorithm 1, to a road link L.
Step 2 Apply counterclockwise loop road algorithm to a road link L.
Step 3 Check whether the same loop road is already stored in the loop road database, for each loop road detected in Step 1 and Step 2.
Step 4 Insert each loop road to the loop road database if the same loop road is not present in the database.

In contrasts to the grid division method, this method can extract all loop roads. A problem with this method is the need to calculate n times for one loop road when a loop road has n road links. Ideally, a loop road must be calculated only once, so the extraction time of a loop road with this method is n times longer compared with the ideal method. For example, when $n = 4$ in the grid road network, the road link method takes four times longer than the ideal method. It is desirable to solve this duplication problem.

3) Road Link Method without Duplication: Finally, we propose a road link method without duplication. This method can extract all loop roads without calculating duplicate loop roads by using the NeighborTbl table.

The algorithm is as follows. We apply the following steps to each road link K, where a link ID of link K is k. The detailed algorithm is shown in Algorithm 2.

Step 1 Select the row for a column, where the link ID is k from the NeighborTbl table. Go to Step 6, if the leftId of this row is not null.
Step 2 Find a loop road R by applying the counterclockwise loop road algorithm based on Algorithm 1, where the loop road ID of the loop road R is r.
Step 3 Go to Step 6 if Step 2 cannot find a loop road.
Step 4 Insert loop road R into the LoopRoadTbl table.
Step 5 For each road K' of the loop road R, update the leftId column in the row of the linkId column in the NeighborTbl table for k' as follows.

```
UPDATE NeighborTbl SET leftId=r WHERE id=k'
```

Step 6 Likewise, apply Step 1 to Step 5 using the clockwise loop road algorithm and rightId.

This method is fast, because this method can build the

Algorithm 2 Road Link Method without Duplications (counterclockwise). $LoopRoad(link)$ function returns the loop road started from $link$ based on Algorithm 1.

```
Require: $Link : k$
$row$ :=select $*$ from $NeighborTbl$ where $id = k$
if $row.leftId = null$ then
    $loop := LoopRoad(row.ID)$
    if $loop \neq null$ then
        for all $link \in loop$ do
            update $NeighborTbl$ set $leftId = loop.ID$
            where $id = link.ID$
        end for
    end if
end if
```

LoopRoadTbl table without duplications by referring to the NeighborTbl table. The number of loop road calculations required this method can be four times smaller than that required by the road link method, when a loop road has 4 road links. This method requires extra disk space for the NeighborTbl table, but the NeighborTbl table is also required for speeding up expanded loop road extraction. For these reasons, we adopted this method.

C. SQL-based Loop Road Extraction

We propose a rapid method for finding the loop road for a target point P by referring to the previously constructed loop road database.

The algorithm of this method is as follows, where the latitude of target point P is $P.lat$, and the longitude is $P.long$. The detailed algorithm is shown in Algorithm 3.

Step 1 Find the candidate loop roads by querying the LoopRoadTbl table using the following SQL query.

```
SELECT * FROM LoopRoadTbl where
  north<P.lat and south>P.lat and
  west <P.long and east>P.long
```

Step 2 Apply a point-in-polygon[6] algorithm to the point P and each loop road candidate to determine the loop road including the point P.

Step 3 The result is a candidate loop road including the point P. If no candidate loop road includes the point P, the result is null.

The advantage of this method is that we can rapidly acquire loop roads using a simple SQL query and without searching any road networks.

D. SQL-based Expanded Loop Road Extraction

Next, we propose a method for finding an expanded loop road for a target point P. The algorithm for this method is as follows, where Set A and Set B are null. The detailed algorithm is shown in Algorithm 4.

Step 1 Find a loop road that surrounds the point P by using the loop road extraction method described in Section III.C.

Algorithm 3 $FastLoopRoad$ Algorithm.

```
Require: $Position : P$
$rows$ :=select $*$ from $LoopRoadTbl$ where $north < P.lat$ and $south > p.lat$ and $west < P.long$ and $east > P.long$
for all $row \in rows$ do
    if $PointInPolygon(P, row)$ then
        return $loop$
    end if
end for
```

Algorithm 4 $FastExpandedLoopRoad$ Algorithm. N means a level of the expanded loop road. $LinksOf(A)$ function returns road links included in loop roads of A.

```
Require: $Position : P, Integer : N$
$A$ :=new $Set()$
$B$ :=new $Set()$
$loop := FastLoopRoad(P)$
if $loop \neq null$ then
    $A.add(loop.ID)$
    for $i = 1 \rightarrow N$ do
        $B.addAll(A)$
        $rows$ :=select $leftId, rightId$ from $NeighborTbl$ where $id$ in (select $id$ from $NeighborTbl$ where $leftId$ in $A$ or $rightId$ in $A$)
        $A.clear()$
        for all $row \in rows$ do
            $A.add(row.leftId)$
            $A.add(row.rightId)$
        end for
        $A.removeAll(B)$
    end for
    return $LoopRoad(LinksOf(A) - LinksOf(B))$
end if
```

Step 2 Add the loop road ID to Set A if the loop road is not null.

Step 3 Repeat the following Step 3.1 to Step 3.4 N times.

Step 3.1 Add Set A values to Set B.

Step 3.2 Find loop roads with road links in loop roads included in Set A by querying the NeighborTbl table with the following query.

```
SELECT leftId, rightId FROM neighborTbl
  WHERE id IN (SELECT id FROM neighborTbl
    WHERE leftId IN ( Set A values ) OR
    rightId IN ( Set A values ))
```

Step 3.3 Update Set A with the values of leftId and rightId based on the SQL results of Step 3.2.

Step 3.4 Remove Set B values from Set A.

Step 4 Remove links in loop roads of Set B from links in loop roads of Set A. Apply LoopRoad method described in Algorithm 1 to the links in order to

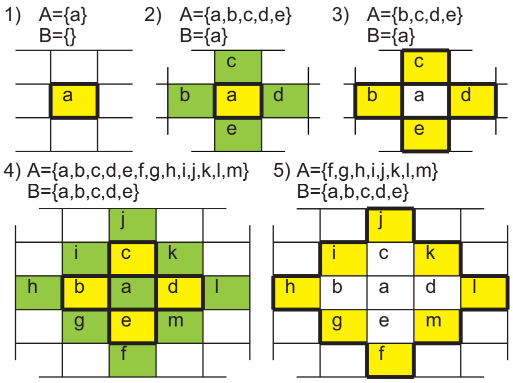


Figure 6. Example of Level 2 expanded loop road algorithm. 1) a is a loop road extracted by FastRoadLink function, 2) Add values of A to B. And, $a, b, c, d, e \in A$ are the loop roads extracted by the SQL query of line 8 of Algorithm 4, which share the edges of a. 3) Remove values of B from A. 4) Add values of A to B. A are the loop roads extracted by the SQL query of line 8, which share edges of b, c, d, e. 5) Remove values of B from A. Remove links in loop roads of B from links in loop roads of A.

connect the links.

The result of Step 5 is the target expanded loop road (Level N).

The advantage of this method is that it is almost completed by the SQL query in Step 3.1. The SQL query is only conducted N times for level N and the size of the NeighborTbl table is smaller than the LoopRoadTbl table, and the road database. Thus, we expect that the processing speed of this method will be lower than other methods.

IV. System Architecture

In order to satisfy Requirement 3 described in Section I, we had to develop a function to communicate with other Web services. Thus, we examined the following two methods. One is a library approach used by Web engineers to install a whole system to a Web server, including the programs and the large road loop road databases. The other method is a Web API approach used by Web services to communicate on-demand with our system using a Web API. The library approach cannot satisfy Requirement 3, because this approach uses large volumes of disk spaces and requires engineers to install programs on Web servers. Therefore, we adopted a Web API approach that more easily allows Web engineers to use the proposed methods.

A. System Architecture

We adopted a server-client architecture, as shown in Figure 7. This system includes an application server and database servers. A database server includes the previously constructed road databases and the loop road database. We adopted MySQL 5 as the database management system. We adopted Tomcat 6 and Java 1.6 in the application server. The method of communication between the server and clients was based on a Web API described in the next section. The adoption of Web API allows clients and other Web servers to find loop roads and expanded loop roads on demand.

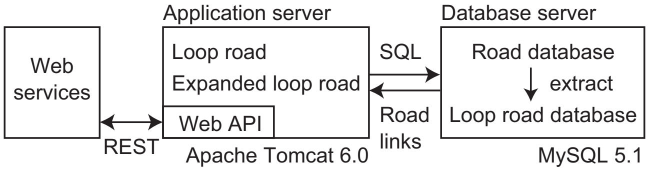


Figure 7. System architecture. We adopted a server-client architecture with Web API functionality.

B. Web API

Web API provides methods such as REST (Representational State Transfer) and SOAP (Simple Object Access Protocol). We adopted the REST method, because this method is easiest for Web services. This method submits parameters by adding them to a request URL and returns results in XML format. A sample URL request is as follows.

```
http://server/api?mode=loop&lat=X&long=Y&level=N
```

In this request, $server$ is the URL of proposed system, while the $mode$ attribute indicates whether the request wants to acquire a loop road or an expanded loop road. The system returns a loop road if the $mode$ attribute is “loop”. The system returns an expanded loop road if the $mode$ attribute is “expand”. We can set $lat/long$ using latitude and longitude attributes. We can set the appropriate level of an expanded loop road using the $level$ attribute. The response contains the road links of the loop road. An example of a response is as follows.

```
<looproad>
  <header>
    <circle lat="" long="" radius=""/>
    <rect top="" bottom="" left="" right=""/>
  </header>
  <roads>
    <road id="" lat1="" lng1="" lat2="" lng2=""/>
    ...
  </roads>
</looproad>
```

The $header$ element contains a summary of the response, which includes a $circle$ element and a $rectangle$ element. The $circle$ element contains a center coordinate and the radius of a circle including the loop road. The $rectangle$ element shows a rectangle containing the loop road. The $roads$ element includes $road$ elements with road links for the loop road, where the $lat1$ and $long1$ attributes denote the starting point of the road link, and $lat2$ and $long2$ indicate the end points. Web services can access loop road information by reading this XML format.

V. Experimental Result

We tested the applicability of the proposed system from the perspective of Web services. Thus, we compared the processing time of the existing method with that of the proposed method.

The experimental environment consisted of a database server and an application server, as described in Section IV. The size of the road database was about 22 GB, so we could not load all roads in the memory. Instead, we loaded road links from the road database on demand. The specification of the database server was as follows: Intel Core i7 2.9 GHz, 12.0 GB memory. The application server contained the programs for the proposed system. The specification of the application server was as follows: Intel Core i7 3.0 GHz, 8.0 GB memory.

A. Loop Road Database Construction

First, we investigated the time required for constructing the loop road database. The loop road database can be constructed automatically in advance from road databases, but the time required to construct the loop road database must be short. We must rebuild the loop road database after the original road databases are updated, so it is better to construct the loop road database as quickly as possible to reduce the lead time.

Therefore, we compared the following three methods. The target area was a central area in a major Japanese city (Nagoya city). The area has 21820 road links.

method 1 We used the grid division method described in Section III.B.1 and divided the map into 50 m, 100 m, or 200 m intervals.

method 2 Road link method, as described in Section III.B.2.

method 3 Road link method without duplication, as described in Section III.B.3.

We investigated the construction time, the number of calculations required, and the number of loop roads with each method. The number of calculations referred to the total number of function calls by the loop road algorithm. The number of loop roads means the actual number of loop roads without duplications, because the same loop road may be counted many times.

Table III
TIME REQUIRED TO CONSTRUCT LOOP ROAD DATABASES USING EACH METHOD. METHOD 3 IS PROPOSED METHOD.

	total time	calculation number	loop road
method 1 (50m)	17291 s	26649	7518
method 1 (100m)	4518 s	6745	4402
method 1 (200m)	1206 s	1828	1407
method 2	1769 s	43640	8082
method 3	459 s	8737	8058

Table III shows the results. First, we compared method 1 with method 2. Method 2 was 0.68-9.8 times faster than method 1. In particular, the total time required for method 1 with 50 m intervals was significant longer than other method, including method 1 with 100 m and 200 m. Method 1 with 50 m intervals detected almost as many loop roads as method 2, whereas method 1 with 50 m and 100 m intervals failed to detect many loop roads. This result suggests that method 1 has a problem with the trade-off between calculation speed and the number of loop roads, which contrasts with methods 2 and 3.

Next, we compared method 2 with method 3. Method 3 was 3.86 times faster than method 2. The loop roads calculation number with method 3 (8737) was 4.99 times higher than method 2 (43640), which suggests that method 3 is effective in avoiding duplications. Since the number of extracted loop roads with method 3 was almost identical to method 2, the extraction rate was almost the same. Method 3 has the disadvantage that it requires extra disk space in the NeighborTbl table, but we consider this problem to be trivial because the NeighborTbl table is also useful for accelerating the expanded loop extraction method. However, method 3 has the advantage that it constructs a stable loop road database rapidly and robustly.

The target area was only one of more than 4800 areas in Japan, so the time required to construct a database for all areas in Japan may be more than 4000 times greater. Therefore, the advantages of the proposed method become more significant when applied to all areas in Japan.

B. Loop Road Extraction

Next, we investigated the loop road extraction time. The extraction time for a loop road included: 1) the time required for accessing databases; and 2) the time required for calculating road networks. Thus, we measures the time required for each process. The target area was the center area of a major city in Japan (Nagoya city). We measures the extraction time for loop roads at 1000 random time points. We compared the following two methods.

method 1 Existing method. The system applied the algorithm described in Section II.A, after loading road links within $1km^2$ of the target point.

method 2 Proposed method described in Section III.C. This system can detect loop roads much more quickly by using the loop road database, rather than the road database.

Table IV shows results. The total time required for method 2 was 45.4 times faster than that with method 1. The reason for this was as follows. Both methods had to access the database once, but the time required for accessing the database with method 2 was 51.0 times faster than with method 1. Method 1 had to acquire road links from the database in a large range (1 km^2), because the range of the target loop road was previously unknown, whereas method 2 had the advantage of acquiring candidate loop roads directly from database by issuing a simple SQL query. Method 1 had to apply a loop road extraction algorithm after acquiring the SQL results, whereas method 2 had the advantage of not applying this algorithm. The average road network calculation time with method 2 was only about 1 ms.

The standard variance with method 2 was much smaller than that with method 1. The reason for this was as follows. The road network calculation time with method 1 varied with the size and complexity of the loop road, whereas this was constant with method 2 which acquired any result by issuing a simple SQL query to the database without any complex processing requirements.

The precision of method 2 (98.0%) was higher than that of method 1 (92.3%). Method 1 searched a loop road starting from only the nearest link of target point, whereas method 2 could search a loop road starting from any of the links. The main reason why the precision of method 2 was not 100 percent was that some areas lacked loop roads, such as the coast. The precision of method 2 was low if the loop road database could not be constructed with high accuracy, which suggests that loop road extraction and construction of the loop road database had high accuracy with the proposed method. Method 2 was more stable and quicker at extracting loop roads than method 1, which suggest that the proposed method is more suitable for Web services.

C. Expanded Loop Road Extraction

Finally, we investigated the extraction time for expanded loop roads. The extraction time for an expanded loop road includes: 1) the time required to access the database; and 2) the time required to calculate data. Thus, we measured the time required for each process. The target area was the center area of a major city in Japan (Nagoya city). We measured the time required to calculate the loop road extraction at 100 random points, for each Level of the expanded loop road. We compared the following two methods.

method 1 Existing method. The system applied the algorithm described in Section II.B to road links, after loading road links within 2 km^2 of the target point.

method 2 Proposed method described in Section III.D. This system used the loop road database to find loop roads more quickly.

Table V shows the results. The total processing time with method 2 was 16.4-25.3 times faster than with method 1. The reason for this was as follows. The detection time with method 1 was higher, because method 1 had to load a very large range of road links from the road database, whereas method 2 had the advantage that it only had to make as many SQL queries as there were loop road levels. Method 1 had to apply the loop road extraction algorithm to road networks for the total number of city blocks, whereas method 1 only had to apply this algorithm once.

The standard variance with method 2 (9.1-39.3) is much smaller than that with method 1 (2096-8321). The reason for this was as follows. The road network calculation time with method 1 varied with the size and complexity of the loop road, while it was constant with method 2 which acquired any result by issuing an SQL query to the database N times, for the level N of the expanded loop road. Thus, method 2 was more stable at detecting loop roads compared with method 1.

These results suggest that the proposed system is better for Web map services than existing systems.

VI. Application

This section describes the application of the loop road extraction method to satisfy Requirement 4 in Section I.

So, we have proposed and developed the Focus+Glue+Context type fisheye view maps Emma [7][8][9]. Emma is the first Fisheye view map system for Web map services. Figure 8 shows that Emma has a Focus to show large-scale maps of target areas, a Context to show a small-scale map, and a Glue that shows the routes connecting Focus with Context. Unlike existing fisheye views [10][11][12][13][14], the Focus and Context have no distortion, because Glue contains all the distortion. Only the Glue must be drawn dynamically, so this system is rapid and suitable for Web map services. Emma users can observe detailed maps of target areas and the geographical relationships between the targets. Users can also adjust the size, scale, and position of the Focus according to target area, by mouse-dragging the edge of the Focus.

The Context and Glue is too small when the Focus is too large when using a small display, so it is important to adjust the size, position, and scale of the Focus to make the Focus as small as possible and fitting the Focus in the target area. Users cannot use a mouse with mobile maps, such as smart phones, which makes it is difficult to adjust the Focus manually. Therefore, we must adjust the Focus automatically based on the target area.

Thus, we propose a “One Click Focusing” function to automatically adjust the size, position, and scale of the Focus, based on the target city blocks. We adopted a server-client model using Web API. The algorithm of this function is as follows.

Step 1 A user clicks any point on the map in the client.

Step 2 The client submits the clicked point to the server using a Web API function.

Step 3 The server calculates a loop road for the clicked point and responds to the client in XML format.

Step 4 The client finds the center position P and the radius R of a circle that contains the loop road based on the XML response.

Step 5 The client adjusts the Focus, based on the position P and radius R.

In Step 5, the center of the Focus, F_P, and the scale of the Focus, F_S, are calculated using the following functions where the radius of the Focus is F_R, which is limited by the display size.

$$F_P = p \tag{1}$$

$$F_S = k \cdot F_R / r \tag{2}$$

ISBN: 978-1-61208-178-6

Table IV
MEAN LOOP ROAD EXTRACTION TIME AND PRECISION FOR EACH METHOD. DATABASE TIME MEANS THE TIME OF ACCESSING DATABASES. CALCULATION TIME MEANS THE TIME OF CALCULATING ROAD NETWORKS. UNIT IS MILLISECOND. METHOD 2 IS PROPOSED METHOD.

	database time		calculation time		total			precision
	mean	std.	mean	std.	mean	std.	factor	
method 1	931.1 ms	258.8	12.2 ms	15.6	943.3 ms	259.9	51.0×	92.3%
method 2	18.5 ms	6.2	0 ms	0	18.5 ms	6.2	1.0×	98.0%

Table V
EXPANDED LOOP ROAD DETECTION TIME USING EACH METHOD, FOR EACH LEVEL OF THE EXPANDED LOOP ROAD. DATABASE TIME MEANS THE TIME REQUIRED TO ACCESS THE DATABASES. CALCULATION TIME MEANS THE TIME REQUIRED TO CALCULATE THE ROAD NETWORKS. UNIT IS MILLISECOND. METHOD 2 IS PROPOSED METHOD.

		database time		calculation time		total		
		mean	std.	mean	std.	mean	std.	factor
level 1	method 1	1212.3 ms	428.9	414.1 ms	2098.6	1626.4 ms	2096.4	14.8×
	method 2	106.6 ms	8.2	3.3 ms	3.1	109.9 ms	9.12	1.0×
level 2	method 1	1174.3 ms	416.7	1540.0 ms	3532.0	2714.7 ms	3684.5	17.2×
	method 2	149.8 ms	8.4	8.2 ms	9.7	158.1 ms	14.8	1.0×
level 3	method 1	1233.9 ms	401.1	2085.2 ms	3977.3	3319.1 ms	3989.4	16.4×
	method 2	190.2 ms	31.62	12.2 ms	20.2	202.4 ms	36.5	1.0×
level 4	method 1	1181.9 ms	369.3	5329.0 ms	8296.7	6510.9 ms	8321.3	25.3×
	method 2	238.0 ms	32.7	19.3 ms	19.5	257.3 ms	39.3	1.0×

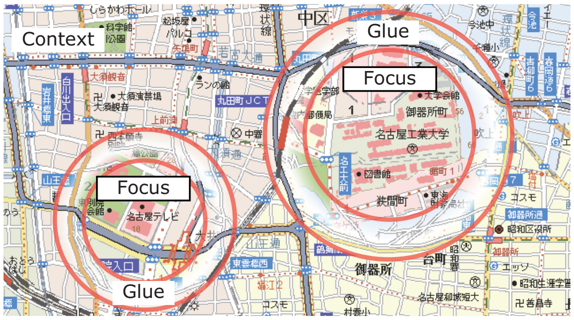


Figure 8. "Focus+Glue+Context" fisheye view maps in Emma. Users can adjust the size, position, and scale of the Focus by mouse-dragging.

k is a constant determined by the display resolution. If users want to expand the Focus, the Focus can be expanded step-by-step by acquiring expanded loop roads, instead of a loop road.

The advantage of the proposed method is that the Focus can be part of target area and also boundary roads of the target area, because the size of the Focus is determined by the loop roads. Therefore, the proposed system can find the routes for the target area of the Focus with certainty, as shown in Figure 9. This suggests that we can develop advanced Web map services by applying the loop road algorithm to existing Web map services.

VII. RELATED WORK

One of the most advanced Web map services is the Open Street Map [15]. Open Street Map enables users to easily edit maps using Web browsers like a Wiki. Users can modify and add roads in the Open Street Map, but users cannot edit and control the map based on city blocks or areas, as described in our current study.

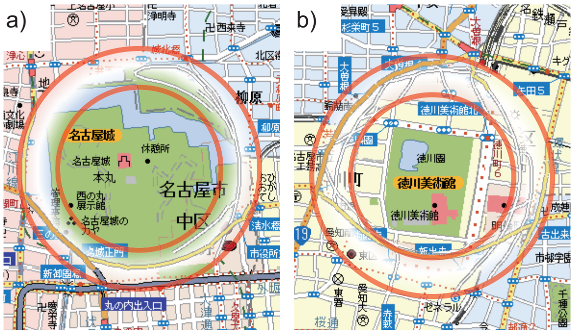


Figure 9. "One Clicking Focusing" function can automatically adjust the size, position, and size of the Focus, based on loop roads. a) The route to the target area is interrupted, because part of a loop road is not shown in the Focus. b) The route to the target area is connected to the target area, because loop roads are shown in the Focus.

Many studies have detected roads by analyzing paper maps or satellite images. For example, satellite images have been used in many studies in the academic fields of geo-science and pattern recognition, particularly in methods for detecting roads by recognizing road edges [16][17], pattern matching [18][19], and methods based on local coidentity of roads [20]. These methods are effective when a user wants to generate new maps for an area that has no road maps. Our method is effective for areas found in road map databases, which contrasts with these image-based methods.

In some field of academic study, such as graph theory [21] and mobile distributed network theory [22], it is popular to detect cycle graphs. In particular, when a network is down due to a loop problem where a data packet goes through the same routes many times if the system dynamically builds the network, such as occurs with cycle graphs in mobile distributed networks. Thus, there are many approaches for generating directed acyclic graphs (DAG) without cycle graphs, which contrasts with our approach. These studies are relevant to our proposed method, but the definition and purpose of the loop road are different from cycle graphs. For example, a cycle graph may not be the smallest network surrounding any point, which contrasts with a loop road.

VIII. Conclusion

This study proposes rapid methods for extracting loop roads and expanded loop roads by constructing a loop road database from an existing road database, for use in Web map services. The proposed method can find a loop road 51.0 times faster than the previous method, and an expanded loop road 16.4-25.3 times faster than previous method. The precision of the proposed method (98.0%) is higher than that of the previous method (92.3%). The proposed method can effectively build a loop road database by avoiding duplications and it constructs a loop road database 3.86 times faster than conventional method. We developed a Web API, which allows Web engineers to develop Web Map services based on city blocks. We also developed a One Click Focusing function as an example application of loop roads. This function can automatically adjust the size, scale, and position of the Focus, according to the target area and city blocks. This system is particularly effective for mobile maps on smartphones.

This system allows Web map services to handle maps based on surfaces, such as city blocks, which contrasts with existing Web map services. We consider that our study contributes to the interface of mobile Web map services, such as One Click Focusing, but also to novel Web map services based on surfaces, such as city blocks. The feature of the proposed system is calculation using SQL technology mainly, so we believe that the processing speed of the proposed system could be further increased by adopting other database technologies, such as replications, which are popular technologies used in Web services. Thus, our system contributes to GIS and novel Web services.

Acknowledgments In the development of the prototype system, we were supported by Yahoo Japan Corporation. We would like to thank Yahoo Japan Corporation. This work was supported by JSPS KAKENHI 20509003.

References

[1] D. Yamamoto, I. Takumi, and H. Matsuo, “Location-based social network services employing student cards for university,” in *Proceedings of the 2009 International Workshop on Location Based Social Networks*, 2009, pp. 21–24.

[2] M. Arikawa, S. Konomi, and K. Onishi, “Navitime: Supporting pedestrian navigation in the real world,” *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 21–29, 2007.

[3] “National land numerical information downalod service,” http://nlftp.mlit.go.jp/ksj-e/index.html, Nov. 8, 2011.

[4] “Tiger data,” http://www.census.gov/geo/www/tiger/, Nov. 8, 2011.

[5] D. Yamamoto, K. Hukuhara, and N. Takahashi, “A focus control method based on city blocks for the focus+glue+context map,” in *Proceedings of the IEEE 24th Internatinal Conference on Adavanced Information Networking and Applications Workshops*, 2010, pp. 956–961.

[6] I. E. Sutherland, R. F. Sproull, and R. A. Schumacker, “A characterization of ten hidden-surface algorithms,” *ACM Computing Surveys*, no. 1, 1974.

[7] D. Yamamoto, S. Ozeki, and N. Takahashi, “Focus+glue+context: An improved fisheye approach for web map services,” in *Proceedings of the ACM SIGSPATIAL GIS 2009*, 2009, pp. 101–110.

[8] N. Takahashi, “An elastic map system with cognitive map-based operations,” *International Perspectives on Maps and the Internet, Michel P. Peterson (Ed.), Lecture Notes in Geoinformation and Cartography*, pp. 73–87, 2008.

[9] “Focus+glue+context map,” http://tk-www.elcom.nitech.ac.jp/demo/fisheye/, Nov. 8, 2011.

[10] M. Sarkar and M. H. Brown, “Graphical fisheye views of graphs,” in *Proceedings of the CHI 92 conference on Human factors in computing systems*, 1992, pp. 83–91.

[11] M. Sarkar, S. S. Snibbe, O. J. Tversky, and S. P. Reiss, “Stretching the rubber sheet: a metaphor for viewing large layouts on small screens,” in *Proceedings of the 6th annual ACM symposium on User interface software and technology*, 1993, pp. 81–91.

[12] L. Harrie, L. T. Sarjakoski, and L. Lehto, “A variable-scale map for small-display cartography,” in *Proceedings of the Symposium on GeoSpatial Theory, Processing, and Applications*, 2002, pp. 8–12.

[13] C. Gutwin and A. Skopik, “Fisheye views are good for large steering tasks,” in *Proceedings of the CHI 2003 conference on Human factors in computing systems*, 2003, pp. 5–10.

[14] C. Gutwin and C. Fedak, “A comparison of fisheye lenses for interactive layout tasks,” in *Proceedings of the Graphics Interface 2004*, 2004, pp. 213–220.

[15] M. Haklay and P. Weber, “Openstreetmap: User-generated street maps,” *IEEE Pervasive*, vol. 7, no. 4, pp. 12–18, 2008.

[16] Y. T. Zhou, V. Venkateswar, and R. Chellapa, “Edge detection and linear feature extraction using a 2-d random field model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 1, pp. 84–95, 1989.

[17] C. Steger, “An unbiased detector of curvilinear structures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 113–125, 1998.

[18] R. Bajcsy and M. Tavakoli, “Computer recognition of roads from satellite pictures,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-6, no. 9, pp. 623–637, 1976.

[19] W. Shi and C. Zhu, “The line segment match method for extracting road network from high-resolution satellite images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 2, pp. 511–514, 2002.

[20] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka, “Road network extraction and intersection detection from aerial images by tracking road footprints,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4144–4157, 2007.

[21] N. Biggs, *Algebraic Graph Theory 2nd Edition*. Cambridge Mathematical Library, 1994.

[22] V. D. Park and M. S. Corson, “A highly adaptive distributed routing algorithm for mobile wireless networks,” in *Proceedings of the Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, 1997, pp. 1405–1413.

Schema Transformation as a Tool for Data Reuse in Web Service Environment

Lassi Lehto
Department of Geoinformatics and Cartography
Finnish Geodetic Institute
Masala, Finland
lassi.lehto@fgi.fi

***Abstract*—The requirement to support the increased demand for geospatial data resources in several application fields, on various technical platforms and in diverse cultural contexts creates a big challenge for data providers. Reusing the same data set in different environments by employing content transformation processes is proposed as a solution. In this paper online schema transformation is evaluated as a tool for supporting data reuse in Web Service-based data delivery architecture. The case implementation is developed in the context of a major EU-project helping the National Mapping and Cadastral Agencies to achieve INSPIRE-compliance.**

Keywords-schema transformation; data reuse; Web processing; Web Services; ESDI

I. Introduction

A. Background and motivation

One of the most important benefits gained by the introduction of Web Services as the means for delivering digital geographic information is the increased use of the data. The extensive national geodata resources can potentially support many different applications and be used in various technical and cultural contexts, both on national level and abroad. Consequently, data providers are faced with increased need for producing ad hoc extracts from their databases and support a number of individual user preferences in the process. The diverse requirements concerning data encodings, schemas, coordinate reference systems and spatial representations are becoming a great challenge for data providers.

The number of use cases for network-accessible digital geodata has multiplied and the data sets are finding usage in many exotic applications and at the same time are in the core of many vital processes of society [1]. Significant additional requirements are set by the need for international co-operation and the subsequent need for data harmonisation, for instance in the context of the European cooperation [2].

Seamless provision of basic geospatial information is a necessity in the increasingly integrated Europe. Especially the challenges related to the protection of environment emphasise the need for providing access to geospatial information consistently across the European borders. The long history of national independence in Europe has contributed to the fact that each Member State (MS) of the European Union (EU) applies rather individual approaches as it regards collection, organisation and maintenance of geographic information. The challenge of integrating these heterogeneous sources of information as a reliable and consistent information service has been tackled in various Pan-European projects and initiatives, aimed at supporting the development of the European Spatial Data Infrastructure (ESDI).

The ESDI is currently under very active development. Various research programmes of the European Commission (EC) have included actions aimed at building components for this infrastructure. One of the most important drivers in this development is the INSPIRE (Infrastructure for Spatial Information in Europe) initiative [3]. The INSPIRE process is initiated by the EC as a European Directive and aims at seamless Pan-European spatial data provision in support of the protection of the environment [4].

A fundamental principle of the INSPIRE process is to build the European level geospatial content services on top of the existing National SDIs (NSDI), without explicitly requiring changes on the Member State data sets. The diversity of the solutions adopted by the European NSDIs makes this approach particularly challenging. One proposed answer to this challenge is to aim at data consistency through transformations on service level [5][6].

B. Related work

Various standard development activities have focused on the topic of schema transformation. These include the work of the Open Geospatial Consortium (OGC) on the concept of Translating Web Feature Service [7] and the results of the ORCHESTRA project as the specification for service type called schema mapping service [8]. The recent examples of the schema transformation-related standardisation include the INSPIRE specification for the generic Transformation Service, published as EC Regulation [9] and the related technical guidance on Schema Transformation Network Service [10].

The HUMBOLDT project has worked extensively on developing mechanisms for schema transformation. The results of this work include a specification for a Web Service called Conceptual Schema Transformer and the development of an interactive schema mapping tool called HALE (HUMBOLDT Alignment Editor) [11].

The problem of schema transformation in the European-wide spatial data provision framework has been studied by Friis-Christensen et al. [12]. Donaubauer et al. describes a mechanism for supporting schema transformations as an extension to the OGC Web Feature Service (WFS), called model-driven WFS (mdWFS) [13]. Curtis and Müller describe a categorisation of atomic schema transformation components in the context of relational database to Extensible Markup Language (XML) translations [14].

ISBN: 978-1-61208-178-6

Recent work related to schema transformation issues include the paper of Wiemann et al. on schema transformation as a component in the SDI development [15] and the work of Foerster et al. on combining schema transformation with another process, model generalisation, in a Web processing work flow [16].

C. *Structure of the paper*

The work described in this paper has been carried out in the context of a major EU-funded development project ESDIN (European Spatial Data Infrastructure Network), aimed at helping the National Mapping and Cadastral Agencies (NMCAs) to attain compliance with INSPIRE regulations [17].

In Chapter II the concept of schema transformation is elaborated and some further background information given. In Chapter III the alternative architectures for schema transformation in the Web Service environment are described. The Chapter IV focuses on the results achieved in the implementation of the approach in the context of a major EU development project. The paper ends with a conclusion in Chapter V.

II. Schema Transformation Concepts

A. *Transformation Categories*

Content transformations dealing with geospatial data can be considered in various levels of abstraction [18]. Transformations that only consider pure *syntactic harmonisation* of content have a well-established tradition and are routinely performed in the existing and upcoming SDIs. Typical examples of these transformations include the use of Geography Markup Language (GML) as the data encoding standard and the internal processes of the Web Feature Service (WFS) implementations that transform data content from the internal representation of the used data store to the GML-encoded output. Integration of two heterogeneous data stores becomes thus technically possible, but as semantic differences are not considered, the combined data set becomes practically useless.

More advanced transformations take into account the *semantics* of the source data and aim at matching semantically equal or close-to-equal data items to each other. The resulting combined data set is internally consistent and can be accessed using a single semantic framework. In a more advanced approach the semantic correspondence between the data items can be indicated in a more flexible way, facilitating the use of semantic similarity as a measure for controlling the data matching process. Ontologies are sometimes used as a tool for managing flexible semantics-aware transformations [19].

The transformation might focus predominantly on restructuring the data content and applying new naming (typically in a different language) to the constructs that encompass individual data items. This kind of data transformation is called *schema transformation*, as the most typical scenario for this process is transforming spatial data set from one schema (for instance a national model) to another (for instance a common European level model). The transformation is expected to maintain the semantics of the data content as far as possible.

Due to the special nature of geospatial data, various *geometric transformations* are essential. An advanced schema transformation may involve geometric operations as a component. These might involve change of the type of the geometric primitive used to depict the position of a spatial object, for instance transforming a small polygon to a point or an elongated polygon to a linear primitive.

B. *Schema Mapping*

An essential part of the transformation process is the establishment of the mapping between data items in the source schema and the corresponding items in the target schema. This mapping can be defined either on the abstract, conceptual level or on a concrete, implementation-dependent schema level. Some research papers recommend the use of schema mapping on conceptual level [20]. The benefits that this approach yields can be listed as follows:

- The person defining the mapping does not need to be knowledgeable on the implementation level details
- The same mapping, done on the conceptual level, can be subsequently implemented in various technical frameworks
- The mapping made on conceptual level can be expected to remain valid for longer period of time when compared with a mapping based on rather fast-changing technology solutions

The weaknesses of this approach include:

- A separate processing step is required to make the conceptual mapping usable in a concrete implementation environment, often requiring manual intervention
- There is no widely adopted mechanism available for recording the conceptual level schema mapping

The technologies proposed as the method for encoding schema mapping on conceptual level include the OMG (Object Management Group) standard for model transformation MOF 2.0 Query/View/Transformation [21] and Ontology Mapping Language (OML). For instance the Geo Ontology Mapping Language (GOML) specified in the HUMBOLDT project is based on OML [22].

III. Architectural Considerations

There are three different approaches for organizing the schema transformation in relation to the Web service-based data delivery workflow.

Firstly, the transformation can be carried out as a completely offline pre-processing step, typically performed while transferring a data set from the original data store to a service database. This approach is most appropriate in cases when the difference between the native schema and the output schema is so significant that the transformation requires long processing time or even manual interventions. In some cases the original data store might be organized in suboptimal way from the efficient service-based content delivery point of view, thus making the use of a separate service database a necessity. This is often also justified from the security concerns' point of view. However, in some cases it might be possible to maintain the transformed data content inside the same database management system with the source data set, for instance by applying mechanisms like database views and triggers.

Secondly, the transformation can be organized as an online process carried out during the interactive request-response dialogue in the Web Service environment. In this approach the transformation is part of the data delivery workflow and has to be performed as an on-the-fly process. The transformation is carried out only on a subset of the data content, retrieved from the database on the basis of the query sentence of the data request. Being a real-time process the on-the-fly transformation must be straightforward enough to be performed as a fully automatic step in a synchronous transaction.

In some cases the right solution would be to combine the two above-mentioned approaches. A separate transformation process is carried out while transferring data from the production environment to the delivery database, accessible by the online services. This transformation is performed as a pre-processing step outside of the online service environment. In addition to this major batch-oriented transformation, the system might still make use of the on-the-fly transformation approach by also including transformation functionality into the online delivery process. This latter transformation will take care of all the remaining modifications that are still needed to make the delivered data set fully compliant with the target schema. The various approaches for including a transformation into the data delivery workflow are depicted in the Fig. 1.

In an operational service-oriented spatial data delivery environment schema transformation can be organised in various different ways. In the following the main approaches are explained, together with a brief analysis of their strengths and weaknesses. The framework has been initially defined in the context of the service development work of the ESDIN project [23].

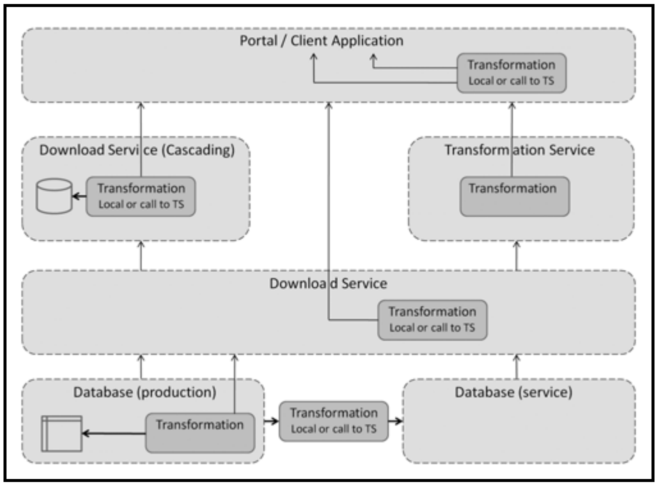


Figure 1. Transformation alternatives. The bold arrows depict offline processes. TS: Transformation Service.

A. *Database-Centric*

On database level simple on-the-fly transformations could be configured as database views. However, in most cases they will not be able to provide the functionality required for achieving full compliancy with the target schema.

In the offline approach various data schemas are maintained inside a single database. Mechanisms available for the transformation include database views, materialized views and tables that are updated by triggers on the source data tables. The transformations might be configured as a set of database scripts (e.g., PL/PGSQL scripts) and can be performed as an integral part of updates to the source data (change propagation).

B. *Service Database-Oriented*

In this approach the data provider decides to set up a separate service database containing data in the target schema. This arrangement can be motivated by the fact that the existing data maintenance platform does not support effective enough data retrieval mechanisms, or does not allow connection with the selected content access service implementation. Security-related concerns might also be the deciding factor.

The transformation from the source schema to the delivery schema is carried out as part of the batch process that transmits data content from the source database to the service database. This process can be based for instance on SQL scripts. Even several separate service databases could be set up, each of them potentially providing content for different services and supporting different schemas. The batch process carrying out the data transmission can be run periodically or it might be triggered for instance by the updates on the source data tables.

C. *Download Service Internal*

One approach for transformations in the currently existing software solutions is to apply transformations as an internal function of the content access service itself. As an example some of the currently available Web Feature Service (WFS) implementations support Coordinate Reference System (CRS) transformations as an integral part of the data access transaction. In a similar way, some WFS products support configurable schema transformation functions to be introduced to the data delivery process, so that the data items of the internal database schema can be mapped to a desired output schema. The WFS implementations that support more extensive transformations are called Translating WFS servers or Transforming WFS servers [24].

D. *Middleware Approaches*

If the transformation is performed somewhere between the source data service and the client application, the approach can be taken as an example of a middleware solution. In this case the transformation gets the input from the data service and provides the transformed data set as an output to be consumed by the user application. Two concrete approaches can be identified for this alternative: a cascading-transforming data service and a dedicated transformation service.

1) Cascading-Transforming Data Service

The concept of cascading services is recognised as a solution for content integration in the geospatial services development community. In this approach a service node works as a client for another service and provides the content of that service as part of its offerings to the real client applications. This way a single WFS node can provide access to resources of various individual WFS services as cascaded content, together with the resources it might be serving from its local data stores. As part of the cascading functionality the WFS can also perform schema transformations.

2) Transformation Service

The transformation could also be performed by a dedicated transformation service. In this approach the transformation service is connected to the source data service for input data and provides the transformed data as an output data set to be consumed by the user application. Unlike in the case of the cascading-transforming data service, the transformation service does not expose itself to the calling applications as content access service, but as a geospatial processing service, enabling the transformation process to be configured by calling applications.

E. Portal-Centric

The predominant role of the portal in an SDI is to integrate distributed services into a single access point. As such a portal is good candidate for also performing other processing tasks, like data transformations. A portal might contain such transforming process as an internal function or rely for the transformation on an outside resource like a transformation service.

IV. Case Implementation

The ESDIN (European Spatial Data Infrastructure Network) project was a Best Practice Network project in the EU eContentplus framework program [17]. The main goal of the ESDIN project was to help the European National Mapping and Cadastral Agencies (NMCAs) in their efforts to fulfil the requirements set by the INSPIRE process. The project Consortium membership included NMCAs from 11 different European countries and it was coordinated by EuroGeographics [25]. The ESDIN project was finished in March 2011.

As one of its main goals the project focused on a coordinated establishment of INSPIRE-compliant Download Services. The services are built according to the service category ‘Direct Access Download Service’, thus implementing the OGC Web Feature Service (WFS) interface. The output data are provided in INSPIRE schemas, encoded in GML version 3.2.1 according to the requirements set in INSPIRE.

One of the primary areas of investigation in the project was to find out, how schema transformations from the national data models to the INSPIRE schemas could be carried out. This included tests both on off-line and on-the-fly transformation approaches. In the final stage, roughly half of the developed services follow at least partially the on-the-fly schema transformation approach, whereas the other half rely on offline processes [26].

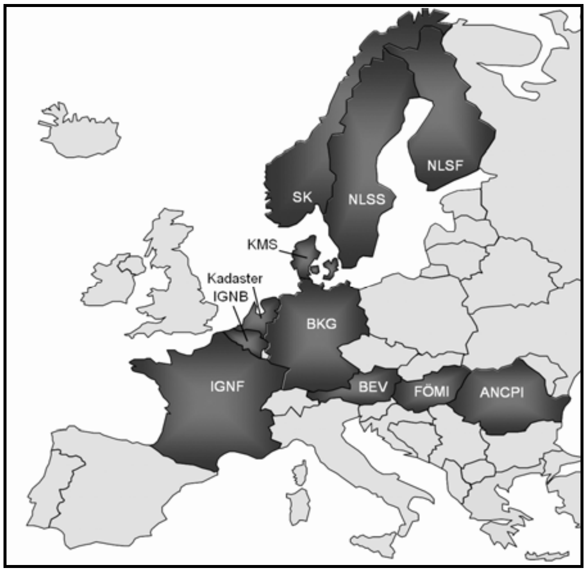


Figure 2. ESDIN countries providing content for the Download Services. Acronyms refer to the NMCA of the country.

As the main result of the service development efforts of the project, roughly 50 data sets are now available as INSPIRE-compliant Download Services. The offering cover five different data themes from the INSPIRE Annex I and has either complete national coverage or is restricted to selected cross-border test areas. The countries providing content for the ESDIN services are depicted in Fig. 2.

In the ESDIN context the proposed approach for schema mapping is to apply an SQL-like pseudo code for the purpose. The proposal includes a method for recording the mapping in a spreadsheet [27].

In the ESDIN project the NMCAs tested various different approaches for schema transformation. The architectural alternatives tested in the project are depicted with numbers 1-4 in the Fig. 3. The four cases are briefly described in the following.

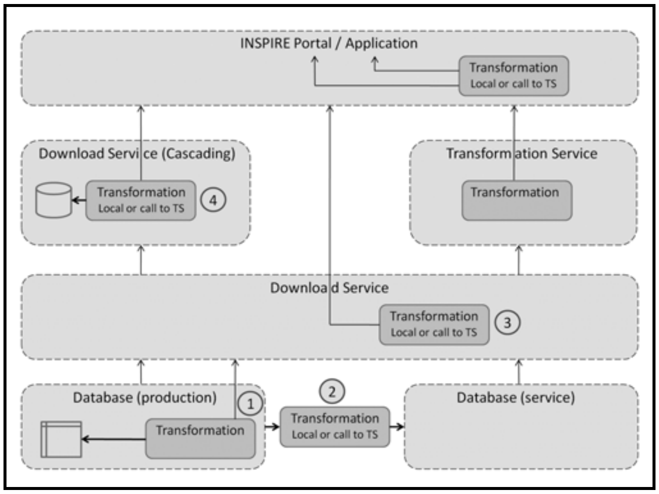


Figure 3. The four transformation architecture alternatives used in the ESDIN project. Numbers refer to cases explained below.

ISBN: 978-1-61208-178-6

Case 1. On-the-fly transformation by database views. In this approach database views are created to transform some aspects of the data to conform to INSPIRE requirements. This approach was implemented by one NMCA.

Case 2. Offline transformation. The process steps include: extract data from the production database, transform it to INSPIRE compliant form, upload it to the service database. This approach follows the general Extract/Transform/Load (ETL) processing model. The transformation tools used include database scripting, XSLT scripting and commercial tools like FME [28][29]. This method was tested by eight NMCAs.

Case 3. Internal on-the-fly transformation carried out by Download Service. In this approach the used Download Service implementation provides support for schema transformations. These can be configured for instance by annotated XML Schema files, XSLT declarations, or by implementation-specific mechanisms. The tools used include deegree and XtraServer [30][31]. This transformation approach was implemented by eight NMCAs.

Case 4. On-the-fly transformation carried out by a cascading-transforming Download Service node. In this alternative the original Download Service provides data in local schema. This service is accessed by another Download Service that uses the first service as its data source. During the data request the cascading-transforming Download Service carries out schema transformations, typically both on query and data streams. This approach was tested by two NMCAs.

V. Conclusion

The experiences gained in the ESDIN service development suggest that many of the transformation types required for INSPIRE compliancy are too complicated or time consuming for the on-the-fly processing approach. Roughly half (12) of the developed services (22) are (at least partially) based on on-the-fly transformation principles. Only eight services are based on pure on-the-fly transformation approach. As a result it can be concluded that on-the-fly transformations did not take as central a role in the project as initially expected. On the other hand, the combined approach, in which both offline and on-the-fly transformations are used together, seems to be rather promising [26]. The final service configuration of the ESDIN project is illustrated in Fig. 4.

A database level solution was raised during the project experiments as a new kind of alternative for content transformations. The concrete implementations developed in the project are based on the use of database views and thus represent on-the-fly approach on schema transformation on the database level.

The experience gained in the ESDIN project suggests that various different approaches for content transformation can be adopted with successful results. It seems that general recommendation cannot be given on any single transformation method, as the best approach always depends on local conditions. One of the most important deciding factors is, how far the national schema is from the related INSPIRE schema. The bigger the difference, the less probable it is that on-the-fly solutions would yield acceptable results.

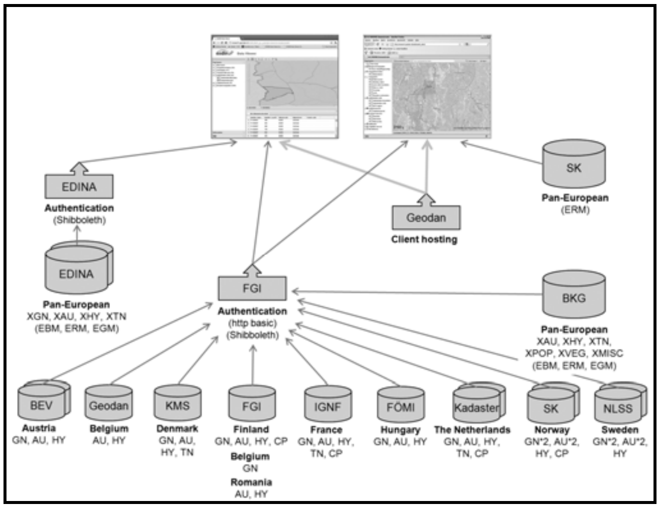


Figure 4. Final service configuration of the ESDIN project. Acronyms below the country names indicate the INSPIRE themes supported. Themes starting with ‘X’ denote schemas developed in the project as extensions to the INSPIRE schemas.

A standardised way to describe the schema transformation is a useful tool for communication among the individuals involved in the schema mapping process. The mechanisms proposed by the ESDIN project have been found to be helpful in this respect.

The number of transformation steps has to be kept to minimum, because each transformation process must be first developed, and further on maintained too. Every transformation also potentially introduces a certain amount of error or information loss to the process.

In most of the cases the off-line approach for schema transformation seems to provide best results. On-the-fly methods can be applied for fine-tuning minor schema nuances or in cases where the differences between the local schema and the corresponding INSPIRE schema are minor. The most prominent advantage in using the on-the-fly approach is the direct connection to up-to-date content.

The main disadvantage in the offline transformation method is the need to have a separate service database and the inherent problem of keeping it up-to-date. On the other hand, this approach also supports easy fusion of the data with other source data sets. Security concerns can also be better taken care of.

The task of setting up an INSPIRE-compliant Download Service is not straightforward. Reasonable amount of resources have to be invested in the process. The main issue to resolve is the transformation of content stored in the national schema to be compliant with INSPIRE schemas. This was successfully tackled in the ESDIN project with solutions developed for both offline and on-the-fly approaches.

Acknowledgement

The ESDIN project was funded by European Commission in the Framework Programme eContentplus (grant agreement number ECP-2007-GEO-317008).

REFERENCES

[1] A. Friis-Christensen et al., "Building service oriented application on top of a spatial data infrastructure – A forest fire assessment example," Proc. AGILE International Conference on Geographic Information Science, 2006, Visegrad, pp. 119–127, Hungary.

[2] L. Lehto, "Real-time content transformations in the European spatial data infrastructure," Proc. International Cartographic Conference, Nov 15-21, 2009, Santiago, Chile, CD-ROM.

[3] JRC, Web site of the INSPIRE process, http://inspire.jrc.ec.europa.eu <retrieved: 11, 2011>

[4] EC, DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), at: http://eur-lex. europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML <retrieved: 11, 2011>

[5] L. Lehto, "Schema transformation as a Web service, Proc. Nordic GIS Conference, Helsinki, Oct 2-4, 2006, pp. 39-44.

[6] L. Lehto, "Schema translations in a web service-based SDI," Proc. AGILE International Conference on Geographic Information Science, 'The European Information Society: Leading the way with geo-information', May 8-11, 2007, Aalborg, Denmark..

[7] Y. Bishr and K. Buehler, "GOS Transportation Portal, Implementation Architecture and Lessons Learned," Internal OGC Interoperability Program Report.

[8] M. Lutz, "Specification of the schema mapping service," OA services specifications, Orchestra project, Deliverable D3.4.3, Joint Research Centre of the EU Commission.

[9] EC, "Commission Regulation (EU) No 1088/2010 of 23 November 2010, amending Commission Regulation (EC) No 976/2009 as regards download services and transformation services," at: http://eur-lex.europa.eu/LexUriServ/ LexUriServ.do?uri=CONSLEG:2009R0976:20101228:EN:PDF <retrieved: 11, 2011>

[10] INSPIRE, "Technical Guidance for the INSPIRE Schema Transformation Network Service," at: http://inspire.jrc.ec.europa.eu/documents/Network_Services/JRC_INSPIRE-TransformService_TG_v3-0.pdf <retrieved: 11, 2011>

[11] T. Reitz, M. de Vries, and D Fitzner, "Conceptual Schema Specification and Mapping", Deliverable A7.0-D3, HUMBOLDT project, at: http://www.esdi-humboldt.eu/press/public_deliverables.html <retrieved: 11, 2011>

[12] A. Friis-Christensen, S. Schade, and S. Peedell, "Approaches to solve schema heterogeneity at the European level," Proc. EC-GI & GIS Workshop, Alghero, Sardinia, Jun 29 – Jul 1, 2005.

[13] A. Donaubauer, F. Straub, and M. Schilcher, "mdWFS: a concept of webenabling semantic transformation," Proc. AGILE International Conference on Geographic Information Science, May 8-11, 2007, Aalborg, Denmark.

[14] E. Curtis and H. Müller, "Schema translation in practice," Snowflake Software White Paper, at: http://www.snowflakesoftware.co.uk/news/papers.htm <retrieved: 11, 2011>

[15] S. Wiemann et al., "Web services for spatial data transformation and exchanges in SDI: a prototypical implementation of the LPIS quality assurance test bed services," International Journal of Spatial Data Infrastructure Research, unpublished.

[16] T. Foerster, L. Lehto, T. Sarjakoski, L. T. Sarjakoski, and J. Stoter, 2010. "Map generalization and schema transformation of geospatial data combined in a web service context," Computers, Environment and Urban Systems, 34(1): 79-88, at: http://dx.doi.org/10.1016/j.compenvurbsys.2009.06.003. <retrieved: 11, 2011>

[17] ESDIN, European Spatial Data Infrastructure Network, project Web site, at: http://www.esdin.eu <retrieved: 11, 2011>

[18] L. Lehto, "Real-Time Content Transformations in a Web Service-Based Delivery Architecture for Geographic Information," Doctoral dissertation, Helsinki University of Technology. Publications of the Finnish Geodetic Institute, N:o 138, 2007, Kirkkonummi, 51 p. + 99 p.

[19] F. Fonseca, G. Câmara, and A. M. Monteiro, "A Framework for Measuring the Interoperability of Geo-Ontologies," Spatial Cognition & Computation, Vol. 6, No. 4, 2006.

[20] H. R. Gnägi, A. Morf, and P. Staub, "Semantic Interoperability through the Definition of Conceptual Model Transformations," Proc. AGILE Conference on Geographic Information Science, 2006, Visegrád.

[21] OMG, Meta Object Facility (MOF) Query/View/Transformation specification (QVT) http://www.omg.org/ spec/QVT/ <retrieved: 11, 2011>

[22] HUMBOLDT, Web site of the Humboldt project, at http://www.esdi-humboldt.eu/home.html <retrieved: 11, 2011>

[23] L. Lehto (ed.), "Best practice for content transformations enabling INSPIRE-compliant data delivery," Deliverable 11.1, Public report, ESDIN project, European Spatial Data Infrastructure Network, Oct 9, 2009, ECP-2007-GEO-317008, 44 p.

[24] E. Curtis, "Schema Translation and Semantics in Data Interoperability," Snowflake Software White Paper, at: http://www.snowflakesoftware.co.uk/news/papers.htm <retrieved: 11, 2011>

[25] EuroGeographics, Association of National Mapping, Land Registry and Cadastral Agencies, Web site at: http://www.eurogeographics.org <retrieved: 11, 2011>

[26] L. Lehto (ed.), "Recommendations for operational deployment of services," Deliverable 11.5, Public report, ESDIN, European Spatial Data Infrastructure Network, Feb 28, 2011, ECP-2007-GEO-317008, 67 p.

[27] L. Lehto and F. Nissen, "Schema transformations in the European spatial data infrastructure," Proc. INSPIRE Conference, 22-25 June, 2010, Krakow, Poland. Poster.

[28] W3C, Extensible Stylesheet Language Transformations (XSLT), at: http://www.w3.org/TR/1999/REC-xslt-19991116 <retrieved: 11, 2011>

[29] Safe Software, The FME Techonology, at: http://www.safe.com/fme/fme-technology/ <retrieved: 11, 2011>

[30] deegree, the web site of the deegree community, at: http://www.deegree.org <retrieved: 11, 2011>

[31] interactive instruments, XtraServer product information (in German language), at: http://www.interactive-instruments.de/index.php?id=xtraserver&L=1 <retrieved: 11, 2011>

Bus Coming: A Service for Tracking Buses in Rural Areas based on Passenger Locations

Arthur Thompson
Department of Computing and Information Technology
University of the West Indies
St. Augustine, Trinidad
arthur.thompson.jr@gmail.com

Wayne Goodridge
Department of Computing and Information Technology
University of the West Indies
St. Augustine, Trinidad
wayne.goodridge@sta.uwi.edu

***Abstract*— The massive influx of programmable smart phones with built in Global Positioning System receivers, provides an opportunity to make the current public transport bus system in Trinidad and Tobago more reliable by providing passengers with information that would allow them to make better travel decisions. A four part system is developed which captures and stores near real time bus location data and provides near real time bus location information visually via a web application and textually via a mobile application. The core communication technology behind the system is based on Web Services and early results demonstrate that this simple approach is very appropriate for rural areas in Trinidad.**

Keywords - ***GPS; smart phones; public transport bus system; webservices***

Transport in the West Indies is difficult for many, especially for those that live in rural areas. Many people depend on Government provided buses to get to and from their homes daily. This bus service, however, is not reliable. Buses are scheduled to run at certain times of day but many issues can and usually do account for delays that then cascade down to other scheduled times causing predefined schedules to be useless. Routes to rural areas are assigned few buses and thus if there is a problem with a rural route bus, there are not many buses that can pick up the slack. Currently, there is no way for an individual in a rural area to access the bus transportation options available to him/her at any given time.

This lack of information can lead to long waits or worst waiting in vain for a bus that never arrives. If passengers were provided with near real time information about the location and status of buses servicing their areas then these issues would be alleviated, improving the passenger's travel experience and allowing passengers to make better more efficient use of their time.

Over the past two to three years, BlackBerry smart phones have been introduced to the West Indies on a large scale. Research In Motion (RIM) [1, 2] the company that manufactures BlackBerry smart phones also provides a developer platform that allows application developers to extend the functionality of the BlackBerry smart phone. This platform provides developers with access to build-in devices such as Global Positioning System (GPS)[3] receivers allowing them to utilize the functionality of these built-in devices in the applications that they develop.

The Bus Coming system proposed in this paper uses the BlackBerry developer platform with the built-in GPS receiver in the BlackBerry smart phone to build a relatively cheap bus tracking system. The system will provide passengers with near real time bus location and status information with respect to the passenger's current location, not bus stops as is usually done. This will improve the travel experience of passengers in rural areas (where there are few bus stops) and allow them to make better and more efficient use of their time.

This paper is organized as follows: Section I gives motivation for the approach taken for development the Bus Tracking system which is based on both GPS mobile devices and the mobile cell phone network; Section II briefly describes the components of the Bus Coming System and how accuracy across multiple platforms is achieved.

I. LACK OF BUS TRACKING

Bus tracking services have recently (within the last 2 years) been introduced in some metropolitan areas. Chicago [4], Washington DC [5] and London [6] are a few cities that have implemented some form of advanced bus tracking services. These services are usually comprised of a GPS equipped fleet of buses and these locations are feed to a central server which then makes the locations available to different readers, web sites (basic lists or graphical representations like map overlays), electronic signs at bus stops, SMS for bus locations at particular bus stops and even mobile applications for smart phones that gives the location of buses on a particular route with respect to bus stops.

In the West Indies, however, electronic bus tracking systems are not yet available. Territories like Jamaica [7], Barbados [8] and Trinidad and Tobago [9] have transport authorities who go as far as providing semi-static scheduling

ISBN: 978-1-61208-178-6

via hard copy or rarely updated websites. These schedules are seldom exact and thus cannot be relied upon. Especially in rural areas, where there is no assistance or information available from transport authority about bus availability or location thus it is difficult to access bus transport in an efficient manner.

In metropolitan cities like Chicago, Washington DC and London; buses strictly stop at predefined bus stops, because of this the bus tracking systems implemented in these cities are focused on bus locations with respect to bus stops. In the West Indies, especially in rural areas this approach will not work, designated bus stops are few and far between. It would not be practical to expect every person living in these areas to walk to the stop nearest them, as the nearest one can be kilometres away. Due to the scarcity and distance between bus stops, buses do not strictly stop at pre-defined bus stops, in many cases a bus driver will stop for passengers at multiple non bus stop locations along the route. The amount of time for the bus to arrive at the next stop in these areas is much less useful than the time for the bus to arrive at the passenger's location along the route. Therefore, building a bus tracking system based on each passenger's current location would be very useful for rural areas.

II. PROPOSED SOLUTION

Internet access is key component of this system, Trinidad and Tobago like many other countries in the West Indies is seeing consistent positive growth in both fixed and mobile internet access. In Trinidad and Tobago, Q1 2011 almost 50% of households have internet access, with 96% of those having broadband access, this is up 11% over the previous year. Also 36% of the population of Trinidad and Tobago have mobile internet access in Q1 2011 up 9% over the previous year. [10] With mobile internet coverage close to 100% in Trinidad and Tobago and general internet usage trending upwards, an internet based solution to the bus tracking problem in the West Indies is very feasible.

A system is needed that takes the passenger's location into consideration. In order to give information about the bus's location with respect to the passenger's location along the route, the passenger's current location along the route is required. The passenger's location along the route would be treated in much the same way as a bus stop would be, without having to erect an unreasonable number of stops along the route. With the passenger's location along the route, the system can tell the passenger the amount of time to the next bus or buses with respect the passenger's current location.

GPS capable smart phones were used to provide near real time bus location information by placing a GPS capable smart phone on a bus and the phone would submit its current GPS location periodically.

A. Reference Points

Google Maps were used to provide a map with near real time bus location overlays. However when the GPS locations generated by the mobile devices were plotted directly onto the Google Map of Trinidad the results were inconsistent. Reference Points were used to help associate the GPS locations generated by the mobile devices with visually consistent positions on the Google Map.

Reference Points are positions along each route that visually correspond to locations on a Google Map of Trinidad. The GPS locations generated by the mobile devices are associated with the reference point nearest it. This allows the system to be consistent with the points that are used to perform calculations like distance to passenger and estimations like time to passenger; while still being able to display visually consistent locations of buses on a Google Map.

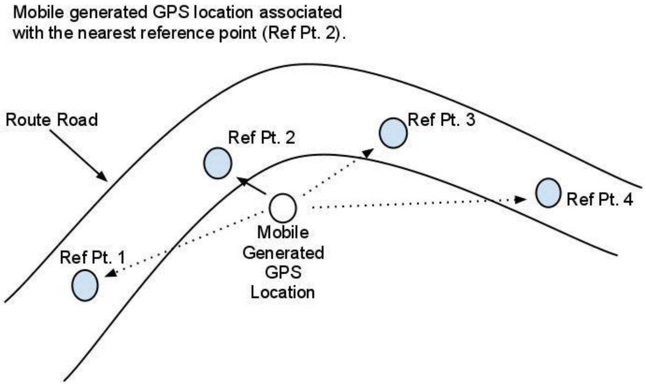

Mobile generated GPS location associated with the nearest reference point (Ref Pt. 2).
Route Road
Ref Pt. 2
Ref Pt. 3
Ref Pt. 4
Ref Pt. 1
Mobile Generated GPS Location

Fig. 1 Mobile GPS location associated with Nearest Reference Point

Reference points provided a solution to the inconsistencies between mobile generated GPS locations and positions on a Google Map. They also provided two benefits as a side effect.

1) Easy calculation of distance between different positions along a route, since a straight line could be drawn from each reference point to the next without going off route. This provided a way of easily calculating distances along winding routes which are common in rural areas.

2) Easily calculate the availability status (coming or gone) of a bus with respect to a passenger along the route. If a passenger is associated with Ref. Pt. 2 and a bus is associated with Ref. Pt. 3, since 3 is greater than 2 we can let the passenger know that that bus has already gone.

The skills and technology required to produce a system that provides both passengers and bus operators with near real time bus location information can be built using readily available and relatively cheap technology involving programmable GPS devices (BlackBerry smart phones in this case, however any programmable GPS capable device would do, eg: Android, IOS, Symbian devices). The availability of fairly cheap programmable GPS enabled BlackBerry smart phones makes a bus tracking system implementable at a relatively cheap cost and in a relatively short period of time with minimal disruption to the current bus operations.

B. *Components of the BusComing System*

The proposed Bus Coming system consists of four components:

1) *Bus Coming Web Service*

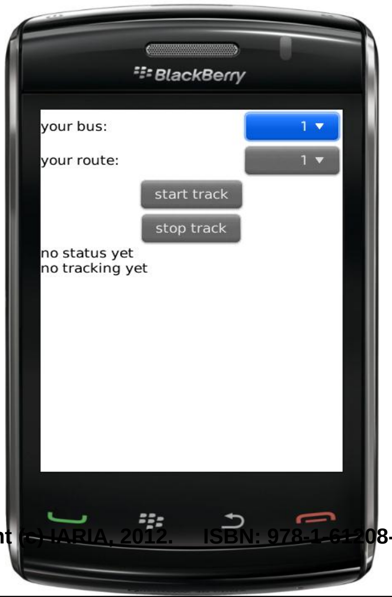
This is a web service [11] that allows all other parts of Bus Coming System to submit information to the server (for example the Bus Tracker component submits information to this web service) or request information from the server (for example the Bus Coming Web and Mobile Views).

Standardised web services were used to provide the ability for multiple platforms to work within and around the system without any major changes. The BlackBerry smart phones used could easily be replaced by one or more of many other web service compliant GPS capable devices.

Parts of the Bus Coming Web Service could be made public, allowing third party developers to use the data in interesting ways.

2) *Bus Tracker*

This is an application running on a GPS capable blackberry, the blackberry is placed on a desired bus to track its location. Figure 1, shows the Bus Tracker application and the interface where the bus driver selects the bus and route. Note that the tracker transmits the bus current location of the bus to the server via a web service periodically.



your bus: 1
your route: 1
no status yet
no tracking yet

Operators did express concern about the bus driver having to initiate the tracking along a journey, so considerations are being made to have the tracking remotely initiated by central operators.

Fig. 2 Bus Tracker Application

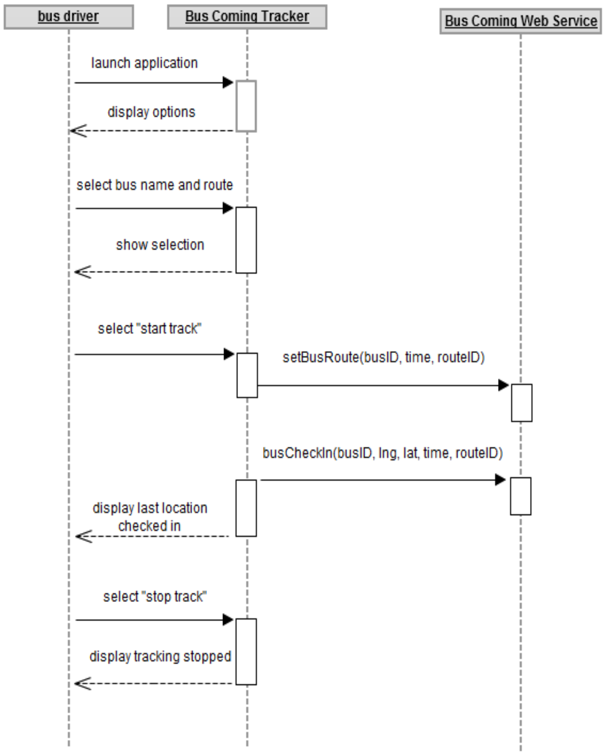


Fig. 3 Bus Tracker Sequence Diagram

Bus Coming Tracker must be given some information before tracking can begin. Figure 2, shows a sequence diagram outlining the steps involved. The user that initiates the tracking (the bus driver) would set the bus name and route before the bus begins to move. Once the bus name and route are set and the bus driver is ready to begin the journey, the driver selects the “Start Track” button and then tracking begins. Bus Coming Tracker lets the Bus Coming Web Service know that it is starting to track by using the ***setBusRoute*** service, which sets the bus’s current route and the time at which tracking has begun.

Once tracking has been initiated, Bus Coming Tracker frequently submits, every five seconds or so, the location of the BlackBerry device and thus the location of the bus, via its built in GPS receiver to Bus Coming Web Service via the ***busCheckIn*** service. These frequent submissions are used by Bus Coming Web Service to provide Bus Coming Web View and Bus Coming Mobile View with information about the bus’s location.

ISBN: 978-1-61208-178-6

3) *Bus Coming Web View*

This is a web view of all routes in the system and all buses on those routes. Figure 3 shows a Google Map web page of this component which is typically used by passengers using a home computer or can be used by the central control station of the bus service department. This web view is interactive and allows the user to view routes of interest. In addition, the site is updated in near real time, periodic AJAX[12] calls are made to the web service to get new data and update the web view accordingly showing the position of buses as received by Bus Tracker.

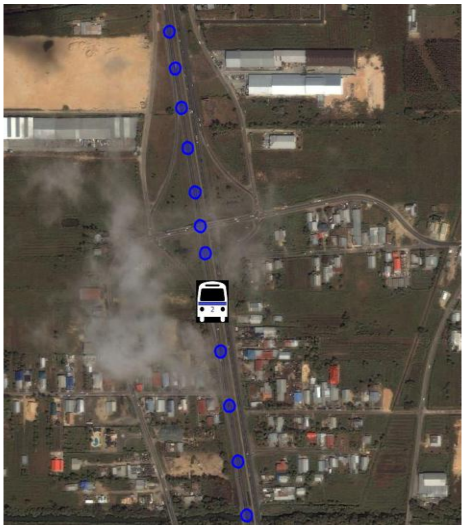


Fig. 4 Google Map showing Bus Locations

Google provides API access to their Google Maps system. Through this API developers can include Google Maps in their web applications. Overlays are added onto the Google Map to show a map with customized information on it, for example reference points along a bus route. Figure 4 shows a sequence diagram which illustrates how the ***getRefPoints*** capability of the web service is used by the web viewer to retrieve latitude and longitude of each reference point along a particular route.

Using the ***getBus*** capability of the Web Service, information containing the latitude, longitude and area of each bus currently on a particular route is retrieved. Google Maps is then instructed to display the bus.

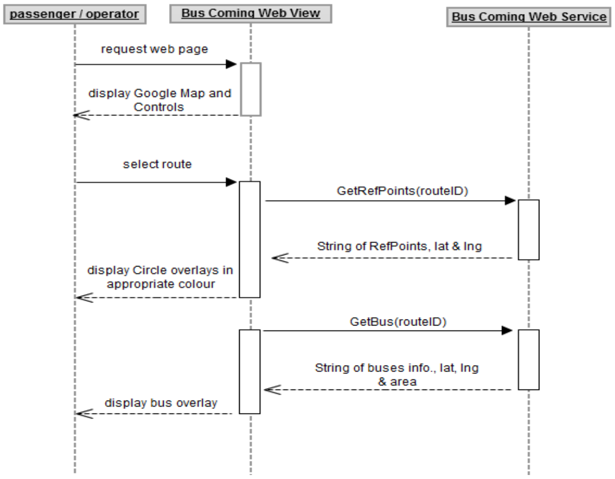
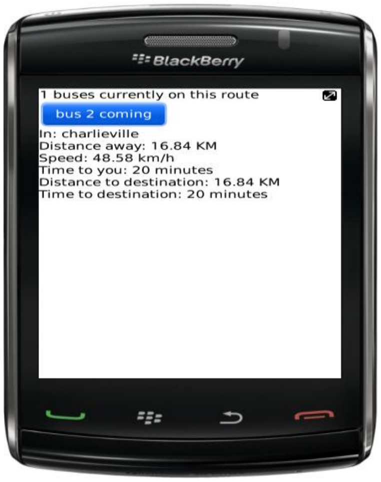

passenger / operator
Bus Coming Web View
Bus Coming Web Service
request web page
display Google Map and Controls
select route
GetRefPoints(routeID)
String of RefPoints, lat & lng
display Circle overlays in appropriate colour
GetBus(routeID)
String of buses info., lat, lng & area
display bus overlay

Fig. 5 Bus Coming Web View Sequence Diagram

4) *Bus Coming Mobile View*

This is an application running a GPS capable blackberry, the application allows the user to select a route and view all buses on the route and see whether or not those buses are coming towards them or if they have already gone past. Figure 5, illustrates that the bus is 20 minutes away from the position of the passenger who requested the service.

From the list of routes the passenger can select the route that the passenger is on. Once the route selection has taken place Bus Coming Mobile View will get the passengers location from the GPS receiver, this location along with the ID of the selected route is submitted to Web Service using a call to ***getBusesOnRoute***. The Bus Coming Web Service then returns a string array containing information about all the buses currently servicing the selected route. This information is initially displayed to the passenger as a button list with the name of the bus and its coming or gone status on the button. If one of the buttons representing a bus is selected more detailed information about the bus will be presented to the passenger with respect to the passenger’s current location. This detailed information includes the current area of the bus, the bus’s distance away from the passenger, the bus’s current speed, the estimated amount of time it will take the bus to get to the passenger, the distance of the bus from its final destination and the estimated amount of time it will take the bus to get to its final destination.



1 buses currently on this route

In: charlieville
Distance away: 16.84 KM
Speed: 48.58 km/h
Time to you: 20 minutes
Distance to destination: 16.84 KM
Time to destination: 20 minutes

Fig. 6. Mobile View for Passengers

The location of the passenger that is calculated by the device’s GPS receiver and submitted to Bus Coming Web Service, is, as with buses, associated with the closest reference point along the route. All calculations with respect to the user are really taking place with respect to a reference point along a selected route. As with buses this is done to account for any inaccuracy of the GPS receiver. If a passenger lives down a street of the route and uses Bus Coming Mobile View, the passenger will get information as if the passenger was standing on the route at the top of the passenger’s street.

III. DISCUSSION

During development and initial testing only two mobile phones were available. The results of this limited testing demonstrate that the Bus Coming System was able to meet all objectives set out. However, it is desirable that a larger scale test be performed.

Currently, requests are being made to mobile operators to supply GPS enabled mobile units to perform larger scale testing of the system. Permission has already been granted by the local Bus Authority of Trinidad and Tobago to perform tests once mobile units are acquired.

REFERENCES

[1] RIM GPS capable BlackBerry smart phone list 2011 http://us.blackberry.com/smartphones/features/gps.jsp

[2] RIM how to deploy on BlackBerry 2011 http://devblog.blackberry.com/2010/06/how-can-i-deploy-my-blackberry-widget/

[3] Global Positioning System 2011, http://www.gps.gov

[4] Chicago Transit Authority 2011 http://www.ctabustracker.com/bustime/eta/eta.jsp

[5] Washington Metropolitan Area Transit Authority 2011 http://www.wmata.com/rider_tools/nextbus/about_nextbus.cfm

[6] Transport For London 2011 http://www.tfl.gov.uk/corporate/projectsandschemes/11560.aspx

[7] Jamaican Urban Transport Company Limited http://www.jutc.com/timetables.php

[8] Barbados Transport Board 2011 http://www.transportboard.com/schedule.php

[9] Public Transport Service Commission, Trinidad and Tobago 2011 http://www.ptsc.co.tt/

[10] http://www.tatt.org.tt/LinkClick.aspx?fileticket=3ORrNw9rKbo%3D&tabid=120

[11] World Wide Web Consortium 2011 Web Services http://www.w3.org/TR/ws-arch/#whatis

[12] http://www.w3.org/standards/webdesign/script.html

Live Cities and Urban Services – A Multi-dimensional Stress Field between Technology, Innovation and Society

Bernd Resch
SENSEable City Lab
MIT
Cambridge, US
berno@mit.edu
IGF
University of Osnabrück,
Germany

Alexander Zipf
Inst. of Geography
University of Heidelberg, GER
alexander.zipf@geog.uni-heidelberg.de

Philipp Breuss-Schneeweis
Wikitude GmbH
Salzburg, AUT
philipp.breuss@wikitude.com

Euro Beinat
Dept. of Geography
University of Salzburg, AUT
euro.beinat@sbg.ac.at

Marc Boher
Urbiotica
Barcelona, ESP
marc.boher@urbiotica.com

Abstract **– Contrary to projections, which stated that the wide-spread distribution of high-speed Internet connections would render geographical distance irrelevant, cities have recently become the centre of interest in academic research. However, especially real-time monitoring of urban processes is widely unexplored. We present the concept of a *Live City*, in which the city is regarded as an actuated near real-time control system creating a feedback loop between the citizens, environmental monitoring systems, the city management and ubiquitous information services. Basically, there are four main barriers towards the implementation of the *Live City*: methodological issues, technical/technological problems, privacy and legislative questions, and quantification of economic opportunities. In this paper, we discuss those challenges and point out potential future research pathways towards the realisation of a *Live City*.**

Keywords-live city, pervasive sensor networks, urban services, real-time information services.

I. INTRODUCTION

Projections stated that the wide-spread distribution of high-speed Internet connections will render geographical distance irrelevant [1], and that cities are not more than mere artefacts of the industrial age [2]. As a side effect, cities were presumed to drastically decrease in importance as physical and social connections, and would play an increasingly ancillary role in socio-technical research.

In reality, the world developed completely differently – cities are back in the focus of academic research. Cities in their multi-layered complexity in terms of social interactions, living space provision, infrastructure development and other crucial human factors of everyday life have re-gained importance in scientific research. This arises from the fact – amongst others – that major developments of scientific and technological innovation took place in the urban context [3],[4].

However in research on urban areas, particularly real-time monitoring of urban processes and digital services are still widely unexplored. These research fields have recently received a lot of attention due to the fast rise of inexpensive pervasive sensor technologies which made ubiquitous sensing feasible and enriches research on cities with uncharted up-to-date information layers through connecting the physical to the virtual world, as shown in Fig. 1.

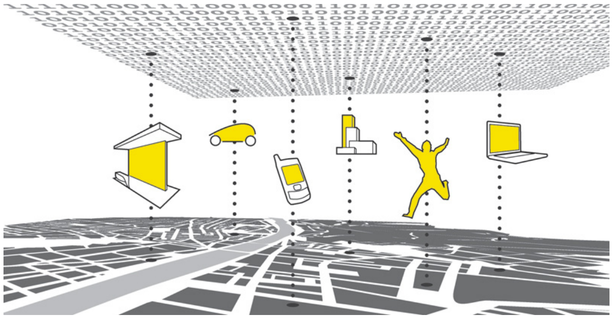


Figure 1. *Live City* – Connecting Physical and Virtual Worlds. [5]

One driver towards this vision is the diminishing digital divide on a global scale. While the digital divide within countries is still strongly affecting the degree of access to information and knowledge, the global digital divide is decreasing due to the fast rise of ICT markets in China, India, South-East Asia, South America and Africa. Mobile phone penetration (i.e. “mobile subscribers per 100 inhabitants”) has been at 76.2% of the world’s population in 2010. This rate is at 94.1% in the Americas and at 131.5% in CIS (Commonwealth of Independent States) [6]. The two fastest growing mobile phone markets China and India currently face a penetration rate of 64% and 70%.

This growth builds the basis for the installation of urban real-time services. In a recent report on Digital Urban Renewal [7], the author states that major demand-side drivers for digital urban projects are the increasing focus on sustainability and emissions reduction, continued pressure on the urban transport infrastructure, and increasing need for citizen services, amongst others. On the supply side, several drivers have been identified including the evolution of the Internet as an underlying framework for services, sensor networks, connectivity technologies, and augmented reality.

However, we are still facing a lack of experience in assessing urban dynamics in real time. One reason is that continuous monitoring is an enormous challenge, and this is particularly true in the urban context, which poses very specific challenges. These comprise technological questions, but also significant economical, societal and political ones, which are rapidly gaining importance.

ISBN: 978-1-61208-178-6

Generally speaking, we are experiencing fast progressing technology development, which is not only moving ahead quickly, but which is moving ahead of society. This development can be compared with a stream moving at high speed, on which we are paddling to remain on the same spot or at least not to drift off too fast. The question, which we have to tackle in this regard, is where our goal for the future lies: down-stream, somewhere near our current spot, or even up-stream?

In this paper, we try to illustrate possible pathways to answering this multi-dimensional question. We incorporate societal, technical, political, privacy and economic issues into our rationale. We are well aware of shortcomings in terms of completeness and technical thoroughness. The paper shall be considered a first leap towards a *Live City* 'Installation Guide'.

This paper is organised as follows: after this introduction we illustrate a few examples on existing approaches towards *Live Cites* in Section II before giving a disambiguation of the term 'live' in Section III. Section IV discusses challenges in current research on the *Live City* and Section V illustrates potential future research avenues, before Section VI summarises conclusions from the paper.

II. STATE-OF-THE-ART – LIVE CITY IN ACTION

One of the first implementations of a 'real-time city' has been done by the MIT SENSEable City Lab [5]. This research group has considerably coined the term '**real-time city**', particularly through visualising the city as a real-time and pulsating entity. In further research initiatives, the SENSEable City Lab investigated human mobility patterns, usage of pervasive sensors to assess urban dynamics, event-based anomaly detection in ICT networks, and correlations between ICT usage and socio-cultural developments.

A new and innovative idea for assessing urban dynamics in real time is the concept of **Living Labs**. According to [8], a Living Lab is a real-life experimentation environment where users and producers co-create innovations. Living Labs are promoted by the European Commission, which characterises them as Public-Private-People Partnerships (PPPP) for user-driven open innovation. A Living Lab is basically composed of four main components: co-creation (co-design by users and producers), exploration (discovering emerging usages, behaviours and market opportunities), experimentation (implementing live scenarios) and evaluation (assessment of concepts, products and services).

Also, much research is performed in the area of **smart cities** (in particular in South Korea also the term 'ubiquitous cities' is popular). For instance, IBM has implemented a number of urban services in the course of their 'Smarter Planet' programme [9]. Research is performed together with cities all over the world to implement applications in the areas of city management, citizen services, business opportunities, transport, water supply, communication and energy. The goal is to seize opportunities and build sustainable prosperity, by making cities 'smarter'.

A sensor-driven approach to ubiquitous urban monitoring is presented in [10] and in [11]. The authors present a measurement infrastructure for **pervasive monitoring** applications using ubiquitous embedded sensing technologies with a focus on urban applications. The system has been conceived in such a modular way that the base platform can be used within a variety of sensor web application fields such as environmental monitoring, biometric parameter surveillance, critical infrastructure protection or energy network observation. Several show cases have been implemented and validated in the areas of urban air quality monitoring, public health, radiation safety, and exposure modelling.

III. A DISAMBIGUATION OF THE TERM 'LIVE'

The term '*Live City*' originates from the modification of the expression 'Real-time City' as definitions and usages of the latter are vague and vary on a quite broad scale.

Anthony Townsend presents a very mobile phone centric definition of a real-time city by stating that 'the cellular telephone […] will undoubtedly lead to fundamental transformations in individuals' perceptions of self and the world, and consequently the way they collectively construct that world' [12]. The author sees the real-time city as a potential platform for dedicated advertising and states that 'accessibility becomes more important than mobility'. This implies that it will be more critical to access urban services rather than moving around physically. This in turn means that the digital (mobile phone) infrastructure will be more important than the physical (transport) infrastructure.

A possible definition of *urban informatics* – a term closely related the real-time city – is 'the collection, classification, storage, retrieval, and dissemination of recorded knowledge of, relating to, characteristic of, or constituting a city' [13]. This definition gives a holistic, but rather general view on the term 'real-time city', which centres around information and knowledge while societal, political and privacy aspects remain greatly untouched.

In these interpretations of the expression 'real-time', its strict definition has been strongly mitigated. The term 'real-time' originated in the field of computer science, where it initially described a process, which is completed 'without any delay'. This broad view was then divided into hard and soft real-time demands. Soft real-time basically defines that deadlines are important, but the whole system will still function correctly if deadlines are occasionally missed. The latter is not true for hard real-time systems. Another term to express non-rigorous temporal requirements is 'near real-time', which describes a delay introduced into real-time applications, e.g. by automated data processing or data transmission [14]. Hence, the term accounts for the delay between the occurrence of an event and the subsequent use of the processed data.

As these definitions of the term 'real-time' have been set up for the domain of computer science it is important to evaluate and re-define the expression in the context of *Live Cities*. Naturally, strict real-time requirements are a central aspect in monitoring applications, whereby these demands are highly application-specific and can vary significantly. Therefore, they are not a fundamental goal in the field of *Live Cities*, as the term 'real-time' is primarily defined by an 'exact point in time', which is the same for all data sources to create a significant measurement outcome. Secondarily, the term defines the possibility to start a synchronous communication at a certain time, which might often be important for geographical monitoring applications, e.g. to enable the generation of an exact development graph for temporal pollutant dispersion over a defined period of time in precise intervals.

Additionally to the suggestion of assessibility of the environment in the 'now', the expression '*Live City*' also implies a feedback loop. The term 'city' does not only define the description of location-aware parameters, but also entails the exploration of causal patterns in these data. In the context of geo-sensor network and monitoring applications, this in turn means that the urban environment is not only analysed remotely by examining quasi-static data, but the procedure of sensing and processing live data offers the possibility of modifying the urban context in an ad-hoc fashion.

In conclusion, it can be stated that the strict term 'real-time' can be interpreted as 'at present' for urban monitoring applications. However, these topicality requirements can vary depending on the application context. For instance, an update on traffic conditions does not have to exceed a delay of a couple of minutes when this information is used for navigation instructions, whereas a 30 minute update interval can well be sufficient for short-term trip planning.

To account for this non-rigorous requirement, the term '*Live City*' seems better suited than 'Real-time City'. In this reflection, 'near real-time' appears to be closest to 'live', as it does not impose rigid deadlines and the expression itself suggests dynamic adaptation of a time period according to different usage contexts.

IV. Challenges in Current Research on the Live City

The urban context poses many challenges to pervasive monitoring and sensing systems. Particular issues arise for the deployment of near real-time information services in the city. These range from physical sensor mounting and other technical challenges to societal and privacy implications. Furthermore, the sensitive urban political landscape has to be accounted for, which might cause unforeseen challenges.

A. *Technological and Technical Issues*

The first essential technological challenge is the integration of different data sources owned by governmental institutions, public bodies, energy providers and private sensor network operators. This problem can potentially be tackled with self-contained and well-conceived data encapsulation standards – independent of specific applications – and enforced by legal entities, as discussed in sub-chapter V.B. However, the adaptation of existing sensors to new interoperability standards is costly for data owners and network operators in the short term, and so increased awareness of the benefits of open standards is required.

From a technical viewpoint, unresolved research questions for ubiquitous urban monitoring infrastructures are manifold. These challenges range from finding a uniform representation method for measurement values, optimising data routing algorithms in multi-hop networks, data fusion, and developing optimal data visualisation and presentation methods. The latter issue is an essential aspect in real-time decision support systems, as different user groups might need different views on the underlying information. For example, in case of emergency local authorities might want a socio-economic picture of the affected areas, while first-response forces are interested in topography and people's current locations, and the public might want general information about the predicted development of a disaster.

In addition, there are a number of well-known technical issues in the establishment of urban monitoring systems (energy supply, sensor mote size, robustness, connectivity, ad-hoc network connections, reliability, connectivity, self-healing mechanisms, etc.). These have to be addressed as the case arises depending on specific requirements of the end application. Thus, they are not part of the presented research.

B. *Various Stakeholders*

Other issues for the installation of a *Live City* are thematic challenges and socio-political concerns, which are rapidly gaining importance. The feedback loop depicted in Fig. 2 is a key factor in designing urban monitoring systems. In practice, various kinds of stakeholders have to be considered including citizens, information providers, research institutions, politicians, the city management, or other influential interest groups. This cycle involves all steps of the deployment process from planning, deployment, customised information provision, and feedback from the citizens and other interest groups. [15]

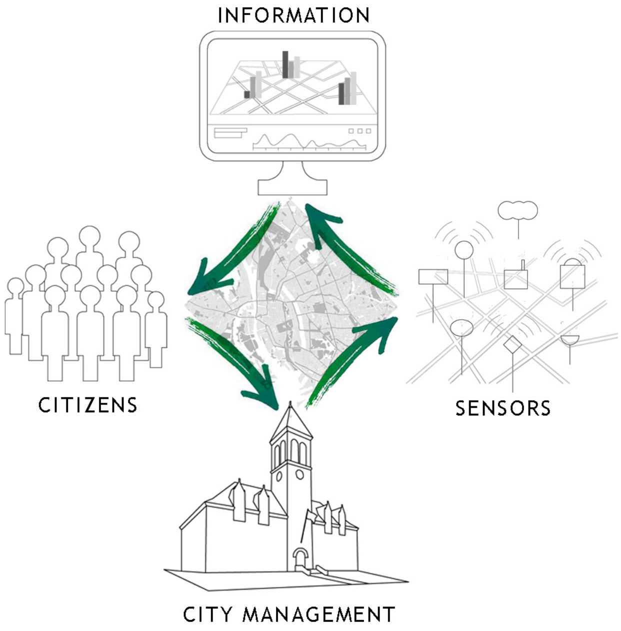


Figure 2. Feedback Loop Enabling the *Live City*.

Another important peculiarity of the urban context is that there are large variations within continuous physical phenomena over small spatial and temporal scales. For instance, due to topographical, physical or radiometric irregularities, pollutant concentration can differ considerably, even on opposite sides of the street. This variability tends to make individual point measurements less likely to be representative of the system as a whole. The consequence of this dilemma is an evolving argument for environmental regulations based on widespread monitoring data rather than mathematical modelling, and this demand is likely to grow.

C. *The Value of Sensing Collective Behaviour vs. Privacy Implications*

Although we experience quickly increasing awareness of the opportunities of digital mobile communication, the question arises how we can engage people to contribute actively being ‘human data sources’. This is necessary in order to leverage collective information in areas such as environmental monitoring, emergency management, traffic monitoring, or e-tourism. One example, in which this kind of volunteered data was of invaluable importance, were the earthquake and the following tsunami in Japan in March 2011. In this case, the *Tweet-o-Meter* [16] application has been used to find anomalies in Twitter activity. Right after the earthquake, people started to post status reports, video streams, and conditions of destroyed areas, which could be interpreted in near real time as an indicator for an extraordinary event. Furthermore, information could be semantically extracted from personal comments and posts.

This development raises the challenge to find the balance between providing pervasive real-time information while still preserving people’s privacy. Strategies to address this stress field are described in sub-section V.C. In addition, it seems self-evident that the provided information has to be highly accurate, reliable and unambiguous. Thus, quality control and error prevention mechanisms including appropriate external calibration will be discussed in sub-section V.A.

In terms of privacy, the claim might arise that we need to be aware of our personal and private data *before* we share them. The essential question in this context, however, is *how* we can raise awareness of ways to deal with that matter. Terms and conditions of digital services and technology are mostly hardly understandable to non tech-experts. Thus, more simple and binding ways of communicating this kind of information have to be found.

Finally, some more unpredictable challenges posed by the dynamic and volatile physical environment in the city are radical weather conditions, malfunctioning hardware, restricted connectivity, or even theft and vandalism. Moreover, there are a number of seemingly obvious but non-trivial challenges such as optimal positioning of sensors, high spatial and temporal variability of measured parameters or rapid changes in the urban structure, which might cause considerable bias in the measurements.

V. Discussion: Future Research Avenues

From the challenges described in Section IV we can derive a number of essential research questions, which have to be tackled in the area of *Live Cities*. These can be divided into methodological aspects, technical and technological issues, questions on privacy and legislation, and the assessment of economic opportunities.

A. *Methodological Research*

Over the last years prospects were made that ‘data would be the new oil’ [17],[18]. It has been stated that – like oil – data cannot be used without first being refined. This means that raw data is just the basic ingredient for the final product of **contextual information** that can be used to support strategic and operational decisions. Thus, a central issue in terms of providing real-time information services is the analysis of data according to algorithmic requirements, representation of information on different scales, context-supported data processing, and user-tailored information provision aligned with the needs of different user groups.

One way to reach this goal is to ‘sense people’ and their immediate surroundings using everyday devices such as mobile phones or digital cameras, as proposed by Goodchild [19]. These can replace – or at least complement – the extensive deployment of specialised city-wide sensor networks. The basic trade-off of this people-centric approach is between cost efficiency and real-time fidelity. The idea of using **existing devices to sense the city** is crucial, but it requires more research on sensing accuracy, data accessibility and privacy, location precision, and interoperability in terms of data and exchange formats.

In terms of geo-data sources, Volunteered Geographic Information (VGI) plays a key role in realising the idea of a *Live City*. We are already experiencing an overwhelming willingness of citizens to contribute their personal observations ranging from opinions posted on Facebook to Tweets about local events or commented photo uploads on Flickr. As mentioned in Section IV, this kind of **collective information** can potentially have a vital impact on operational real-time strategies in areas such as emergency management, dynamic traffic control or city management.

A central issue in VGI is the representativeness of volunteered information [19],[20]. We argue that defining or deriving consistent semantics in user-generated content possibly requires the combination of bottom-up and top-down approaches. In bottom-up approaches, user communities build their own semantic objects and connections between those by using their own personal taxonomies. In contrast, top-down approaches – mostly academically driven – try to define semantic rules and ontological connections in a generic way prior to and independently of the end application. Only the combination of those using Linked Data concepts (rather than rigid and inflexible ontology approaches) can lead to **domain-independent and comprehensive semantic models**, which are needed to cover the whole breadth of topics, users and applications in the *Live City*. In this regard, semantic search will be an essential concept to extract knowledge and information from user-generated data.

An aspect, which is strongly connected to availability of data sources, is **openness of data**. As argued by Jonathan Raper [21], quality of decision-support is increasing with the quality and the quantity of available data sources. We are currently facing a situation that in most cases, too little data are available to support well-informed decisions in near real time. This brings up the question how data owners such as companies in the environmental sector, energy providers or sensor network operators can be animated to open their data repositories for public use.

On the contrary, we might face a vast amount of data freely available in the near future, contributed by a variety of different data producers. This of course raises the concern of trustworthiness of these data. Thus, **automated quality assurance** mechanisms have to be developed for uncertainty estimation, dynamic error detection, correction and prevention. In this research area, we are currently seeing different approaches including Complex Event Processing (CEP) [10] for error detection, standardisation efforts for representing uncertainty in sensor data (e.g. Uncertainty Markup Language – UncertML) [22], or proprietary profiles to define validity ranges for particular observations. Only when these questions are solved, reliability and completeness of recommendations can be ensured.

Furthermore, measurements are only available in a quasi-continuous distribution due to the high **spatial and temporal variability** of ad-hoc data collection. Addressing this issue will require complex distribution models and efficient resource discovery mechanisms in order to ensure adaptability to rapidly changing conditions.

B. Technical and Technological Research

Apart from technical sensor network research on energy supply, miniaturisation, connectivity, etc., **standardisation and interoperability** are vital prerequisites for establishing pervasive and holistic monitoring systems. As current sensor networks are mostly built up in proprietary single-purpose systems, efforts to develop a uniform communication protocol will be needed [23]. One promising approach in this field is the Sensor Web Enablement (SWE) initiative [24] by the Open Geospatial Consortium (OGC). SWE aims to make sensors discoverable, accessible and controllable over the Internet. SWE currently consists of seven standards and interoperability reports, including the Sensor Observation Service (SOS) for observation data retrieval, Observations and Measurements (O&M) for sensor data encoding, Sensor Markup Language (SensorML) for platform description and the Sensor Alert Service (SAS) for event-based data transmission.

Furthermore, sensor fusion algorithms are a vital prerequisite to combine data stemming from different heterogeneous sensor networks. **Sensor fusion** basically stands for the harmonisation of data in terms of units of measure, time zones, measurement models and observation semantics. To be compliant with the requirements of a *'Live' City*, the fusion process has to happen in near real time. One approach for on-the-fly integration of measurements coming from different SOS instances using the free open-source server GeoServer (http://geoserver.org) is presented in [25]. The system harmonises measurements in real time and provides them on the fly via standardised OGC web service interfaces such as the Web Feature Service (WFS) and the Web Map Service (WMS). Although this implementation is still improvable in terms of fusion capabilities, it demonstrates a seminal approach towards sensor fusion.

The geo-analysis of real-time data sources can be implemented using the OGC Web Processing Service (WPS) in a standardised way. But the WPS architecture is very generic in its current version so that the developments of further specialised (domain-specific) application profiles are necessary as argued in [26] and [27]. The power of using WPS for implementing more complex analysis functionality for urban models has for instance been shown in [28].

Another technological issue is the availability of **ubiquitous communication media**. Today we are used to a functioning Internet to transmit information. However, in case of emergency, this layer is potentially not available, as we experienced for instance during hurricane Katrina in 2005 in New Orleans. Possible alternate solutions comprise long-range ad-hoc networks or a robust communication core network, which can withstand harsh external influences such as tsunamis, earthquakes, avalanches or even vandalism.

C. Privacy and Legislation Measures

A crucial question in the context of *Live Cities* is how we can **preserve people's privacy** dealing with ubiquitous information and partly personal data. One possible solution to address this issue is to make use of new 'collective sensing' approaches. This methodology tries not to exploit a single person's measurements and data, but analyses aggregated anonymised data coming from collective sources, such as Twitter, Flickr or the mobile phone network [29]. Like this, we can gain a coarse picture of the situation in our environment without involving personal details of single persons. In case of tracking applications or services, in which personal data are involved, people have to have an opt-in/opt-out possibility. This means that users can decide themselves whether they want to use the application – and also withdraw their consent – being aware of the type and amount of data that is collected and transmitted.

Another central issue in deploying monitoring systems in the city is the personal impact of fine-grained urban sensing, as terms like 'air quality' or 'pollutant dispersion' are only surrogates for a much wider and more **direct influence** on people, such as life expectation, respiratory diseases or quality of life. This raises the demand of finding the right level of information provision. This again can potentially entail a dramatic impact in a very wide range of areas like health care, the insurance sector, housing markets or urban planning and management.

A central question in this context is: can we actually achieve a system, in which transactions are not tracked or traced? Thinking about mobile phone calls, credit card payments or automated toll collection, each of the underlying systems has to have some kind of logging functionality in order to file payments and generate automated reports. In these cases it is probably just not possible prevent storage – at least for a short time. Thus, **legal frameworks** have to be developed on national and global levels. The dominant limiting factor in this regard is the varying interpretation of 'privacy' in different parts of the world. For instance, privacy can be traded like a good by its owner in the USA, whereas it is protected by law in the European Union. This means that supra-national legislation bodies and initiatives are called upon to set up appropriate world-wide regulations.

This also includes the critical question of **data ownership** – who owns the data: the data producers (i.e. the citizens or a mobile phone network operator), the institutions that host a system to collect data, or the data providers? Furthermore, if sensitive data is analysed to produce anonymised information layers, who is responsible if decisions that are based on this information are wrong due to lacking quality of the base data?

D. Assessing the Economic Value of Live Urban Services

Basically, the economic value of *Live City* services and applications can be either defined in **concrete revenues or as an after effect of improved quality of life**. The Economist Intelligence Unit's liveability ranking [30] assigns a score for over 30 qualitative and quantitative factors across five broad categories: stability, healthcare, culture and environment, education, and infrastructure. These five categories basically sum up 'what people want'.

The technologies that have been developed in the few last years, like pervasive sensors to assess urban dynamics and especially mobile technologies, offer new opportunities to 'tune' and 'fine-tune' urban processes. These processes can be transportation related, to monitor and direct traffic in real time, optimise parking spaces and navigate to available parking, or simply to help people with their daily tasks, finding jobs, finding housing, connecting people in spare time. Tools that bring the feedback loop directly to people make it easy to **promote events and give people instruments** to rate the attractiveness of these happenings.

Mobile technologies offer great **opportunities for young start-ups** to build GPS-enabled, crowd-sourced, location-based apps. Just one example is the Wikitude World Browser [31], which is tailored to individual needs. Igniting and funding a start-up scene can be the starting point for any government to build a connected *Live City*: start-ups create jobs and apps, which in turn – if tailored for locals – benefit people in the city and improve quality of life.

The improved economic **value of a 'tuned' city** can be enormous. On one hand there can be cost saving advantages, for instance in considerable fuel savings if available parking spaces are reserved on a first-come-first-served policy and the driver is routed to this parking space rather than having to circle looking for a parking space.

On the revenue side Google successfully leverages Internet advertisement by matching the search terms people enter in the Google search engine with ads. One key to generating revenue in the field of *Live Cities* may be to apply what Google did with the Internet to the real world, offering information and **search services** that focus on time, location, context and people rather than on simply search terms.

VI. CONCLUSION

Contrary to projections, which stated that the wide-spread distribution of high-speed Internet connections would render geographical distance irrelevant, cities have recently gained importance in academic research. Especially real-time monitoring of urban processes enriches research on cities with uncharted up-to-date information layers.

Hence, within this vision of a *Live City*, the city is not only regarded as a geographical area characterised by a dense accumulation of people or buildings, but more as a multi-layered construct containing multiple dimensions of social, technological and physical interconnections. Through this viewpoint of urban areas as an actuated **multi-dimensional conglomerate of dynamic processes**, the city itself can also be seen as a complex near real-time control system creating a feedback loop between the citizens, environmental monitoring systems, the city management and ubiquitous information services.

In the *Live City*, the everyday citizen is empowered to monitor the environment with sensor-enabled mobile devices. This feedback of 'sensed' or personally observed data, which is then analysed and provided to citizens as decision-supporting information, can change people's behaviour in how they use the city and perceive their environment by supporting their short-term decisions in near real time. This again requires promotion of the user sensitisation of information through awareness of limitations.

Basically, we identified four main barriers towards the implementation of the *Live City* concept: methodological issues, technical/technological problems, privacy and legislative questions, as well as quantification of economic opportunities. We discussed these challenges and future research avenues in Section IV and V.

We believe that promoting the *Live City* concept will trigger a profound rethinking process in collaboration and cooperation efforts between different authorities. Also, a people-centric view of measuring, sharing, and discussing urban environments might increase agencies' and decision makers' **understanding of a community's claims** leading to proactive democracy in urban decision-making processes.

In terms of privacy and personal data collection, it is evident that everybody has to have the right to decide what kind of personal data is collected by whom, and for which purposes these data are used. In this context, people have to have an **opt-out possibility** to withdraw their consent to personal data collection. One new paradigm to tackle the issue of privacy is 'collective sensing', which tries not to exploit single people's measurements and data, but analyses aggregated anonymised data coming from collective sources such as Twitter, Flickr or the mobile phone network.

As mentioned in the introductory section, we are experiencing a fast progressing technology development, which is already moving ahead of society. The deciding final question can be: If we compare this development with a **stream moving at high speed**, on which we are paddling to remain on the same spot or at least not to drift off too fast, where does our goal for the future lie: down-stream, somewhere near our current spot, or even up-stream? We argue that the issues of privacy, data ownership, accessibility, integrity and liability have to be tackled thoroughly all at once and not separately from each other. In the end, legislation bodies are called upon to set the legal stage for leveraging *Live City* technologies, exploit economic opportunities, but still preserve citizens' privacy.

ACKNOWLEDGMENT

This paper has originated from the panel discussion 'The Real-time City: Technology – Innovation - Society' at GI Forum and AGIT conferences in July 2011 in Salzburg, Austria. The authors would like to thank the audience for their valuable contributions to the discussion, which partly served as input for this publication. In particular, the authors would like to thank Prof. Dr. Manfred Ehlers and Prof. Dr. Petra Staufer-Steinocher their fruitful comments.

REFERENCES

[1] Cairncross, F. (1997) The Death of Distance: How the Communications Revolution Will Change Our Lives. Harvard Business School Press, Boston, MA, USA, 1997.

[2] Gilder, G. and Peters, T. (1995) City vs. Country: The Impact of Technology on Location. Forbes ASAP, 155(5), pp. 56-61, 27 February 1995.

[3] Dierig, S., Lachmund, J., and Mendelsohn, A. (2000) Science and the City. http://vlp.mpiwg-berlin.mpg.de, Workshop, Max Planck Institute for the History of Science, Berlin, Germany, 1-3 December 2000. (10 September 2011)

[4] Netherlands Organization for Scientific Research (2007) Urban Sciences. http://www.urbansciences.eu, Interdisciplinary Research Programme on Urbanization & Urban culture in The Netherlands, 2007. (26 August 2011)

[5] SENSEable City Laboratory (2009) MIT SENSEable City Lab. http://senseable.mit.edu, September 2011. (29 July 2011)

[6] International Telecommunication Union (2010) Key Global Telecom Indicators for the World Telecommunication Service Sector. http://www.itu.int, 21 October 2010. (31 August 2011)

[7] Green, J. (2011) Digital Urban Renewal - Retro-fitting Existing Cities with Smart Solutions is the Urban Challenge of the 21st Century. http://www.cisco.com, Ovum Report OT00037-004, April 2011. (11 September 2011)

[8] ENoLL (2011) Open Living Labs | The First Step towards a new Innovation System. http://www.openlivinglabs.eu, September 2011. (11 September 2011)

[9] IBM (2009) A Vision of Smarter Cities - How Cities Can Lead the Way into a Prosperous and Sustainable Future. http://www.ibm.com, IBM Global Business Services Executive Report, 2009. (04 September 2011)

[10] Resch, B., Lippautz, M. and Mittlboeck, M. (2010) Pervasive Monitoring - A Standardised Sensor Web Approach for Intelligent Sensing Infrastructures. Sensors - Special Issue 'Intelligent Sensors 2010', 10(12), 2010, pp. 11440-11467.

[11] Murty, R., Mainland, G., Rose, I., Chowdhury, A., Gosain, A., Bers, J., and Welsh, M. (2008) CitySense: A Vision for an Urban-Scale Wireless Networking Testbed. 2008 IEEE International Conference on Technologies for Homeland Security, Waltham, MA, May 2008.

[12] Townsend, A.M. (2000) Life in the Real-time City: Mobile Telephones and Urban Metabolism. Journal of Urban Technology. (7)2, pp.85-104, 2000.

[13] Foth, M. (Ed.) (2009) Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City. ISBN 978-1-60566-152-0, Hershey, PA: Information Science Reference, IGI Global..

[14] General Services Administration (1996) Federal Standard 1037C. Telecommunications: Glossary of Telecommunication Terms, http://www.its.bldrdoc.gov, 7 August 1996. (11 September 2011)

[15] Resch, B., Mittlboeck, M., Lipson, S., Welsh, M., Bers, J., Britter, R. and Ratti, C. (2009) Urban Sensing Revisited – Common Scents: Towards Standardised Geo-sensor Networks for Public Health Monitoring in the City. 11th International Conference on Computers in Urban Planning and Urban Management - CUPUM2009, Hong Kong, 16-18 June 2009.

[16] UCL Centre for Advanced Spatial Analysis (2011) Tweet-o-Meter - Giving You an Insight into Twitter Activity from Around the World!. http://www.casa.ucl.ac.uk/tom, 12 September 2011. (12 September 2011)

[17] Palmer, M. (2006) Data is the New Oil. http://ana.blogs.com, 3 November 2006. (12 September 2011)

[18] Kennedy, J. (2011) Data is the New Oil. http://www.siliconrepublic.com, 23 June 2011. (29 July 2011)

[19] Goodchild, M.F. (2007) Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research, vol. 2, pp. 24-32, 2007.

[20] Craglia, M., Goodchild, M.F., Annoni, A., Camera, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S. and Parsons, E. (2008) Next-Generation Digital Earth: A Position Paper from the Vespucci Initiative for the Advancement of Geographic Information Science. International Journal of Spatial Data Infrastructures Research, vol. 3, pp. 146-167.

[21] Raper, J. (2011) Realising the Benefits of Open Geodata: Lessons from London's Experience. Keynote at AGIT 2011, 6 July 2011, Salzburg, Austria.

[22] Williams, M., Cornford, D., Bastin, L and Pebesma, E. (2008) Uncertainty Markup Language (UncertML). OGC Discussion Paper 08-122r2, Version 0.6, 8 April 2009. (14 August 2011)

[23] Resch, B., Blaschke, T. and Mittlboeck, M. (2010) Live Geography - Interoperable Geo-Sensor Webs Facilitating the Vision of Digital Earth. International Journal on Advances in Networks and Services, 3(3&4), 2010, pp. 323-332.

[24] Botts, M., Percivall, G., Reed, C. and Davidson, J. (Eds.) (2007) OGC Sensor Web Enablement: Overview And High Level Architecture. http://www.opengeospatial.org, OpenGIS White Paper OGC 07-165, Version 3, 28 December 2007. (17 August 2011)

[25] Resch, B. (in press) On-the-fly Sensor Fusion for Real-time Data Integration. In: Proceedings of the Geoinformatics 2011 Conference, 28-30 March 2012, Braunschweig, Germany, pp. pending.

[26] Lanig S. and A. Zipf (2010) Proposal for a Web Processing Services (WPS) Application Profile for 3D Processing Analysis. 2nd International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2010), St. Maarten, Netherlands Antilles, 10-15 February 2010, pp. 117-122.

[27] Resch, B., Sagl, G., Blaschke, T. and Mittlboeck, M. (2010) Distributed Web-processing for Ubiquitous Information Services - OGC WPS Critically Revisited. 6th International Conference on Geographic Information Science (GIScience 2010), Zurich, Switzerland, 14-17 September 2010.

[28] Stollberg, B. & Zipf, A. (2009): Development of a WPS Process Chaining Tool and Application in a Disaster Management Use Case for Urban Areas. UDMS 2009. 27th Urban Data Management Symposium, Ljubljana , Slovenia.

[29] Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (in press) Estimating Origin-destination Flows Using Opportunistically Collected Mobile Phone Location Data from One Million Users in Boston Metropolitan Area. IEEE Pervasive Computing, 2011.

[30] Economist Intelligence Unit (2011) Liveability Ranking and Overview 2011. http://www.eiu.com, February 2011. (4 September 2011)

[31] Wikitude GmbH (2011) Wikitude World Browser | Wikitude. http://www.wikitude.com, September 2011. (29 August 2011)

Comparison of Stacking Methods Regarding Processing and Computing of Geoscientific Depth Data

Claus-Peter Rückemann
Leibniz Universität Hannover,
Westfälische Wilhelms-Universität Münster (WWU),
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

***Abstract*—This paper presents a comparison of different stacking methods available for processing and computing of geoscientific depth data for use with various high level applications and computing architectures. These methods are used with seismics and comparable geophysical techniques for example. Todays resources enable to use these methods for more than batch processing. For these cases it is important to analyse the algorithms regarding strengths and implementation benefits. The algorithms presented have been successfully implemented and evaluated for their purpose with application scenarios using High End Computing (HEC) resources. The focus is on integrating stacking algorithms in information and computing systems, utilising Distributed Computing and High Performance Computing (HPC) from integrated systems for use in natural sciences disciplines and scientific information systems.**

***Keywords*–Processing; Scientific Computing; Stacking; Comparison; Seismics; Geosciences; Depth Data.**

I. Introduction

Processing of geoscientific depth data does have a long and successful history and evolution. Decades of methods development, numerical algorithms, and implementation of processing and visualisation systems have been needed to understand how to reveal and analyse some essential information of depth sequences and profiles, from the work built by the complex geological and geophysical processes of millions of years. Processing of depth data has always been very data and computing intensive, so there is no focus on geo-data processing itself in this paper. Computing architectures have been available for decades but these have been quite limited regarding computing power and therefore resulted in long processing time intervals, in many cases to weeks and months, even for single profiles. This has restricted applications and algorithms to batch processing and rarely interactive applications have been reasonable. Classical use and applications are known from published use cases [1]. With the modern parallel architectures many interactive and dynamical applications, Active Source, and InfoPoints get into the focus [2]. In the early 1990s the advanced superstack algorithm has been developed [3]. Even the full vectorisation of the algorithms has resulted in days of processing time on VAX mainframe and Unix machines, even for small parts of profiles, and with increased resolution and data size computing times have not been reduced today. With parallel architectures several processing algorithms are undergoing parallelisation efforts, meaning parallelisation regarding data sets, algorithms, iterations, and so on. These algorithms are for example part of stacking methods, migration methods, Fresnel Section [4] calculation as well as elementary algorithms like Fast Fourier Transformations (FFT) and many more, for seismic software [5], [6] as well as for dynamical information system components [7], [8].

This paper is organised as follows. Section two presents motivation and complexity with the implementation. Section three introduces stacking, basic terms, and challenges. Section four shows the different stacking methods and algorithms. Section five and six discuss and evaluate the methods regarding application and Sections seven and eight summarise the lessons learned, conclusions, and future work.

II. Motivation

In most cases geoscientific and geophysical algorithms are used in conventional batch applications. None of these have been integrated into interactive information system components so far. The reason is that computing power is limited for up-to-date applications and resolutions and that parallelisation would not be possible for local or standalone computing systems. Parallelisation, loosely and embarrassingly parallel, of geoscientific algorithms will help to support new application scenarios, for example dynamical interactive information and computing systems for geosciences and environmental sciences. With the implementation use cases for Information Systems the suitability of Distributed and High Performance Computing resources supporting processing and computing of geoscientific data have been studied. These use cases have focus on event triggered communication, dynamical cartography, compute and storage resources. The goal has been, to bring together the features and the experiences for an integrated information and computing system. An example that has been implemented is a spatial information system with hundreds of thousands of ad-hoc simulations of interest. Within these interactive systems depth information may play an important role as "next

ISBN: 978-1-61208-178-6

informations of interest", being dynamically calculated in parallel. Due to the complexity of integrated information and computing systems, we have applied meta-instructions and signatures for algorithms and interfaces. For these cases, envelopes and IPC has been used to provide a unique event and process triggered interface for event, computing, and storage access.

III. Stacking, Terms, Goals, and Challenges

Stacking is an essential part of seismic data processing. The primary goal of seismic stacking methods is the enhancement of the Signal-to-Noise Ratio (SNR) of the data material. Increasing demands for high resolution and true amplitudes, and allowing interpretation of amplitude ratios have led to seismic stacking methods [9]. The basic stacking methods and references have been collected and described [3]. Stacking and migration are the central methods for discovering and defining slanted crustal boundary layers [10]. Stacking is done in order to to reduce Common Mid-Point (CMP) gathers into one trace. The appropriate corrections for statics and Normal-Moveout (NMO) should be done in all cases where advanced precision is necessary. The description of standard NMO correction and CMP methods can be found in all common textbooks. In some cases stacking methods are used in order to combine groups of traces other than CMP groups. For example in vertical stacking from repeated shots traces of sequences of depth points are combined. Nearly all stacking methods commonly used are phase stacks. The term Maximum Coherency Stack is sometimes used to point out explicitely that maximum coherence of different traces is achieved by moving them in direction of the time axis, for example by NMO correction. Only in rare case stacking in the frequency domain is used, for example the envelope stacking. All stacking methods obtain their significance by experience, not primarily by theoretical deliberations [9]. Along with improving the SNR, stacking reduces disturbing events and energy in the data. These disturbances can be called "unwanted energy" [9]. Although this term depends on the situation, in most cases it means the following effects:

- Multiples energy,
- Refracted energy,
- Uncorrelated noise,
- 'Noise bursts' with large amplitude,
- Cable noise.

In many cases randomly distributed energy is the target. Suitable assumptions for noise regarding stacking strongly depend from the preprocessing and much less from the post-stack processing. For that, stacking methods are most desirable, that make a distinction between signal and coherent noise. It is also desirable to retain the signal form and amplitude with the stacking algorithm as these contain physical and geological information but in many cases the focus is on recognition of primary signals especially with a very small SNR. Besides the goals there are various challenges making use of stacking algorithms with different application scenarios. The consequences of the properties and strengths of different stacking methods as well as the very different processing and computing requirements has made it neccessary considering operational areas with complex systems, suitable for scientific and educational purposes. Which methods can be applied for the purposes of dynamical processing of depth data or only used for precomputation of depth data? Different application scenarios need different stacking methods, especially this holds true for informations system and expert system components for which individual processing decisions and dependencies are not feasible in most cases. With this we need to know the architecture of the algorithms, specialisation, and strengths.

IV. Comparison of different stacking methods

The following sections give a short comparison of available stacking methods and show their strengths and possible field of operation in the context of complex application scenarios. Although each principle is characteristic, variations of the methods are applicable.

A. *Straight Stack, Mean Stack*

This stacking method is the most simple one [11]. This method is a special, simple case of more common stacking methods like the Superstack or the Trimmed Mean Stack. Nevertheless, many higher level considerations are based on it [12]. The Straight Stack sums up the sample amplitude values at the isochrone locations and divides by the number of values, for all channels to be processed:

$$a_t^{StraightStack} = \frac{1}{N} \sum_{i=1}^{N} S_i \qquad (1)$$

N is the number of isochone values, S_i the amplitude at a sample location, and $a_t^{StraightStack}$ is the amplitude of the stacked trace at a respective time.

B. *Stacking with predefined weighting*

1) Weighting by muting: The amplitude values at every sample in the gather are assigned with a weighting, a value 1 or 0. If this hold true for a threshold value this is a mute function. For strong multiples like the ocean bottom multiples an inner-trace-mute can be reasonable.

2) Weighting as function of the offset: This is comparable to an inner-trace-mute and is used to increase the stacking response of primary reflections. Far offset traces get a higher weighting where multiples get out of phase. Weights are calculated from velocity ratios.

3) Weighting as function of the array response: Based on the ratio of overall response of the array system relative to response of the primary signals plus multiples. Used in cases of water coverage absorbing too much energy, important for strong multiples scenarios

C. Stacking with data adaptive weighting

1) Optimum Weighted Stack (OWS), Weighted Stack: Based on optimum criteria this algorithm is used before summation, including optimum stacking filters. The algorithm for the stacking value s_t^{OWS} is based on:

$$r_{j,t} \quad \text{with} \quad j = 1, 2, \ldots, J \quad \text{and} \quad t = 1, 2, \ldots, T \tag{2}$$

$$s_t^{OWS} = \sum_{j=1}^{J} (w_j r_{j,t}) \tag{3}$$

where $r_{j,t}$ represents the traces to be stacked, w_j the weights, T the number of samples per trace and J the number of traces per CMP gather.

2) Diversity Stack (DS): The result of the DS is the amplitude variation of the input data:

- Subdivide trace into time windows.
- Calculate the overall energy per window:

  $$E = \sum_{\Delta T} \left(a_t^2\right) \tag{4}$$

  with amplitude a_t at a sample and window length ΔT.
- Calculate scale factor $D = C/E$ with $C = const.$ for each window,
- Determine gain trace, scaling factor D, this can be a selected trace or every trace for itself.
- Scale trace by application of gain trace to the original trace using cross multiplication.
- Summation for scaled traces and gain traces. The sum of scaled traces is divided by the sum of gain traces.

3) Coherency Stack (CS): The application of the CS is: Choose windows for NMO corrected CMP gather. calculate coherency values for windows using coherency measure like semblance, coherency model trace can be calculated by sorting and interpolation relative to subsequent central values. Coherency stacking is applied by addition of choosen percentages of the coherences model trace on the conventional stacked trace.

D. Iterative and comparable methods

The term iterative stacking is often used synonym to the term Superstack. It is especially important with iterative methods to take care of reflected signals with phase reversal on far offset. This will require preprocessing or reduction to nearer offset when optimising the SNR.

1) Iterativer Stack, Superstack (IS, SS): Iterative stacking with the Superstack [13] is based on separating Amplitudes, positive a_j and negative b_j, for all reflection times with amplitude r_j. The Number of iteration is n, after the first iteration for the data matrix holds $n = 1$. The norm factor M for sums after an iteration is called multiplicity. In basic form the algorithm is described by:

$$a_j^n = \begin{cases} r_j & \text{for} \quad r_j > 0 \\ 0 & \text{for} \quad r_j \leq 0 \end{cases} \tag{5}$$

$$b_j^n = \begin{cases} 0 & \text{for} \quad r_j \geq 0 \\ r_j & \text{for} \quad r_j < 0 \end{cases} \tag{6}$$

$$s_+^n = \frac{1}{M} \sum_{j=1}^{J} a_j^n \qquad \text{and} \qquad s_-^n = \frac{1}{M} \sum_{j=1}^{J} b_j^n \tag{7}$$

with sums s_+^n and s_-^n of isochrone positive and negative amplitudes at a respective time and NMO corrected amplitudes r_j in the CMP gather, with

$$r_{j,t} \quad \text{with} \quad j = 1, 2, \ldots, J \quad \text{and} \quad t = 1, 2, \ldots, T \tag{8}$$

The weighting of a single amplitude value at a time sample is done using:

$$\begin{aligned} &a_j^{n+1} = s_+^n \text{ for } a_j^n > s_+^n \quad \text{and} \quad a_j^{n+1} = a_j^n \text{ for } a_j^n \leq s_+^n \\ &b_j^{n+1} = s_-^n \text{ for } b_j^n < s_-^n \quad \text{and} \quad b_j^{n+1} = b_j^n \text{ for } b_j^n \geq s_-^n \end{aligned} \tag{9}$$

If I is the overall number of samples $i = 1, \ldots, I$ of a trace, stacking can be described using the matrix of the original data set $C_{I,J}$, utilising n iterations for the modified matrix $C_{I,J}^n$ to the resulting stacked trace $S_{I,1}$ with the following formula. Separation in positive and negative amplitude sums is done for every iteration.

$$\begin{bmatrix} c_{1,1} & \cdots & c_{1,J-1} & c_{1,J} \\ c_{2,1} & \cdots & c_{2,J-1} & c_{2,J} \\ c_{3,1} & \cdots & c_{3,J-1} & c_{3,J} \\ \vdots & \ddots & \vdots & \vdots \\ c_{I-1,1} & \cdots & c_{I-1,J-1} & c_{I-1,J} \\ c_{I,1} & \cdots & c_{I,J-1} & c_{I,J} \end{bmatrix} \tag{10}$$

$$\Big\downarrow \; n \text{ iterations}$$

$$\begin{bmatrix} c_{1,1}^n & \cdots & c_{1,J-1}^n & c_{1,J}^n \\ c_{2,1}^n & \cdots & c_{2,J-1}^n & c_{2,J}^n \\ c_{3,1}^n & \cdots & c_{3,J-1}^n & c_{3,J}^n \\ \vdots & \ddots & \vdots & \vdots \\ c_{I-1,1}^n & \cdots & c_{I-1,J-1}^n & c_{I-1,J}^n \\ c_{I,1}^n & \cdots & c_{I,J-1}^n & c_{I,J}^n \end{bmatrix} \overset{\sum_{j=1}^{J}}{\longrightarrow} \begin{bmatrix} s_{1,1}^n \\ s_{2,1}^n \\ s_{3,1}^n \\ \vdots \\ s_{I-1,1}^n \\ s_{I,1}^n \end{bmatrix} \tag{11}$$

For reduction of the polarity of a small number of amplitudes a Higher Degree Stacking (HDS) can be done, separating positive and negative amplitudes and exponentiating it with the degree of the stack.

2) Single Trace Iterative Stack (STIS): This is an alternative application of the above Superstack algorithm [14]. With many applications the term Near Trace Iterative Stack is a good description as increased weighting is on near offset traces and not on a single trace.

3) Iterative Weighted Stack (IWS): The IWS algorithms is used the following way:

- A CMP gather is stacked (Straight Stack) into one trace (pilot trace).
- At every sample the amplitude variance along the gather is calculated. Often the mean amplitude value from the pilot trace is used.

- Weights are calculated (Gauß distribution) for every sample, the mean value from b) is used. For every trace the weighting is smoothed with the time. For every sample time weights are normed along the gather.
- The weighted stacking is calculated and 1) used as finally stacked trace or 2) defined as new pilot trace and the process is iterated from b) on.

4) The N^{th}-Root Stack (NRS): The NRS is not an iterative method but it is based on a comparable principle. In the most simple form the NRS can be written in the following form for calculating a stacking element s_t^{NRS}.

$$s_t^{NRS} = \left[\frac{1}{J} \sum_{j=1}^{J} sgn(s_{j,t}) \mid s_{j,t} \mid^{\frac{1}{N}} \right]^N \tag{12}$$

$$\text{for} \quad t = 1, 2, \ldots, T \quad \text{with} \quad sgn(s_{j,t}) = \frac{s_{j,t}}{\mid s_{j,t} \mid} \tag{13}$$

with the number of traces J in the CMP gather and N a number 2^P, with $P \in$ IN. In most cases $N = 2$ or $N = 4$ are used. There are various ways of application [15].

E. Other Methods

1) Median Stack (MS), Alpha-Trimmed Mean Stack (ATMS): The median amplitude values from traces to be stacked are picked. The stacked trace contains for every sample the median value at the same time with the amplitudes along the CPM gather. Thus the MS stack does not result from summed up values. Given amplitudes a_i with $i = 1, 2, \ldots, J$, and J is an odd integer value, the ATM a_α can be calculated as:

$$a_\alpha = \frac{1}{J - 2L} \sum_{i=L+1}^{J-L} a_i \tag{14}$$

Trimming is done by choosing the value $L = \alpha J$, where α is the trimming parameter ($0 \leq \alpha < 0.5$). This means with the ATMS it is possible to do a step wise and weighted combination of Straight Stack for $\alpha = 0$ and Median Stack for $\alpha \rightarrow 0.5$ for stacking. The MS result can appear like adding high-frequency noise. This is reduced by summing up more than one amplitude, which after resorting the input values follow in rising sequence around the central position. It can be used to exclude extreme amplitude value groups from the stack. This is done using the ATM method.

2) Trimmed Mean Stack (TMS): The TMS can be described by the following algorithm: The amplitudes of a gather are sorted by value, numbered, and summed up at a time using the values up to a defined amplitude number. The summation for non symmetrical elimination of extreme amplitudes can be performed as:

$$a^{TMS} = \frac{1}{N - K} \sum_{i=\frac{K}{2}+1}^{N-\frac{K}{2}} S_i \tag{15}$$

with number N of samples, overall number K of excluded sample values, the amplitude S_i at the respective sample, and the amplitude a^{TMS} of the stacked trace at the respective time. The TMS is a generalisation of the Straight Stack.

3) Minimum/Maximum Sample Value Exclusion Stack: The “Min/Max Stack” excludes samples with the largest and smallest amplitudes from the stacking.

4) Signed Bit Stack (SBS): The SBS adds $+1.0$ to the stacked sum if the absolute amplitude value at a sample is positive or zero and -1.0 if the amplitude value is negative.

5) Random Stack (RS): The RS can be described with:

a) Given a NMO corrected gather:

$$r_{j,t} \quad \text{with} \quad j = 1, 2, \ldots, J \quad \text{and} \quad t = 1, 2, \ldots, T \tag{16}$$

For determination of random sample trace values for every discrete sample time a random value $k(t)$ is picked from the J values of the corrected gather. The amplitude of the random value is $r_{k(t),t}$. The result of the process is a Constant Velocity Gather (CVG).

b) The traces of the CMP gather are combined in one trace: For every sample time amplitudes from the random traces are stacked, if they show they same sign. In the other case the output trace is set to zero. The result is a Constant Velocity Stack (CVS).

V. Discussion

There are several consequences and conditions for geoscientific interpretation and integrated systems, resulting from the different stacking algorithms described.

Precondition of the Straight Stack is uncorrelated noise and amplitudes and SNR comparable in magnitude over the traces. Otherwise Optimum Weighted Stack is an option. Normalisation the results using a scalar is used in praxi, e.g., root power scalar for stack normalisation. It is of common use, computation depends on size and dimensions of data but is easy to implement and can be widely parallelised.

Weighted stacking methods are important for very long source arrays, whose effect can be considered by calculation of the special geometrical parameters for the ‘predetermined’ weighting. They rely on data analysis a part of the processing. The result of the OWS depends on two main criteria, the adaption of the model to the data and the precision of calculating or estimating the signal amplitudes and the SNR. The DS is best suited for excluding high noise levels, especially noise bursts. In use are Diversity Power Stack and Diversity Amplitude Stack. It can reduce interference noise and is mostly used on land data. The CS reduces the effect of strong, non-coherent amplitudes in the final stack. It is mostly used on post-stack (Straight Stack). Relative amplitudes and therefore resolution can be disturbed.

The Superstack can be most useful for very low SNR. Very large or small abnormal amplitudes, that rarely occur in lateral view regarding isochrone samples, are reduced. In tendency amplitudes are reduced towards the smallest absolute value, the stacking converges. With increasing number of iterations frequency spectra of single traces result in higher frequencies. The number of iteration should be suitably small in order to minimise disturbance of the wave form, in case the wave form is relevant. This is less relevant with a high noise percentage on longer travel times, when looking for appearent velocities near infinite. A large number of traces present can help correct this. Due to the algorithm of the Superstack/HDS, the large computational requirements can be encountered by vectorisation and parallelisation. STIS is mostly used when primary events in recorded traces to be summed up are not equal. The IWS is used for enhancing the velocity discrimination and reduction of multiples. The NRS is used in order to eliminate false alarms from strong noise amplitudes in single channels. NRS is most helpful being used on reflection data. The effect and the modification of the wave form is comparable to that of the Superstack. With most data there is less need for parallelisation but efficiency will profit nevertheless.

The significance of the MS/ATMS results from the fact that it removes all abnormal strong noise amplitudes that may occur with a small number of input traces and it is only less impacted by partly coherent signals that appear at the same time with the primary signal in less than half of the number of traces. With the TMS a given percentage of extreme amplitude values are excluded from the stack. This can be used to reduce the noise quota, as from interfering seismics events or experiments at the same time. Thus the TMS can, in extreme situations, result in decreased SNR compared to conventional stacking, when applied on data without significant noise. The Min/Max Stack is mostly used depending from the data and array facts, it most suitable for manual application. The SBS is a simple algorithm for excluding extreme values from stacking, without considering the real amplitudes. This method is only suitable with a large number of stackable traces. In other case even low noise quota will result in loss of information for interpretation. The SBS is most suitable for manual application. The RS is an alternative to conventional NMO corrected stacking. It can be used for velocity analysis. The ides of the RS is that is all primary amplitudes are the same, they should reside where the NMO corrections for the primary reflections are ideal. If signals are not all in phase the signal form can be destroyed. The RS destroys the noise signal form significantly whereas the primary reflections are conserved.

VI. Evaluation

Table I shows the matrix resulting for the investigated stacking methods and applicability regarding processing and computing with integrated systems for defined qualities:

- A: common method, batch use;
- B: combined use, mostly depending on other methods and pre- and post-processing;
- C: increased processing and resources requirements;
- D: parallelisation for the overall application;
- E: automation with integrated systems and workflows.

Table I
Stacking methods and qualities.

Stacking Method	*A*	*B*	*C*	*D*	*E*
Straight Stack, Mean Stack	✓	–	–	✓	✓
Stacking with predefined weighting					
by Muting	✓	✓	–	–	–
as Function of the Offset	✓	✓	–	–	–
as Function of the Array Response	✓	✓	–	–	–
Stacking with data adaptive weighting					
(Optimum) Weighted Stack	✓	✓	✓	–	–
Diversity Stack	✓	✓	✓	–	–
Coherency Stack	✓	✓	✓	–	–
Iterative and comparable methods					
Iterative Stack, Superstack, HDS	✓	–	✓	✓	✓
Single Trace Iterative Stack	✓	–	✓	✓	✓
Iterative Weighted Stack	✓	✓	✓	✓	✓
The N^{th}-Root Stack	✓	–	–	–	–
Other methods					
Median Stack, Alpha-Trimmed Mean	✓	✓	–	–	–
Trimmed Mean Stack	✓	✓	–	–	–
Min/Max Stack	✓	✓	–	–	–
Signed Bit Stack	✓	✓	✓	–	–
Random Stack	✓	✓	–	–	–

The stacking methods provide various algorithms handling signals and noise, suitable for the different nature of data. This goes along with different array properties as well as with the special attributes of the measurement. For all stacking methods discussed, errors in the calculated weighting can result in stacks that can be less suitable than the Straight Stack. They do have different geophysical focus and should not be seen competitive. The named stacking methods are mostly empirical and not thought as strong mathematical consequence of presumptions made [9]. Nevertheless some methods like the iterative Superstack algorithm use statistical and empirical information for the processing. The third category besides the “qualities” with integrated systems and their suitablity for geo-conditions is the order of magnitude for the number of instances (Table II) that will be used with one interactive application.

Table II
Number of instances with one interactive application task.

Instances	*Straight*	*Weighted*	*Iterative*	*Other*
Single	✗	✗	✗	✗
Several	✗	✗	(✗)	✗
Many	✗	(✗)	–	(✗)

The case studies have shown which compute intensive methods can be used with several instances per application, simpler methods may easily be used with thousands.

VII. Lessons Learned

Regarding the results, we can divide the methods into several main groups for use with different application scenarios. Common use: Batch use, this will work for mostly all methods. Human interaction and semi-manual use, like with the SBS and Min/Max Stack. This group is of common interest. Combined use: Methods demanding for combined use with additional methods like migration or various pre- and post-processing. These methods will be difficult to be integrated in automated processes or workflows. Integration into special application scenarios: Focus on iteration parallelisation, Straight Stack, Superstack, Fresnel Section support. These are especially interesting for future multi-dimensional measurement and processing. If the complexity is demanding large data or compute intensive processing this can be solved using the appropriate HEC architectures with integrated systems. The integration of secondary –geological and environmental– information with the workflow does provide benefits for the interpretation efficiency.

VIII. Conclusion and future work

The analysis of the different stacking methods has demonstrated that the applicability of the algorithms is very different for application scenarios regarding processing and computing with integrated systems and can be categorized by properties and purpose. Stacking can be used for integrated information systems as with dynamical concepts. For example if there is only uncorrelated noise the Straight Stack or Optimum Weighted Stack is most effective. If the interpretation needs to process data regarding increased sharpness in time in order to localise reflection elements, for deep structures this increasingly correlates with sharpness in space, then a Superstack can be the means of choice, even when it goes along with lateral smoothing of the trace domain. The application allows a smooth segmentation for using distributed resources. Some implemented methods have proved useful for application of dynamical processing. The Straight Stack can be efficiently integrated for standard information, loosely coupled parallel Superstack and HDS for depth analysis and low SNR, not concentrating on signal form. The implementation is computing intensive and has been vectorised for vector architectures and parallelised for parallel architectures and is as well suitable for automation and integration if appropriate HEC and HPC resources are available. Most of the other methods are suitable with stepwise workflows. In the future, the integration of suitable methods for depth data processing with information systems will be done according to these results.

Acknowledgements

I am grateful to all my scientific colleagues at the Westfälische Wilhelms-Universität Münster and the Leibniz Universität Hannover for their support as well as to international academic and industry partners in GEXI collaborations for innovative constructive test beds, IRI, HLRN, ZIV Münster, D-Grid, and the participants of the postgraduate EULISP Programme for prolific discussion and participation in case studies on integrated systems over the last years.

References

[1] K. Waters, *Reflection Seismology, A Tool for Energy Resource Exploration*, 3rd ed. A Wiley-Interscience Publication, John Wiley and Sons, N. Y., Chilchester, Brisbane, Toronto, Singapore, 1987.

[2] C.-P. Rückemann, “Implementation of Integrated Systems and Resources for Information and Computing,” in *Proceedings of the International Conference on Advanced Communications and Computation (INFOCOMP 2011), October 23–29, 2011, Barcelona, Spain*. XPS, Xpert Publishing Solutions, 2011, Rückemann, C.-P. (ed.), 6 pages, ISBN: 978-1-61208-161-8.

[3] C.-P. Rückemann, “Anwendung verschiedener Stapelmethoden auf gering überdeckende reflexionsseismische Daten aus der Heimefrontfjella, Antarktis,” Diplomarbeit, Institut für Geophysik, Westf. Wilhelms-Universität Münster, 1994.

[4] C.-P. Rückemann, “Application and High Performant Computation of Fresnel Sections,” *Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of the 9th International Conference on Numerical Analysis and Applied Mathematics (ICNAAM), September 19–25, 2011, Halkidiki, Greece, AIP Conference Proceedings 1389, Proceedings of the American Institute of Physics (AIP)*, 2011, ISBN: 978-0-7354-0956-9, URL: `http://proceedings.aip.org/` [accessed: 2011-08-21].

[5] C.-P. Rückemann, “Software Seismic Workshop (SSW),” *[Internet]*, 1995, URL: `http://www.user.uni-hannover.de/cpr/x/rprojs/en/index.html#SSW` (Project information).

[6] “Seismic Unix,” 2011, Center for Wave Phenomena, URL: `http://www.cwp.mines.edu/cwpcodes/` [accessed: 2011-08-21].

[7] “Applications with Active Map Software, Screenshots,” *[Internet]*, 2005, URL: `http://wwwmath.uni-muenster.de/cs/u/ruckema/x/sciframe/en/screenshots.html` [accessed: 2011-02-20].

[8] “Geo Exploration and Information (GEXI),” 1996, 1999, 2010, 2011, URL: `http://www.user.uni-hannover.de/cpr/x/rprojs/en/index.html#GEXI` [accessed: 2011-08-21].

[9] O. Naess and L. Bruland, *Stacking Methods other than simple Summation, in: Developments in Geophysical Exploration Methods*, F. (ed.), Ed. Elsevier Applied Science Publishers, 1985, vol. 6.

[10] K. Bullen and B. Bolt, *An Introduction To The Theory Of Seismology*. Cambridge Univ. Press, Cambridge, 1963, 1985.

[11] W. Mayne, “Common reflection point horizontal data stacking techniques,” *Geophysics, (Classic Papers), GE 50-11-1856*, vol. 50, no. 11, 1962.

[12] W. Mayne, “Practical considerations in the use of common reflection point techniques,” *Geophysics, GE 32-02-0225*, vol. 32, no. 2, p. 225, 1967.

[13] O. Naess, “Superstack – an iterative stacking algorithm,” *Geophysical Prospecting, GP 27-01-0016*, 1979.

[14] O. Naess, “Single-trace processing using iterative cdp-stacking,” *Geophysical Prospecting, 30, GP 30-05-0641*, pp. 641–652, 1981.

[15] E. Kanasewich, C. Hemmings, and T. Alpaslan, “Nth-root stack nonlinear multichannel filter,” *Geophysics*, vol. 38, pp. 327–338, 1973.

Automatic Classification of Points-of-Interest for Land-use Analysis

Filipe Rodrigues, Francisco C. Pereira, Ana Alves
Centre for Informatics and Systems of the University of Coimbra
University of Coimbra
Coimbra, Portugal
{*fmpr,camara,ana*}*@dei.uc.pt*

Shan Jiang, Joseph Ferreira
Department of Urban Studies and Planning
Massachusetts Institute of Technology
Boston, U.S.A.
{*shanjang,jf*}*@mit.edu*

***Abstract*—This paper describes a methodology for automatic classification of places according to the North American Industry Classification System. This taxonomy is applied in many areas, particularly in Urban Planning. The typical approach is to manually classify places/Points-of-Interest that are collected with field surveys. Given the financial costs of the task some semi-automatic approaches have been taken before, but they are still based on field surveys and official census. In this paper, we apply machine learning to fully automatize the classification of Points-of-Interest collected from online sources. We compare the adequacy of several algorithms to the task, using both flat and hierarchical approaches, and validate the results in the Urban Planning context.**

***Keywords*-machine learning; space analysis; points-of-interest; urban planning; GIS.**

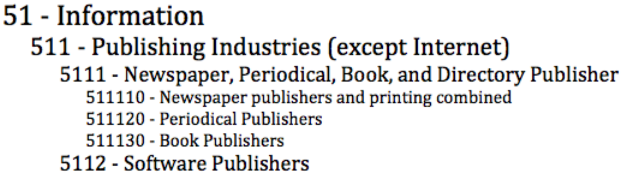
I. Introduction

A Point-of-Interest (or POI for short) is a specific point location that a considerable group of people find useful or interesting. POIs can be used in navigation systems, characterization of places, context-aware systems, city dynamics analysis, geo-referencing of texts, etc.

Despite its usefulness, the production of POIs is scattered across a myriad of different websites, systems and devices, thus making it extremely difficult to obtain an exhaustive database of such wealthy information. There are hundreds, if not thousands, of POI directories in the Web like Yahoo.com, Manta.com and YellowPages.com, each one using its own taxonomy of categories or tags. It is therefore essential to unify these different sources by mapping them to a common taxonomy, otherwise their application as a whole becomes impractical.

In this paper, we propose the use of machine learning techniques to automatically classify POIs from different sources to a standard taxonomy such as the North American Industry Classification System (NAICS) used in the U.S., Canada and Mexico, or the International Standard Industrial Classification (ISIC) used in the United Nations. Doing so is essential to allow a proper analysis of the POI data, especially when coming from different sources. A good example is the land-use analysis, which is a crucial task in Urban Planning. If the POIs do not share a common taxonomy we are not be able to determine, for instance, how many POIs of universities exist in a given area, since a POI source might classify them as "schools" and the other as "higher education". Although our approach would be similarly applicable to other classification standards, in this paper we are only interested in classifying POIs according to the North American Industry Classification System (NAICS).

The NAICS is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy [1]. The NAICS was adopted in 1997 to replace the old Standard Industrial Classification (SIC) system. It is a two to six-digit hierarchical classification code system, offering five levels of detail. Each digit in the code is part of a series of progressively narrower categories, and more digits in the code signify greater classification detail. The first two digits designate the economic sector, the third digit designates the sub-sector, the fourth digit designates the industry group, the fifth digit designates the NAICS industry, and the sixth digit designates the national industry. A complete and valid NAICS code contains six digits [2]. Figure 1 shows part of the NAICS hierarchy.



51 - Information
- **511 - Publishing Industries (except Internet)**
  - **5111 - Newspaper, Periodical, Book, and Directory Publisher**
    - 511110 - Newspaper publishers and printing combined
    - 511120 - Periodical Publishers
    - 511130 - Book Publishers
  - **5112 - Software Publishers**

Figure 1. Example of the NAICS hierarchy

After comparing several classification methods, we apply the results to the urban modeling task of estimating employment size at a disaggregated level. This task is traditionally made at a coarser level (Traffic Analysis Zone, Census Tract or Block Group level) than what could be now possible.

To the authors best knowledge, there is no previous work that automatically classifies POIs into the NAICS taxonomy. This is our main contribution.

The rest of this paper is organized as follows. Section II presents previous related studies. Section III explains our data analysis and modeling methodology, from data preparation to model generation and validation. Section IV shows the obtained results. In Section V we describe an application of this methodology to the field of Urban Planning. We finish the paper with some conclusions and further work.

II. State of the Art

The applications of machine learning algorithms in classification tasks are vast and cover diverse areas that range from Speech Recognition to Medicine, including forecasting in Economics and Environmental Engineering or Road Traffic Prediction. On the other hand, in Urban Planning, land-use/land-cover information has long been recognized as a very important material [3]. However, as Fresco [4] claimed, accurate data on actual land-use cannot be easily found at both global/continental and national/regional scales. In order to cope with these problems, automatic approaches to classify land-use are being developed using different techniques usually based on machine learning algorithms.

A common approach to infer land-use/land-cover is to use satellite imagery. However, while these approaches have already proven to get good results, they are more suited to land-cover inference, which is considered somehow different from land-use by many authors. Campbell [5], for example, considers land-cover to be concrete whereas land-use is abstract. That is, land-cover can be mapped directly from images, while land-use requires land-cover and additional information on how the land is used. Danoedoro [6] tries to improve land-use classification via satellite imagery by combining spectral classification, image segmentation and visual interpretation. Although he showed that satellite imagery could be used for generating socio-economic function of land-use at 83.63% accuracy, he is the first to recognize that applying such techniques to highly populated areas would be problematic.

Li et al. [7] use data mining techniques to discover knowledge from GIS databases and remote sensing image data that could be used for land-use classification. Using the C5.0 algorithm they get an accuracy of 89% in land-use classification.

An alternative to satellite imagery is the POI data. Using a large commercial POI database, Santos and Moreira [8] create and classify location contexts using decision trees. They identify clusters by means of a density-based clustering algorithm, which allow them to define areas (or regions) through the application of a concave hull algorithm they developed to the POIs within each cluster. Finally, making use of the C5.0 algorithm, they classify a given location according to such characteristics as the number of POIs in a cluster, the size of the area of the cluster and the categories of the POIs within the cluster.

In order to use POI data for the classification of places and land-use analysis, POI classification is an essential task. Griffin et al. [9] use decision trees to classify GPS-derived POIs. However, they refer to POIs as "personal" locations to a given individual (i.e., home, work, restaurant, etc.). The main goal of their approach is then to automatically classify trips. In their approach, they start by determining clusters of trip-stops (i.e., stops that took more than 5 minutes) using a density-based clustering algorithm (Dbscan). Then, they make use of the C4.5 algorithm to classify the generated clusters as being "home", "work", "restaurant", etc., based on the time of the day and the length of the stay. However, no previous approaches have been made to classify POIs to a classification system such as NAICS. The latter is widely used for industry classification and has already been used, for instance, to classify Web Sites through machine learning techniques [10].

Spatial analysis has long been a topic of interest for researchers, who seek a comprehensive understanding on how the city behaves in different perspectives and its impact in the economy. Methods for analyzing spatial (and space-time) data have already been well developed by statisticians [11] and econometricians [12]. An interesting example is provided by Currid et al. [13], who try to understand the importance of agglomeration economies as a backbone to urban and regional growth, by identifying clusters of several "advanced" service sectors (professional, management, media, finance, art and culture, engineering and high technology) and comparing them in the top ten populous metropolitan areas in the U.S.

III. Approach

In this section we describe our approach, particularly what are the sources of our POI data, how we generate the training data, what methods we use for classification and how we perform validation.

A. POI Sources

Our data consists of a large set of POIs extracted from Yahoo! through their public API, another set acquired from Dun & Bradstreet (D&B) [14], a consultancy company that specializes in commercial information and insight for businesses, and a third one from InfoUSA.com provided by the Harvard Center for Geographic Analysis (ESRI Business Analyst Data). In the first data set (from Yahoo!), the database is essentially built from user contributions. In the other two the data acquisition process is semi-automatic and involves integration of official and corporate databases, statistical analysis and manual evaluation [14]. The POIs from D&B and InfoUSA have a NAICS code assigned (2007 version), which is not present in Yahoo!. However, each POI from Yahoo! is assigned, in average, roughly two arbitrary categories from the Yahoo! categories set. These categories are specified by the user, through a textfield and can be rather disparate since Yahoo! forces no restrictions over them, thus they can be seen as mere tags. Considering that every POI

ISBN: 978-1-61208-178-6

source provides either some categories or tags associated with their POIs, we take advantage of this information to classify them to the NAICS, where a single unifying code is assigned to each POI.

Our dataset contains 156364 POIs from Yahoo!, 29402 from D&B and 196612 from InfoUSA for the greater metropolitan area of Boston, Massachusetts. We also used 331118 POIs from Yahoo! and 16852 from D&B for the New York city area to see how our previously trained model would perform in a different city. We estimate that the Yahoo!'s categories taxonomy has more than 1300 distinct categories distributed along a 3-level hierarchy. On the other hand, NAICS has a total of 2332 distinct codes distributed along their 6-level hierarchy (1175 only in the sixth level).

Given its nature, the growth of the Yahoo! database (or any other user-content platform) is considerably faster than D&B and InfoUSA, and the POI categorization follows less strict guidelines, which in some cases, as mentioned before, may become subjective. This dynamic nature of these internet POI sources, together with the fact that they are publicly available to anyone and usually cover entire countries, make them extremely attractive. Our hypothesis is that there is considerable coherence between Yahoo! categories and NAICS codes, such that a model can be learned that automatically classifies incoming Yahoo! POIs.

B. POI Matching and Data Preparation

In order to generate training data for the machine learning algorithms we use a *POI Matching* algorithm, which compares POIs according to their name, Web Site and distance. It makes use of the JaroWinklerTFIDF algorithm [15] to identify close names, ignoring misspelling errors and some abbreviations. We set the similarity thresholds to high values in order to get only high confidence matches. By manually validating a random subset of the POI matches identified (6 sets of 50 random POIs assigned to 6 volunteers), we concluded that the percentage of correct similarities identified was above 98% ($\sigma = 1.79$). Differently to validations later mentioned in this paper, this is an extremely objective one, not demanding external participants or a very large sample[1].

After matching Yahoo! POIs to D&B and InfoUSA, we built two different geographic databases, where each POI contains a set of categories from Yahoo! and a NAICS classification provided by D&B and InfoUSA respectively. From this point on, we shall refer to the initial dataset, which results from POI matches between Yahoo! and D&B, as dataset A, and to the dataset resultant from the POI matching between Yahoo! and InfoUSA as dataset B. The later is six times larger than the former, due to larger coverage of InfoUSA in Boston.

Table I shows some statistic details of both datasets used.

Table I
SOME STATISTICS OF DATASETS A AND B

Dataset	A	B
NAICS source	D&B	InfoUSA
Total POIs	7289	44634
Distinct NAICS	504	689
Distinct Yahoo! categories	802	1109
Distinct Yahoo! category combinations	569	1002
Category combinations that appear only once	136	92
Categories that appear only once	181	107
NAICS that appear only once	115	96

The dataset A contains 7289 POIs for Boston and Cambridge and 2415 for New York. In comparison with the original databases, these are much smaller sets due to a very conservative POI matching approach (string similarity of at least 80%, max distance of 80 meters). However the POI quantities are high enough to build statistically valid models. We performed a detailed analysis of this data and identified 569 different category combinations, which included only 802 distinct categories from the full set (of over 1300). From D&B, our data covers 504 distinct six-digit NAICS codes. However, the 2007 NAICS taxonomy has a total of 1175 six-level categories, meaning that our sample data only covers some of the most common NAICS codes, which only represents about 43% of the total number of the NAICS categories. Nevertheless, the remaining ones are more exotic in our context and hence less significant for posterior analyses.

Further analyses on the coherence between NAICS and Yahoo! showed that only in 80,2% of the POIs in dataset A the correspondent NAICS was consistent with the most common one for that given set of categories, which means that about one fifth of the POIs are incoherent with the rest of the sample. This fact highlights the problem of allowing users to add arbitrary categories to their POIs without restrictions. For different NAICS levels, particularly for two-digit and four-digit NAICS, the same analyses showed, as expected, a higher level of coherency. For the two and four-digit NAICS, 87,1% and 83,4% of the POIs, respectively. Therefore, by having the same set of Yahoo! categories mapping to different NAICS codes in different occasions, it is not expectable that we obtain a perfect model that correctly classifies all test cases. In order to understand the impact of these inconsistencies in the results, we also modified the POI dataset so that the NAICS code of a given POI would match the NAICS codes of the other POIs with the same category set, assigning to each POI the most common NAICS code for that given category set in the dataset. The results of this separate experiment are also presented in Section IV.

Tables II and III show, respectively, the five most common NAICS and Yahoo! categories we identified in dataset A.

Regarding dataset B, we identified 689 distinct NAICS

[1] Using the central limit theorem, the standard error of the mean should be near 0.73. Assuming an underestimation bias for n=6 of 5% (by the [16]), accuracy keeps very high, yielding a 95% confidence interval of [96.5%, 98.7%]

Table II
MOST COMMON NAICS IN THE DATASET A

NAICS code	Description	Occurrences
423730	Warm Air Heating and Air-Conditioning Equipment and Supplies Merchant Wholesalers	707
446130	Optical Goods Stores	200
314999	All Other Miscellaneous Textile Product Mills	193
493120	Refrigerated Warehousing and Storage	136
332997	Industrial Pattern Manufacturing	123

Table III
MOST COMMON YAHOO! CATEGORIES IN THE DATASET A

Yahoo! category	Occurrences
Salons	157
All Law Firms	129
Government	116
Trade Organizations	115
Architecture	86

Table IV
BRIEF DESCRIPTION OF THE ALGORITHMS TESTED

Implementation	Description
ID3	Unpruned decision tree based on the ID3 algorithm.
C4.5	Pruned or unpruned C4.5 decision tree.
C4.5graft	Grafted C4.5 decision tree.
RandomForest	Forest of random trees, i.e., trees with K randomly chosen attributes at each node.
JRip	Propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), as proposed by W. Cohen as an optimized version of IREP.
IBk	K-nearest neighbors classifier that can do distance weighting.
IB1	1-nearest-neighbor classifier. Simplification of IBk.
K*	K* is an instance-based classifier. The class of a test instance is determined from the class of similar training instances . It uses an entropy-based distance function.
BayesNet	Bayesian Network
NaiveBayes	Naive Bayes model

codes and 1109 distinct categories of the more than 1300 that we found in Yahoo!. The latter are in larger number than the ones from dataset A (only 802) and therefore dataset B provides a better coverage of the source taxonomy. The number of distinct category combinations almost doubled when compared to dataset A, which leads to more diversity in the training data and probably to more accurate classifiers.

C. Flat Classification

The “flat classification” task corresponds to directly assigning a NAICS code to a POI given its “bag” of Yahoo! categories. It is “flat” because the inherent hierarchy of the NAICS is not taken into account in the classification model. Each NAICS code is simply seen as an isolated string “tag” that is assigned to a POI.

We experimented various machine learning algorithms for this particular classification task. Table IV provides a brief description of the algorithms we tested. It is beyond the scope of this paper to describe any of the algorithms in detail. The interested reader is redirected to dedicated literature (e.g., [17], [18]).

In our experiments we built classifiers for different NAICS levels (i.e., NAICS categories with different granularities), particularly two, four and six-digit NAICS codes. This choice is typical in Urban Planning depending on the study at hand (e.g., level 2 allows to analyze economic sectors, while level 6 goes to the level of the establishment specificities).

For validation purposes we use ten-fold cross-validation [17], [18]. We also performed validation with an external test set (data from a another city, New York) to understand the dependency of the generated models on the study area.

D. Hierarchical Classification

In this approach we take advantage of the hierarchical structure of the NAICS, thus the overall classifier is itself a hierarchy of classifiers. In this hierarchy each classifier decides what classifier to use next, narrowing down the NAICS code possibilities on each step, until a final 6-digit code (or 4-digit code, depending on the goal) is achieved. Figure 2 depicts one possible hierarchy.

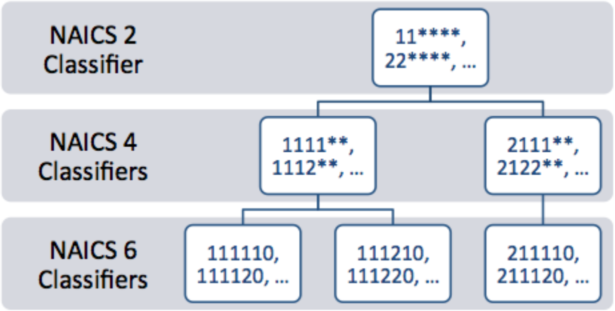


Figure 2. A possible hierarchy of classifiers

By looking at the hierarchy above, we can see that it has 3 levels (2, 4 and 6-digit NAICS). The first level always consists of a single classifier that decides which NAICS economic sector (2-digit code) the POI belongs to. Taking the sector into account, the algorithm then decides which classifier to use next at the second level. After that, the same process repeats until a leaf node is achieved in the tree structure of the hierarchy of classifiers. To provide an example consider a POI that has the following NAICS code: 111110. According to Figure 2 the top-level classifier will decide that it belongs to sector 11 (”Agriculture, Forestry, Fishing and Hunting”) and the left-most level 2 classifier will be used next. Then, this classifier will determine that the 4-digit NAICS code of the POI is 1111 (”Oilseed and Grain Farming”) and, based on this decision, the left-most classifier in the third level of the figure will be used, and will supposedly classify the POI with the NAICS code 111110 (”Soybean Farming”). Of course along this top-down process a mistake can be made by one classifier. In this case, the error would propagate downwards and there would be no way to recover from it, and hence the final NAICS code would be wrong.

Our hypotheses is that by using a hierarchy of classifiers, the classification task will be divided into several classification models, each one less complex, more accurate and dealing with a simpler problem. If we consider, for example, the ID3 algorithm, the entropy values for the different features will be computed according to a smaller class subset, and therefore the selection of the next feature to use (which is based on the entropy calculation) will be different and the resulting tree will also be different. Hopefully, the generated classifier will be more suited to that particular classification (like deciding for a POI if it belongs to the subcategory 531, 532, etc, knowing that it belongs to the NAICS sector 53).

In our experiments we use three different hierarchies of classifiers, two with 2 levels:

- NAICS 2 and NAICS 4
- NAICS 2 and NAICS 6

and other one with 3 levels:

- NAICS 2, NAICS 4 and NAICS 6

As we did for the flat classification, we also tried to test different types of machine learning algorithms: bayesian networks, tree-based learners, instance-based learners and rule-based learners. Neural networks were not possible to test due to their computational demands, both in processing power and memory.

For the hierarchical approaches we also perform ten-fold cross-validation, but the data splitting between training/testing is more prone to biased results than with standard flat classification. As in normal ten-fold cross-validation, we also start by leaving 10% of the data out for test and use the remaining 90% for training, repeating this process ten times. However, each classifier in a given level only receives the part of those 90% of training data that respects to it. For instance, a level two classifier for deciding which subcategory of the NAICS sector 53 a given POI belongs to would only be trained with POIs that belong to that NAICS sector. Hence, the only classifier that receives all the training data (90%) would be the top-level classifier (i.e., the one that decides which NAICS sector a POI belong to). After the training phase, the hierarchy is tested with the 10% of the data left out. This process is repeated ten times, and the average accuracy over the ten iterations is determined.

IV. Results

Table V shows the accuracies obtained using different machine learning algorithms in a “flat” setting for different NAICS levels (two, four and six-digit codes) for dataset A. We can see that the tree-based (e.g., ID3, RandomForest) and instance-based learning approaches (e.g., IBk, K*) are the ones that perform better in this classification task, especially the latter. Notice that at the sixth-level only 80,2% of the NAICS codes in the data were assigned in a totally non-ambiguous way. The most successful algorithm is IBk (with k=1), which essentially finds the similar test case and assigns the same NAICS code. The difference in accuracy between tree-based and instance based approaches is too small to conclude which one outperforms the other, however we could expect that instance based models bring better results since the distribution of the different Yahoo! categories is relatively even among examples of the same NAICS code (implying no clear “dominance” of some categories over others). Understandably, the Naive Bayes algorithm performs badly because the assumption that different Yahoo! categories for the same NAICS classification are independently distributed is obviously false (e.g., “Doctors & Clinics, Laboratories, Medical Laboratories” are correlated). Such assumption is not fully necessary in Bayesian Networks, which actually brings better results. Unfortunately, we could not find a model search algorithm that performs in acceptable time (less than 72 hours) and produces a more accurate model. We used Simulated Annealing and Hill Climbing.

Table V
Accuracies obtained by different machine learning algorithms with POIs from dataset A for the Boston area

Algorithm	NAICS2(kappa)	NAICS4(kappa)	NAICS6(kappa)
ID3	85.495 (0.842)	77.955 (0.776)	74.015 (0.737)
C4.5	84.241 (0.828)	77.630 (0.772)	73.071 (0.727)
Random Forest	86.174 (0.849)	79.298 (0.789)	74.753 (0.744)
JRip	81.334 (0.795)	74.340 (0.737)	69.264 (0.686)
IB1	82.736 (0.812)	74.266 (0.738)	68.644 (0.683)
IBk	86.646 (0.854)	79.475 (0.791)	75.343 (0.750)
K*	85.702 (0.844)	79.726 (0.794)	75.387 (0.751)
BayesNet	80.950 (0.790)	56.721 (0.554)	45.064 (0.438)
NaiveBayes	74.399 (0.715)	40.446 (0.382)	30.264 (0.283)

As expected, we obtained better results classifying POIs with two-level NAICS codes than with the six-level NAICS codes, since the noise due to ambiguous NAICS codes assignments in the POI dataset is smaller (we now have 87.1% of non-ambiguous cases; see Section III-B).

In Table VI we can see the results obtained by changing the POI dataset A so that the NAICS codes of POIs where ambiguities arise are grouped together in the same “super-category”, eliminating the inconsistencies.

Table VI
Accuracies obtained by different machine learning algorithms using a re-classified version of dataset A

Algorithm	NAICS2	NAICS4	NAICS6
ID3	92.975	89.728	88.680
RandomForest	93.609	90.805	89.846
IBk	94.170	91.189	89.979

By comparing the results in Table VI with the results in Table V, we realize that the NAICS labeling inconsistencies in the POI data have a major negative effect in the performance of the machine learning algorithms, reducing the accuracy in more than 16% in some cases for the six-level NAICS codes. This also gives indications for future versions of the NAICS, where some categories may become aggregated according to these “super-categories”.

It would be expectable to obtain accuracies closer to 100% for the results in Table VI. However, that does not happen since 115 of the 514 NAICS codes covered by our dataset A only occur once. Therefore, when we split the dataset to perform the ten-fold cross-validation, a significative number of the test cases will have NAICS codes that were not observed during training, causing the algorithm to incorrectly classify them.

Table VII shows the results we obtained by training the machine learning approaches with dataset A from Boston and Cambridge and testing them with New York POI data. As we can see in the results, if we apply the generated model to a different city, it still performs well, even though the accuracy drops a small amount in some cases. This is understandable since even the Yahoo! taxonomy differs slightly from city to city.

Table VII
ACCURACIES OBTAINED BY DIFFERENT MACHINE LEARNING ALGORITHMS USING POI DATA FROM BOSTON FOR TRAINING AND POI DATA FROM NEW YORK FOR TESTING

Algorithm	NAICS2	NAICS4	NAICS6
ID3	85.061	75.586	70.209
RandomForest	85.488	76.867	71.318
IBk	85.360	76.909	71.276

Table VIII shows the results obtained for the different machine learning algorithms using dataset B. By analyzing these results we can see that the classification accuracies have significantly improved over dataset A, which shows the importance of the training data size in the performance of the machine learning algorithms.

Table VIII
ACCURACIES OBTAINED BY DIFFERENT MACHINE LEARNING ALGORITHMS WITH POIS FROM DATASET B FOR THE BOSTON AREA

Algorithm	NAICS2(kappa)	NAICS4(kappa)	NAICS6(kappa)
ID3	90.567 (0.897)	85.459 (0.852)	82.091 (0.819)
C4.5	90.113 (0.800)	85.085 (0.849)	81.831 (0.816)
RandomForest	90.758 (0.899)	85.710 (0.855)	82.436 (0.823)
JRip	85.748 (0.844)	80.998 (0.807)	78.495 (0.780)
IB1	87.224 (0.861)	81.495 (0.812)	76.826 (0.766)
IBk	91.024 (0.902)	85.974 (0.858)	82.553 (0.824)
K*	90.227 (0.893)	85.849 (0.856)	82.522 (0.824)
BayesNet	88.961 (0.880)	77.964 (0.776)	67.877 (0.675)
NaiveBayes	87.910 (0.868)	70.250 (0.696)	56.052 (0.554)

Finally Tables IX to XI show the results obtained using the different hierarchical classification schemes for various types of machine learning algorithms. There are some missing results in the cases where the algorithm took over 72 hours to run.

Table IX
COMPARISON BETWEEN THE RESULTS FOR DATASET B USING FLAT CLASSIFICATION (4-DIGIT NAICS) AND HIERARCHICAL CLASSIFICATION WITH 2 LEVELS (NAICS 2 AND 4)

Algorithm	Flat classification accuracy	Hierarchical classification	
		Level1 acc.	Level2 acc.
ID3	85.459	90.659	85.620
C4.5	85.085	90.172	84.901
RandomForest	85.710	90.959	85.969
JRip	80.998	85.806	80.440
IB1	81.495	87.637	81.126
IBk	85.974	91.080	86.097
K*	85.849	90.305	85.244
BayesNet	77.964	88.002	74.243
NaiveBayes	70.250	30.688	20.091

Table X
COMPARISON BETWEEN THE RESULTS FOR DATASET B USING FLAT CLASSIFICATION (6-DIGIT NAICS) AND HIERARCHICAL CLASSIFICATION WITH 2 LEVELS (NAICS 2 AND 6)

Algorithm	Flat classification accuracy	Hierarchical classification	
		Level1 acc.	Level2 acc.
ID3	82.091	90.659	82.100
C4.5	81.831	90.173	81.484
RandomForest	82.436	90.959	82.477
JRip	78.495	85.806	76.398
IB1	76.826	87.637	76.826
IBk	82.553	91.080	82.551
K*	82.522	90.305	81.661
BayesNet	67.877	89.059	69.336
NaiveBayes	56.052	88.002	59.885

Table XI
COMPARISON BETWEEN THE RESULTS FOR DATASET B USING FLAT CLASSIFICATION (6-DIGIT NAICS) AND HIERARCHICAL CLASSIFICATION WITH 3 LEVELS (NAICS 2, 4 AND 6)

Algorithm	Flat classification accuracy	Hierarchical classification		
		Level1 acc.	Level2 acc.	Level3 acc.
ID3	82.091	90.659	85.620	82.111
C4.5	81.831	90.172	84.901	81.341
Random Forest	82.436	90.959	85.969	82.398
JRip	78.495	85.806	80.440	76.889
IB1	76.826	87.637	81.126	76.826
IBk	82.553	91.080	86.097	82.539
K*	82.522	90.305	85.244	81.486
BayesNet	67.877	-	-	-
NaiveBayes	56.052	-	-	-

Our intuition was that hierarchical classification would perform generally better than standard flat classification. However, only in some algorithms the results improved. Therefore, we will not argue that hierarchical classification of POIs into the NAICS is always a better solution. In fact, as shown before by comparing the datasets A and B, the quality and the dimensions of the dataset seems to have a much bigger impact on the results than whether we apply hierarchical or flat classification.

Another interesting fact in the results from the hierarchical classification is that the accuracies vary considerably with the hierarchy type used. For instance, when classifying POIs with 6-digit NAICS codes, we can see that using a two-level hierarchy the RandomForest algorithm improved over the flat classification, while using a three-level hierarchy it became worse (although the differences in accuracy are small). One of the possible cause for this, is that the hierarchy type used directly affects the number of training instances at each node of the hierarchy tree, and depending on the machine learning algorithm, the number of training instances will have different impacts on the results.

V. An application in Urban Planning

In this section we describe a practical application of Yahoo! POIs classified to the NAICS using a non-hierarchical approach with the k-nearest neighbor classifier (see Section III-C for more details).

In the field of Urban Planning, urban simulation models have evolved significantly in the past several decades. For instance, the travel demand modeling approach has been evolveing from the traditional Four-Step Model (FSM) to the Activity-Based Model (ABM) [19]. Consequently, requirements for disaggregated data increase greatly, ranging from population data, employment data, to travel survey data. The employment data (on the travel destination side) is usually obtained from proprietary sources, which adds another layer of barriers to widely applying the Activity-Based Modeling approach, let alone the expensive travel-survey data acquisition. In order to study this issue, researchers are trying to develop new methods of estimating disaggregated employment size and location by category.

In our case, we intend to develop a set of new methods and demonstrate their applications for estimating activities, incorporating them into travel demand and urban simulation models. This will be beneficial for cities that lack detailed survey data for building Activity-Based Models but wish to test the sensitivity of travel behavior to policy changes such as Intelligent Transportation Systems (ITS) implementations that are likely to alter activity patterns. An important step to achieve these goals is to obtain a disaggregated employment distribution by POIs of an area. For the case of Cambridge, MA, we have official data at the Block Group (BG) level (obtained from the U.S. Census Transportation Planning Package 2000), which essentially describes the total size of employees by economic sector at that spatial resolution. We need to distribute these totals into Block or Parcel level.

For demonstration purposes we only use POIs from the "Retail Trade" sector of the NAICS taxonomy, i.e., categories whose code starts by 44 or 45. Figures 3 and 4 show the aggregated retail employment density at the Block Group level and distribution of our POI data from Yahoo! at the Census Block level for Cambridge, respectively.

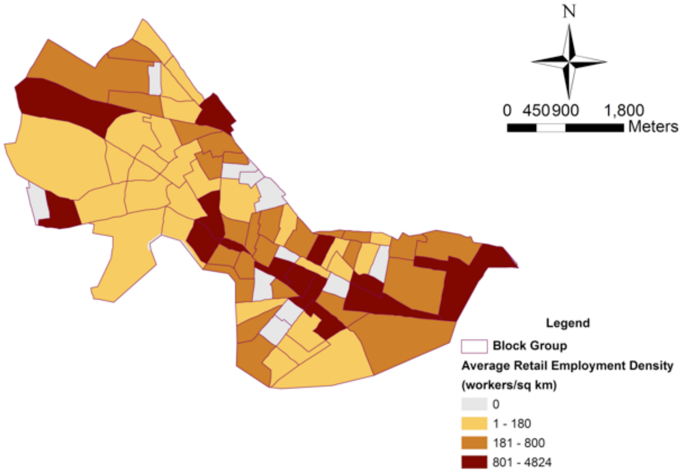


Figure 3. Aggregated retail employment density at the Block Group level (pl/sq km= employed people per square kilometer).

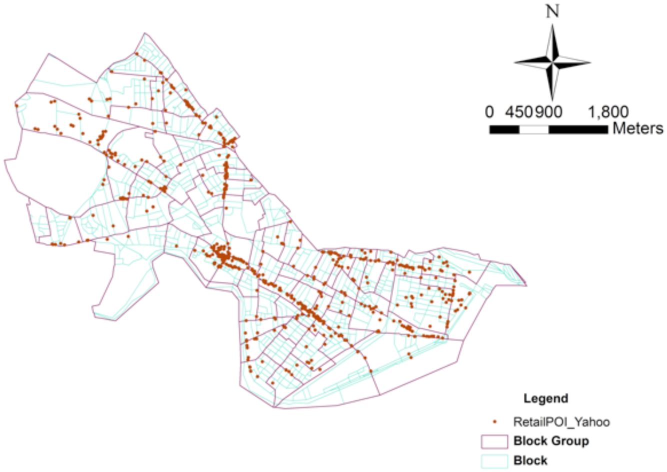


Figure 4. Cambridge retail POI distributions from Yahoo!

By using the business establishment survey data (from InfoUSA, 2007), which is believed to be close to the population, we are able to obtain a benchmark estimate of employment size by category at the Census Block level for the study areas. This will function as a ground truth to test our algorithm. Notice however that the dates for each of the databases are quite distinct (2000 for Census, 2007 for InfoUSA and 2010 for Yahoo!) therefore some error is expected to happen.

We employ the local maximum likelihood estimation (MLE) method as described below to derive the disaggregated destination estimation at Block level.

1) We calculate the total number of POIs (destinations) by category c in each Block b.

2) We assume that the employment size at destination d in Block Group g of category c is proportional to some function f of its associated block area $a_{d,c,g}$, which means the effective area of the destination d in Block Group g of category c. The form of function f will be explored based on the empirical data, and we also allow the possibility that $f(a_{d,c,g}) = a_{d,c,g}$, which is the natural benchmark case. Mathematically, assume that for employment category c, there are $n_{c,g}$ destinations at Block Group g. For $d = 1, 2, \ldots, n_{c,g}$, let the random variable $e_{d,c,g}$ be the employment size of category c at destination d in Block Group g.
3) We assume that $e_{d,c,g}(d = 1, 2, \ldots, n_{c,g})$ are i.i.d.$(f(a_{d,c,g}) \cdot \alpha_{c,g}, \sigma^2_{c,g})$, where $\alpha_{c,g}$ is the employment size of category c per unit of effective area at Block Group g; $\alpha_{c,g}$ and $\sigma_{c,g}$ are positive constants independent of d. $E(e_{d,c,g}) = f(a_{d,c,g}) \cdot \alpha_{c,g}$ and $Var(e_{d,c,g}) = \sigma^2_{c,g}$. We then estimate $\alpha_{c,g}$ by employing the maximum likelihood method locally at Block Group g for employment category c. Thus we obtain an estimate of employment size $e_{d,c,g}$ of category c at destination d in Block Group g.
4) Finally, we sum up the employment size in category c in Census Block b in Census Block Group g.

By employing the same local maximum likelihood method described above and using the business establishment survey data (e.g., ESRI Business Analysis package), which is believed to be close to the population POIs, we obtain a benchmark estimate of employment size by category at the Block level for the study area, $E^*_{b,c,g}$. By using the derived POI information (obtained from the machine learning algorithm), we obtain an estimate of employment size by category c at Block b for the study area, $\hat{E}_{b,c,g}$.

Then the mean squared error (MSE), weighted mean squared error (WMSE), and the relative weighted mean squared error (RWMSE) can be calculated to evaluate the goodness of fit of the model (see Equations 1, 2, 3, and 4).

$$MSE(\hat{E}_{b,c,g}, E^*_{b,c,g}) = \sum_{b,c,g} (\hat{E}_{b,c,g} - E^*_{b,c,g})^2 \tag{1}$$

$$WMSE(\hat{E}_{b,c,g}, E^*_{b,c,g}) = \sum_{b,c,g} w_{b,c,g} (\hat{E}_{b,c,g} - E^*_{b,c,g})^2 \tag{2}$$

$$RWMSE(\hat{E}_{b,c,g}, E^*_{b,c,g}) = \frac{\sum_{b,c,g} w_{b,c,g} (\hat{E}_{b,c,g} - E^*_{b,c,g})^2}{\sum_{b,c,g} w_{b,c,g} (\bar{E}_{b,c,g} - E^*_{b,c,g})^2} \tag{3}$$

$$\bar{E}_{b,c,g} = \frac{w'_{b,g} \sum_q E^*_{q,c,g}}{\sum_q w'_{q,g}} \tag{4}$$

Weights $\{w_{b,c,g}\}$ are normalized to reflect the proportion of each Census Block in the whole map. In Equation 2, when we take the weight $w_{b,c,g} = 1$ for any subscripts b, c, and g, the corresponding WMSE becomes MSE. In Equation 4, $w'_{b,g}$ = area of Block b in Block Group g, and $\bar{E}_{b,c,g}$ is the estimated employment size in Block b of category c, using the traditional disaggregation approach, assuming that the employment is uniformly distributed across blocks in each Block Group g.

If RWMSE is less than 1, it means that the quality of the derived POIs is reliable, so is the new method; the smaller the RWMSE, the more accurate is the method. If WMSE or RWMSE equals to 0, it means that the derived POIs from the Internet match exactly with the trusted proprietary POIs (treated as the population POIs). However, if RWMSE is greater than 1, it means that the derived POIs cannot well reflect the distribution of the population POIs.

Figures 5 and 6 show the estimation results of the disaggregated retail employment density at Block level in Cambridge, MA, by using POIs from infoUSA and Yahoo! respectively. By comparing the estimation results, we find that the disaggregated employment estimations by using the POIs captured from the Internet using Yahoo! and those obtained from the proprietary source (infoUSA 2007) are very close.

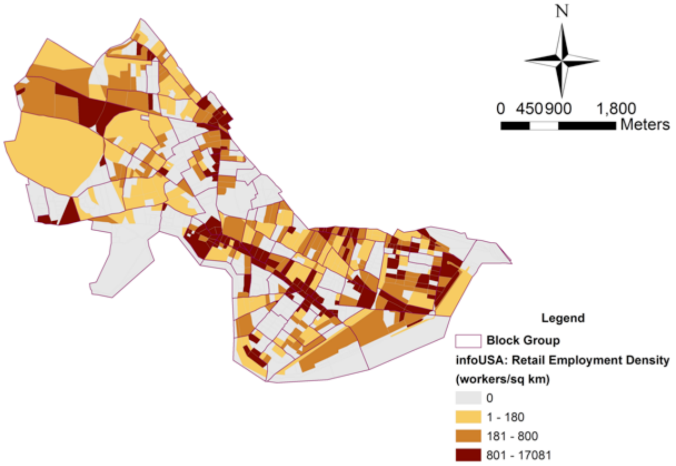


Figure 5. Disaggregated retail employment densities at the Block level, in Cambridge, MA, by using POIs from infoUSA

Employing Equation 3, the disaggregated employment estimation at the Block level using Yahoo! POI gives RMSE = 0.312. The RMSE is significantly smaller than 1, which means that using the extracted Yahoo! online POIs to estimate the disaggregated employment sizes at the Block level has reduced the mean squared error by around 69% compared to the traditional average disaggregation approach.

VI. Conclusion

In this paper, we showed that it possible to classify POI to the widely used NAICS system with several different machine learning algorithms using only the categories or tags that are commonly associated with them. We matched two different POI databases (InfoUSA and Dun & Bradstreet) to Yahoo!, in order to build two reliable training sets that have POIs with user provided bags of categories

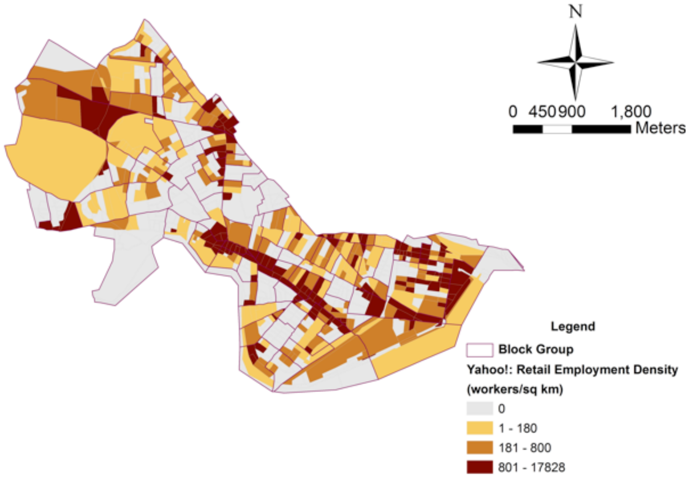


Figure 6. Disaggregated retail employment densities at the Block level, in Cambridge, MA, by using POIs from Yahoo!

classified with NAICS codes. We tested several classification algorithms and the results show that the best approaches for this particular task are inductive based algorithms, namely instance based and tree based learning. These allow for an accuracy as high as 82% in the most complex task (classification with 6-digit NAICS codes). We also tried to perform classification in a hierarchical way, however the results did not showed many improvements over the flat approaches, leading us to the conclusion that the size of the training set and its consistency/quality can have a larger impact on the results than the classification algorithm itself (except maybe for Bayesian approaches).

The classified POIs were applied to the urban modeling task of employment size and location disaggregation from Block Group level to Block level and the results show encouraging quality. This strengthens the idea that well classified POI data to a convenient taxonomy like the NAICS is of great use and can have many distinct applications.

To the authors best knowledge, this is the only work that proposes an automatic approach for classifying POIs to the NAICS, and therefore a comparison with other works is not possible. Thus, we contribute with a novel approach to this important problem that has high impact in urban modeling and space classification.

References

[1] U. C. Bureau. North american industry classification system (naics): Introduction, <Retrieved: November 2011>. http://www.census.gov/eos/www/naics/.

[2] N. Association. NAICS association: FAQ, <Retrieved: November 2011>. http://www.naics.com/faq.htm.

[3] D. T. Lindgren. *Land-use Planning and Remote Sensing*. Martinus-Nijhoff, Boston, MA, 1985.

[4] L. O. Fresco. *The Future of the Land – Mobilizing and Integrating Knowledge for Land-use Options*. John Wiley & Sons, Chichester, 1997.

[5] J. B. Campbell. *Mapping the Land – Aerial Imagery for Land use Information*. Association of American Geographers, Washington, D.C., 1983.

[6] P. Danoedoro. Extracting land-use information related to socio-economic function from quickbird imagery: A case study of semarang area, indonesia. *Map Asia 2006*, 2006.

[7] D. Li, K. Di, and D. Li. Land use classification of remote sensing image with GIS data based on spatial data mining techniques. *Geo-Spatial Information Science*, pp. 30-35, 2000.

[8] M. Santos and A. Moreira. Automatic classification of location contexts with decision trees. *CSMU-2006 : Proceedings of the Conference on Mobile and Ubiquitous Systems, Guimares, Portugal*, pp. 79-88, 2006.

[9] T. Griffin, T. Huang, and R. Halverson. Computerized trip classification of GPS data. *Proceedings of 3rd International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA 2006)*, pp. 22-30, 2006.

[10] J. Pierre. On the automated classification of web sites. *Linkoping Electronic Articles in Computer and Information Science*, 6, pp. 1-12, 2001.

[11] R. P. Haining. *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge, 1990.

[12] L. Anselin and R. Florax. *New directions in spatial econometrics*. Springer, New York, 1995.

[13] E. Currid and J. Connolly. Patterns of knowledge: The geography of advanced services and the case of art and culture. *Annals of the Association of American Geographers*, pp. 414-434, 2008.

[14] D. . Bradstreet. D & B Website, <Retrieved: November 2011>. http://www.dnb.com/.

[15] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03), Acapulco, Mexico*, pp. 73-78, 2003.

[16] J. Gurland and R. Tripathi. A simple approximation for unbiased estimation of the standard deviation. *American Statistician*, pp. 30-32, 1971.

[17] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[18] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[19] M. McNally and C. Rindt. *The Activity-Based Approach*. Handbook of Transportation Modeling. Elsevier, Amsterdam, London, 2008.

An Algorithm Based Methodology for the Creation of a Regularly Updated Global Online Map Derived From Volunteered Geographic Information

Marcus Goetz, Johannes Lauer, Michael Auer
Chair of GIScience, Institute of Geography
University of Heidelberg
Heidelberg, Germany
{m.goetz, jlauer, auer}@uni-heidelberg.de

Abstract— **Global online maps are an important tool and data sets such for such maps are normally provided by commercial providers or public authorities. Nevertheless, the ever expanding trend of collaboratively collected geodata by hobbyists, namely Volunteered Geographic Information (VGI), increases regarding both data quantity and quality. Therefore, VGI can be considered as a real alternative data source for the provision of a global online map service, similar to those provided by Google Maps or Bing Maps. Therefore, an online map service needs to be created, whereby relevant data comes from a regularly updated database containing VGI. Due to the dynamic and fast changing nature of VGI sources, the workflow for processing VGI data needs to be automated on a regular base. The particular innovation of the here presented approach is that after an initial data import all required processing steps for transforming VGI data into a map-optimized data structure, is done internally with SQL database functions. That is, the processing is purely based on database technology and no additional software is required. The developed system uses standards and open-source software and is publicly available at www.osm-wms.de. The data can be consumed by either using a user adaptable standardized WMS, or via a high-performance web map application with partly pre-rendered map tiles. With the here presented approach, an regularly updated map based on open data can be provided.**

Keywords-Open Geospatial Consortium; OpenStreetMap; PostgreSQL; Volunteered Geographic Information; Web Map Service

I. INTRODUCTION

Maps are an important tool for diverse planning activities such as route planning, urban planning or agricultural planning. Moreover, services such as Google Maps or Bing Maps additionally push the usage of online maps, thus global online maps are omnipresent in professional and public areas. The before mentioned services are based on data collected by commercial data providers such as Teleatlas or Navteq.

Trying to push and evolve the spirit of the Web 2.0 approach, i.e., collaboratively providing and sharing information over the web by a broad mass, a new kind of geographic data source has arised during the last couple of years. This data source, namely called Volunteered Geographic Information (VGI), describes an ever expanding group of users, which collects geographic data in a voluntary and collaborative manner [1]. Thereby, different users with different levels of skills create spatial data by either performing personal measurements via GPS etc. or by tracing publicly available aerial images such as those provided by Bing. Afterwards, this data is uploaded to a Web 2.0 community and shared with other users, which are also allowed to reedit existing data or to use existing data at no charge. Additionally, also other geo-referenced information such as geo-referenced pictures or place locations can be considered as VGI. There are many different communities and portals sharing and collecting VGI, and there is an enormous potential arising from six billion humans acting as remote sensors [2]. The OpenStreetMap (OSM) community can be considered as the most prominent example of such a VGI community. With more than 400,000 registered users [3], i.e., more than 400,000 potential contributors, the OSM community grew rapidly considering the available data (Cf. chapter 3 for more details on the data structure). General statements about the quality of the OSM data (regarding both accuracy and completeness) from a global perspective are hard to tell, because both amount and quality vary between different regions. Nevertheless, diverse evaluations performed by Zielstra & Zipf [4], Haklay [5], Neis et. al. [6] or Ludwig et. al. [7] have proven that, especially in urban regions, OSM is able to compete against or even surpass data provided by commercial providers or governmental authorities. This is also the main motivation for providing an online map service similar to those of Google Maps or Bing Maps, whereat the data is purely based on VGI data from OSM. Therefore, the main contribution of this paper is the presentation of a workflow for creating a regularly updated database with VGI for the automated provision of a global online map, whereby the main processing steps are performed inside the database. The key characteristic of the presented approach is the innovative processing methodology for processing fast changing and dynamic geodata from VGI sources.

The rest of this paper is organized as follows: First, there is a brief overview about related work, followed by an introduction to OSM. Afterwards, a workflow for the provision of a regularly updated VGI database for map services is presented. Thereafter, a short introduction of the system architecture is given. Concluding, the developed methodology for materializing database views is presented and the workflow is demonstrated. Finally, a summary of the presented work is provided, and future work on this urging topic is discussed.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-178-6

II. RELATED WORK

Online maps, no matter whether they are focused on urban regions or on a global perspective, are omnipresent in the internet and it seems as if they have displaced ordinary paper maps. Some of the most famous examples for global online maps are Google Maps [8] or Bing Maps [9], whereby these are both based on data provided by commercial data providers such as Navteq, Tele Atlas, etc.

In contrast to those "commercial maps", there is a global online map available on the OpenStreetMap project page [10], whereby this map is purely based on collaboratively and voluntarily collected geodata (i.e., VGI). The OSM map illustrates different map features such as streets, naturals (e.g., forests or water areas), Point-of-Interests (POIs), rails, waterways etc, thus provides a detailed overview about both urban and rural areas. As stated above, diverse research approaches demonstrated that OSM is able to compete against commercial data providers, thus a map based on VGI from OSM can also be utilized for official processes such as urban planning. However, the architecture of the OSM project page is not based on Open Geospatial Consortium standards such as Web Map Services (WMS). It is not possible to set a user style (like SLD – Styled Layer Descriptor for WMS). The Mapnik renderer, which is used to produce the OSM-tiles, is configurable on server side but not on client side. They only provide server side rendered tiles in discrete zoom levels (the WMS is flexible and has continuous zoom levels). Further, there's no option to get a feature info and there's no option (without requesting the OSM-API) for getting the features themselves.

Nevertheless, sometimes it is required to operate an own online map service (e.g., for personal map styling or personal requirements). In achieving this goal, it is necessary to import VGI data from OSM into a database and to provide access to this database for a map service. Generally, there are two tools, which can be utilized for OSM data import, namely osm2pgsql [11] and OSMOSIS [12]. The former one converts OSM data into a format that can be loaded into a PostgreSQL database and is often used in combination with the map renderer mapnik. However, osm2pgsql does not import all available data, but does some kind of preselection according to keys and values. In contrast, OSMOSIS performs a whole data import, so that every kind of information is available in the database. Hence an import with OSMOSIS is more comprehensive, thus afterwards the application is more adjustable to personal requirements and desires, which makes the application more flexible and adaptable.

Another possibility is the processing of the OSM data with an ordinary programming language. In the past, our OSM-WMS was based on such a processing file, but the rapidly increasing OSM data resulted in long processing times. Actually, the processing of an OSM Europe file took about 14 days on a workstation, which is not acceptable.

The tool OSM-in-a-Box provides an out-of-the-box application set for the automated provision of a free world map [13]. The solution is based on a PostgreSQL/PostGIS database, a Geoserver with additional OGC conform Web-Feature-Service (WFS) and Web-Map-Service on top. For a fast map provision, the solution also contains a GeoWebCache, so that map-tiles are rendered beforehand. This solution allows a fast and easy-to-install out-of-the-box solution, but it is not adjustable to personal requirements and desires.

When investigating related research, it became apparent, that there are little publicly available solutions for a personalized provision of a global online map. On the one hand, the available solutions are too specialized and not generalizable for individual requirements. On the other hand, the processing and computation mechanisms inside the solutions are hidden and not described, so it cannot be said, how the solutions exactly work (e.g., the online map of the OSM project).

To complete, it should also be mentioned that there is of course different work on low level data model formats (like that of OSM, Cf. next Section) available. However, the data format of OSM has been designed with a special focus on VGI purpose, thus scientific models such as [14] or [15] do not suite the requirements of crowdsourced geodata. That is, such models will not be discussed in detail within this paper.

III. DATA SOURCES AND THE DATA STRUCTURE OF OSM

As mentioned above, the developed map service shall be purely based on VGI data. Since OSM is one of the most popular VGI projects, it is assumed that it is also the best suitable data source for this intention.

By mid of September 2011, inside the OSM database there were nearly 1,200,000,000 tagged points, whereby every point describes a distinct geographic location with distinct latitude and longitude values. These points can be furthermore combined into ways (currently nearly 110,000,000), whereby these can be either closed (i.e., an area) or non-closed (i.e., a line). Being able to map complex geometries such as polygons with holes etc., there is furthermore the concept of relations inside OSM. A relation (currently about 1,100,000) is a collection of different ways, nodes or relations, whereby these so called relation members belong together to some extent. Relations can be especially utilized for mapping complex polygonal geometries, whereby one or more outer elements contain several inner elements (e.g., a closed outer ring describing an area with several closed inner rings describing holes in the area). For adding different semantic information on top of those geometries, OSM adapts a concept of open key-value pairs. This concept allows that OSM users can tag their geometries (single nodes, ways or relations) with different key-value pairs, whereby the key describes a distinct information domain or condition and the value describes the corresponding information or information refinement. For example, a way with the key *highway* and the value *residential* generally describes (according to the key) a street for vehicles and/or humans and additionally specifies (according to the value), that this street is a residential street inside a city. Thereby, the amount of key-value pairs for an element is not limited, because additional information can be attached by further key-value pairs (e.g., tagging the before mentioned way with key *maxspeed* and value *30* for describing that the speed limit of this street segment is 30). Similar to streets, this key-value schema can also be utilized for mapping natural areas such as forests and seas, for mapping a Point-of-Interest (POI) such as an ATM or letter box, or for mapping the outer shape of buildings. The key-value pair concept in OSM is very flexible, because there are no limitations for the keys and values. There are indeed diverse best-practices and recommendations for mapping distinct map features such as the keys *amenity*, *boundary*, *building*, *natural*, *place*, *waterway* (refer to the OSM wiki [16] for more information) with corresponding values, but in general a user is able to add keys and values however liked. A complete list of all currently used keys is available at Tagwatch [17].

IV. System Architecture and Processing Workflow

The developed system architecture can be generally characterized as a classic 3-tier architecture [18] as depicted in Fig. 1. At the bottom there is a data tier (blue layer) with different data sources and data types combined in a database. On top of that, there is the processing tier (orange), which processes raw data of the data tier and stores the processed data in a database. On top, there is the presentation tier (green) representing graphical user interface (GUI), which allows a user to consume and interact with the provided data. Since the aim of this work is the development of an online map, the GUI consists of a HTML webpage, which can be accessed all over the world with an ordinary web browser or by utilizing standardized OGC web services (e.g., WMS).

The architecture itself is distributed, meaning that different system components such as the database or the webpage are located on different servers. This procedure allows for the distribution of work load onto different servers, thus is likely to increase the overall performance of the system. In particular, this distribution allows that a server can be configured according to its purpose (e.g., a database server can be equipped with much RAM and fast hard disks). The workflow for the data processing and data supply and the whole computation chain is depicted in Fig. 2.

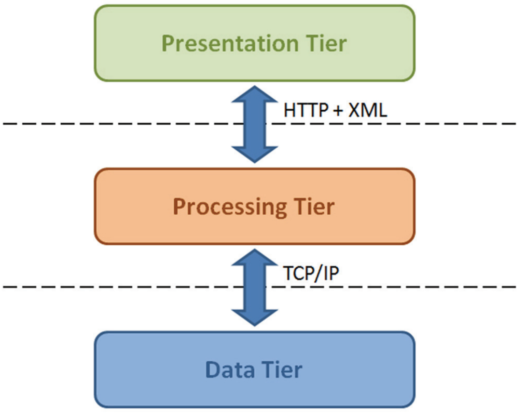


Figure 1. 3-tier architecture.

As described, an initial OSM planetfile (can be retrieved from [19]) import can be performed by using the processing tool OSMOSIS [12], which will be described in more detail in the next chapter.

A. OSMOSIS Import

OSMOSIS is a command line tool written in JAVA, which provides many functions to extract and convert OSM data. The tool is stream based, so there is always an input and an output stream. The tool is able to read almost all different OSM Formats such as XML, compressed XML, pbf etc., and furthermore it is capable to store the extracted OSM data into different data formats and data storage types (e.g., XML, pbf, mySQL, PostreSQL, diff-files, etc.). It also creates the former mentioned data schema that represents the whole OpenStreetMap Data. It is possible to create geometries for points and linestrings while importing the OSM data into a PostgreSQL/PostGIS relational database. Furthermore the tool can import diff-files (files that represent the changes between two OSM data sets). Within the here presented work, the OSMOSIS tool is used to create the database schema (namely called OSMOSIS simple schema) for PostgreSQL/PostGIS. Thereby, OSMOSIS separates the data into different tables such as nodes, ways etc. This allows a relational perspective on the OSM planetfile and with a few simple SQL queries, all data (i.e., key-value pairs or other information) for a distinct node, way or relation, can be retrieved. After the initial import, diverse indices are created on several database tables for increasing the performance of the database queries. The OSMOSIS database can be kept up-to-date by regularly performing diff-file updates. These diff-files are available for minutely changes, hourly changes or daily changes on the same webpage as the planetfile (and some mirror pages) and can be imported into the database by using the OSMOSIS parameter *--read-xml-change*. You can find further information about the tool at the OSM-Wiki pages [20]. Based on these OSMOSIS tables, different database views (i.e., virtual perspectives on the current data sets) have been created. These views were designed in a way, that they contain relevant data for a distinct information domain. For more information on the database views themselves, please refer to the next Section.

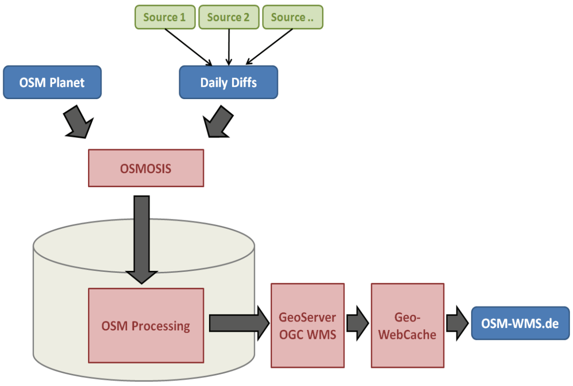


Figure 2. OSM data processing workflow

B. View Creation

As stated above, the developed system is based on an OSMOSIS database schema. That is, all VGI data from OSM is included in the database and separated into the tables *nodes*, *ways*, *way_nodes*, *relations*, and *relation_members*. The tags of the different elements are added to the corresponding data tuples by using a so called *hstore* column, which allows the storage of multiple key-value pairs in one single database row.

Depending on what type of element (i.e., what type of geometry) shall be included into the view, the selection of required OSMOSIS tables varies. So e.g., when creating a view, which contains single point-geometries (e.g., for POIs), the database view only considers the *nodes* table. In contrast, when concentrating on simple linestring geometries (e.g., streets or rails), the database view considers the *ways* tables. Since the OSMOSIS import is performed with activated linestring creation (i.e., OSMOSIS creates linestring geometries for all ways while importing the data), the table *way_nodes* is not required at all in the process. For more complex geometries such as the extraction of complex-shaped natural areas (e.g., forests or water areas) it is necessary to join different tables with each other (e.g., the *relation* table with the *way* table). Moreover, some geometries in OSM consist of several non-closed ways, whereby the collection of all of them results in one closed linestring. That is, when trying to create a polygon for such a relation, one must be aware of this issue, thus the corresponding view must be designed accordingly.

Depending on what type of geometry shall be created and what type of information shall be described, there is one or more database views required. For example, for a table containing all water elements of the world map, one would require a view for polygons created via one single closed way, one view for elements with several non-closed ways, and one view for polygons created via several (non-closed) ways whereby the resulting polygon consist of several non-overlapping elements.

The view itself is created via a SQL script, which iterates over a distinct OSMOSIS table (e.g., the *ways* table) and selects different attributes. Also by using *WHERE* conditions, it is possible to filter the results, so that for example only ways with special conditions or constraints are included in the materialized view.

Since database views are virtual, and for accessing them, they first need to be computed, access times on the views are not satisfying. Additionally, it is not possible to create indexes on views. For overcoming this issue, it has been decided to regularly materialize those views into real physical database tables (i.e., once the views are computed, their content is stored in a real database table). In doing so, all views are computed and all data tuples from the views are stored into physical tables. For further improving access times, several indexes for different table columns are created and stored in the database. These materialized tables are the data input source for the Geoserver WMS, which is used by the web-frontend.

However, this approach requires some further methodology for OSM updates, because an OSMOSIS update will not affect the materialized database views. That is, after an OSMOSIS update (i.e., a diff-file import), it is additionally required to redo the view materialization. This is achieved by recomputing the database view and storing the new view data into a physical database table. As long as the view is not computed successfully and all the data is stored in the materialized table, the system still utilizes the old tables. After the completion of the materialization process, the old tables are renamed (for backup possibilities), and the new ones are plugged into the system architecture.

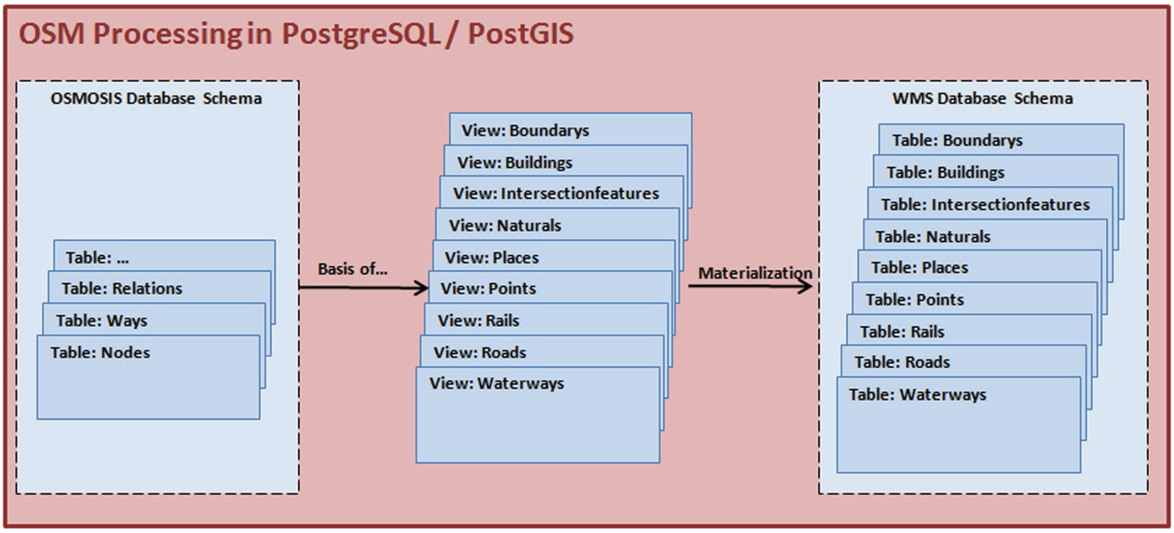


Figure 3. The Workflow of processing raw OSM data for a WMS

C. WMS and GeoWebCache

After importing, processing and storing the data, the created database tables are ready for the rendering part of the visualization pipeline. A very widespread method for visualizing web maps is the use of a rendering software, which on the one hand can interpret geographic data formats and on the other hand provides the results through a standardized web mapping interface. Such a standardized interface for rendering and delivering map images from geodata is described by the well known WMS Standard from the OGC. There are several open source and also commercial software tools, which implement this standard. In our case we are using the open source software Geoserver, which can be connected to the before mentioned database tables. With a standardized WMS Request (including different parameters like Bounding Box, Spatial Reference System etc.) it is possible do draw a map extent on the fly and deliver it to the client. The appearance of the resulting map image can be influenced by appending a user style in OGC SLD/SE or by predefined SLD styles on the server. This on-the-fly rendering of geodata provides a great flexibility with respect to the adaptation of maps to different user requirements, but is related to a high consumption of processing resources. Especially when rendering low zoom levels with many features, the rendering process is very slow. A dataset with such a big amount of features like the OSM dataset cannot be rendered on-the-fly in an acceptable period of time without reduction of the data by generalization or omitting certain feature types. This is particular the case if multiple users request map images from the server at the same time.

To avoid the problem of rendering time in worldwide datasets, it is possible to preprocess the map images in a regular grid of map tiles for several discrete zoom levels (scales). Assuming that the client always requests the map tiles in the same manner (size and origin) it is only necessary to process the tiles once and store the result on the server. That is, the tiles can be reused every time a new request from a user arrives, thus the usage of computation performance can be decreased. Such a tile caching can speeds up the performance of large scale map tiles from minutes to milliseconds. For caching map tiles there are also several open source solutions available like the TileCache from MetaCarta Labs [21] or GeoWebCache [22]. For providing the worldwide OSM map the GeoWebCache is used in a mixed application. The first 12 zoom levels from the whole world on one tile to the scale of approx. 1:200 000 (Manhattan on a tile) is preprocessed resulting in 5.5 million tiles of 256 by 256 pixels using 30 GB of disk space. Tiles in further zoom levels can be processed in an acceptable time so it isn’t necessary to preprocess all the tiles. Nevertheless, since it is unlikely that all parts of the world (especially the sea) are of interest in every zoom level for the user, it has been decided, that all further tiles will be rendered on-the-fly on the first request. For all further requests for these distinct tiles, they will also be stored in the cache.

This demand driven approach of tile rendering saves a lot of storing and processing resources as each zoom level has 4 times the amount of tiles than the one before.

V. Application of the workflow for a global dataset

A. Hardware Resources

The developed framework is applied on a global dataset of OSM. In the real system architecture there are two different servers: one is a processing and database server, which is responsible for the processing of OSM VGI data and for the data management. It is a dedicated server with 35 GB RAM and an eight double-core processor with 2.9 GHz (each processor). There are four 400GB hard disc drives in a virtual RAID 0 assemblage for the database itself and an additional 500 GB hard disc for the operating system (OS). The other server contains an open-to-the-public Geoserver and a webpage representing the GUI for the user. When requesting different map data via either the web-frontend or the standardized OGC WMS interface, the server connects to the database server and retrieves the relevant data tuples (often with the help of a spatial database index). The WMS server is also a dedicated one with 12 GB RAM and a four double-core processor with 2.5 GHz (each processor). It has three hard discs, one of them with 1 TB and two with 512 GB each.

B. Issues on Data Transformation

The database server is set up with an initial OSMOSIS planet import. The WMS database contains 13 tables, whereby nine of them (namely *boundaries, buildings, intersectionfeatures, naturals, places, points, rails, roads, waterways*) are updated regularly. The other ones (o*sm_coastlines*, *sea_all*, *world*, *world_ocean*) contain data that hardly changes, thus we decided that a regular update for them is not required. The process of computing the raw OSM data (gathered from OSMOSIS) for the WMS is depicted in Fig. 3. For creating the nine updated tables, 13 database views have been created – Table 1 contains an overview of them. The column view describes the view name, the column table describes, which WMS table the view is for, the column OSM type describes whether the source of the corresponding geometry results from a node, a way or a relation, and the column geometry type describes how the geometry is created (e.g., by collecting several non-closed ways).

The SQL scripts for the different views are simple SQL scripts containing nested SELECT-statements, CASE-WHEN statements (these are used for deciding between different relevant OSM keys) and WHERE-statements. The geometries are created via SQL by using different PostGIS functions such as *ST_MakePolygon*. Fig. 4 depicts a quite simple SQL script for the database view *intersectionfeatures_v*. Intersectionfeatures are kind of crossing-points in the street-network, thus the geometry is a single point. Therefore, the SQL script iterates the nodes table and only selects these rows with relevant OSM key-value pairs (e.g., *highway = mini_roundabout*).

TABLE I. DATABASE VIEWS WITH CORRESPONDING MATERIALIZED WMS TABLES AND OSMOSIS SIMPLE SCHEMA SOURCE TABLES

View	WMS table	OSMOSIS source tables	Geometry type
boundarys_oneout	boundaries	ways, relations	one closed linestring or relation with a closed linestring as outer way
boundarys_sevout	boundaries	relations	one relation with several distributed closed linestring
boundarys_sevout_nc	boundaries	relations	one relation with several non-closed linestrings
buildings_v	buildings	ways, relations	one closed way or relation with one outer closed way
intersectionfeatures_v	intersectionfeatures	nodes	one point geometry
naturals_oneout	naturals	ways, relations	one closed linestring or relation with a closed linestring as outer way
naturals_sevout	naturals	relations	one relation with several distributed closed linestring
naturals_sevout_nc	naturals	relations	one relation with several non-closed linestrings
places_v	places	nodes	one point
points_v	points	nodes	one point
rails_v	rails	ways	one way
roads_v	roads	ways	one way
waterways_v	waterways	ways	one way

```
CREATE OR REPLACE VIEW intersectionfeatures_v AS

SELECT
nodes.id AS osm_id,
(
   CASE WHEN defined(nodes.tags, 'junction') THEN nodes.tags -> 'junction'
   ELSE nodes.tags -> 'highway'
END
) AS type,
geom AS the_geom

FROM nodes

WHERE
defined(nodes.tags, 'junction')
   OR
   (nodes.tags -> 'highway' IN ('mini_roundabout', 'stop', 'give_way', 'traffic_signals', 'crossing', 'roundabout', 'motorway_junction'))
;
```

Figure 4. SQL view definition for *intersectionfeatures_v*

The SQL script for the view *buildings_v* (Cf. Fig. 5) is a bit more complicated. Buildings can be either modeled via a single closed linestring or (in a more complex way) as a relation with one outer closed linestring and several inner closed linestrings (representing holes in the building). Generally, the SQL script iterates the *ways* table and only selects those tuples with a valid building tag in the corresponding *hstore* column (i.e., the column tags must contain a key *building* with any kind of value). Additionally, the SQL script also checks whether the way has enough points (it is not possible to create a polygon with a line with less than 4 points, start and end-point have to equal). Also the area of the polygon is calculated, because buildings with an area of zero shall not be included in the database (i.e., *ST_Area(ST_MakePolygon(linestring)) != 0*). For those tuples that pass the WHERE-statement (i.e., which are buildings), several key-value pairs (e.g., height, building:roof etc.) are collected and stored in view columns. For the geometry it needs to be figured out, whether the current way is an outer member of a relation, because in that case the polygon must be created by cutting out the inner holes of the polygon (this is realized with the CASE-WHEN-statement). If the current way is not a member of a relation, the polygon is created by simply using *ST_MakePolygon*. The SQL view definitions are similar to those described beforehand.

The geometries that are mapped with several non-closed ways (i.e., *boundarys_sevout_nc* and *naturals_sevout_nc*) are kind of a special case. Before the polygon can be created, it needs to be figured out, whether the collection of all those linestrings leads to one closed linestring. This is realized with the PostGIS functions *ST_Collect* and *ST_Linemerge*. These functions form a polygon by sewing together several linestrings. If it is absolutely not possible to form one single linestring, the methods return a multi-linestring. However the SQL-script cannot deal with multi-linestrings, thus relations that lead to a multi-linestring are skipped during processing. If *ST_Collect* and *ST_Linemerge* result in one single linestring, a polygon can be created by using *ST_MakePolygon*. When creating a polygon with *ST_Collect* and *ST_Linemerge*, one strange effect has been discovered. For some geometries, *ST_Linemerge* did not result in a single linestring, although all linestrings shall be connected with each other (as could be seen when viewing the geometry in the well-know-text format WKT with the Method *AsText*). This issue has been solved by transforming the multi-linestring (resulting from *ST_Linemerge*) into WKT, then retransforming the WKT into a geometry (with *ST_GeomFromText*) and then performing again the *ST_Linemerge* method. This work-around was necessary because of a probable PostgreSQL rounding bug. It seems that the converting function (geometry to WKT) cuts the decimals of the coordinates, so that after the conversion the point matching works better. Additionally, this effect might also be caused by slight and minor inaccuracies within OSM. However, with this chain of several PostGIS functions it is possible to overcome the described problem.

```sql
CREATE OR REPLACE VIEW buildings_v AS

SELECT
ways.id AS osm_id,
(
   CASE WHEN defined(ways.tags,'name') THEN (ways.tags -> 'name')
   WHEN defined(ways.tags,'name:de') THEN (ways.tags -> 'name:de')
   WHEN defined(ways.tags,'ref') THEN (ways.tags -> 'ref')
   ELSE NULL
END
) AS name,
(
   CASE WHEN defined(ways.tags,'building:type') THEN (ways.tags -> 'building:type')
   WHEN defined(ways.tags,'building:use') THEN (ways.tags -> 'building:use')
   ELSE NULL
END
) AS type,
(
   CASE WHEN defined(ways.tags,'levels') THEN (ways.tags -> 'levels')
   ELSE NULL
END
) AS levels,
(
   CASE WHEN defined(ways.tags,'building:levels') THEN (ways.tags -> 'building:levels')
   ELSE NULL
END
) AS "building:levels",
...
(
   CASE WHEN EXISTS (SELECT 1 FROM relation_members WHERE member_id = ways.id AND LOWER(TRIM(member_role)) = 'outer')
   THEN
      (ST_MakePolygon(
         (linestring),
         ARRAY(
            SELECT linestring FROM ways AS w2
            WHERE w2.id IN (
               SELECT member_id
               FROM relation_members
               WHERE
                  LOWER(TRIM(member_role)) = 'inner' AND
                  member_type = 'W' AND
                  relation_id IN (SELECT relation_id FROM relation_members WHERE member_id = ways.id AND member_role = 'outer')) AND
                  ST_IsClosed(linestring) AND ST_NumPoints(linestring) > 3
          )
      ))
   ELSE ST_MakePolygon(linestring)
END
) AS the_geom

FROM
   ways

WHERE
   defined(ways.tags, 'building')
   AND
   ST_IsClosed(linestring)
   AND
   ST_NumPoints(linestring) > 3
   AND
   ST_Area(ST_MakePolygon(linestring)) != 0
;
```

Figure 5. Shortened SQL view definition of *buildings_v*

ISBN: 978-1-61208-178-6

C. *Results*

As a result of the described workflow, an amount of 24 GB is processed and distributed in nine WMS optimized database tables (Cf. Table 2).

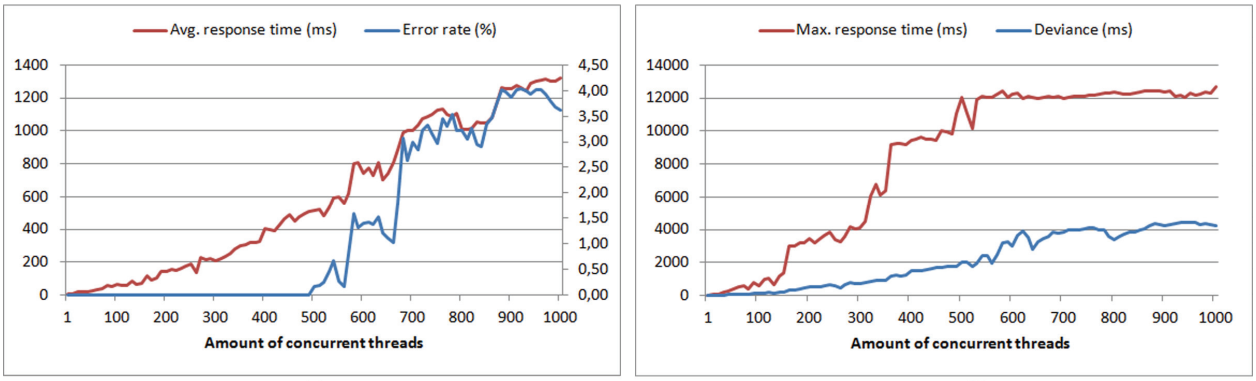
An initial import of OpenStreetMap data takes about 8 hours. The daily OSM update (about 50 MB in a zip file) for the database (including the download) requires less than 3 hours. The amount of time required for an update is that high, because of the required search operations and linestring computations. Especially the search operations, which are not required for an initial import, have high computation costs. The processing inside the database (i.e., the table creation, view materialization, index creation) requires 20 hours. The subsequent preprocessing of the tiles for the first 12 zoom levels by the WMS for storage into the GeoWebCache takes about 6 hours. That is, in total a ready-to-run startup system (initial OSM import, database processing and tile preprocessing) can be realized in about one and a half day.

TABLE 2. QUANTITATIVE FIGURES OF DIFFERENT DATABASE TABLES (DATA FROM 2011-09-11)

Table	Number of tuples	Table size	Index size
boundaries	612,114	293 MB	61 MB
buildings	42,187,016	8,683 MB	3,935 MB
intersectionfeatures	740,166	51 MB	78 MB
naturals	9,843,763	3,833 MB	942 MB
places	2,245,932	168 MB	302 MB
points	13,563,055	976 MB	1,287 MB
rails	833,872	199 MB	82 MB
roads	44,178,883	10,487 MB	4,394 MB
waterways	5,215,237	2,005 MB	478 MB
Total	119,419,988	~26 GB	~11 GB

For keeping the system up-to-date, several methodologies have been invented to drive an automated update. The database server downloads every day the daily-diff files of OSM and imports them into the database. Once a week (at the weekend) the database processing is performed, resulting in new and updated WMS tables. It has been decided that a weekly update of the WMS tables is enough, although a daily update could be feasible (OSM update plus processing would require about 23 hours). Additionally, once in a week (after the database processing), new tiles (with new features) for the GeoWebCache are generated.

For demonstrating the scalability of the system architecture, different performance analyses for the cached WMS data have been conducted. A varying amount of concurrent threads simulate the concurrent usage of the WMS by different users. Thereby, the amount of concurrent threads can be interpreted in two different ways: on the one hand, for example 100 concurrent threads can be considered as one user requesting 100 tiles from the cached WMS. On the other hand, it can also be regarded as for example 10 users, whereby each of them concurrently requests 10 tiles. The analyses were conducted in the range between one concurrent thread and 1.000 concurrent threads, which ought to be realistic numbers for the operative system's usage. For the tiles it has been decided to utilize a medium zoom level and a resolution of 256 * 256 pixels, resulting in a filesize of 7kb for each individual tile. The results of the analyses are depicted in Fig. 7. Fig. 7 (a) depicts the average response time of the WMS in *ms* (the red, upper line), as well as the error rate (the blue, lower line), i.e. the percentage of WMS requests which could not be processed due to timeout issues (in the current system, a timeout occurs after 20 seconds processing time for an individual thread). Fig. 7 (b) visualizes the maximum response time in *ms* (the read, upper line), as well as the deviance of the response times of all requests in *ms* (the blue, lower line). These two diagrams show, that the current system is able to handle up to 500 concurrent threads with acceptable response times. For more than 500 concurrent threads, some requests receive a timeout (about *3.5%* timeout rate for 1.000 concurrent threads). With increasing amount of concurrent threads, also the maximum response time increases, converging an upper bound of about *12.700 ms*.



(a) (b)

Figure 6. Average response times in *ms* and error rate in *percent* (a), deviance in *ms* and maximum response times ins *ms* (b) for the cached WMS with varying amount of concurrent threads

VI. CONCLUSIONS AND FUTURE WORK

Crowdsourced geoinformation from web community platforms can serve as a powerful and huge data source for online map services. Following a background literature review and investigation of existing web map services, an introduction to OpenStreetMap as one of the most popular examples of crowdsourced geodata is given. Afterwards a workflow for the automated processing of OSM data for the provision of a global online map has been described. See Fig. 7 for an exemplary excerpt of the map, showing the city of Heidelberg. The workflow has been textually described and a proof of concept was given. Additionally, quantitative figures about the current database, as well as qualitative performance analyses were provided.

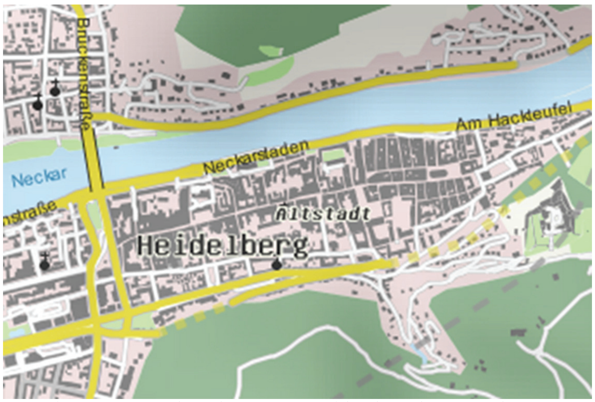


Figure 7. WMS based OSM map with adapted User Style

As a future step, the different views of the processing chain could be refined and optimized, so that performance can be further increased. Additionally, by adding further OSM key-value pairs, other different types of map objects can be integrated into the online map. Although not directly connected with the work described in this paper, the improvement of the data quality inside OSM is also an important issue. More proper and qualitative data in OSM is likely to have a positive effect on the OSM-WMS itself, because the higher the quality and correctness of the data is, the more competitive is the OSM-WMS compared to commercial map providers such as Google Maps.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for providing valuable comments towards the improvement of this paper. Furthermore we would also like to thank all members of the Chair of GIScience for their proofreading and helpful hints. This research has been partially funded by the Klaus-Tschira Foundation (KTS) Heidelberg.

REFERENCES

[1] Goodchild, M.F. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4): p. 211-221.

[2] Goodchild, M.F. 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research, 2: p. 24-32.

[3] OSM. 2011. Stats - OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Statistics. (Accessed 09/11/2011).

[4] Zielstra, D. and A. Zipf. 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. AGILE 2010. The 13th AGILE International Conference on Geographic Information Science: Guimarães, Portugal.

[5] Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B-Planning & Design, 37(4): p. 682-703.

[6] Neis, P., D. Zielstra, A. Zipf, and A. Strunck. 2010. Empirische Untersuchungen zur Datenqualität von OpenStreetMap - Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste, AGIT 2010 Symposium für Angewandte Geoinformatik.: Salzburg. Austria.

[7] Ludwig, I., A. Voss, and M. Krause-Traudes. 2010. Wie gut ist OpenStreeetMap? Zur Methodik eines automatisierten objektbasierten Vergleichs der Straßennetze von OSM und Navteq in Deutschland. GIS.Science, 4: p. 148-158.

[8] Google. 2011. Google Maps. http://maps.google.de/ (Accessed: 11/14/2011)

[9] Bing. 2011. Bing Maps - Driving Directions, Traffic and Road Conditions. http://www.bing.com/maps/ (Accessed: 11/14/2011)

[10] OSM. 2011. OpenStreetMap. http://www.openstreetmap.org. (Accessed: 11/14/2011)

[11] Osm2pgsql. 2011. Osm2pgsql. http://wiki.openstreetmap.org/wiki/Osm2pgsql. (Accessed: 09/11/2011).

[12] Osmosis. 2011. Osmosis - OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Osmosis (Accessed: 09/11/2011).

[13] Hof, R., M. Huber, F. Renggli, and S. Keller. 2009. OpenStreetMap-in-a-Box - Geo-webservices for the free world map, Computer Science. HSR: Rapperswil, Suisse.

[14] Morehouse, S. 1992. The ARC/INFO geographic information system. Computers&Geosciences, 18(4): p. 435-441

[15] Hoel, E.G., Menon, S., and Morehouse, S. 2003. Building a Robust Relational Implementation of Topology. SSTD, 2750: p. 508-524

[16] OSM. 2011. OpenStreetMapWiki. (Accessed: 09/11/2011).

[17] Tagwatch. 2011. Tagwatch Planet-latest. http://tagwatch.stoecker.eu/Planet-latest/En/tags.html. (Accessed: 09/11/2011)

[18] Eckerson, W.W. 1995. Three Tier Client/Server Architecture: Achieving Scalability, Performance, and Efficiency in Client Server Applications. Open Information Systems, 10(3): p. 1-20.

[19] OSM. 2011. Planet OSM. http://planet.openstreetmap.org/ (Accessed 11/14/2011)

[20] OSMOSIS. 2011. OSMOSIS - OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Osmosis. (Accessed 09/11/2011).

[21] MetaCarta. 2011. Geographic Search and Reference Solutions - Meta Carta - At the Forefront of the GeoWeb. http://www.metacarta.com/. (Accessed 06/30/2011).

[22] GeoWebCache. 2011. GeoWebCache. http://geowebcache.org/. (Accessed 09/11/2011)

Integrated Geoprocessing for Generation of Affected Assets and Rights Reports for High Voltage Electrical Infrastructures

Federico-Vladimir Gutiérrez-Corea,
Miguel-Ángel Manso-Callejo
Universidad Politécnica de Madrid
Madrid, Spain
fv.gutierrez@upm.es, m.manso@upm.es

Francisco-Javier Moreno-Marimbaldo
Department of the Environment
Red Eléctrica de España (REE)
Madrid, Spain
fmoreno@ree.es

Emerson Castaneda-Sanabria
Universidad Politécnica de Madrid
Madrid, Spain
emecas@ieee.org

***Abstract*— The development of linear construction projects, such as roads, railways, gas pipelines and electricity installations among others, whose execution on the territory impact a number of real estates, implies calculation of the effects over the plots in order to rewards landowners in terms for the influence of these new constructions on their properties as the case of Assets and Rights. This paper is about the high voltage electrical installations projects, focussed on the optimization GeoProcessing workflow for the generation of affected Assets and Rights Reports (ARR) for landowners due to the impact of affections for building electrical installations. This calculation and the generation of reports are carried out in different stages, requiring geoprocessing of different types of electrical constructions affections over the land, whose data come from different local and remote (Web services) sources. These overall characteristics make the process complex, time-consuming and requiring many resources. This paper shows the study case of the Integrated Geoprocessing for calculation of the affected ARR as a result of the new electrical conveyance infrastructures in Spain. The challenge consists of complying with the planning objectives as defined by the Ministry of Industry, Tourism and Trade (MITYC) for 2016, and reviewed for 2020, which implies the new creation of circuits and substations. In order to assisting in the accomplishment of those objectives and with the purpose of automating part of that workflow, an application software has been developed that integrates parts of the stages of that flow. The integrated stages are: (a) encoding of the plots susceptible to be affected; (b) calculation (geoprocessing) capable of processing 15 types of affections; (c) generation of an alphanumerical report with information about the owners of the affected plots and the ARR; (d) storage and Internet publishing of the affections caused by the projects through OGC Web Services and an ISO TC211 standard-based server for result dissemination in other formats. Up to five different data sources (two local, three remote) may come into play. The use of this application in the creation of the ARR has allowed improving productivity by optimising the amount of ARR generated per annum, the amount of calculated affections, the reliability of the calculations and the terms, thus reducing the cost of this activity.**

Keywords - Geoprocessing, SDI, Assets and Rights Report, Electrical installations, Landowners, Real estates.

I. Introduction

The development of linear construction projects whose execution on the territory touches a number of real estates implies calculation of the effects over the plots in order to rewards landowners in terms for the influence of these new constructions on their properties as the case of Assets and Rights. This paper is about the high voltage electrical installations projects, focussed on the optimization GeoProcessing workflow for the generation of affected Assets and Rights Reports (ARR) for landowners due to the impact of affections for building electrical installations; examples are the constructions of streets/roads [1], railways [2], gas pipelines [3], and electrical installations [4]. Spanish Electrical Network (REE) is the private/public company in charge of the maintenance and construction of the high voltage distribution network in all over Spain, which is also responsible for the calculation of affections and the ARR derived from the electrical installations. This calculation and the generation of reports are carried out in different stages, requiring geoprocessing of different types of electrical constructions affections over the land, whose data come from different local and remote (Web services) sources. These overall characteristics make the process complex, time-consuming and requiring many resources. This paper shows the study case of the Integrated Geoprocessing for calculation of the affected ARR as a result of the new electrical conveyance infrastructures in Spain.

REE owns the entire high voltage electricity network on the Spanish territory; it is also the operator of the conveyance system, responsible for the management, development, extension, maintenance and improvement of the network [5]. Fig. 1 shows both the present state of the high voltage network and the planning for 2016 on the Spanish peninsular territory.

The dimensions of the Spanish high voltage electricity network (up to 2010) are: 35,875 Km of electrical circuits [6], with approximately 70,000 electricity pylons and 450 substations [6][7]. The planning worked out on May 2008 by the Ministry of Industry, Tourism and Trade (MITYC) [8], whose objective is looking ahead to the future energy needs based on the increment of the demand experienced in recent years [6] that allows an estimation for 2016 of approximately 15,000 Km of circuits, 35,000 new electricity pylons and 200 new substations [9].

A. Background

The conventional method to calculate the affections caused by the high voltage electrical installations of REE and the generation of the ARR involved the use of different programmes and the manual execution of several tasks, partially automated in each stage of the process.

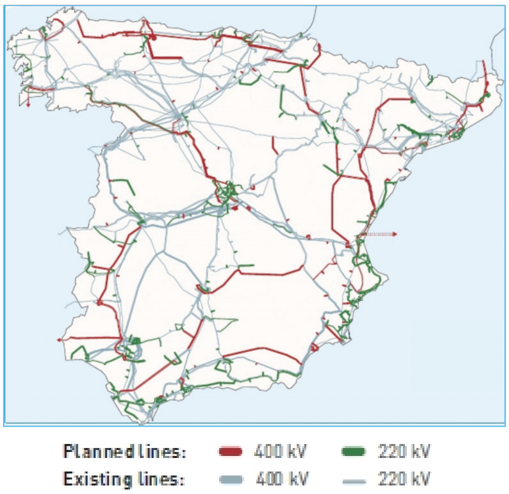


Figure 1. High voltage electrical energy distribution throughout the Spanish peninsular territory (adapted from [10]).

Stage 1 consisted of manual encoding of the plots likely to be affected by a new installation. Stage 2 consisted of the calculation of only four types of affections over the plots. In this stage, an AutoCAD script was used to calculate the affection due to the flight area of the high voltage electrical cable produced for the wind influence, the other three affections (felling area, permanent occupation due to the electrical installation infrastructures, and flight linear length of the high voltage electrical cable) were calculated manually. The script showed a dialogue window that visualised the value of the area affected by the overhead easement as a plot was selected. That value was used in stage 3: generation of the ARR in alphanumerical format as the values identifying the plot and the affected surface were being registered. This calculation method was useful for the REE until 2008. If we consider 250 the average amount of plots that may be affected by a new electrical line of 50 Km length and four the possible types of affections that may exist at present, the time needed to carry out these iterative processes with the traditional method would be approximately 200 hours.

B. *Motivating Scenario*

The ambitious planning of the MITYC for the electrical sector [8][9] and the traditional way of carrying out the calculation of affections and the subsequent generation of the ARR allow identifying the following critical points in the processes: (1) need of producing more ARR in less time, (2) excessive duration of the process (200 hours in our example), (3) high cost of data preparation for each programme, (4) questionable reliability of data transfer between programmes without fully automated procedures, and (5) variety of affection types to be managed (up to 15). These five critical points warrant the design of a working methodology and the creation of a software application tool supporting optimisation of the processes of calculation of the affections and generation of the ARR through automating of the largest part of the workflow. The software application called RBD-MercatorREE has been developed so that it will implement the identified requirements to face the challenges of the new planning, integrating and automating the workflow. The integrated stages are: (a) encoding of the plots susceptible to be affected, (b) calculation (geoprocessing) of up to 15 possible types of affections, (c) generation of an alphanumerical report with owners' information and the ARR of affected properties, and (d) storage of the history of the affections on a server with geospatial standards (ISO-OGC) [11] for results dissemination in other formats. Five data sources (two local, 3 remote – Web services) may come into play along the process.

II. Calculation of Affections and ARR Generation: Processes and Information Characterisation

The processes and data needed for the development and maintenance of high voltage electrical installations share generic characteristics with other types of lineal construction projects and infrastructures [12]. Some of the shared characteristics are: (1) affected properties on the territory, this fact implies the second characteristic, (2) the use of cadastral information, (3) one or more areas of direct affections on the plots by the elements to be built, and (4) one or more indirect affections on the plots by pathways which will allow access of the needed mechanical means for building and for future maintenance.

The source/destination of the information to calculate the affections and ARR involves several agencies. Table 1 resumes the actors involved in the workflow; the relationship between them are shown in Fig. 2.

TABLE I. Actors for affections and ARR.

Player	*Player Description*	*Main Source*
SEC-PR	Web access to the Cadastral Cadastre Electronic Headquarters (SEC) [13] to obtain protected information from the plot owners.	GOV
SEC-CA	Web access to the cartography of the Cadastral	GOV
GOV	Administrations and Spanish media in which the affections are communicated: Official State Gazette (BOE) [14][4], autonomic media [15], provincial media [2] and widely-read newspapers among others. for publishing on the BOE	GOV
DMA	Department of the Environment	REE
DIL	Department of Line Engineering	REE
DTR	Department of Handling	REE
REE/ UPM	A set of Web services temporarily on the Technical University of Madrid (UPM) which are conformant with the Spatial Data Infrastructure (SDI) standards in order to provide interoperability [16].	REE / UPM
CTC	Consulting Service for Surveying and Field	NGOV
CTO	Consulting Services for Surveying and Processes	NGOV
CMA	Consulting Service for the Environment	NGOV
GEN	public in general –not operating as yet	NGOV

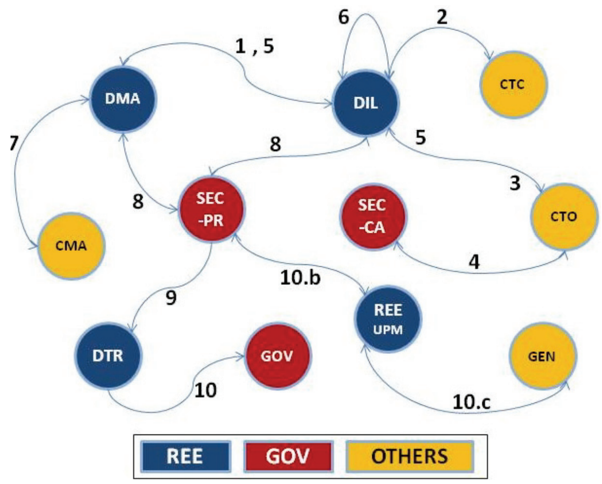


Figure 2. Relationship between actors involved in the calculation of affections and the generation of ARR

In Table 1, the main source column indicates a classification of actors corresponding to REE departments, the government agencies (GOV), and the non-governmental agents external to the REE (NGOV).

The relationship between actors is the following: DMA and DIL interact to define the passageways and alignments of the new installation, with the CTC carrying out the field revisions (links 1 and 2 in Fig. 2). This information is then used by the CTO to get the cadastral data from SEC-CA (links 3 and 4 in Fig. 2) and they encode the plots likely to be affected. With the cadastral cartography, DIL and DMA define the distribution of the electricity pylons (link 5 in Fig. 2) so that DIL will subsequently define and calculate the affections due to the influence of the trace (link 6 in Fig. 2) and DMA, interacting with CMA, will define and calculate the affections due to accesses (link 7 in Fig. 2) to the installation for building and future maintenance. After the affections have been calculated by DIL and DMA respectively, the protected information of the owners of the affected plots is retrieved from the SEC-PR (link 8 in Fig. 2), regardless of the type of affection. With the protected information, an alphanumerical ARR draft is generated (link 9 in Fig. 2), used by DTR to process and generate the final ARR. This report is then submitted to GOV for approval and publishing on the BOE and media (link 10). Simultaneously the information of the affections and all their geographic components are submitted to REE/UPM for OGC-standard-based web publishing (link 10.b in Fig. 2) which allows alternative publication as maps; at the same time the geometric history of the affections is saved for future access by the general public (link 10.c in Fig. 2).

The information exchanged by the players may be characterised as follows:

A. *Based on its format*

The fact of having several actors involved in all the ARR process implies having to deal with heterogeneous data formats. Five groups of formats are identified as follows: (1) twenty one GIS layers, (2) at least one XML protected file with the landowners information, the land price and more private data for each project that will be processed, (3) one alphanumerical database (DB) of provinces and municipalities (all of them are necessary input data for the processes). All of these three data formats are used as input for the ARR process. As part of the output data, there are sixteen more GIS layers containing information about the affected surfaces on the plots, and the next two data formats: (4) several spreadsheets files, containing the assets and rights report (prior to its publication on the BOE and pertinent media), and (5) a set of geospatial Web services (OGC) for storage and publishing over Internet all the historical ARR projects and its results, such as the graphic reviews of the affected plots, the pieces of geometric calculations and so on, all the different kind of produced output. Table 1 shows the different data formats and its main characteristics.

B. *Based on its purpose*

a) Plots: It represents the real estates involved in each electrical installation project. The encoding that identifies ordinally each plot, according to the direction of the electrical trace and accesses, are registered in addition to the values of the surface calculated for each type of affection. This information is generally obtained from the Cadastre through its SEC [13].

b) Layers with the affections: REE carries out studies for every new electrical installation project to take decisions and to process the information by affections. The data may be categorised according to affection on the ground as follows: (a) intangible affections, e.g. area affected by the flight of the high voltage electrical cable due to the wind influence, (b) tangible affections, e.g. area of permanent occupation by each electricity pylon. On the basis of their attributes, the data may be categorised: (c) with codes, e.g. every electricity pylon has an identification code, (d) without a code, e.g. area affected by tubing. According to the type of their geometry, the data may be: (e) polygons, and (f) lines.

TABLE II. DATA FOR AFFECTIONS AND ARR.

Data + (number)	*Group*	*Format*	*Source*	*Source Type*
Plots (1)	1	GIS	SEC	Remote
Layers for Encdoding (6)	1	GIS	REE	Local
Layers for Affections (15)	1	GIS	REE	Local
Landowners Data (1)	2	XML	SEC	Remote
Prov/Municipalities DB	3	BD	IDEE	Remote/ Local
ARR Draft	4	XLS	SEC	Remote
Histoy of Affections (16)	5	GML/ WFS-T	UPM/ REE	Remote

Improving the number of affections that may influence plots from four to fifteen was one important issue in order to accomplish the MITYC requirements. In the previous working way it was infeasible to work with more than 3 manually made affections. These 15 affections are: (1) aerial trace (superficial electrical cable generally suspended between pylons), (2) underground trace (part of the electrical cable is buried), (3) flight (safety area represented by the possible maximum movement of the electrical cable due to the influence of wind with a velocity of 120 Km/h perpendicular to the axis of the electrical line), (4) tube (representing the vertical projection of the cables on the ground with wind of 0 Km/h), (5) Felling (area that should be felled around the trace for safety purposes), (6) permanent occupation area (area occupied by the pylons), (7) permanent underground occupation (surface the underground power line will occupy permanently), (8) temporary occupation area (needed area for the building of the electricity pylons and other materials), (9) temporary underground occupation (needed occupied surface for underground electric wiring), (10) splicing chamber (area occupied permanently by concrete boxes where cable splicing is carried out), (11) telecommunication boxes (surface permanently occupied by the boxes used for telecommunication equipment associated to underground cable for remote maneuvering of the line), (12) landmarks (surface occupied by concrete posts in place to indicate on the surface the underground channelling of electrical cables), (13) accesses (easement needed for access from the electrical installations for building and maintenance), (14) auxiliary 1 and (15) auxiliary 2 (two generic affections available in the future if needed).

c) Layers for encoding: There is a dataset that helps in the encoding of the project plots; these must be identified for future field calls. These layers are: encoding trace; access to the pylons; buffer by trace; buffer for access and supports. This information is worked out by the REE.

d) Landowners data: Information such as address or identity card is protected as a part of the real estate register system. It is obtained from the Cadastre through its SEC [13]. This service requires users to identify themselves and to have the appropriate privileges to obtain this information.

e) History of affections: The geographic information of the polygons included in a project concerning the ARR (affections and cadastral plots) is duly saved into a server compliant with the standards of an SDI by using Web Map Services (WMS) [17] and WFS-T [18].

III. APROACH

A. Workflow automation and Integrated Geoprocessing

Some of the processes to calculate the affections and ARR generation such as the field calls for definition of alignments by passageways or the creation of traces and affection polygons are not in the first instance susceptible to be automated in this context. However, other parts of the processes does allow being optimised by automation and the creation of a new workflow; such is the case for the following four processes: (1) encoding of the plots, (2) process of calculation of up to 15 types of affections on the plots, (3) generation of the ARR combining the plots with their affections and the owners' protected information, and (4) saving of the history of affections and its publishing on the Internet by means of SDI standards.

Fig. 3 shows every one of the above-mentioned four processes following a gear metaphor. This figure is made up of three rows and four columns and represents the automated workflow. In each column is shown the name of the automated process, the needed input data to carry out the process and an output are graphically exemplified underneath. Next we describe each process (columns) briefly.

a) Plot encoding: This part of the workflow is comprised of seven GIS layers, one for the parcels and the other six for encoding. All parcels intersecting with the buffers of influences by affections are encoded; encoding consists of assignment of an identification value to each plot likely to be affected by a new installation, taking into account the direction of advancement of the trace (numbered with integers); at the intersections with the accesses, the access path is followed (numbered with decimal values). To reach this requirement, techniques and software libraries are used for the handling of the linear reference systems. The output is the same layer of plots to which a new attribute for encoding and its values is added.

b) Calculation of affections: This part of the workflow consists of 16 GIS layers, i.e. the plot layer and up to 15 layers of affections. The process carries out integrated geoprocessing on the parcel geometries as a function of the affections, taking into account the codes identifying them. The geoprocessing operations realised are: aggregation by attributes, intersections and clipping among others. As the software application carries out the operations, the areas and perimeters are calculated for each type of affection and saved on the same plot layer which again serves as the output of this process.

c) Generation of the assets and rights report: This part of the workflow uses two input data, the plot layer with all the values of the affections for each plot and an XML file with the protected information of the affected plot owners. The process combines the two data sources and generates an ARR output as a spreadsheet. This is the initial document to be reviewed by the Department of Handling before being submitted to the Administrations for its publication on the BOE and other media.

d) Saving and SDI publishing of historical data: The plot with all its affection values as well as the geometric details of each affection are saved in an SDI standard-complying server. Data are converted to the interchange Geography Markup Language (GML) format [19][20] and they are submitted through transactional operations to the server which implements the WFS-T. Once stored in this server, they may be queried in a standard fashion by using the WMS and WFS services.

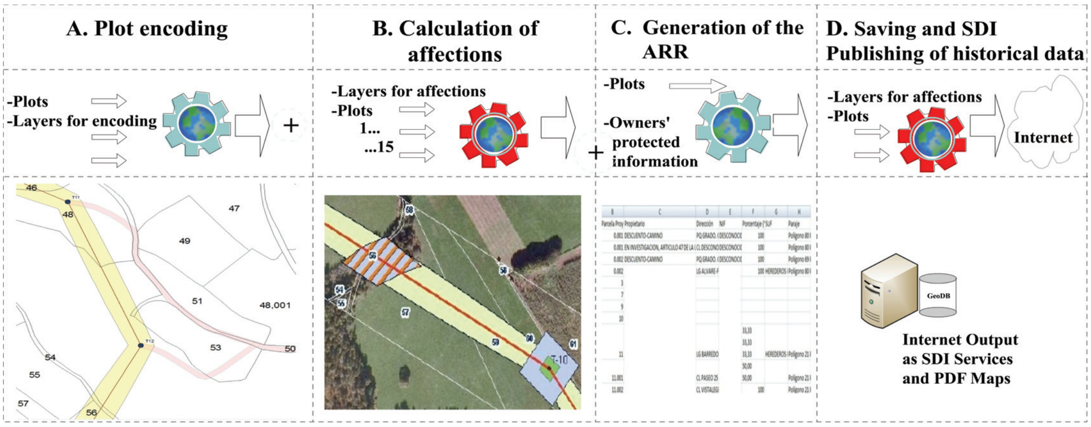


Figure 3. Automated workflow for calculation of affections and ARR generation

B. System architecture and design

The RBD-MercatorREE software application tool has been designed following the principles of object-oriented programming. This has allowed defining the objects that recap the necessary logic to comply with the software requirements and functionalities. The definition of those objects (aka business logic) allowed dividing the software into 3 uncoupled layers as independent and interrelated projects: (1) business objects, (2) user interface, and (3) remote persistence server. Fig. 4 shows the entire RBD-MercatorREE software application as a blue rectangle containing three gray rectangles inside. Every one represents a software logic layer (the upper rectangle is the desktop user interface, the central rectangle is the business logic layer and the lower rectangle represents the persistence layer). The dark blue circles on the central rectangle represent the main business objects making up the software; the sky blue circles of the rectangle on top represent forms and controls typical of a desktop user interface that instance the business objects. Fig. 4 also shows an overview of the system architecture for the integrated geoprocessing through the RBD-MercatorREE software application. Six rectangular areas are shown representing the involved players. The upper left rectangle represents the REE and its departments, with access to all the software functionalities. The upper right rectangle represents the remote server REE/UPM, used for storage of the affections; access to it is achieved from the software persistence layer through WFS-T, GML, XML and HTTP. The two lower left rectangles with the CMA, CTC and CTO labels represent the consulting firms that realise and process cartography for the REE; the circles 1 and 2 inside indicate that they have access to the option of plot encoding and calculation of affections through the local installation of the software. The lower central rectangle –under the Internet cloud– represents SEC and its different ways of access to the cadastral data (WFS, Web app and Web service).

The lower right rectangle represents the general public who can access the affections and assets and rights reports (ARR) published by REE/UPM conformant with SDI standards and Web applications.

C. Technical Resources

The RBD-MercatorREE software application has been developed using C# as programming language and .NET 4.0 [21]. The basic spatial operations have been carried out with the ArcObjects libraries for ArcGIS of ESRI [22]. The server for storage and publication conformant with SDI standards of historical data of affections and ARR is GeoServer 2.0.2 [23]. The XML encoding of the WFS-T requests Insert/Delete, etc. as well as the conversion of the data to GML format and the communication with the server through HTTP have been entirely carried out with the .NET basic libraries for C#.

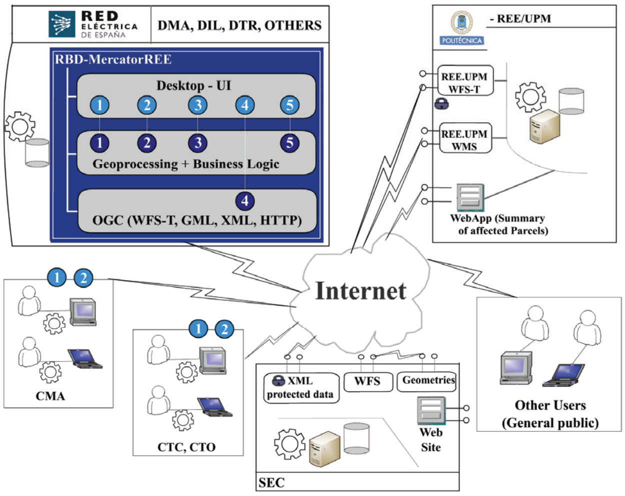


Figure 4. System architecture

IV. RESULTS

The possibility of obtaining the current cadastral cartography and the protected data of the affected parcels quickly and reliably as well as linking these to the data of each project, measuring the affection of the new installation on the territory automatically, has meant for REE the shortening of terms –in some processes over 90% reduction; in addition, it is possible now to count on the reliability of the published information, i. e. the official information managed by the Ministry of Economy and the Treasury.

In view of this advance, the departments involved in the technical drafting of projects, have established a line of work whose information flow is quick and smooth, avoiding the choke points that occurred with the old procedures. This fact brings about an increased capability of technical drafting of projects; it also makes it easier to comply with the planning approved by the MITYC on May 2008.

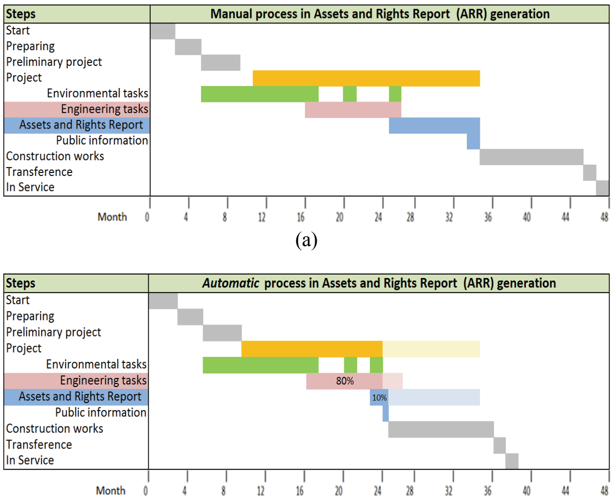
This positive impact on the run times is reflected on Fig. 5 where a chronogram represents the average project timescales of a project of a 50 Km long electrical line.

Fig. 5(a) shows the times according to the old procedures and Fig. 5(b) according to the new line of work. By comparing both figures it appears that the project activity represented by the yellow bar has been shortened considerably. This reduction represents about 40%, the engineering activity has been reduced in 20% and the largest optimisation corresponds to the ARR, with more than 90% reduction.

The availability of digital geographic information together with the analysis and optimisation of the workflow have been the key elements in the improvement of the methodologies. In the second place, automation of the processes has allowed a reduction in the terms and costs and an increase in the reliability of the results and consequently an improvement in the efficiency and productivity of the technicians. Thus we can present the following comparative Table with the conventional methodology and with the developed tools that automate the new methodology put in place.

TABLE III. COMPARISON OF THE PREVIOUS METHOD AND THE CURRENT METHOD FOR THE CALCULATION OF AFFECTIONS AND ARR GENERATION

Concept	*Old method*	*Current method*
Software Development Cost	0 €	30000 € (once)
# calculated affections	4	15
# ARR per annum	10-15	>200
Terms	200 hours	1 hour
Data preparation cost	14000 €	1000 €
Total ARR	50000 €	5000 €



(a)

(b)

Figure 5. Result values

V. CONCLUSION AND FUTURE WORK

The use of the RBD-MercatorREE software application that automates the workflow processes for the calculation of affections and generation of ARR and integrates geoprocessing, has improved process productivity optimising the following variables: number of ARR generated per annum, number of calculated affections, reliability of the calculations, terms for completion and consequently reduction in the costs of this activity.

Characterisation of the data has allowed us determining that the affections present well defined patterns (lines, polygons, affections without/with codes). These patterns will enable us to develop a new software version that might be personalised for another type of linear construction project such as gas pipelines, roads, railways, etc., thus generalising the developed solution.

Presently and since the Cadastre through its SEC allows the users identified with a digital certificate to download non-protected cadastral data, it is possible to automate other aspects of the workflow such as downloading of cadastral cartography. Likewise the protected information of the plots can also be automated since it is electronically accessible for registered, identified users.

Advances are being achieved in the cartographic representation that will afford a better insight into the connivance of a new installation with the environment. On the one hand it will provide the technicians with an accurate tool for virtual visualisation of the environment and on the other hand it will ensure that the owners affected by a new installation may be able to fully and easily understand the project with the help of the conventional project plans.

Considering the design and architecture of the RBD-MercatorREE software application in three logic layers, the business logic functionalities (calculation of affections and ARR generation) will be exposed on the Internet in the future through programming of other software layers on top of the business logic; those layers may be Web service SOA type interfaces or Web user interfaces through web pages.

REFERENCES

[1] Gobierno de Cantabria - Boletín Oficial de Cantabria (BOC), "Expropiación Forzosa: RBD Por ocupación de terrenos necesario para la ejecución del ensanche de la calle Aguayos," BOC núm. 51, CVE-2011-3347, Mar. 2011, pp. 8944-8945

[2] Diputación de Castelló - Boletín Oficial de la Provincia de Castellón (BOP), "Exposición Pública Anexo Canal De Drenaje del Ferrocarril: Proyecto Barranco de Fraga," BOP núm. 87, 07226-2011-U, Jul. 2011, pp. 1-2

[3] Ministry of the Presidency - State Official Gazette (BOE), "Resolución de la Delegación del Gobierno en Cataluña por la construcción: Gasoducto de conexión al almacenamiento subterráneo Castor," BOE núm. 91, 13199, Apr. 2011, pp. 43118-43120

[4] Ministry of the Presidency - State Official Gazette (BOE), "Resolución de la Delegación del Gobierno en Galicia por la que se convoca el levantamiento de actas previas a la ocupación de las fincas afectadas por la construcción de la subestación a 400 kV, denominada:Xove, en la provincia de Lugo." BOE núm. 113, 15864, May. 2011, pp. 52322-52324

[5] Red Eléctrica de España (REE), "Company profile, "http://www.ree.es/quien_es/presentacion.asp, Last visited 08/23/2011.

[6] Red Eléctrica de España (REE), "The Spanish electricity system 2010," 2011, pp. 1-147

[7] Moreno, F.J., Gutierrez F.V., and Bernabé, M., "Estándares OGC en el flujo de trabajo para la implantación de nuevas instalaciones eléctricas," 1° Congreso Internacional de Ordenamiento Territorial y Tecnologías de la Información Geográfica, Oct. 2010, http://faces.unah.edu.hn/ctig/sitios/congreso/programdetall2.html

[8] MITYC, "Planificación de los sectores de electricidad y gas 2008-2016," May. 2008,

[9] MITYC, "Informe de Sostenibilidad Ambiental de la Planificación de los Sectores de Electricidad y Gas 2007-2016," Jul. 2007, pp. 1-394

[10] Red Eléctrica de España (REE), "The Spanish electricity system 2010. Summary," 2011, pp. 1-15

[11]Open Geospatial Consortium Inc. (OGC), "OGC® Standards and Specifications," http://www.opengeospatial.org/standards, Last visited 09/07/2011.

[12] Rinaldi S.M., Peerenboom J.P., and Kelly T.K., "Identifying, understanding, and analysing critical infrastructure interdependencies," Control Systems Magazine, IEEE, vol. 21, Dec. 2001, pp. 11-25, doi: 10.1109/37.969131.

[13] Ministerio de Economia y Hacienda, "Cadastre Electronic Headquarters," http://www.catastro.meh.es/esp/sede.asp, Last visited 08/24/2011.

[14] Ministry of the Presidency. "Official State Gazette (BOE)," http://www.boe.es/index.php?id=en, Last visited 08/24/2011.

[15] Comunidad de Madrid - Boletín oficial de la Comunidad de Madrid (BOCM), "RESOLUCIÓN de 4 de enero de 2010, de la Dirección General de Industria, Energía y Minas...," BOCM núm 117, 21, May. 2010, pp. 50-51

[16]L.L. Alves, and C.A. Davis Jr, "Interoperability through Web services: evaluating OGC standards in client development for spatial data infrastructures," Proc. VIII Brazilian Symposium on GeoInformatics (GEOINFO 2006), Nov. 2006, pp. 173-188.

[17] Open Geospatial Consortium Inc. (OGC), "OpenGIS Web Map Server Implementation Specification," OGC 06-042, Mar. 2006, pp. 1-85

[18] Open Geospatial Consortium Inc. (OGC), "Web Feature Service Implementation Specification," OGC 04-094, May. 2005, pp. 1-117

[19] International Organization for Standardization (ISO), "Geographic information - Geography Markup Language (GML)," ISO 19136:2007, Dec. 2010, http://www.iso.org/iso/catalogue_detail.htm?csnumber=3254, Last visited 02/11/2011.

[20] Open Geospatial Consortium Inc. (OGC), "OpenGIS Geography Markup Language (GML) Encoding Standard," http://www.opengeospatial.org/standards/gml, Last visited 02/11/2011.

[21] Microsoft, "Introduction to the C# Language and the .NET Framework," http://msdn.microsoft.com/en-us/library/z1zx9t92.aspx, Last visited 09/07/2011.

[22] ESRI, "What is ArcObjects?," http://resources.esri.com/help/9.3/arcgisdesktop/com/shared/ao_foundation/what_is_ao.htm, Last visited 06/15/2010.

[23] GeoServer, “What is GeoServer,” http://geoserver.org/display/GEOS/What+is+Geoserver, Last visited 09/07/2011.

ISBN: 978-1-61208-178-6

Towards Identifying the Best New Connection in a Spatial Network:

Optimising the Performance of Hole Discovery

Femke Reitsma
The Department of Geography,
The University of Canterbury,
Christchurch,
New Zealand
femke.reitsma@canterbury.ac.nz

Tony Dale, William Pearse
BlueFern Supercomputing and Services Facility,
The Univesrity of Canterbury,
Christchurch,
New Zealand
tony.dale@canterbury.ac.nz

Abstract—Networks are used to represent phenomena such that we can measure their structure. Spatial networks are a special class of networks that reflect the embedding space within which the network is contained, incorporating the property of spatial autocorrelation and its impact on network measures. These measures of spatial network structure have thus far largely focused on the structure of the network, as opposed to the absence of that structure. This paper builds on past work in identifying an aspect of the absence of structure, that of identifying holes or chordless cycles in a network. It is the first step in identifying the best new connection to make in the network, identifying where the absence of structure is having the most significant impact. This paper presents the implementation of optimisations in discovering network holes.

Keywords-network; spatial network; chordless cycle; hole.

I. INTRODUCTION

Networks (or graphs) are used widely to model and analyse features in the world that either appear network like, such as infrastructure and river networks [1][2], or that we can model as a network, such as the relationships among groups of people in a social network [3]. Identifying the best new connection to create in a spatial network for the purposes of increasing the connectivity of that network has many relevant geographic applications. For example, identifying the best place to put a new cycle path that maximally increases the overall connectivity of the network, or identifying the best new connection in a power supply network that reduces the potential for power outages. Part of the process of identifying the best new connection is to first determine where the gaps in connectivity are and to recognise which would be the best gap to fill.

This paper presents the optimised implementation of a part of the process of identifying the best new connection in a spatial network. It improves past research that has developed the methodology but which was limited to very small networks due to an inefficient algorithm implementation [4]. This paper does not explore the application context where the discovered holes could be prioritized for developing a new connection across them, as this is the scope of future work.

The paper begins by reviewing some of the background literature on spatial networks, which is followed by a description of how holes in a network are discovered, presenting search optimisation methods in Section 2. Efficient methods for storing the input and output of the search method are described in Sections 3 and 4 respectively. And in Section 5, performance optimisations are discussed, followed by concluding comments in Section 6.

II. BACKGROUND

Work on spatial networks has developed in many different fields, including transportation engineering, hydrology and social geography. Research has repeatedly found that spatial networks have properties that are distinct from other kinds of networks due to spatial autocorrelation. Consequently they require unique measures [6][7]. For example, Ravasz and Barabasi [7] have found that scale-free patterns that are found in a large range of networks, such as the world wide web, do not exist in geographic networks, such as internet routers and power grid structures.

Barthélemy [6] thoroughly reviews measures for spatial networks, all of which focus on local and global measures that characterise a network, such as the work by Cardillo *et al.* [8], who describe the structural properties of urban street networks. Little thought has thus far been given to the absence of that structure, which is the focus of this paper, other than comparing the measures of network structure when new links or nodes are added or removed. For example, Buhl *et al.* [9] consider the robustness of street networks, measuring the robustness of a network by studying how it becomes fragmented as an increasing number of nodes are removed, where the impact on street network reliability varies whether nodes of high degree or low degree are removed.

ISBN: 978-1-61208-178-6

An algorithm for identifying one kind of lacking structure, namely holes, or chordless cycles, has been developed by Chandrasekharan, Lakshmanan and Medidi [5]. We build upon this early work and present an alternative implementation and the details of simple steps to optimize the performance of the algorithm presented below.

III. DETECTING HOLES

A hole in a network is mathematically defined as a chordless cycle. A chordless cycle is a cycle in a graph, a path such that the first node is connected to the last node, which includes at least four nodes with no chord connecting those nodes. A minimum size can also be defined, such that it is composed of more than four nodes.

For the future purposes of the application of this method to spatial data, the relationship between each node and its spatial identifier can be stored in an external spatial data file. The discovered holes can then be highlighted in the spatial data by joining the appropriate tables.

A. Search method

To detect a chordless cycle a vertex **v** from the graph is selected. This vertex and all outbound edges from this vertex are considered a tree structure, with the root of the tree being vertex **v**. A depth-first search (traversal) of this tree structure is performed. When the search returns to the vertex **v** we have discovered a cycle. This method can be described by the following recursive algorithm:

```
dfs (graph, startVertex, currentVertex,
visitedVertices) {
    if(visitedVertices.contains(currentVertex)){
        if (currentVertex == start){
            // we have found a cycle! do something...
        }
        return;
    }
    visitedVertices.add(currentVertex);
    for each (outboundChildVertex from
                                  currentVertex){
          dfs(graph, startVertex,outboundChildVertex,
                                  visitedVertices)
    }
    visitedVertices.remove(currentVertex);
}
```

Each cycle found must be checked for the following conditions:

- The cycle contains at least four vertices.
- Any non-adjacent vertices in the cycle are not connected by a single graph edge.

B. Search method optimisation

The disadvantage of the basic search method described in earlier is that each vertex will be visited many times before the search completes. This leads to very high computational complexity. The computational complexity can be significantly reduced by removing vertex **v** from the graph after a depth-first search from vertex **v** has been completed. If the depth-first search begins a vertex $\mathbf{v}_n$ this optimization can be implemented by instructing the search algorithm to ignore any vertices $\mathbf{v}_m$ where $m < n$.

In almost all instances, this significantly reduces the number of edges that must be traversed to complete subsequent searches. Removing vertex **v** from the graph does not reduce the number of cycles found, because after the search (beginning at vertex **v**) has been completed all cycles beginning and ending at vertex **v** have been found. Furthermore, all cycles beginning and ending at vertex **v** are actually all cycles containing vertex **v**. This is due to the nature of a cycle. Removing these vertices from the graph also has the positive effect of removing duplicate cycles from the search results, as all cycles beginning and ending at a vertex actually contain that vertex.

IV. INPUT NETWORK FORMAT

A network can be expressed as an adjacency matrix, which we use as the input structure for the method presented. The matrices are read and written through plain text files, with formatting based on the UCINET full matrix format [10].

The basic format of a UCINET full matrix text file is a two-line header, followed by the data. The data values must be separated by at least one space and commas or other punctuation symbols are not allowed. We have simplified this for our purposes, where the first line in the file is the number of columns, and therefore rows, in the matrix. The remainder of the file must contain a number of 1 or 0 characters. The exact number depends on the size of the matrix. The second and subsequent lines in the file represent each node in the graph; and its connection via edges to other nodes in the graph are indicated by a 1. For example, Figure 1 is a 5 x 5 matrix in a simplified UCINET matrix format:

```
5
 0 1 1 1 1
 1 0 1 0 0
 1 1 0 0 1
 1 0 0 0 0
 1 0 1 0 0
```

Figure 1. Sample matrix format

The corresponding graph for the matrix above is depicted below in Figure 2.

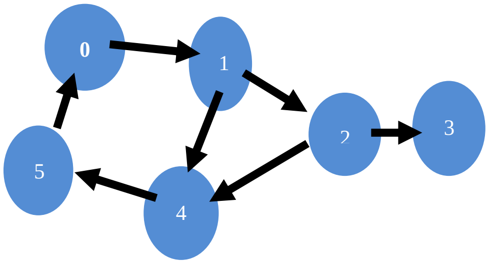


Figure 2. Network corresponding to simple adjacency matrix

The same matrix can also be represented in this compact representation:

```
5
011111010011001100001 0100
```

V. RECORDING SEARCH RESULTS

Consider the graph depicted in Figure 2 above. This graph has 6 vertices and 7 directed edges. Whether the edges are directed or not makes no difference to the problem, however directed edges significantly simplify the problem explanation.

This graph can be represented by the adjacency matrix shown below in Figure 3. Rows and columns are labeled 0-5 for convenience. A value of 1 in an adjacency matrix element indicates the presence of an edge. Row i in the adjacency matrix depicts the outgoing edges from node i. Similarly, column j contains the incoming edges for node j.

$$\begin{array}{ccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 \\
2 & 0 & 0 & 0 & 1 & 1 & 0 \\
3 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 0 & 1 \\
5 & 1 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Figure 3. Input adjacency matrix

A simple back-stepping depth-first search beginning at each node in this graph will reveal the presence of four chordless cycles within this graph. These four cycles are:

```
0-1-4-5, 1-4-5-0, 4-5-0-1, 5-0-1-4
```

It is immediately obvious that these four cycles are actually one cycle. If, however, vertex 0 is removed from the search space after the first cycle `0-1-4-5` is found (as proposed in Section 2), the three following duplicate cycles will not be discovered.

It is important, especially in the case of very large networks, that the solution is stored in a compact format. This can be done by storing the edges of each chordless cycle in a second adjacency matrix. For every cycle found, each edge in the cycle is stored in the adjacency matrix. Any existing edge in this second adjacency matrix may be overwritten. Following the discovery of the four chordless noted above, they are stored in an adjacency matrix, which contains only the edges from these chordless cycles:

```
0-1, 1-4, 4-5, 5-0
```

Despite the identification of the same chordless cycle four times, there is no duplication of edges in the solution adjacency matrix as they are overwritten (Figure 4).

$$\begin{array}{ccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 0 & 1 \\
5 & 1 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Figure 4. Solution adjaceny matrix with chordless cycles

To reconstruct the chordless cycles the original back-stepping depth-first search is repeated on the graph depicted by the solution adjacency matrix.

This method of storing chordless cycles has a number of advantages. First, it requires a known storage capacity, in contrast to other storage techniques such as flat-file format, or linked-lists, which require non-deterministic storage capacity. The storage capacity required is compact, where duplicate cycles do not cause the required storage capacity to ‘balloon’. By comparing elements of the first and second adjacency matrices the nodes that do (or do not) appear in chordless cycles can easily be determined (Figure 5). Very little computational effort is required to store data in the matrix and retrieve it. The adjacency matrix containing chordless cycles is typically sparse compared to the original adjacency matrix, making retrieval of cycles relatively effortless. The main disadvantages of storing chordless cycles in this manner is that it may take a significant amount of time to obtain a list of chordless cycles from the matrix because the reconstruction algorithm is the same as the algorithm used to find the cycles.

$$\begin{array}{ccccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 1 & 1 & 0 \\
3 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Figure 5. Adjacency matrix delta contains edges: 1-2, 2-3, 2-4

ISBN: 978-1-61208-178-6

VI. Performance optimization

Initial performance testing was performed on an Apple MacBook Pro with a 2.66 GHz Intel Core 2 Duo and 4 GB of memory. The code was compiled using GCC 4.2. The code was then further tested on an IBM p575 supercomputer (now replaced). The code was run under AIX 5.3 on a node with 16 1.9 GHz dual-core CPUs and 32 Gbytes of memory.

A. Compiler optimization

Before optimization, total execution time for a test matrix of 146 nodes, where the maximum cycle length searched is limited to 9 nodes, was 1 minute and 51 seconds. After removing assertions (such as ensuring given matrix coordinates were inside the bounds of the matrix) execution time was reduced to 1 minute 17 seconds. Using the compiler flags -funroll-loops -Os execution time was further reduced to 48 seconds. Without rewriting any code the program execution time was reduced to 43% of the original execution time.

B. Code tuning

The ‘node_create’ method allocates a block of memory for a node structure, and the malloc function is known to be time consuming. To reduce memory allocations, the node_create function, which consumes 29% of execution time, was removed from the code. This was done by replacing the node structure with a single unsigned long data type. After refactoring the total execution time for the test matrix was reduced to 19 seconds.

The code was then transferred to an IBM p575 computer at the BlueFern facility [11] and compiled to optimize the output and allow code profiling. Figure 6 shows that the **dfs** routine, split into two calls by compiler optimization, now account for the majority of CPU time consumed, and should be our next target for tuning. The **dfs** code is currently single-threading, and could benefit from a parallel implementation such as one of those surveyed by Freeman [12]. Future work will focus on tuning a parallel implementation of the network code and especially of the depth-first search routine.

VII. Conclusion and Future Work

Using the methods described above, the performance of the hole discovery algorithm has been significantly improved, where this optimisation will enable the analysis of much larger spatial networks. Given it remains constrained by processing power, however, future work will also consider parallelising the code. While we can clip spatial networks to a manageable size within a GIS, there will invariably be scenarios where it would be useful to consider much larger networks that represent a larger geographical system, such as the road network of California which includes 1,965,206 nodes and 5,533,214 edges [13].

The next phase of this research is to apply this optimized algorithm to urban cycle networks. Identifying holes in urban cycle networks will aid in the determination of the best new cycle path to make while considering factors such as the origin and destinations of cyclists, existing flows, and demographics.

Acknowledgment

The University of Canterbury’s Summer Scholarship Scheme is gratefully acknowledged.

References

[1] Lammer, S., B. Gehlsen, and D. Helbing (2006). “Scaling laws in the spatial structure of urban road networks”, Physica A 363(1) pp. 89-95

[2] Tarboton, D. G. (1996). “Fractal river networks, Horton’s laws and Tounaga cyclicity”. Journal of Hydrology 187: 105-117.

[3] Scellato, S., A. Noulas, R. Lambiotte, and C. Mascolo (2011). “Socio-spatial Properties of Online Location-based Social Networks”. Proceedings of ICWSM 11, 329-336 .

[4] Reitsma, F. and S. Engel (2004). “Searching for 2D Spatial Network Holes”. Computational Science and its Applications -ICCSA 2004 Conference, Assisi, Italy, May 14 Ð 17 2004, Proceedings, Part II: Lecture Notes in Computer Science 3044 Springer: pp. 1069-1078

[5] Chandrasekharan, N., V. S. Lakshmanan and M. Medidi (1993). “Efficient Parallel Algorithms for Finding Chordless Cycles in Graphs”. Parallel Processing Letters 3(2): 165-170.

[6] Barthélemy, M (2011). “Spatial Networks”. Physics Reports 499:1-101.

[7] Ravasz, E. and A. Barabasi (2003). “Hierarchical Organization in Complex Networks”. Physical Review E 67(026112).

[8] Cardillo, A., S. Scelllato, V. Latora, and S. Porta (2006). “Structural properties of planar graphs of urban street patterns”. Physical Review E, 73: 066107.

[9] Buhl, J., J. Gautrais, N. Reeves, R. V. Sole, S. Valverde, P. Kuntz, and G. Theraulaz (2006). “Topological patterns in street networks of self-organized urban settlements”. The European Physical Journal B, 49: 513-522.

[10] http://www.analytictech.com/networks/dataentry.htm last accessed 1/10/2011

[11] http://www.bluefern.canterbury.ac.nz last accessed 1/10/2011

[12] Freeman, J. (1991). “Parallel Algorithms for Depth-First Search”. University of Pennsylvania Technical Report MS-CIS-91-71.

[13] http://snap.stanford.edu/data/ last accessed 1/10/2011

Development of an Acoustic Transceiver for Positioning Systems in Underwater Neutrino Telescopes

Giuseppina Larosa[a], Miguel Ardid[a], Carlos D. Llorens[b], Manuel Bou-Cabo[a],
Juan A. Martínez-Mora[a], Silvia Adrían-Martínez[a]
[a] Institut d'Investigació per a la Gestió Integrada de les Zones Costaneres (IGIC) – Universitat Politècnica de València,
C/ Paranimf 1, 46730 Gandia, València, SPAIN
e-mail: giula@doctor.upv.es
[b] E.P.S. Gandia, Universitat Politècnica de València, C/ Paranimf 1, 46730 Gandia, València, SPAIN

Abstract— **In this paper, we present the acoustic transceiver developed for the positioning system in underwater neutrino telescopes. These infrastructures are not completely rigid and need a positioning system in order to monitor the position of the optical sensors of the telescope which have some degree of motion due to sea currents. To have a highly reliable and versatile system in the infrastructure, the transceiver has the requirements of reduced cost, low power consumption, high intensity for emission, low intrinsic noise, arbitrary signals for emission and the capacity of acquiring and processing the received signal on the board. The solution proposed and presented here consists of an acoustic transducer that works in the 20-40 kHz region and withstands high pressures (up to 500 bars). The electronic-board can be configured from shore and is able to feed the transducer with arbitrary signals and to control the transmitted and received signals with very good timing precision. The results of the different tests done on the transceiver in the laboratory are described here, as well as the change implemented for its integration in the Instrumentation Line of ANTARES for the in situ tests. We consider the transceiver design is so versatile that it may be used in other kinds of marine positioning systems, alone or combined with other marine systems, or integrated in different Earth-Sea Observatories, where the localization of the sensors is an issue.**

Keywords-acoustic transceiver; underwater neutrino telescopes; calibration; positioning systems.

I. INTRODUCTION

The acoustic transceiver presented in the article has been developed to be used in the acoustic positioning system of the neutrino telescopes in the Mediterranean Sea KM3NeT [1], and it is going to be tested on the ANTARES neutrino telescope. ANTARES is currently the biggest underwater neutrino telescope in the world and in operation in the Northern Hemisphere [2–3]. The detector is located in the Mediterranean Sea on a marine site 40 km SE offshore from the city of Toulon (France), at about 2400 m depth. Its construction was completed in May 2008, and now it is collecting data, which is analyzed in order to bring insights into different scientific problems related not only to astroparticles, but also in several fields of Earth-Sea Sciences. On the other hand, KM3NeT is a European Consortium that aims to design, build and operate a cubic kilometre neutrino telescope in the Mediterranean Sea [1–4]. The project is now in the Preparatory Phase for the Construction funded by the VII Framework Program of the European Union.

Undersea neutrino telescopes have become very important tools for the study of the Universe. Moreover, these infrastructures are also abyssal multidisciplinary observatories with the installation of specialized instrumentation for biology, seismology, gravimetry, radioactivity, geomagnetism, oceanography and geochemistry offering a unique opportunity to explore the properties of a deep Mediterranean Sea site over a period of many years. The main elements of a neutrino telescope are an array of optical sensors (photomultipliers in glass spheres) located in flexible structures deployed in the deep sea and maintained vertical with buoys. For KM3NeT, the array will cover large volumes, of the order of a cubic kilometre, to have adequate sensitivity for the expected fluxes of neutrinos in these processes. It is able to detect the Cherenkov light from the muons produced by neutrino interactions with matter around the detector. The arrival times of the light collected by the optical detectors can be used to reconstruct the muon trajectory, and consequently that of the neutrino. The accuracy of the reconstruction of the muon track depends on the precision in measurement of light arrival time and on precise knowledge of the positions of the optical detectors [5–7]. The positioning system is necessary because marine currents may produce inclination of the structures, and thus displacement the optical sensors of the telescope by several metres from the nominal position. Then, the precise knowledge of the relative positions of all optical sensors is essential for a good operation of the telescope, and must be known with ∼10 cm accuracy. On the other hand, the absolute geo-referenced positions are needed to point back to astronomical sources. In order to know and monitor with precision the relative positions of the optical modules an triangulation method is applied in the acoustic positioning system constituted of receiving hydrophones attached to the structures and of emitting transceivers in fixed positions near the sea bottom.

A considerable effort has been made by the KM3NeT Collaboration for the development of such system [8–10]. Here, we will present the work done in order to develop, test and integrate the solution for the KM3NeT acoustic transceiver.

ISBN: 978-1-61208-178-6

The transceiver design is very versatile, and thus, it can be easily adapted to other kinds of marine positioning systems, alone or combined with other marine systems, or integrated in different Earth-Sea Observatories, where the localization of the sensors is an issue. Therefore, we think that it can be a very useful tool in geo-processing applications in marine environment.

In Section II, the acoustic transceiver is described. The laboratory tests performed on it are discussed in Section III. Section IV shows the activities for the integration of the system in the ANTARES telescope for the in situ tests. Finally, the conclusions are presented, as well as the different possibilities of the system for being used in other marine positioning or localization systems.

II. THE ACOUSTIC TRANSCEIVER

The Acoustic Positioning System (APS) for the future KM3NeT neutrino telescope consists of a series of acoustic transceivers distributed on the sea bottom and receivers located on the lines near the optical modules. Each of these acoustic transceivers is composed of a transducer and one electronic board named '*sound emission board*'. Next, we will present these two parts of our system.

A. *The acoustic sensor*

The acoustic sensor has been selected to attend to the specifications needed for the KM3NeT positioning: withstand high pressure, good receiving sensitivity and transmitting power, nearly omnidirectional, low electronic noise, high reliability, and affordable for the units needed in a cubic kilometer. Among different options we have selected the Free Flooded Ring (FFR) transducers, model SX30 manufactured by Sensor Technology Ltd. The FFR transducers have geometrical forms such as rings, and then the hydrostatic pressure is the same on the inside and the outside. This characteristic form reduces the change of the properties of piezoelectric ceramic under high hydrostatic pressure. So they are a good solution to the deep submergence problem [11]. The SX30 FFRs are efficient transducers that provide reasonable power levels over wide range of frequencies and deep ocean capability. They work in the 20–40 kHz frequency range and have dimensions of 4.4 cm outer diameter, 2 cm inner diameter, 2.5 cm height.

They have unlimited depth for operation (already tested up to 440 bars [12]) with a transmitting and receiving voltage response of 133 dB Ref. 1µPa/V at 1m and -193 dB Ref. 1V/µPa, respectively. The maximum input power is 300 W with 2% duty cycle. These transducers are simple radiators and have omnidirectional directivity pattern in the plan perpendicular to the axis of the ring (plane XY), while the directivity in the other planes depends on the length of the cylinder (plane XZ), 60º for the SX30 model [13]. The cable on the free-flooded rings is 20 AWG, TPE (Thermoplastic elastomer) insulated. The cable is affixed directly to the ceramic crystal. The whole assembly is then directly coated with epoxy resin. Both the epoxy resin and the cable are stable in salt water, oils, mild acids and bases.

The cables are therefore not water blocked (fluid penetration into the cable may cause irreversible damage to the transducer). For this reason the FFR hydrophones have been over-molded with polyurethane material to block water and to facilitate its fixing and integration on mechanical structures. Figure 1 shows pictures of the FFR transducer without and with over-molding.

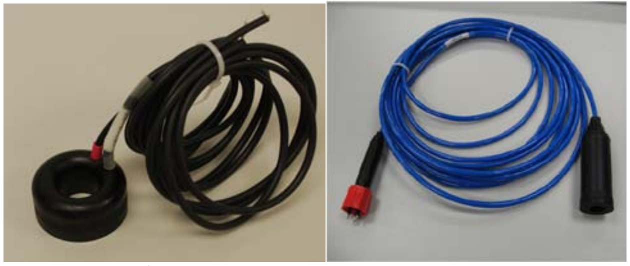


Figure 1: View of the Free Flooded Ring hydrophone (without and with over-molding).

In the next plots (Figures 2 to 5), we present the results of the tests carried out in our laboratory to characterize the transducers in terms of the transmitting and receiving voltage responses as a function of the frequency and as a function of the angle (directivity pattern). For the tests omnidirectional transducers, model ITC-1042 and calibrated RESON-TC4014 have been used as reference emitter and receiver, respectively. Particularly, Figure 2 and Figure 3 show the Transmitting Voltage Response and the Receiving Voltage Response, respectively, of the FFR hydrophones as a function of the frequency (measured in the plane XY, that is, in the perpendicular of the axis of the transducer); Figure 4 and Figure 5 show the Transmitting Voltage Response and the Receiving Voltage Response, respectively, of the FFR hydrophones as a function of the angle using a 30 kHz tone burst signal (measured in the plane XZ, 0º, which corresponds to the direction opposite to cables).

B. *The Sound Emission Board*

We have developed dedicated electronics, Sound Emission Board (SEB), in order to be able to communicate, configure the transceiver and control the emission and reception. Relative to the emission, it is able to feed the signals for positioning and amplify them in order to have enough acoustic power so they could be detected from acoustic receivers at about 1 km away from the emitter.

Moreover, it stores the energy and gives enough power for the emission and to switch between emission and reception modes. The solution adopted is specially adapted to the FFR transducers and is able to feed the transducer with high amplitude short signals (a few ms) with arbitrary waveform. It has as well the capacity of acquiring the received signal. The diagram of the board prototype is shown in Figure 6. It consists of three parts: the communication and control which contains the micro-controller dsPIC (blue part), the emission part constituted by the digital amplification plus the transducer impedance matching (red part) and the reception part (green part). In the reception part a relay controlled by the dsPIC switches the mode and feeds the signal from the transducer to the receiving board of the positioning system.

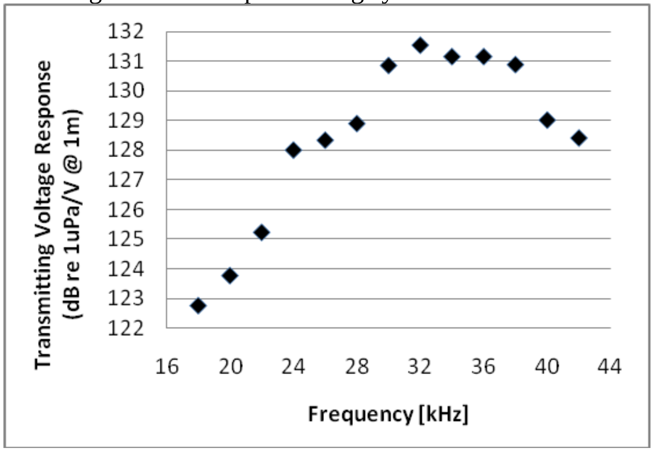


Figure 2: Transmitting Voltage Response of the FFR hydrophones.

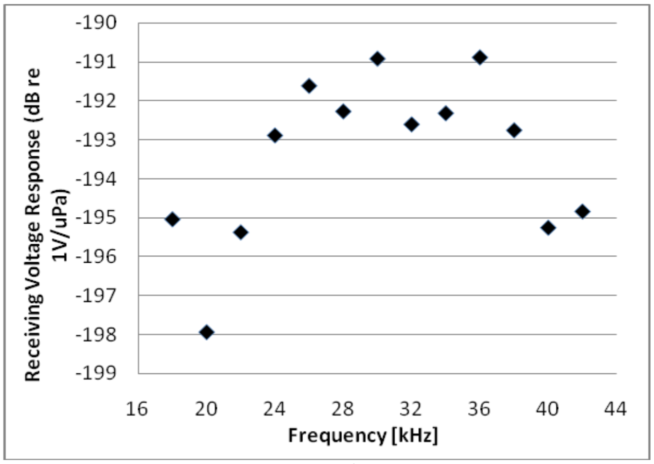


Figure 3: Receiving Voltage Response of the FFR hydrophones.

The SEB has been designed for low-power consumption and it is adapted to the neutrino infrastructure using power supplies of 12 V and 5 V with a consumption of 1 mA and 100 mA respectively, furnished by the electronic of the neutrino telescope. To avoid initial high currents, there is a current limit of 15 mA when the capacitor starts to charge, but few seconds later the current stabilizes at 1 mA. With this, a capacitor with a very low equivalent series resistance and 22mF of capacity is charged allowing storing the energy for the emission. The charge of this capacitor is monitored using the input of the ADC of the micro-controller. Moreover, the output of the micro-controller is connected through 2x Full Mosfet Driver and a MOSFET full bridge; this is successively connected at the transformer with a frequency and duty cycle programmed through the micro-controller. The transformer is able to convert the voltage of the input signal of $24V_{pp}$ to an output signal of about $500V_{pp}$.

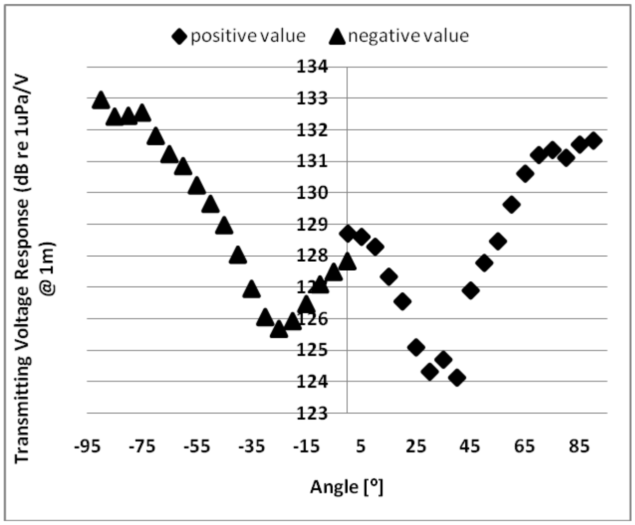


Figure 4: Transmitting Voltage Response of the FFR hydrophones.

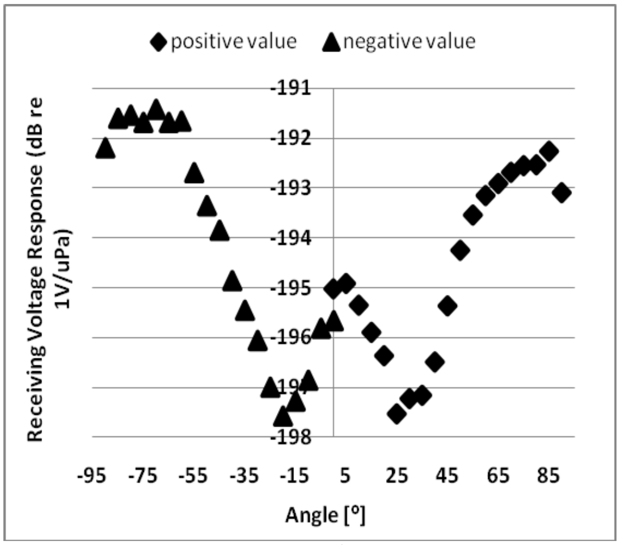


Figure 5: Receiving Voltage Response of the FFR hydrophones.

Besides, concerning the reception part of the board, the board has the possibility to directly apply an anti-aliasing filter and return the signal to an ADC of the microcontroller.

This functionality may be very interesting not only in the frame of the neutrino telescopes, but also to have the receiver implemented in different underwater applications, such as affordable sonar systems or echo-sounds.

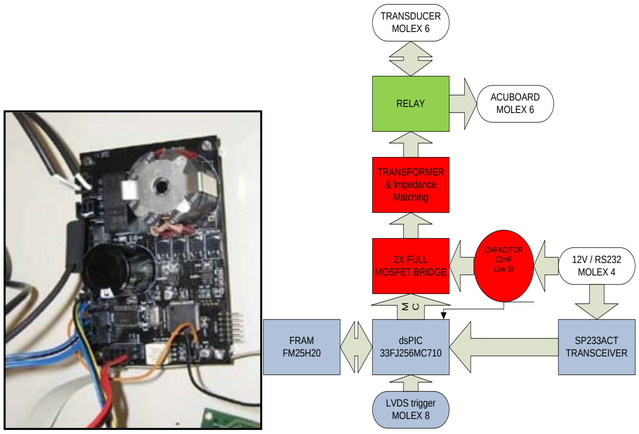


Figure 6: View and diagram of the Sound Emission Board.

The micro-controller contains the program for the emission of the signals and all the parts of control of the board. The carry frequency of the emission signal is 400 kHz and has tested up to 1-1.25MHz. The signal modulation is done with Pulse-Width Modulation technique which permits the emission of arbitrary intense short signals [14].

The basic idea of this technique is to modulate the signal digitally at a higher frequency using different width of pulses and the lower frequency signal is recovered using a low-pass filter. In addition, it will have an increase in the amplitude of the signal using a full H-Bridge. The communication of the board with the PC is established through the standard protocol RS232 using an adapter SP233 in the board. In order to have very good timing synchronization the emission is triggered using a LVDS signal.

In summary, the board, designed for an easy integration in neutrino telescope infrastructures, can be configured from shore and can emit arbitrary intense short signals or act as receiver with very good timing precision (the measured latency is 7 μs with a stability better than 1 μs), as shown in the joint tests of the INFN-CNRS-UPV acoustic positioning system for KM3NeT [15].

III. Laboratory Tests of the Transceiver

The transceiver has been tested in the laboratory and it has been integrated in the instrumentation line of the ANTARES neutrino telescope for the in situ tests. Next, we describe briefly the activities and results of these tests.

The measurement tests in the laboratory have been performed firstly in a tank of 87.5 x 113 x 56.5 cm^3 with fresh-water, and secondly in a pool of 6.3 m length, 3.6 m width and 1.5 m depth. We have tested the system using the FFR hydrophone over-molded and the SEB. The molding of the transducer has been done by McArtney-EurOceanique SAS which over-molded completely the back of the transducer. Moreover, 10 meters of the cable type 4021 has been molded onto free issued hydrophones plus one connector type OM2M with its locking sleeves type DLSA-M/F. The moldings are done in polyurethane, the connector body in neoprene and the locking sleeve is in plastic.

Besides, some changes in the SEB board have been done to integrate the system in the ANTARES neutrino telescope and to test the system in situ at 2475 m depth. For simplicity and limitations in the instrumentation line, it was decided to test the transceiver only as emitter, the functionality as receiver will be tested in other in situ KM3NeT tests. The changes done in the SEB are the following: to eliminate the reception part, to adapt the RS232 connection to RS485 connection and to implement the instructions to select the kind of signals to emit matching the procedures of the ANTARES DAQ system.

To test the system we have used the transceiver in different emission configurations in combination with omnidirectional transducers, models ITC-1042 and a calibrated RESON-TC4014, used as emitter and receiver respectively. Different signals have been used (tone burst, sine sweeps, MLS signals, etc.) to see the performance of the transducer under different situations.

Figure 7 shows the Transmitting Acoustic Power of the transceiver as a function of the frequency (measured in the plane XY, that is, in the perpendicular of the axis of the transducer). The Transmitting Acoustic Power of the transceiver as a function of the angle (directivity pattern) using a 30 kHz short tone burst signal (measured in the plane XZ, 0 º corresponds to the direction opposite to cables) is shown in Figure 8.

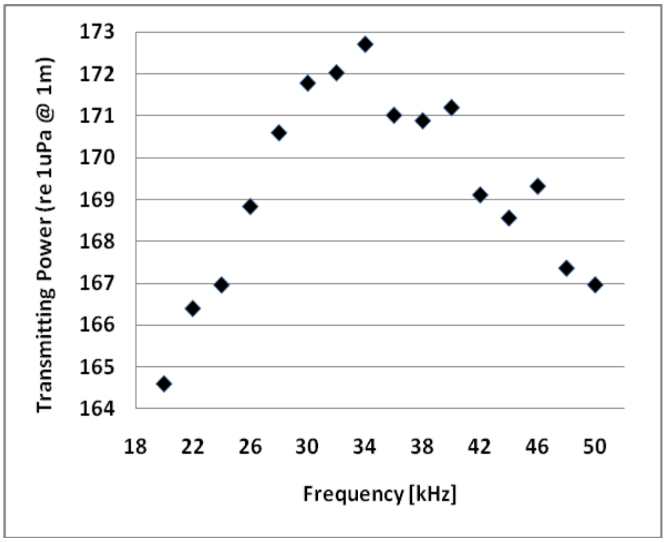


Figure 7: Transmitting Acoustic Power of the transceiver.

If we compare the Receiving and Transmitting Voltage Response of the FFR over-molded with the FFR without over-molding a loss of ~1-2 dB is observed. Figures 7 and 8 show that the results for the transmitting acoustic power in the 20-50 kHz frequency range is in the 165-173 dB re. 1μPa@1m range, in agreement with the electronics design and the specifications needed. Despite this, acoustic transmitting power may be considered low in comparison with the ones used in Long Base Line positioning systems, which usually reach values of 180 dB re. 1μPa@1m, the use of longer signals in combination with a broadband frequency range and signal processing techniques will allow us to increase the signal-to-noise ratio, and having an acoustic positioning system with about 1 μs accuracy (~ 1.5 mm) over distances of about 1 km, using less acoustic power, that is, minimizing the acoustic pollution.

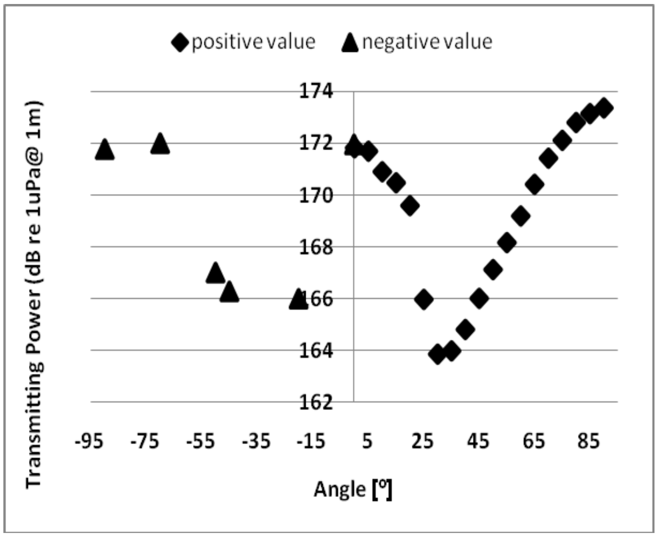


Figure 8: Transmitting Acoustic Power of the transceiver.

IV. INTEGRATION IN THE ANTARES NEUTRINO TELESCOPE

The system tested was finally integrated in the active anchor of the Instrumentation Line of ANTARES through the Laser Container used for timing calibration purposes. In fact, a new functionality for the microcontroller was implemented (to control the laser emission as well). The FFR hydrophone was fixed in the base of the line at 50 cm from the standard emitter transducer of the ANTARES positioning system with the free area of the hydrophone looking upwards. It has been fixed through a support of polyethylene designed and produced at the *Instituto de Física Corpuscular*, Valencia (Spain). The SEB was located inside a titanium container holding other electronic parts and the laser. Figures 9 and 10 show some pictures of the final integration of the system in the anchor of the Instrumentation Line of ANTARES. Finally, the Instrumentation Line was successfully deployed at 2475 m depth on 7th June 2011 at the nominal target position. The connection of the Line to the Junction Box will be in autumn 2011, when the ROV will be available, and afterwards the transceiver will be fully tested in real conditions.

V. SUMMARY AND CONCLUSIONS

We have discussed the needs of the acoustic positioning system in underwater neutrino telescopes, and presented the acoustic transceiver developed at UPV for the positioning system of KM3NeT. We have shown the results of the tests and measurements done to the FFR hydrophones and to the SEB associated, concluding that the transceiver proposed can be a good solution with the requirements and accuracy needed for such a positioning system. The transceiver, with low power consumption, is able to have a transmitting power above 170 dB ref. 1μPa@1m that combined with signal processing techniques allows to deal with the large distances involved in a neutrino telescope. Moreover, the changes performed in the transceiver, particularly in the SEB, show the capacity to adapt the electronic parts to the situation and available conditions.

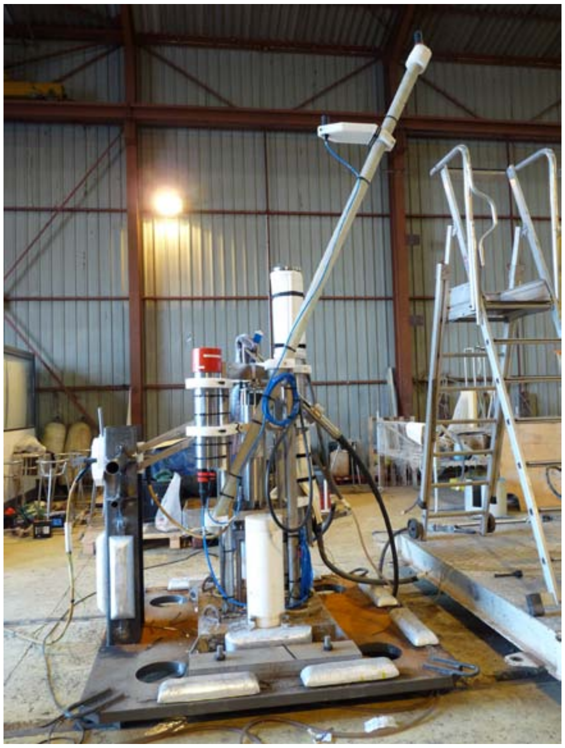


Figure 9: Picture of the anchor of the Instrumentation Line of ANTARES with the final integration of the transceiver.

The system has been integrated in the ANTARES neutrino telescope and now, we are waiting for the connection of the Instrumentation Line to test the transceiver in situ.

Finally, we would like to remark that the acoustic system proposed is compatible with the different options for the receiver hydrophones proposed for KM3NeT and it is versatile, so in addition to the positioning functionality, it can be used for acoustic detection of neutrinos studies or for bioacoustic monitoring of the sea.

Moreover, the transceiver (with slight modification) may be used in other marine positioning systems, alone or combined with other marine systems, or integrated in different Earth-Sea Observatories, where the localization of the sensors is an issue. In that sense, the experience gained from this research can be of great use for other possible applications.

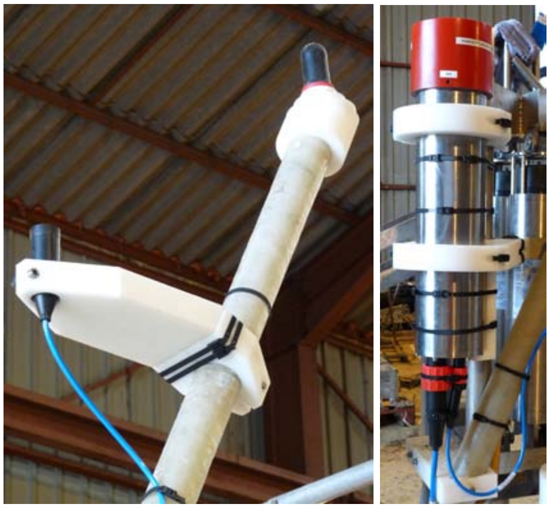


Figure 10: Views of the FFR hydrophone with the support and of the titanium laser container that contains the SEB.

ACKNOWLEDGMENTS

This work has been supported by the Ministerio de Ciencia e Innovación (Spanish Government), project references FPA2009-13983-C02-02, ACI2009-1067, AIC10-00583, Consolider-Ingenio Multidark (CSD2009-00064). It has also been funded by Generalitat Valenciana, Prometeo/2009/26, and the European 7th Framework Programme, grant no. 212525.

REFERENCES

[1] The KM3NeT Collaboration, KM3NeT Technical Design Report (2010) ISBN 978-90-6488-033-9, available on www.km3net.org.

[2] M. Ageron et al.(ANTARES Collaboration), ANTARES: the first undersea neutrino telescope, Nucl. Instr. And Meth. A, vol. 656, 2011, pp. 11-38.

[3] M. Ardid, ANTARES: An Underwater Network of Sensors for Neutrino Astronomy and Deep-Sea Research, Ad Hoc & Sensor Wireless Networks, vol. 8, 2009, pp. 21-34.

[4] C. Bigongiari, The KM3NeT project for a Very Large Submarine Neutrino Telescope, Ad Hoc & Sensor Wireless Networks, vol. 8, 2009, pp. 119-140.

[5] J.A. Aguilar et al., Time calibration of the ANTARES neutrino telescope, Astropart. Phys. 34 (2011) 539.

[6] M. Ardid, Positioning system of the ANTARES neutrino telescope, Nucl. Instr. and Meth. A, vol. 602, 2009, pp. 174-176.

[7] S. Toscano et al., Time calibration and positioning for KM3NeT, Nucl. Instr. and Meth. A, vol. 602, 2009, pp. 183-186.

[8] F. Ameli et al., R&D for an innovative acoustic positioning system for the KM3NeT neutrino telescope, Nucl. Instr. and Meth. A, vol. 626-627, 2011, pp. S211-S213.

[9] M. Ardid et al., R&D towards the acoustic positioning system of KM3NeT, Nucl. Instr. and Meth. A, vol. 626-627, 2011, pp. S214.

[10] H. Motz, Position calibration for the future KM3NeT detector, , Nucl. Instr. and Meth. A, vol. 623, 2010, pp. 402-404.

[11] C.H. Sherman and J.L. Butler, Transducers ad Array for Underwater Sound, The Underwater Acoustic Series, Springer, 2007.

[12] M. Ardid et al., A prototype for the acoustic triangulation system of the KM3NeT deep sea neutrino telescope, Nucl. Instr. and Meth. A, vol. 617, 2010, pp. 459-461.

[13] http://www.sensortech.ca.

[14] M. Barr, Introduction to Pulse Width Modulation, Embedded Systems Programming 14 No. 10, 2001,p. 103.

[15] F. Simeone et al., Design and first tests of an acoustic positioning and detection system for KM3NeT, Nucl. Instr. and Meth. A, vol. 662, 2012, pp. S246-S248.

Data Cleansing and Selective Query Processing in Sensor Networks

Maria Drougka, Theodore Tsiligiridis
Division of Informatics, Mathematics and Statistics,
Department of General Science, Agricultural University of Athens,
Athens, Greece, EU
{*mdrougka, tsili*}@*aua.gr*

***Abstract*—Wireless sensor networks have been widely used in numerous monitoring applications and real life phenomena. Due to the low quality of sensors and random effects of the environment, the collected sensor data is subject to several sources of errors. Such errors may seriously impact the answer to any query posed to the sensors and therefore, it is very critical to clean the sensor data before using them to answer queries or conduct data analysis. Well known data cleansing approaches, are used to meet the requirements of both energy efficiency and quick response time in many sensor related applications. Several energy saving methods based on clustering sensors, so that sensors communicate information only to cluster-heads and then the cluster-heads communicate the aggregated information to the processing center are reviewed, and discussed. We focus on the distribution of sink operation among sensor nodes and specify the criteria of selecting effective cluster-heads that enable both balanced and low energy consumption in data query and collection. As a result of the effective selection of cluster-heads, both balanced and decreased energy consumption are superior compared with the conventional retrieval model, which is not using cluster-heads, because their use provide maximum data aggregation among various sensors, and it alleviates the heavy energy consumption near the sink.**

***Keywords*-Data cleansing, selective query, Wireless Sensor Network.**

I. Introduction

Wireless Sensor Networks (WSN) have become an important source of data with numerous applications in monitoring various real-life phenomena as well as agro-environmental and industrial applications. Data delivered by sensors is typically noisy and unreliable. The task of improving, correcting and filtering the incoming data is usually referred as *data cleansing*. Such errors may seriously impact the answer to any query posed to the sensors, since they may yield imprecise or even incorrect and misleading answers, which can be very significant if they result in immediate critical decisions or activation of actuators.

A. Dimensionality of Data Cleansing

Data cleansing has different dimensions, which are related to the different kinds of noise.

- *Noise Reduction or Smoothing:* Stochastic noise or noise caused by interferences can often be treated by some kind of smoothing (like a moving average) or other noise reduction methods. The idea is to extract the relevant portion of the incoming signal. The incoming signal is modified but not removed. Thus the risk of loosing relevant information by this kind of cleansing is very low.
- *Filtering:* Faulty data due to sensor failures or human errors can often be recognized by cross-checking information from different sources. This is often based on physical models. In certain places for example one could use temperature and smoke sensors to detect fire. Cross-checking the signals from spatially close sensors of both kinds allows to detect unlikely or impossible measurements of a single sensor and could help to avoid false-alarms. However filtering has to be used carefully as it might be vulnerable for manipulations.
- *Generation:* Data cleansing may also try to fill gaps in the data due to missing sensors or transmission errors. This task is likely to employ physical models. Filling up missing data can be very useful as it allows to draw an overall picture of the state of an infrastructure. Such data should be marked as generated, virtual or simulated since it is not as reliable as real measurements.

B. Methods for Data Cleansing

In order to remove noisy sensor data or at least reduce the effect that brought about by noises is a key issue to answer queries or detect events accurately. Statistical and probabilistic modelling techniques have been used to solve the issues we discussed earlier. Modelling usually involves two phases: training and testing. In training, the parameters of the characteristic function representing the data are learned. Sometimes held-out data is used for validation to further improve the accuracy of the training process by preventing over-fitting. In the next phase, predictions are made about the testing data. Training is frequently done off-line while testing can be done either off-line or on-line.

Sensor data is temporal and spatial in nature. In general, a reading is usually of the format <*sensor-id, location, time, value*>. If the sensors are static, then the location field is usually omitted. Individual observations are assumed to be independent. Traditional data cleansing techniques cannot be applied to sensor data as they do not take into account the strong spatial and temporal correlations typically present in sensor data. Nevertheless, well known data modelling methods like Kalman filters and regression have shown good results in capturing spatio-temporal correlations. The Kalman filter is an efficient recursive filter, which estimates the state of a dynamic system from a series of incomplete and noisy measurements. Regression usually involves fitting the best curve for a given set of points. In the case of time-varying and spatial data the use of regression is mainly to find the best curve approximating the readings. This curve can be used not only to find missing or unknown data but also to reduce noise.

So far, interesting prior work has been done on data cleansing in the context of sensor network data. Many toolkits for cleansing noisy and incomplete sensor data are available, supporting interpolation and extrapolation functionalities. In addition, they provide data analysis tools like visualization and a step utility to examine the actual training process. Comparative studies of two methods shows promising results for the Kalman filter for most of the physical attributes modelled.

C. Cleansing and Query processing

If cleansing is performed at the sensors, there would be significant communication cost in sending the parameters of the model to the individual sensors. Furthermore, there is a storage cost associated with storing the parameters. In addition to communication and storage costs, performing the actual cleansing at the sensors would incur a processing cost on the resource-constrained sensors. These problems do not arise if cleansing is done at the sink, given the typical processing power and storage capacity of sinks. Moreover, there would be huge savings by not having to communicate the model parameters to each individual sensor. Given that the lifetime of the sensors is heavily dependent on the amount of communication that they do, communication savings is very important. Having the data and the model at the sink is advantageous when it comes to query processing, as answers to user queries can be easily computed.

We define a monitoring query as a continuous data collection task that requests sensed values from nodes fulfilling selection criteria based on certain physical conditions. For instance, queries that monitor object movements in a field would report the sensed values only from the nodes that have recently sensed these movements. The following are key aspects of monitoring queries:

- *Monitoring queries are selective:* Typically, a WSN can cover an area much larger than the area of interest at any given point in time. For instance, in the monitoring example presented above, although nodes are present all over the field, the object movements may only be found within a limited area. We argue that for energy efficient optimization of such queries, the data collection task should be selectivity-aware.
- *Monitoring queries are continuous:* A monitoring task, by design, is expected to query readings from sensor nodes over an extended period of time. The mainstream WSN database systems have realized the need for continuous queries and provide SQL clauses to define such queries.
- *Monitoring queries select spatially correlated nodes:* Physical phenomena are characterized by their spatial correlation; hence, when monitoring a physical phenomenon, sensor nodes at proximal locations tend to have similar values. Therefore, this spatial correlation, coupled with the notion of selectivity, results in clustered participation. For instance, if a node is selected by a query based on its sensed temperature value, there is a high probability that neighbouring nodes will also be selected by the same query.

Cleansing and query processing can be performed either at the individual sensors or at the sink. Performing the cleansing at the sensor level and query processing at the sink has no clear advantages. This is because communicating a single noisy reading to the sink and performing the cleansing work there incurs less communication cost than communicating all the model parameters over to the sensor itself. The latter, as we have mentioned earlier, imposes unnecessary processing and storage overhead on the sensors.

The remainder of the paper is structured as follows: Section II summarises the background information related with the existing cleansing and querying mechanisms, Section III analyses, in terms of minimization of the total energy spent in the system, two schemes proposed for data collection and querying process, and finally, Section IV concludes the present work with some discussion on the results and a future work.

II. Material and Methods

A. Cleansing Mechanism

1) Weighted moving average algorithm: A well known approach to remove noise in random samples and compute the monitoring values is to use the moving average [1], [2]. Note that moving average in sensor networks has two dimensions. Sensor data are averaged temporally within one sensor, and also spatially among neighbouring sensors. For example, at any time t, the algorithm first averages a sequence of samples $x_{i,t-k+1}, \ldots, x_{i,t}$ at each sensor i, and gets $\bar{x}_i = (x_{i,t-k+1} + \ldots + \bar{x}_{i,t})/k$. We then average values of neighbouring sensors, $\bar{\bar{x}}_i = \sum_{j \in R(i)} \bar{x}_j / |R(i)|$, where $R(i)$ is a set of neighbouring sensors of sensor i.

The above approach is not suitable for sensor network applications, mainly because there is a trade-off between energy efficiency and query response time. For example, to improve the efficiency sampling rates should be low, namely, the interval between two consecutive samples should be long. Thus, any change takes a long time to be reflected in the moving average. On the other hand, if the sampling rate is high, the response to a change is short and therefore more samples need to be taken. However, in practice, since sampling is one of the costly operators for sensors, the energy efficiency of high sampling rate is lowered. To address these two aspects together a weighted moving average algorithm has been proposed [3], that collects confident data from sensors and computes the weighted moving average.

In particular, the *temporal moving average* is defined as:

$$\bar{x}_i^t = (w_{i,t-k+1}x_{i,t-k+1} + \ldots + w_{i,t}x_{i,t})/h \qquad (1)$$

where $w_{i,t}$ is the weight of value $x_{i,t}$ at sensor i, which is related to the confidence of this value and $h = w_{i,t-k+1} + \ldots + w_{i,t}$ is the accumulated temporal weight. In addition, the *spatial moving average* of the data coming from spatially correlated sensors is defined as:

$$\bar{x}_i^s = \frac{1}{m} \sum_{j \in R(i)} b_{j,t} w_{j,t} x_{j,t} \qquad (2)$$

where $m = \sum_{j \in R(i)} b_{j,t} w_{j,t}$ is the accumulated weights and weight $b_{j,t}$ is a digit value according to whether or not a neighbouring sensor j reports the confident value $x_{j,t}$ to the sink. We can greatly achieve low sampling cost when $x_{i,t}$ is a smooth and predictable value. Data cleaning is performed in two places. On the sensor side, multiple sampling is used to remove random noises in data, whereas on the sink side, the weighted moving average is used in both dimensions to further smooth the data.

From equation (1) the temporal part is a weighted version of the normal moving average $\bar{x}_i^t$ provided that $w_{i,t} = \hat{w}_{i,t} \forall i, t$ and k is a given window size. The spatial part includes value $x_{j,t}$ from sensors j, where $x_{j,t}$ is in the error range $[x_{j,t} - e, x_{j,t} + e]$ of $x_{i,t}$, and it has a high confidence with its weight $w_{j,t} > 1$. That means $x_{j,t}$ is in $x_{i,t}$'s error range and has high confidence to improve the moving average at $x_{i,t}$. In a similar way the spatial part of moving average is provided by equation (2). Finally the weighted moving average is the combination of the temporal and spatial parts: $\bar{x}_{i,t} = (h\bar{x}_i^t + m\bar{x}_i^s)/(h + m)$.

2) k-Means Algorithm: The latest sensor network techniques enable a sensor to sense multiple measures simultaneously. Therefore, multidimensional data analysis such as clustering is needed for analyzing sensor network data. For example, for temperature sensor network the dimension is one, for a sensor network measuring both temperature and humidity the dimension is two, whereas for a sensor network measuring temperature, humidity, and density at the same time the dimension is three, etc. k-means algorithm proposed in [4] and [5], is an old, simple, and well known method for analysing multidimensional data by separating data into different clusters. It converges very fast when the dimension of data is small, such as in the cases of environmental problems.

In mathematical terms the problem is defined as follows. We consider a set of points U, where U is assumed to be an r-dimensional space. At a time instant i, the value of a point $u \in U$ is $u^i = (u_1, \ldots, u_r)$. In other words, a point in U can be regarded as a moving object in an r-dimensional space. We are interested in k-means clustering of the current values of the points at each time instant. At an instant i, let $c_1, \ldots, c_k$ be k points, which may or may not be in U. The points in U can be partitioned into k exclusive subsets $U_1, \ldots, U_k$ according to their values at instant i: a point $u \in U$ is assigned to cluster U_i if:

$$dist(u^i, c_i) = \min_{1 \le j \le k} \{dist(u^j, c_j)\} \qquad (3)$$

where $dist()$ is the distance function in question. Note that the points $c_1, \ldots, c_k$ are the k-means of U if:

$$\min \left\{ \sum_{i=1}^{k} \sum_{u \in U_i} dist(u^i, c_i) \right\} \qquad (4)$$

To lower down the communication cost a hierarchical clustering structure is proposed in [6] and [7]. A set of sensors are grouped together, and one of them becomes a Cluster-Head (CH). In data collection, each sensor in the cluster sends its data to the CH, and the CH reports the aggregated data to the sink. A distributed randomized algorithm was proposed to cluster the sensors. Each sensor takes a probability to become a CH, and broadcasts itself to other sensors within certain hops. The sensors tha t are not CHs join the closest CH. The optimal parameters of the clustering, which minimize the communication cost are also derived. However, as time goes by, the status of each sensor may change, and thus the so-built hierarchical structure may not always be optimal. For example, some sensors may use more energy to collect data, so they are dying faster than the others. If we use such sensors as CHs, the lifetime of the whole cluster decreases. The problem has been studied in [8], where it periodically recomputes the CHs based on the residual energy of each sensor and its relationship to other sensors.

To maximize the lifetime of the sensor network a hierarchical model is used that utilizes data aggregation and in-network processing at two-levels of the network hierarchy. First, a set of sensor nodes called Local CHs (LCHs) are elected to form a fixed virtual routing architecture on, which the first level of aggregation and routing is performed. Then, the problem is that of finding an optimal subset of LCHs, called Master CHs (MCHs), which are selected to perform the second level of aggregation under the objective to maximize the network lifetime. Clearly, the problem of optimal selection of MCHs is NP-complete since it is equivalent to the p-median problem in graph theory, which has been shown to be NP-complete [9].

To implement the k-means algorithm one needs to initialize mean values with k, say, random MCHs from the set of LCHs. To initialize k MCHs we choose k LCHs at random from the set of sensors S, $|S| = s \in \mathbf{N}$ while making sure that the pairwise distance (number of hops) between these k MCHs is large enough. One way to do that is to choose m LCHs at random from S (where $m >> k$ but $m < s$) and then perform k-means clustering on those m LCHs. To initialize this sub-problem, we arbitrarily assign LCHs to different k MCHs(classes). The algorithm simply iterates until a termination condition is met. In every iteration, each LCH in S is assigned to a chosen MCHs such that the distance from LCHs to sink through that selected MCHs is minimized. Then, for each class, we recalculate the means of the class based on the local nodes that belong to that class. Theoretically, k-means should terminate when no more LCHs are changing classes; however, in practice, this may require a large number of iterations.

3) Weighted data cleansing: The above approach introduces a periodic multilevel data cleansing algorithm aiming to optimize the volume of data transmitted thus saving energy consumption and reducing bandwidth on the network level. It is based on a tree network where sensed data needs to be aggregated on the way to their final destination. In particular, a frequency filtering technique is applied, which exploits the ordering of measurements according to their frequencies, by means of the number of occurrences of this measure in the set. Since sensor nodes are deployed randomly, it is most likely that neighbouring nodes generate similar sets of data.

Data aggregation works in two phases, the first one at the LCHs, where each node compacts its measurements set according to a link function. The objective is to identify similarities between neighbouring sensor nodes, and integrate their sensed data into one record while preserving information integrity. This first level of cleansing process is called *in-sensor process periodic cleansing*. A second cleansing process is applied on the level of the MCH itself, where the frequency filtering technique will be applied. The cleaned data is finally sent from the MCHs to the sink.

In periodic sensor networks, at each period T each node sends its aggregated data set to its proper LCH, which subsequently aggregates all data sets coming from different sensor nodes and sends them to the sink. We consider that each sensor node i at each slot takes a new measurement y_t^i. Then node i forms a new set of sensed measurements M_i with period T, and sends it to the aggregator. Note that, a sensor node can take different kind of measures (e.g., temperature, humidity, light, etc), making of y_t^i a vector instead of a scalar. For the sake of simplicity, in the rest of the paper we shall consider that $y_t^i \in \mathbf{R}$. It is likely that a sensor node takes the same (or very similar) measurements several times especially when t is too short. In this phase of aggregation, we are interested in identifying duplicate data measurements in order to reduce the size of the set M_i. Therefore, to identify the similarity between two measures, we define the link function between two measures as:

$$link(y_{t_j}^i, y_{t_k}^i) = \begin{cases} 1 & : \text{ if } \|y_{t_j}^i - y_{t_k}^i\| \leq \delta \\ 0 & : \text{ otherwise} \end{cases} \quad (5)$$

where δ is a threshold determined by the application. Furthermore, two measures are similar if and only if their link function is equal to 1. The frequency of a measurement y_t^i is defined as the number of the subsequent occurrence of the same or similar (according to the link function) measurements in the same set. It is represented by:

$$f(y_t^i) = \sum_{j=t_i+1}^{T} link(y_{t_j}^i - y_{t_k}^i)$$

4) Greedy Algorithm: In the virtual graph of the set of sensors S, $|S| = s \in \mathbf{N}$, LCHs are numbered sequentially from 1 (left-upper corner) to s starting from 1 and then from left to right proceeding row by row. The greedy algorithm starts with the first node in S and proceed sequentially through the whole topology in a left-to-right and top down fashion [10]. We assume that each LCH has global information about network topology (shortest path from each LCH - LCH and from LCH - sink can be obtained by running all-pairs shortest path algorithm) and this information is broadcasted before the greedy algorithm is executed. To construct the MCHs graph, the shortest path is first established for the first LCH source node acting as the first MCH to the sink. Each subsequent LCH source is incrementally connected to the MCHs graph either as a MCH itself or by selecting a MCH from the set of nodes that are already been allocated as MCHs. In the latter, a LCH selects the MCH that results in the least power consumption to reach the sink. Then, LCH sends the MCH its group number. A MCH holds a registry for its constituent LCH and for those LCHs that exist in multiple groups to distinguish data coming common LCHs. The process is iterated until all LCHs are covered by MCHs and until there is no change in the value of the objective function, which is to obtain the least total power consumption to reach the sink.

B. Query Resolution Mechanism

When the sink receives a query message, it resolves the name of the query to the corresponding sensor IDs, according to the resolution table. After a query′s name is translated into an ID group corresponding to sensors, the sink calculates the query area by deriving a rectangle, in which all corresponding sensor nodes reside. A rectangle area is calculated based on the locations (x_i, y_i) of each sensor $i = 1, 2, \ldots, s$ that is in the ID list obtained from the query resolution: $R_{xy} = R_x \times R_y = [min\{x_i\}, min\{y_i\}, max\{x_i\}, max\{y_i\}] \;\; \forall i = 1, \ldots, s$ with $s = |S|$ be the number of all sensors in the network. It is assumed that all sensors can reach the sink, using multi-hop communication. In addition, whenever a sensor $j = 1, \ldots, s$ is activated it emits a constant energy signal E_j in the surrounding environment. The measured signal is inversely proportional to the distance from the activated sensor raised to some power $\gamma \in R^+$, which depends on the environment. As result, the measurement of a CH $h \in H, H \subseteq S$ is given by the equation:

$$z_h = min \left\{ E^{ch}, \sum_{j=1}^{s} \frac{E_j}{r_{hj}^{\gamma}} \right\} + w_h \tag{6}$$

where E^{ch} is the maximum energy, which can be recorded by a CH, r_{hj} is the radial distance of CH h from the sensor $j \in S$, $r_{hj} = \sqrt{(x_h^{ch} - x_j)^2 + (y_h^{ch} - y_j)^2}$ and w_h is additive Gaussian noise with zero mean and variance σ_h^2. Note that the neighbourhood of a CH h is defined as the set of all sensors that are located at a distance less than or equal to r_c. Therefore the neighbourhood of a CH $h \in H$ is the set of all sensors in S that are in the disk centred at $\vec{x}_h$ with radius r_c, or,

$$CH_{r_c}(h) = \{j : \|\vec{x}_h - \vec{x}_j\| \leq r_c, \forall j \in S, j \neq h\},\ \forall h \in H$$

Thus, in case r_c is the communication range of the sensor, then the set $CH_{r_c}(h)$ defines all sensors that are one hop away from the CH h.

Generally, sensing data is collected at appropriate CHs, at which is aggregated and forwarded to the sink. There are two types of CH selection schemes. One is *fixed selection scheme*, in which an identical CH will be selected at different times for the queries that have the same query resolution result. The next is *dynamic selection scheme*, in which various CHs will be selected at different times for the queries that have the same query resolution result. Note that for the fixed selection scheme a CH is selected to be the node nearest to the centre location of the query area, whereas for the dynamic selection scheme a rotation operation on the CH is utilised.

Localised data query distribution consists of query unicast and query geocast. The unicast distribution is used to deliver query messages from a sink to the CH. Based on the nodes' location there are many approaches for the sink to calculate a source route to the CH. For example, in the Sink-CHs-Sensors scheme [11], the sink first selects and includes in the source route the sensor that is nearest the CH among one-hop neighbours of the sink. Then the sink calculates the next sensor in the source route, by selecting the sensor nearest to the CH among the neighbours of the previous selected sensors in the source route. This process continues until a source route is found to the CH. As soon as the query message is delivered by the CH in the query area, it is geographically broadcast to all sensors inside the query area. The CH forwards the query message to its one-hop neighbours.

Localised sensing data collection consists of local data delivery, data aggregation, and aggregated data delivery to the sink. On receiving a query message, the corresponding sensors send the sensing data back to the CH in a local region using the reverse path obtained from the query geocast initiated by the CH. The CH collects the sensing data locally before forwarding it to the sink and aggregates the sensing data by placing multiple sensing data into one packet.

In the following an analysis and discussion is provided on the energy bottleneck in data collection of both, the Sink-Sensors-Sink and the Sink-CH-Sensors schemes, by means of the fixed and dynamic selection schemes, suggested above. As it will be seen, under certain conditions, the effective selection of CHs and MHs are comparable over the conventional Sink-Sensors-Sink retrieval model in terms of both balanced and saving energy consumption.

III. Energy Model Analysis

A. Selection of CHs in Data Query

For simplicity we assume the query area is a square consisting of $|S| = s$ nodes. The average energy consumption of a one-hop transmission of a packet is assumed to be E_0. In the conventional model flooding is typically adopted for disseminating a query to sensors in the network. The energy cost of a data query is given by $E_{dq}^c = s * E_0$. In the Sink-CHs-Sensors information retrieval model, the operation consists of unicasting from sink to CH, and geocasting from CH to sensors in the geocast area, which is a combined rectangle area that contains both CH and the query area. Assuming $p \in (0, 1]$ presents the ratio of the number of sensors in the geocast to the total sensors s, then $p * s$ equals the number of sensors in the query geocast area. If S_{hops} denotes the number of hops of transmission in the unicast from the sink to the CH then, the energy cost of Sink-CHs-Sensors query is $E_{dq}^{ch} = p * s * E_0 + S_{hops} * E_0$. As a result, the ratio λ_{dq} of data query cost of the Sink-CHs-Sensors model to the conventional model is given by $\lambda_{dq} = E_{dq}^{ch} / E_{dq}^c = (p * s + S_{hops})/s$, which means that greater energy savings is obtained for smaller values of λ_{dq}. Thus, when $\lambda_{dq} \leq 1$ or equivalently $S_{hops} < s(1 - p)$ a Sink-CHs-Sensors query saves energy compared with the conventional scheme. If a query interest area is defined, both values of p and S_{hops} can be determined by the position of a CH.

B. Selection of CHs in Data Collection

In the Sink-Sensors-Sink conventional scheme of data collection, the energy cost can be approximately given by the sum cost of the replied data unicast from each sensor to the sink. Let $p_1, p_2 \in (0, 1]$ be the ratio of sensors that will reply a query message, and the ratio of aggregated data size to the non-aggregated data size, respectively. Then $p_1 * s$ equals the number of sensors corresponding to the Sink-CHs-Sensors query and $p_2 * (p_1 * s)$ equals the number of sensors corresponding to the aggregated data size. In addition, let $\bar{L}_0$ and $\bar{L}_{ch}$ be the average route length from each corresponding sensor to the sink and the average route length from each corresponding sensor to the CH, respectively. Then, the energy cost of the conventional data collection scheme, denoted by E^c_{dc}, is given by $E^c_{dc} = p_1 * s * E_0 * \bar{L}_0$, whereas the energy cost of the Sink-CHs-Sensors data collection information retrieval scheme, denoted by E^{ch}_{dc}, is given by $E^{ch}_{dc} = p_1 * s * E_0 * \bar{L}_{ch} + p_2 * (p_1 * s) * S_{hops} * E_0$. Thus, the ratio of data collection cost of Sink-CHs-Sensors to the conventional model will be given by $\lambda_{dc} = E^{ch}_{dc}/E^c_{dc} = (\bar{L}_{ch} + p_2 * S_{hops})/\bar{L}_0$. Obviously, in case $\lambda_{dc} \leq 1$ or equivalently $\bar{L}_0 > \bar{L}_{ch} + p_2 * S_{hops}$, there is energy saving in Sink-CHs-Sensors data collection scheme compared with convention data collection scheme.

IV. Conclusions

In this work we reviewed some interesting clustering based approaches employing data cleansing, data aggregation, and data query in order to extend the lifetime of sensor networks. Further, we have discussed a distributed algorithm for organizing sensors into a hierarchy of clusters with an objective of minimizing the total energy spent in the system to communicate the information gathered by these sensors to the information-processing center (sink). As it was expected we indicated (a rigorous proof remains an open question) that the sensors, which become the CHs in the proposed architecture spend relatively more energy than other sensors because they have to receive information from all the sensors within their cluster, aggregate this information and then communicate to the higher level CHs or the information processing center. However, cluster-based algorithms along with data aggregation and in-network processing can achieve significant energy savings in the sensor networks.

Data aggregation and in-network processing techniques is performed at two levels. We introduced a method to select master/local CHs such that the network lifetime is maximized. Clearly, data aggregation was affected by several factors, such as the placement of aggregation points, the aggregation function, and the density of the sensors in the network. In this framework the determination of an optimal selection of aggregation points, by means of reducing the number of redundant data sent to the end user while preserving data integrity, was crucial and thus very important. In the analysed schemes, sensing data was collected at appropriate CHs, at which data was aggregated and sent to the sink. A query's name was resolved into the IDs and locations of corresponding sensor nodes before being distributed to the network. According to the location of sensor nodes, query distribution and data collection were performed in a corresponding local area. The query message was efficiently unicasted to the CH in a query area, and was then forwarded to a localised area of the network. Sensing data were collected at a CH, at which data were aggregated and sent to the sink. The energy bottleneck analysis and discussion on the criteria of selecting effective CHs show that the discussed schemes were promising.

References

[1] S. Jeffery, G. Alonso, M. Franklin, W. Hong, and J. Widom, "Declarative support for sensor data cleaning," in *Proc. of the 4th IEEE International Conference on Pervasive Computing and Communications (PerCom'2006)*, Piza, Italy, Mar. 2006.

[2] J. Hellerstein, W. Hong, S. Madden, and K. Stanek, "Beyond average: toward sophisticated sensing with queries," in *Proc. of the 2nd International Conference on Information Processing in Sensor Networks (IPSN'2003)*. Berlin, Heidelberg, Germany, EU: Springer-Verlag, 2003.

[3] Y. Zhuang, L. Chen, X. Wang, and J. Lian, "A weighted moving average-based approach for cleaning sensor data," in *Proc. of the 27th International Conference on Distributed Computing Systems (ICDCS'2007)*. Washington, DC, USA: IEEE Computer Society, 2007.

[4] J. Macqueen, "Some methods of classification and analysis of multivariate observations," in *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkley, UC, USA, 1967, p. 281297.

[5] S.Z.Selim and M. Ismail, "k-means-type algorithms: A generalized convergence theorem and characterization of the local optimality," *IEEE Trans. on Pattern Analysis*, vol. 6(1), pp. 81–86, 1984.

[6] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. on Wireless Communications*, vol. 1(4), pp. 660–670, 2002.

[7] K.Padmanabhan and P.Kamalakkannanand, "Energy efficient adaptive protocol for clustered wireless sensor networks," *International Journal of Computer Science Issues (IJCSI)*, vol. 8(5), pp. 296–301, 2011.

[8] O. Younis and S. Fahmy, "Heed: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Trans. on Mobile Computing*, vol. 3(4), pp. 366–379, 2004.

[9] M. Garey and D. Johnson, *Computers and Intractability. A guide to the Theory of NP-Completeness*. W.H. Freeman and Company, 1979.

[10] R. Haider, M. Javed, and N. Khattak, "Design and implementation of energy aware algorithm using greedy routing for sensor networks," *International Journal of Security and its Applications*, vol. 2(2), pp. 71–86, 2008.

[11] R. Teng, B. Zhang, and Y. Tan, "Design and analysis of intelligent sink for the information retrieval in sensor networks," in *Second International Conference on Sensor Technologies and Applications (SENSORCOMM'08)*, USA, Aug. 2008, pp. 348–353.

e-Report Generator Supporting Communications and Fieldwork: A Practical Case of Electrical Network Expansion Projects

Emerson Castaneda-Sanabria
, Miguel-Ángel Manso-Callejo
Universidad Politécnica de Madrid
Madrid, Spain
emecas@ieee.org, m.manso@upm.es

Francisco-Javier Moreno-Marimbaldo
Department of the Environment
Red Eléctrica de España (REE)
Madrid, Spain
fmoreno@ree.es

Federico-Vladimir Gutiérrez-Corea
Universidad Politécnica de Madrid
Madrid, Spain
fv.gutierrez@upm.es

Abstract— **In this piece of work we present a simple way to incorporate Geographical Information System tools that have been developed using open source software in order to help the different processes in the expansion of the electrical network. This is accomplished by developing a novel fieldwork tool that provides the user with automatically generated enriched e-reports that include information about every one of the involved private real estates in a specific project. These reports are an eco-friendly alternative to paper format, and can be accessed by clients using any kind of personal device with a minimal set of technical requirements.**

Keywords - Geoprocessing, Assets and Rights Report (ARR), Government and public GIS, Geo-modeling.

I. Introduction

Electric power supply is essential for the functioning of society since it plays a prominent role in maintaining the standard of living [1]. Electric power is the engine of growth of any developing country, so no major economic activity can be sustained without adequate and reliable power [2]. For this reason the electric power transport and distribution is considered a public utility service.

The Spanish Electrical Network (REE) is the owner of the entire high voltage electric energy transport network in the Spanish territory; it is responsible for its expansion, maintenance and management. The high voltage electric lines located on the territory induce different affections on public and private property land. These affections may be of different types of occupation or easement and should be thoroughly identified on the pertinent documentation to be included into the execution project of these types of infrastructures. Every project is endorsed, approved and processed according to the responsible body corresponding to every particular case; 400 kV installations should be processed according to the requirements of the State Administration; 220 kV installations are handled complying with the requirements of the involved regional governments. This fact implies that the way the projects are presented are not homogeneous since requirements are variable in the Communities and are continuously evolving. This changing situation concerning the requested documentation compels the REE to handle the project data in a structured, standardized way, integrating new processes within consolidated workflow.

The project information is handled as a whole and in the case of possible affections on real estate, lists are generated with the cadastral data of every plot and the result of the calculation of affections due to the new electrical installations. In Spain these lists are affected Assets and Rights Reports (ARR) for landowners. Due to the current high demand, the creation of ARR has been turned into an evolving process with the aim of reaching the goals defined by the strategic plans of the Ministry of Industry, Tourism and Trade. These ARR always go hand in hand with a set of plans containing the pertinent cadastral cartography and the information about the new installation where geographic information about affections is incorporated. The association ARR – set of plans becomes an indivisible unit needed for understanding of the new situation that will take place after building and commissioning of the new installation. From the point of view of the technical drafting of the project, its deliverables, and the competent Administration in the handling of each project, that information is sufficient for the understanding of the project. However, the data of affections should be made known through the public information, so that according to the law the owners of the affected assets have knowledge of the future situation and are able to plead as they see fit.

GIS technologies have an invaluable number of applications in most of the areas related to human development. The electricity market is one of the candidates highly likely to be improved by adopting GIS technologies. Some interesting examples of applications of GIS technologies in the electricity market are: a mobile computing system for repairing and patrolling of electrical power facilities as described in [3], a decision support system based on spatial information of land resources [4], and the development and application of GIS in power marketing and expansion [5].

This paper introduces a simple way to incorporate GIS open source tools in order to assist different processes in electrical infrastructure projects. Some of the processes that are targeted in this work are the group of activities derived from an expansion of electrical networks.

Our contribution can be summarized as a fieldwork tool that supports technical aspects during the development of negotiation process activities. It provides a clear and human understandable preview of future affections that will be produced because of an expansion of the electrical network.

ISBN: 978-1-61208-178-6

It also offers a global graphical vision of the susceptible elements included in the project. This vision can be distributed using a tool that allows getting enriched automatically generated reports. Economic, statistical, and geometric information about each individual part of the project can be included with any kind of useful data and descriptors.

On one hand, the automatically generated reports can be taken on as an official communication tool that involves REE, parcels owners, public administration organisms and the citizens. On the other hand, a remarkable purpose is to reduce the impact on natural resources (paper) by focusing on the adoption of e-information policy to expand information society. These kind of e-reporting tools are easily developed with open source resources on an inexpensive hardware infrastructure, which can be accessed by clients using almost any kind of personal device with a minimal set of technical requirements.

The remainder of this paper is structured as follows: section II provides a complete definition of the issue and its solutions. A review of the pertinent literature is presented in Section III.. Section IV presents the proposed methodology followed by the technical aspect of our proposal. The obtained results are discussed in section V. Finally, section VI includes conclusions and future development.

II. Definition Of Issue

The issue is the need for a graphical representation of the future affections that will be brought about by the expansion of the electrical network. The purpose is for involved citizens, government, and the electrical network company to be offered as an available digital system supporting official communications (e-reports) able to incorporate economic, statistical, geometric information, or any kind of useful data about the expansion projects. We provide a system to give a solution to the lack of this type of communication so far.

Public utility administrative recognition implies a number of obligations which include the formal legal notification to the owners of the affected assets. The government gets closer to citizens by showing a growing concern for the difficulties of the plots owners affected in understanding the technical information, especially old age citizens for whom cadastral codes and interpretation of maps is a problem. The legal notification to owners of the affected plots is carried out through publication on the Official State Gazette in every case. The projects depending on the State government are published on the Official State Gazette (BOE), Official Province Gazette (BOP) and on two of the most widely read newspapers. The projects whose competence rests with the regional government are published on the Official Autonomous Community Gazette, the Official Province Gazette and two of the most widely read newspapers. Thus the owners of the affected plots will learn about the new situation that will take place in their properties.

Pari passu the Local government has opened a new way of communication with the owners consisting of sending them specific information about the affections of the new installation on their assets. This information should be generated by the REE and this in turn could bring about important delays in the preparation of the project and in the terms established with the different players of the process.

The solution to this new requirement should be integrated into the consolidated workflow, generating a new result using the available data. To that effect, we have designed two types of tools. The first one enabling visualisation of the project's geographic information using maps publishing standards (WMS-OGC) [6]. The second type of tool carries out automatically the necessary tasks to generate the individualized report of the affections of the plots involved in every project. The generation of the reports comprises the creation of a main page made up of a map with the geographic information of the surveyed plot as well as a page for each one of the possible affections.

The manual procedure for generation of these documents comprises different manual processes as follows: (1) *Generation of PDF page* with a quantitative summary and an ancillary graphic of the affections of the new installation on the surveyed plot followed by the generation of another PDF page for each type of affection sustained by the plot. The time consumed by the generation of all the pages making up a single document varies as a function of the number of affections sustained by the studied plot. (2) *Generation of PDF document* by putting together all the pages corresponding to the same plot that have been generated in the previous process. The average time for the generation of a final PDF document is 860 seconds (14 minutes). The average time consumed in the automated procedure is 5 seconds. Therefore a considerable reduction in the terms of execution of this task is achieved.

The new procedure put in place not only influences execution times but it also has a bearing upon the homogeneity of the generated information and its reliability and accessibility. It is of consequence to indicate that it is about a Web application developed with open source software; these facts multiply the possibilities of document generation cutting down the costs derived from software licences.

III. Related Work

The topics covered in this section are: GIS technology applications in the electricity market, main principles of electrical network expansion, tools and methodologies for official communications (e-participation and e-government), and reduction of impact on natural resources.

A. *GIS technology applications in the electricity market*

Shin et al. [3] present the development of a mobile computing system for repair and patrol of electric power facilities. Shin gives special importance to the integration of a GIS database system with a distribution mobile repair

ISBN: 978-1-61208-178-6

work based mobile computing in order to contribute to modernisation of mobile repair work by implementing a system that can track the location of repair vehicles moving all along, so as to dispatch the vehicles to the failed place quickly. There are three main components of Shin's system: the mobile repair server, the terminal for vehicles (Mobile Data Terminal) a PDA device with GPS-based services and the wireless communication network.

A second application is a decision support system based on spatial information of land resources [4]. In this application, Xu et al. highlight the technical support provided by the spatial information technology to the spatial decision support system. They claim the importance of obtaining quickly and accurately the use of land resources and its real-time monitoring by using GIS powerful management and spatial analysis function with the purpose of setting up and honing the whole monitoring system, comprehending land market trends and providing a basis for the government to timely adjust relevant policies.

Finally, the subject of development and GIS application on power marketing and its expansion are described in [5]; the authors state that this application of GIS should not only realise the effect of visual display, quick inquiry, and data management and update of power facilities and circuits, but should also be connected to other existing system and graphic data. They stress improvement of the working efficiency of the system, avoiding repeated investment and saving resources and operating costs.

From the 3 pieces of work above highlighted, we have identified some concepts and elements to be taken into account in our proposal: (1) the mobile computing component with the adoption of a modular architecture based on layers, with clearly identified functionalities, (2) the ability to support fast and accurate decision making based on GIS, spatial systems and new technologies, even in real time scenarios, and (3) the interconnection of GIS applications not only focusing on the display screen, quick reference, management and updating tasks of facilities and circuits, but also for a strategic investment planning focused on saving costs and resources and working efficiency.

B. *The main principles of electrical network expansion*

An interesting reference about one of the topics related to our work that takes into account the principles of electrical network expansion after a long-term period of planning expansion of the Unified Energy System in Russia is [7]. In that piece of work Voropai starts from the principles of essential updating of electrical network expansion carrying on to a full description of how the requirements for expansion of the Main Electrical Network (MEN) and development of new technologies for power production, transmission and consumption affect the principles of the MEN formation. Furthermore the author includes information about the structure of electrical network expansion planning problems.

In this paper we highlight the principles related to planning and required tools for modelling every single stage of the expansion planning process, namely: Uncertainty Planning: in order to decrease the uncertainty of future conditions of the MEN expansion, it is necessary to develop a technology for planning and monitoring this expansion and the adjustment of its planning. Stages of Expansion Planning Problems: a general structure of the MEN expansion planning problems includes three stages: generation of the MEN expansion options, analysis of their operating conditions, and selection of the best option. The functionality of our system contributes to improve these planning and modelling processes by giving a comprehensive description of the involved areas.

C. *Tools and methodologies for official communications*

In this section, the concepts of e-participation using GIS and e-Government tools are analysed in order to provide a context about the adoption of methodologies and tools to implement official communications. Our proposal is being presented as an innovative case study regarding the electrical network expansion projects, involving citizen participation in responsibilities of REE, as well as national and regional government authorities.

The term e-participation [8][9][10] is referred to the use of information and communication technologies to broaden and deepen political participation by enabling citizens to connect with one another and with their elected representatives. This definition includes all stakeholders in democratic decision-making processes and not only citizen-related top-down government initiatives. Thus e-participation can be viewed as part of e-democracy, where e-democracy means the use of ICT by governments in general, by elected officials, media, political parties and interest groups, civil society organisations, international governmental organisations, or citizens/voters within any of the political processes of states/regions, nations and local and global communities.

On the other hand, the term e-government [8][11][12] is referred to digital interaction between a government and citizens (G2C), government and businesses (G2B), and between government agencies (G2G). This digital interaction consists of governance, business process re-engineering (BPR), and e-citizens at all levels of government (city, state/province, national, and international). Essentially, the term e-Government,Digital Government, refers to ‘How government utilises IT, ICT and other telecommunication technologies to enhance the efficiency and effectiveness in the public sector’.

A study involving an evaluation of GIS tools and e-participation is [13]. In this piece of work, Loukis et al. performed a systematic evaluation of an e-participation platform based on GIS tools. The evaluation methodology was based on the Technology Acceptance Model (TAM) [14]. It was worked out and adapted to this specific types of information system taking into account the specific

ISBN: 978-1-61208-178-6

objectives and capabilities of this platform. The main evaluation dimensions were usage, ease of use, functional usefulness, political usefulness and importance of discussion topic; each of them was analysed into a number of sub-dimensions. Using this methodology five pilot applications of this platform in ‘real-life’ situations and problems were evaluated with both quantitative and qualitative techniques. Finally, the work concludes that the use of GIS tools can provide significant value in the area of e-participation, which however depends on a number of context factors such as citizens’ computer literacy and familiarisation, trust in the political system, interest of the sponsoring public authorities, appropriate promotion, importance of the topic under discussion, and quantity and quality of reference information appended on the digital maps by public authorities.

One additional reference related to the improvement of Web content delivery in e-government application is [15]. This piece of work focuses on creating a Web platform for accessing, processing and delivering statistical data in an effective way. The platform consists of a set of integrated tools specified in the organisation, validation, maintenance and safe delivery of statistical information resources on the Web. The idea is the enhancement of a system providing direct dissemination functionalities to authorized international users. Datasets are tested and validated on the national level, then they will be stored in the Transmission System Database, from which it can be downloaded by Eurostat [16]. The work presents all motivations and technology background for the implementations and it also includes a full case study with the consideration of the implementations. Finally, it concludes that the statistics, production, and dissemination process are improved because the validation verifications take place at an earlier stage as it is performed by the National Statistical Institute itself whereas the transmission is performed by Eurostat.

D. Reduction of impact on natural resources

The topic about the reduction of impact on natural resources is mainly handled from two perspectives. One of them is based on the reduction/replacement of paper use, focusing on the adoption of an e-information policy. An additional benefit and motivation for the adoption of our proposal is that these e-information tools are easily developed with open source software on a low-cost hardware infrastructure that can be accessed by customers with almost any personal device with a minimum set of technical requirements.

A second perspective is related to the adoption of a ubiquitous e-custom service such as the system described in [17]. Razmerita and Bjorn-Andersen address the strategic goals for future custom systems such as simplified paperless trade procedures, the advantages of using novel technologies and the inclusion of an innovative e-custom system architecture. The main technologies for innovative trade procedures considered by the authors are the Web Services (WS), Services Oriented Architectures (SOA) and The Tamper Resistant Embedded Controller (TREC). These technologies are directly involved in our proposal, as described in the following section.

On the other hand, adopting the redesign of administrative procedures for international trade is presented by focusing on the ITAIDE Project (Information Technology for Adoption and Intelligent Design for e-Government, www.itade.org) which has two objectives: reducing the administrative costs implied in the international trade transactions while increasing security and controlling trade procedures. Our work is the first approach to the redesign of administrative procedures to make them more efficient and useful. This affirmation is based on the solicitations for the adoption of our tool received from regional government authorities, in order to notify all people possibly affected by electrical network expansion projects and during field work.

IV. Description Of The Proposal

This section presents the proposed methodology and the technical aspect of our proposal including details about system architecture, inputs, procedures, outputs, and web user interface.

A. System Architecture

Fig. 1 shows the proposed solution for the implemented architecture; an prototype has been developed by using an incremental methodology.The main parts of this architecture and a brief description of the architecture are as follows: (1) Web Server: it is responsible for publishing the WebApp on the Internet. This server can redirect the requests to the application server or provide the service as a web container. (2) GeoDatabase: it is the repository that stores the entire spatial information related to the expansion project. (3) Data Services: data services are required to provide additional alphanumerical information and to establish additional connections to other database or GeoDatabase. (4) Map Server: it provides the service of maps using the OGC standard Web Map Server (WMS). (5) Web Reporting Application: it is responsible for handling the clients' requests and for generating the e-reports that integrate spatial and alphanumerical information, through interaction between Data Services and Map Server. It is the core of our proposal making geoprocessing tasks to integrate alphanumerical and spatial information and incorporating novel technologies and standards for web development and GIS. (7) Clients: users of the Web Reporting Application such as mobile devices, PCs, or any device with minimum requirements to establish connection with the Web Reporting Application and to handle answers as e-reports.

B. Inputs

The inputs for the process of generation of e-reports are the Data Tables and the 16 Layers (both are contained in the GeoDatabase Fig. 2.). One of the layers represents the real estates involved in each electrical installation project (Base Layer); this information is generally obtained from the Cadastre. The rest of the layers represent the 15 affections (Affection Layers), which come from studies for every new electrical installation project to take decisions and to estimate the information by affections.

The number of affections that may influence the real estates is 15: aerial trace (superficial electrical cable generally suspended between pylons); underground trace (part of the electrical cable is buried); flight (safety area represented by the possible maximum movement of the electrical cable due to the influence of wind with a velocity of 120 Km/h perpendicular to the axis of the electrical line); tube (representing the vertical projection of the cables on the ground with wind of 0 Km/h); felling (area that should be felled around the trace for safety purposes); permanent occupation area (area occupied by the pylons); permanent underground occupation (surface the underground power line will occupy permanently); temporary occupation area (needed area for the building of the electricity pylons and other materials); temporary underground occupation (needed occupied surface for underground electric wiring); splicing chamber (area occupied permanently by concrete boxes where cable splicing is carried out); telecommunication boxes (surface permanently occupied by the boxes used for telecommunication equipment associated to underground cable for remote manoeuvring of the line); landmarks (surface occupied by concrete posts in place to indicate on the surface the underground channelling of electrical cables); accesses (easement needed for access from the electrical installations for building and maintenance); auxiliary 1 and auxiliary 2 (two generic affections available in the future if needed).

C. Procedures

In a previous stage, information for a new expansion project has been SDI standard-complying and the data have been converted to the interchange Geography Markup Language (GML) format [18]. Once the data have been submitted, using transactional operations to the server which implements the WFS-T, the *GeoDatabase* is ready to serve as data source from the input layers and data tables. The layers may be queried in a standard way through the Map server by using the WMS and WFS services to get spatial information, and data tables through the Data Services which provide the alphanumerical information.

The core of our geoprocessing consists of the integration of both kinds of data sources in order to get a single and portable document format through a real time procedure. On one hand, the procedure identifies, builds, and executes all needed requests to the map server to obtain the graphical information subset. The identification part defines which layers are involved in map construction, the building part integrates the different elements related with bounding box, styles, colours, labels, scale, and the executing part is related to the control of process to request every single petition and manage these graphic temporal results for its future integration into an e-report. On the other hand, the procedure identifies and recovers through the Data Services all the alphanumerical required information to enrich every single e-report. Finally, the process join both kinds of data to build an e-report by filling templates previously structured and designed. This same geoprocessing used to get a single e-report is executed in an iterative way to get all the e-reports of a specific project.

D. Outputs

Fig. 3 shows the e-report of a project of electrical network expansion that includes the ARR of a particular rural area affected by permanent occupation, power lines passing across and deforestation. In this case the e-report is made up of 6 pages; in our system this kind of e-report could contain up to 16 pages.

The first page shows an outline of the zones of every potential impact on the area susceptible to affection by the expansion project. The next 5 pages detail each one of the areas of real estate that could be affected and the way the affections would be distributed into and around the property. This particular example includes 5 kinds of affection: a. accesses, b. flight, c. felling, d. permanent occupation area, and d. splicing chamber. In addition, other output of the tool is a zip file that contains all e-reports related to a specific expansion project in case that the PDF massive generation option has been used.

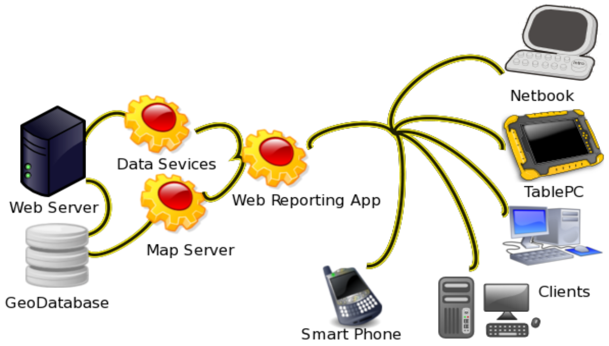


Figure. 1. Proposed Architecture

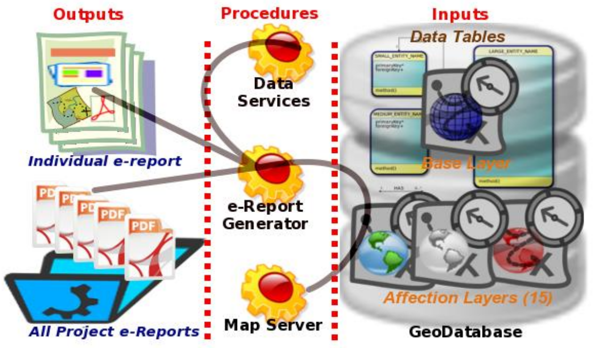
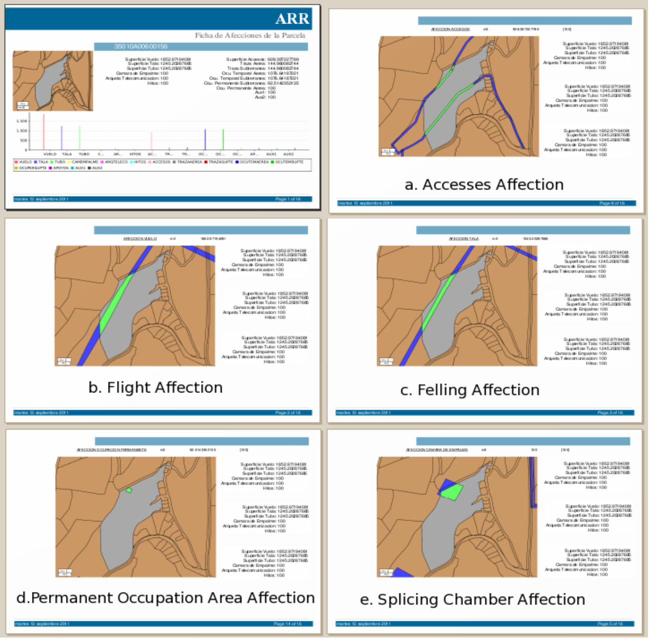


Figure 2. Inputs, Procedures, and Outputs involved in the geoprocessing



a. Accesses Affection

b. Flight Affection

c. Felling Affection

d.Permanent Occupation Area Affection

e. Splicing Chamber Affection

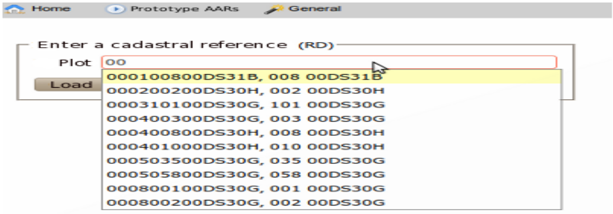
Figure 3. e-Report of affected assets and rights

E. Web User Interface

A Web User Interface is developed using broadly extended Web technologies (HTML and Ajax with the Apache Click API); the system works to assure that almost any mobile device interacts with the Web Applications handling the answers as e-reports (usually PDF).

A Web interface can be appreciated in Fig. 4. It includes part of the main menu and shows auto-complete option that allows the user to search for an individual real estate stored in the spatial database. If the user gets a valid record it will be displayed showing its graphical and alphanumerical information before it generates the e-report. Subsequently the e-report may or may not be requested by the user. Tool includes the option to generate all the e-reports as part of an expansion project and to select the project code and see the list of all the affected plots.

Fig. 5 shows the last confirmation before starting the process of generation of all e-reports in a project. Included is a checking box option to exclude all the real estates that are part of the project but will not be affected. The project shown in Fig. 5 is made up of 1101 realties with the possibility to exclude 858.

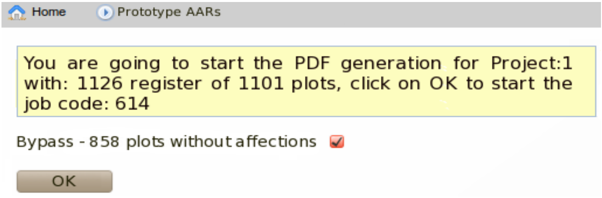


Enter a cadastral reference (RD)

Plot 00

000100800DS31B, 008 00DS31B
000200200DS30H, 002 00DS30H
000310100DS30G, 101 00DS30G
000400300DS30G, 003 00DS30G
000400800DS30H, 008 00DS30H
000401000DS30H, 010 00DS30H
000503500DS30G, 035 00DS30G
000505800DS30G, 058 00DS30G
000800100DS30G, 001 00DS30G
000800200DS30G, 002 00DS30G

Figure 4. Web user interface searching properties



You are going to start the PDF generation for Project:1 with: 1126 register of 1101 plots, click on OK to start the job code: 614

Bypass - 858 plots without affections ☑

Figure 5. Web user interface before e-report generation for a full project

In this example the generation of 255 e-reports has taken approximately 24 minutes and a total of 1478 images have been generated to produce 255 PDF files (a total size of 80 Mb); that means 1478 successful requests from the Web Reporting Application to the Map server using WMS and 255 requests from the Web Reporting Application to the Data Services to make the geoprocessing tasks to integrate alphanumerical and spatial information.

V. RESULTS

After many iterations of the prototype we have got relevant functionalities about e-reporting that allows us integrating in a single document all the necessary information for every vulnerable real estate to be affected during the electrical network expansion project.

During the first iterations of the project some inconveniences occurred related to the time needed to get the output as an e-report in PDF format. They were cleared up by incorporating a new manner of creating the output that allows handling the process on a low level and at the same time supporting characteristics to produce all the e-reports associated to a specific project on a large scale; the technologies taken on for that solution were JasperReports API, data services implemented with Apache Cayenne API, and an image generator using WMS requests to the Map Server (Geoserver).

The adoption of all these technologies results in a scalable architecture that can be strengthened with new elements whenever the prototype requires to incorporate any new functionality or be improved in order to go on to a new stage.

It should be taken into account as a relevant advantage that every one of these technologies can be manipulated as a plug-in which, depending on its performance, can be kept as a part of the proposed architecture or be replaced with a better element. We want highlight that all the technologies incorporated into the prototype stick to the philosophy of open source development and are under GNU general public licences.

Fig. 6 shows a comparison of the times needed for the manual and automated procedures for e-report generation; the automated method is approximately 175 times faster than the manual method. The manual method is able to generate 1000 e-reports in 14000 minutes (around 10 days) while automated method would need just 80 minutes.

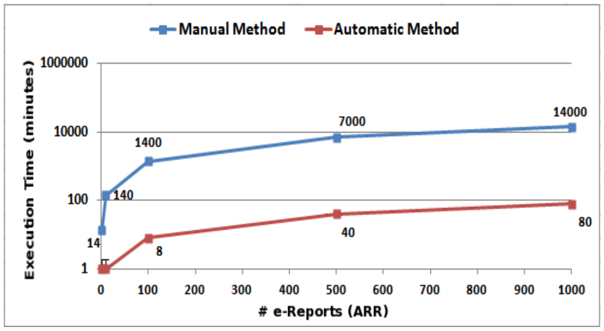


Figure 6. Times needed by traditional and automated methods

As said above, the average time consumed in the automated procedure is 5 seconds per e-report. An important reduction in the terms of execution of this task is achieved. The new procedure influences not only execution times but also has a bearing upon the homogeneity of the generated information and its reliability and accessibility.

VI. CONCLUSION AND FUTURE WORK

The results obtained reflect a favourable situation for all parts involved. On one hand REE incorporates new automated processes into the consolidated procedures providing greater and better information for interested users without warning about delays in the committed terms. On the other hand, the Government satisfies its need to ensure that the rights and obligations of a public utility of an installation are respected, and finally the owners of the affected assets will have a specific, detailed knowledge of the affections the new installations will generate on their real estates.

The new procedure put in place not only influences the execution times but also bears on the homogeneity and reliability of the generated information, and since we are dealing with a Web application, accessible from any point with Internet connection, the possibilities of documentation generation are multiplied and the software licence costs are reduced.

As future work we have planned the incorporation of a security mechanism similar to the one described in [17]. We have taken a step toward the implementation of an e-custom system with characteristics that respond to the new requirements of REE and users. We expect the model to be adopted as an official communication tool that should be recognized by government authorities and any additional social player participating in or being affected by the expansion project.

Along this line the first requests from the different regional government are being formalised to implement the tool during the process of notification to people affected by the network expansion project and by the necessary field work; thereby a second stage is anticipated for continuation of further development of the tool.

REFERENCES

[1] T.Ramachandra, V.George, K.Vamsee, and G.Purnima, Decision support system for regional electricity planning. Energy Education Science and Technology, 17(1): 2006 pp.17-25.

[2] A. Kumar , S.D. Bhatnagar, and P.K. Saxena Integrated multimedia based intelligent group decision support system for electrical power network. AJIS (Australasian Journal of Information Systems). 9(2). May. 2002.

[3] J. Shin, B. Yi, J. Song, J. Kang, J. Lee, and S. Cho, “A Development of the Mobile Computing System for Repair and Patrol of Electric Power Facilities,” in Fourth Annual ACIS International Conference on Computer and Information Science, 2005 pp. 622–627.

[4] S. Xu, P. Yan, and X. Wang, “Application research of land resources decision support system based on spatial information,” in 2010 18th International Conference on Geoinformatics, 2010 pp. 1-5.

[5] X. Du, L. Wang, and A. Sun, “The development and application of GIS on power marketing and expansion,” in Information Science and Engineering (ICISE), 2010 2nd International Conference on, pp. 4143 -4146, 2010.

[6] Open Geospatial Consortium Inc. (OGC), "OpenGIS Web Map Server Implementation Specification" OGC 06-042, Mar. 2006, pp. 1-85

[7] N.Voropai, “Main Principles of Electrical Network Expansion in a Market Environment of Russia,” in 2006 International Conference on Power System Technology, pp. 1-6, 2006.

[8] Wikipedia contributors, Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/ (accessed September 16, 2011).

[9] MOMENTUM. Project to monitor the existing and on coming eParticipation projects co-funded by European Commission, http://www.ep-momentum.eu/

[10] PEP-NET. European network of all stakeholders active in the field of eParticipation. http://pep-net.eu/

[11] eGovernment at W3C: Better Government Through Better Use of the Web. http://www.w3.org/egov

[12] Open Directory Project: Government Computerization Section.http://www.dmoz.org/Society/Issues/Science_and_Technology/Computers/Government_Computerization/

[13] E. Loukis, A. Xenakis, R. Peters, and Y. Charalabidis, “Using GIS tools to support e_participation - a systematic evaluation,” in 2nd IFIP WG 8.5 international conference on Electronic participation, 2010 pp. 197–210.

[14] R. J. Holden and B. Karsh, “The Technology Acceptance Model: Its past and its future in health care,” Journal of Biomedical Informatics, 43(1), Feb. 2010. pp. 159-172.

[15] K.Markellos, M. Katsis, and S. Sirmakessis, “Improving Web Content Delivery in eGoverment Applications,” in Tools and Applications with Artificial Intelligence, 2009, pp. 181-197.

[16] Eurostat: Detailed statistics on the EU and candidate countries. http://epp.eurostat.ec.europa.eu

[17] L. Razmerita and N. Bjorn-Andersen, “Towards Ubiquitous e-Custom Services,” in IEEE/WIC/ACM International Conference on Web Intelligence, pp. 833–837, 2007.

[18] International Organization for Standardization (ISO), "Geographic information - Geography Markup Language (GML)," ISO 19136:2007, Dec.2010.

ISBN: 978-1-61208-178-6

Level Sets and Voronoi based Feature Extraction from any Imagery

Ojaswa Sharma
Dept. of Computer Science
IIT Mumbay
Mumbai, India
Email: ojaswa@gmail.com

François Anton
Dept. of Informatics and Mathematical Modelling
Technical University of Denmark
Kongens Lyngby, Denmark
Email: fa@imm.dtu.dk

Darka Mioc
National Space Institute
Technical University of Denmark
Kongens Lyngby, Denmark
Email: mioc@space.dtu.dk

***Abstract*—*Polygon features are of interest in many GEOProcessing applications like shoreline mapping, boundary delineation, change detection, etc. This paper presents a unique new GPU-based methodology to automate feature extraction combining level sets, or mean shift based segmentation together with Voronoi skeletonization, that guarantees the extracted features to be topologically correct. The features thus extracted as object centerlines can be stored as vector maps in a Geographic Information System after labeling and editing. We show application examples on different sources: paper maps, digital satellite imagery, and 2D/3D acoustic images (from hydrographic surveys). The application involving satellite imagery shown in this paper is coastline detection, but the methodology can be easily applied to feature extraction on any king of imagery. A prototype application that is developed as part of this research work.*

***Keywords*-*Feature Extraction; Imagery; Segmentation; Level sets; Voronoi; Mean Shift.*

I. Introduction

Polygon features are of interest in applications like shoreline mapping, boundary delineation, change detection, etc. This paper presents a unique new GPU-based methodology to automate feature extraction on any imagery (not only raster images) by deformation of level sets or mean shift based segmentation / Voronoi skeletonization that guarantees the extracted features to be topologically correct. The features thus extracted as object centerlines can be stored as vector maps in a Geographic Information System after labeling and editing. We show application examples on different sources: scanned paper maps, digital satellite imagery, and 2D/3D acoustic images (from hydrographic surveys). This paper presents new results closely related to previous work of the authors on Voronoi [26] and level set evolution [25] based feature extraction.

The application involving satellite imagery shown in this paper is coastline detection, but the methodology can be easily applied to feature extraction on any king of imagery. For example, a coastline is defined as the boundary between land and water. Coastline mapping is important for coastal activity monitoring, resource mapping, navigation, etc. A lot of work on coastline extraction from SAR (Synthetic Aperture Radar) and multi-spectral imagery has been done. A technique for coastline extraction from remotely sensed images using texture analysis is described in [3]. The delineation of the complete coastline of Antarctica using SAR imagery is shown in [17]. A morphological segmentation based automated approach for coastline extraction has been suggested in [2]. Di et al. use the image segmentation algorithm by [8] to segment an image and detect the shoreline [9]. Our work in progress is an extension to the work by Gold and Snoeyink [12]. Our boundary (crust) extraction algorithm converts detected image features into connected sets of vectors that are topologically equivalent to the segmented objects. We claim topological equivalence of the extracted features since these are obtained as subsets of the Voronoi diagram or the Delaunay triangulation. This method can be applied on imageries, that have a high signal to noise ratio (scanned maps, aerial photographs, optical satellite imagery). For lower signal to noise imagery, we need level sets or continuous deformations.

The application involving acoustic images shown in this paper is 3D reconstruction of fishes from large acoustic datasets. Level set based methods have been shown to successfully restore noisy images [24]. Malladi and Sethian [18] have shown image smoothing and enhancement based on curvature flow interpretation of the geometric heat equation. In a more recent approach to use level set methods for acoustic image segmentation, Lianantonakis and Petillot [16] provide an acoustic image segmentation framework using the region based active contour model of Chan and Vese [4]. Here we focus on using the level set methods for simultaneous suppression of noise and 3D reconstruction of relevant features. We limit features of interest to fishes from acoustic images and provide a level set based framework for acoustic image segmentation. Image restoration techniques based on level set evolution are generally oriented to segment the image or to remove noise from it. Work by Lianantonakis and Petillot [16] is closest to our approach since they use active contours using Mumford-shah functional for seabed classification, but together with extraction of Haralick feature set for textural analysis.

Since acoustic data resulting from marine surveys can result in gigabytes of information, we employ GPU (Graphics Processing Unit) based computations for 3D reconstruction. The GPU is not very suitable for data intensive applications due to unavailability of large memory on commodity hardware. A number of publications suggest schemes to circumvent this situation by performing computations in a streaming manner [14], but most of the implementations process 2D sections to generate a 3D reconstruction. We present a new Level Set method implementation with computations performed entirely in 3D using the 3D textures (read only) available to the new CUDA (Compute Unified Device Architecture) 2.0 framework.

This paper is organized as follows. Section II presents the new methodology for image segmentation and feature extraction. Section III presents the results obtained with the new methodology presented in Section II. Finally, Section IV concludes this paper.

II. Image segmentation and feature extraction

Edge detection produces global edges in an image. This means that there is no object definition attached to the edges. Therefore, it is required to somehow define the objects first and then obtain edges from them. This can be achieved by using *image segmentation*. The main goal of image segmentation is to divide an image into parts that have a strong correlation with objects or areas of the real world depicted in the image [28, chap. 5]. Thus, image segmentation divides the whole image into homogeneous regions based on color information. The regions can be loosely defined as representatives of objects present in the image. While edge detection is very sensitive to noise, level sets based segmentation tolerates noise quite well.

A. Mean Shift Algorithm Segmentation

The segmentation method adopted here is the one provided by [7] which is based on feature space analysis. Feature space analysis is used extensively in image understanding tasks. [7] provide a comparatively new and efficient segmentation algorithm that is based on feature space analysis and relies on the *mean-shift algorithm* to robustly determine the cluster means. A *feature space* is a space of feature vectors. These features can be object descriptors or patterns in case of an image. As an example, if we consider a color image having three bands (red, green, and blue) then the image we see as intensity values plotted in Euclidean XY space is said to be in *image space*. Consider a three dimensional space with the axes being the three bands of the image. Each color vector corresponding to a pixel from the image can be represented as point in the feature space.

Given n data points x_i, $i = 1, \ldots, n$ in the d-dimensional space $\mathbb{R}^d$, a *flat kernel* of a location x that is the characteristic function of a λ-ball in $\mathbb{R}^d$ is defined as

$$K(x) = \begin{cases} 1 \text{ if } \|x\| \leq \lambda \\ 0 \text{ if } \|x\| > \lambda \end{cases} . \tag{1}$$

The *mean shift* vector at x is defined using the kernel of radius r as

$$M_\lambda(x) = \frac{\displaystyle\sum_{r \in \mathbb{R}^d} xK(r - x)}{\displaystyle\sum_{r \in \mathbb{R}^d} K(r - x)} - x \tag{2}$$

Cheng [6] shows that the mean shift vector, the vector of difference between the local mean and the center of the window $K(x)$, is proportional to the gradient of the probability density at x [6]. Thus mean shift is the steepest ascent with a varying step size that is the magnitude of the gradient. Further, Comaniciu and Meer use mean shift vector in seeking the mode of a density by shifting the kernel window by the magnitude of the mean shift vector repeatedly [8]. The authors also prove that the mean shift vector converges to zero and eventually reaches the basin of attraction of that mode.

In their research work, Comaniciu and Meer state a simple, adaptive steepest ascent mode seeking algorithm [7].

1) Choose the radius r of the search window (i.e, radius of the kernel).
2) Choose the initial location of the window.
3) Compute the mean shift vector and translate the search window by that amount.
4) Repeat until convergence.

The mean shift algorithm gives a general technique of clustering multi-dimensional data and is applied here in color image segmentation. The fundamental use of mean shift is in seeking the modes that give regions of high density in any data.

The method described in [7] provides an autonomous segmentation technique with only the type of segmentation to be specified by the user. This method emphasizes the importance of utilizing the image space along with the feature space to efficiently perform the task of segmentation. The segmentation has three characteristic input parameters:

- Radius of the search window, r,
- Smallest number of elements required for a significant color, N_{min}, and
- Smallest number of connected pixels necessary for a significant image region, N_{con}.

The size of the search window determines the resolution of the segmentation, smaller values corresponding to higher resolutions. The authors use square root of the trace of global covariance matrix of the image, σ, as a measure of the visual activity in the image. The radius r is taken proportional to σ. Later, Comaniciu and Meer provide an improvement [8] over this segmentation algorithm by merging the image domain and the feature (range) space into a joint spatial-range domain of dimension $d = p + 2$, where p is the dimension of the range domain. This gives an added advantage of considering both the spaces together and gives good results in cases where non-uniform illumination produces false contours when the previous segmentation algorithm is used. Therefore, the new algorithm is particularly useful to segment natural images with man-made objects. An added computational overhead to process higher dimensional space is inevitable here. The simple mean shift based segmentation algorithm provides satisfactory results in the case of scanned maps as shown in the results section.

Segmentation provides us with definite boundaries of objects that are used to extract sampling points around an object. This methodology does not allow one to guarantee that the topology of the segmented objects matches the topology of the imaged objects.

B. Feature extraction after mean shift algorithm

Anton et *al.* [1] suggest a new algorithm for skeleton extraction. This is based on the concept of *Gabriel Graph* [10]. A Gabriel graph G (highlighted in Figure 1) is a connected subset of the Delaunay graph $\mathscr{D}$ of points in set S such that two points p_i and p_j in S are connected by an edge of the Gabriel graph, if and only if, the circle with diameter p_ip_j does not contain any other point of S in its interior. In other words, the edges in G are those edges from $\mathscr{D}$ whose dual Voronoi edges intersect with them.

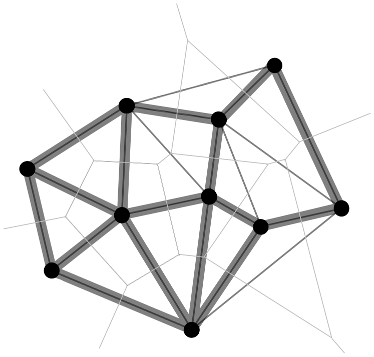


Fig. 1. Gabriel graph highlighted in a Delaunay triangulation.

Given the Delaunay triangulation $\mathscr{D}$ and the Voronoi diagram V of sample points S from the boundary of an object, the algorithm for centerline extraction in [1] proceeds by selecting all the Gabriel edges in graph G. Each dual Voronoi edge v of the Gabriel edge g from G is inserted in the skeleton K if the following condition is met:

$$\begin{gathered} g.Origin.Colour \neq g.Destination.Colour \\ \text{Or} \\ g.Origin.Colour \neq v.Origin.Colour \\ \text{Or} \\ g.Origin.Colour \neq v.Destination.Colour \\ \text{And} \\ \|g.Origin.Colour - g.Dest.Colour\| \geq \\ \|v.Origin.Colour - v.Dest.Colour\| \end{gathered} \tag{3}$$

Here, $Origin.Colour$ and $Destination.Colour$ are color values from the gray scale image corresponding to the location of the origin and the destination of an edge respectively.

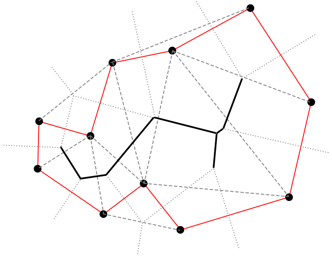


Fig. 2. Anti-crust from the crust.

1) Obtaining Anti-crust from the Voronoi diagram: The anti-crust of an object, as described above, forms a tree like structure that contains the skeleton. Once all the Delaunay edges belonging to the border set or the crust are identified using the condition given by Gold [11], it is easy to identify the Voronoi edges belonging to the anti-crust. In Figure 2, consider the Delaunay triangulation (dashed edges), the corresponding Voronoi diagram (dotted edges) and the crust edges (solid red edges).

Navigation from a Delaunay edge to its dual Voronoi edge can be achieved by using the $Rot()$ operator in the quad-edge data structure [15]. A Voronoi edge $e.Rot()$ of the dual Delaunay edge e is marked as an edge belonging to the anti-crust if the following conditions are satisfied:

1) $e \notin Crust$
2) $e.Rot().Origin \in I$
3) $e.Rot().Destination \in I$,

where $e.Rot().Origin$ is the origin of the edge $e.Rot()$, $e.Rot().Destination$ is the destination of the edge $e.Rot()$ and I is the selected object. This marks all the Voronoi edges belonging to the anti-crust that fall inside the selected object. Negating conditions (2) and (3) so that the points do not fall inside the object will give us the exterior skeleton or the *exoskeleton*. Once the anti-crust is identified, an appropriate pruning method can be applied to get rid of the unwanted edges.

2) Pruning: Gold [11] also discusses the “hairs” around the skeleton that result due to the presence of three adjacent sample points whose circumcircle does not contain any other sample point - either near the end of a main skeleton branch or at locations on the boundary where there is minor perturbation because of raster sampling. Gold and Thibault [13] suggest a skeleton retraction scheme in order to remove the hairs that also results in smoothing of the boundary of the object. Ogniewicz [20] presents an elaborate skeleton pruning scheme based on various residual functions. Thus, a hierarchic skeleton is created which is good for multi-scale representation. Sharma et *al.* [27] suggest the use of ratio based pruning in order to simplify a network of skeletons for extracting linear features from satellite imagery.

The problem of identifying skeleton edges now reduces to reasonably prune the anti-crust. We next present an optimal criterion for pruning by successively removing leaf edges from the anti-crust.

3) Pruning by Removing Leaf Edges: Gold and Thibault [13] present a retraction scheme for the leaf nodes in the anti-crust. The skeleton is simplified by retracting the leaf nodes of the skeleton to their parent nodes. Gold and Thibault [13] recommend performing the retraction operation repeatedly until no further changes take place. An observation reveals that an unwanted branch in a skeleton may be composed of more than one edge (see Figure 3). Therefore, single retraction may not be sufficient to provide an acceptable skeleton.

A similar simplification can be achieved by pruning the leaf edges instead of retracting the leaf nodes. Leaf edge pruning produces satisfactory results and requires only two or three

ISBN: 978-1-61208-178-6

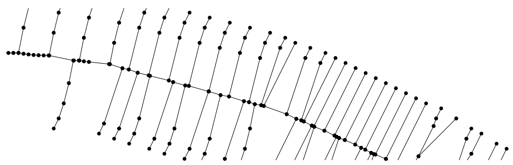


Fig. 3. Hair around the skeleton composed of multiple edges.

levels of pruning. Before pruning the leaf edges, they must be identified in the anti-crust. An edge e from a tree of edges $T \in V$, where V is the Voronoi diagram, is marked as a leaf edge if the following condition is satisfied:

$$\begin{array}{c} e.Oprev() \notin T \text{ And } e.Onext() \notin T \\ \text{Or} \\ e.Sym().Oprev() \notin T \text{ And } e.Sym().Onext() \notin T \end{array} . \quad (4)$$

This condition essentially selects all the Voronoi edges belonging to the anti-crust that have at least one end point free (i.e., connected to an edge not belonging to the anti-crust). This condition is used to locate leaf edges followed by their removal from the skeleton. Experiments show that removing leaf edges two to three times simplifies the skeleton to a major extent for linear features. We find an optimal criterion for removal of extraneous hair from the skeleton by pruning the leaf edges (see results in next section).

C. Level sets based segmentation

Let an image $I(x, y)$ be defined on a bounded open subset $\Omega : \{(x, y)|0 \leq x, y \leq 1\}$ of $\mathbb{R}^2$, with $\partial\Omega$ as its boundary. I takes discrete values between 0 and $(2^n - 1)$ where n is the number of bits used to store intensity. The basic idea in active contour model is to evolve a curve $C(s) : [0, 1] \rightarrow \mathbb{R}^2$ by minimizing the following energy functional [21]:

$$E(C) = \alpha \int_0^1 |C'|^2 \, ds + \beta \int_0 |C''| \, ds - \lambda \int_0^1 |\nabla I(C)|^2 \, ds,$$

where α, β, and λ are positive parameters, and ∇I denotes the gradient of I. In the above energy functional, the evolution of curve C is controlled by the internal energy (first two terms that define the smoothness of the curve) and the external energy (the last term that depends on the edges present in the image). More intuitively, the curve evolves to minimize the differences of intensity between the points in the interior of the curve, and at the same time maximizes the differences between the points in the interior and the points in the exterior of the curve, thus, clustering the image. The curve C can be represented by an implicit function ϕ, $C = \{(x, y)|\phi(x, y) = 0\}$, where the evolution of C is given by the zero level curve at any time t of the function $\phi(x, y, t)$.

With this formulation, an edge detector is defined as a positive decreasing function $g(\nabla I)$ based on the gradient of image [23] such that

$$\lim_{|\nabla I| \to \infty} g(\nabla I) = 0$$

Therefore, the zero level curve evolves in the normal direction and stops at the desired boundary where g vanishes.

Evolving the curve C in normal direction amounts to solving the partial differential equation (PDE) [22]

$$\frac{\partial \phi}{\partial t} = |\nabla \phi| F \quad (5)$$

with the initial condition $\phi(x, y, 0) = \phi_0(x, y)$, where $\phi_0(x, y)$ is the initial contour. Motion by mean curvature allows for cusps, curvature and automatic topological changes [22], [5]. This results in the speed function $F = div\left(\frac{\nabla\phi}{\|\nabla\phi\|}\right)$ in terms of the curvature of ϕ

$$\frac{\partial \phi}{\partial t} = |\nabla \phi| div\left(\frac{\nabla \phi}{|\nabla \phi|}\right), \phi(x, y, 0) = \phi_0(x, y)$$

where $div(\cdot)$ is the divergence operator.

1) Minimising the Mumford-Shah Functional in Image: Chan and Vese [5] provide an alternative approach to the edge based stopping criterion. The authors suggest the stopping term based on Mumford-Shah segmentation techniques [19]. The motivation behind using this alternative stopping term is that in many cases, the edges in an image are not very well defined. Either it is ambiguous to position the edges across the gradient due to smoothly varying intensities [5] or it is difficult to select prominent edges due to presence of noise (as in the case of acoustic images). The method of Chan and Vese [5] is minimization of an energy based segmentation. Assuming that the image I is composed of two regions of piecewise constant intensities of distinct values I^i and I^o, and that the object of interest is represented by I^i, we define the curve C to be its boundary. Using the Heaviside function H, and the Dirac-Delta function δ_0,

$$H(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0 \end{cases}, \ \delta_0(z) = \frac{d}{dz} H(z)$$

the energy functional is formulated as

$$\begin{aligned} E(c_1, c_2, C, t) =& \mu \int_\Omega \delta_0(\phi(x, y, t)) |\nabla \phi(x, y, t)| \, dx \, dy \\ &+ \nu \int_\Omega H(\phi(x, y, t)) \, dx \, dy \\ &+ \lambda_1 \int_\Omega |I(x, y) - c_1|^2 \, dx \, dy \\ &+ \lambda_2 \int_\Omega |I(x, y) - c_2|^2 \, dx \, dy \end{aligned} \quad (6)$$

where, $\mu \geq 0$, $\nu \geq 0$, $\lambda_1, \lambda_2 > 0$ are fixed parameters. c_1 and c_2 are average intensity values inside and outside C. The constants c_1 and c_2 can also be written in terms of I and ϕ

$$c_1 = \frac{\int_\Omega I(x, y) H(\phi(x, y, t)) \, dx \, dy}{\int_\Omega H(\phi(x, y, t)) \, dx \, dy}, \quad (7)$$

$$c_2 = \frac{\int_\Omega I(x, y)(1 - H(\phi(x, y, t))) \, dx \, dy}{\int_\Omega (1 - H(\phi(x, y, t))) \, dx \, dy} \quad (8)$$

The variational level set approach gives the following Euler-Lagrange equation [5]

$$\frac{\partial\phi}{\partial t} = \delta_\epsilon(\phi)\left[\mu\nabla\cdot\frac{\nabla\phi}{|\nabla\phi|} - \nu - \lambda_1(I-c_1)^2 + \lambda_2(I-c_2)^2\right] \tag{9}$$

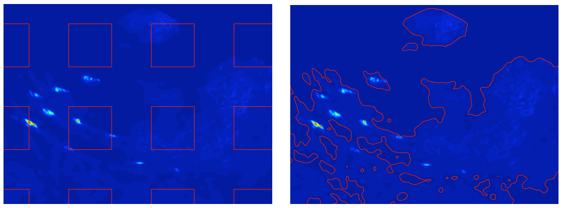
with the initial condition, $\phi(x,y,0) = \phi_0(x,y)$ and

$$\delta_\epsilon(z) = \frac{\partial}{\partial z}H_\epsilon(z) = \pi^{-1}\epsilon^{-1}\left(1+\frac{z^2}{\epsilon^2}\right)^{-1} \tag{10}$$

where, the regularised one-dimensional Heaviside function is given by:

$$H_\epsilon(z) = \frac{1}{2}\left(1+\frac{2}{\pi}\tan^{-1}\left(\frac{z}{\epsilon}\right)\right).$$

The acoustic images considered by Lianantonakis and Petillot [16] are of the seabed. Such images show strong textural variations of the bottom surface of the sea. In this paper, we restrict ourselves to acoustic images of freely swimming fishes. While such images are also corrupted by speckle noise, they do not show specific textural patterns. Figure 4(a) shows part of such an image where the fish cross sections are discriminated by very high intensities compared to the background. The presence of reflectance from air bubbles mixing into water, also contribute to the noise. While working with level sets, a standard procedure is to keep ϕ to a signed distance function [21]. A direct application of the level set equation given by equation (9), with $\phi(x,y,0) = 0$ initialized to set of squares regularly distributed over the image, shows that the evolution of the level set eventually stops at the wrong place (see Figure 4(b)). We used specific multi-beam acoustic noise removal techniques.



(a) Initialization contour. (b) Result at convergence.

Fig. 4. Application of the level set equation (9).

2) Noise suppression model: Considering the image I to be time varying, the basic idea behind noise suppression is to solve the following equation as an update step to the level set equation resolution in a single pass:

$$\frac{\partial I(x,y,t)}{\partial t} = k\cdot\max(0,\hat{c}-I(x,y,t)) \tag{11}$$

where k is a constant and $\hat{c}$ is a scalar parameter that is computed as an optimal threshold at any time step t based on $\phi(x,y,t)$.

The computation of $\hat{c}$ is based on the bounded subset I^i given by

$$I^i(x,y,t) = I(x,y,t)\cdot H_\epsilon(\phi(x,y,t)).$$

The values given by the set I^i are used to compute the weighted median [29] as shown in algorithm 1 which is used as $\hat{c}$ at that particular time step t.

```
**Input**: $I(x,y,t)$, $H_\epsilon(x,y,t)$
**Output**: $\hat{c}$
$V = \{v_i : v_i = I(x,y,t),\ x\in[1,l],\ y\in[1,m],$
        $i\in[1,n],\ n = l\cdot m\}$
$W = \{w_i : w_i = H_\epsilon(x,y,t),\ x\in[1,l],\ y\in[1,m],$
        $i\in[1,n],\ n = l\cdot m\}$
Sort $V$ in ascending order
$W \leftarrow W\backslash\{w_z\}\ \forall w_z = 0$
$V \leftarrow V\backslash\{v_z\}, \{v_z : v_z\in V,\ \forall z$ where $w_z = 0\}$
$S \leftarrow \sum_{k=1}^{n} w_k,\ w_k\in W$
Find index $i$ such that $\sum_{k=1}^{i} w_k \le \frac{S}{2}, w_k\in W$
Find index $j$ such that $\sum_{k=j}^{n} w_k \le \frac{S}{2}, w_k\in W$
Median $M = \{v_i, v_j\}$
$\hat{c} \leftarrow \min(v_i, v_j)$
```

Algorithm 1: Computation of weighted median

We now show that the estimate of $\hat{c}$ based on the weighted median is a good approximation for the grey-level threshold that separates the noise from the signal, and is robust in a way that the evolution of the level set converges with increasing t.

$H_\epsilon(z)$ attains values close to zero for regions outside C and values close to one inside C. In fact, $\lim_{z\to\infty} H_\epsilon(z) = 1.0$ and $\lim_{z\to-\infty} H_\epsilon(z) = 0.0$. At the start of level set evolution, I^i covers most of Ω and therefore, $H_\epsilon(z)$ attains values close to one for most of the intensity values. This results in computation of $\hat{c}$ which is equivalent to an unweighted median for values in I^i. A median is the central point which minimizes the average of absolute deviations. Therefore, a median better represents the noise level when the data contains high intensity values that are fewer in number, and a majority of intensity values that correspond to the noise. As a result, the initial iterations of the solution suppress the intensity values that are less than the median to a constant level (the median itself). One should expect the median value to increase as the level set contracts, but since we use a regularized Heaviside function as weight for the intensity values, the weighted median converges to zero since most of I contains intensity values of zero with near-zero weight.

Other variations of estimation of $\hat{c}$ are certainly possible, but we find that a weighted median based approach results in effective noise removal with very small information loss. For instance, a value of $\hat{c}$ taken to be c_1, the mean intensity inside C, does a similar suppression but with a high signal loss compared to the former.

3) CUDA Implementation for 3D Reconstruction: Equation (9) can be solved by discretization and linearization in

ϕ [5]. Discretization of equation (11) in I gives

$$\begin{aligned}\frac{I_{n+1}(x,y) - I_n(x,y)}{\Delta t} &= k \cdot \max\left(0, \hat{c} - I_n(x,y)\right) \\ &= \begin{cases} 0, \text{if } I_n(x,y) \geq \hat{c} \\ k \cdot (\hat{c} - I_n(x,y)), \text{otherwise} \end{cases}\end{aligned} \tag{12}$$

with $k = \frac{1}{\Delta t}$, and $t_{n+1} = t_n + \Delta t$. The above time discretization yields the following

$$I_{n+1}(x,y) = \begin{cases} 0, & \text{if } I_n(x,y) \geq \hat{c} \\ \hat{c} - I_n(x,y), & \text{if } I_n(x,y) < \hat{c} \end{cases} \tag{13}$$

Acoustic images captured by echo-sounders are generally taken as planar image scans by moving the echo-sounder in one direction, thereby sweeping a volume. Let us denote individual images as $I(x, y, \tau)$ for images taken after every $\delta\tau$ time interval. A volume is constructed by stacking these individual images in sequence and applying geometric correction for distance $\delta\tau(v)$ between individual slices, where v is the instantaneous speed of the instrument (the current data was captured with constant unidirectional instrument velocity). It must also be noted that the individual acoustic images are obtained from a set of acoustic intensity signals along beams by a polar transformation. The level set equations for curve evolution in $\mathbb{R}^2$ extend uniformly to surface evolution in $\mathbb{R}^3$. The second differential equation also holds true for noise suppression in a volume. Therefore, it is possible to reconstruct 3D moving fishes with the level set evolution of these equations combined.

Processing a huge dataset demands that a minimum of memory is consumed. We propose to keep two volumes in the host memory, one for the intensity values (I) and the other for the signed distance function (the implicit function, ϕ). The CPU manages the memory scheduling by dividing the volumes into small subvolumes that can be processed on the GPU. We keep two small 3D textures of size $128 \times 128 \times 128$, I_{GPU} and ϕ_{GPU}. A complete level set update is divided into a set of subvolume updates. Each subvolume in the two volumes is fetched to the GPU via 3D textures (read only, but with good cache coherence). Results of computations are written to CUDA memory and then transferred back to the CPU volumes. A simplified diagram of this is shown in Figure 5.

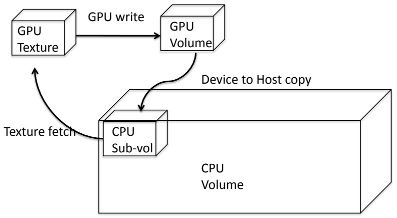


Fig. 5. Parallelization using the GPU.

CUDA exposes a set of very fast 16KB shared memory available to every multi-processor in a GPU. However, a 16 KB memory chunk is shared only between a thread block, and thus to make use of it the application must load different data for different blocks. Furthermore, the 16 KB limit poses a restriction on the amount of data that can be loaded at any point of time. Here, we use 3D textures for reading the data. Since we do not want to write back to the same texture (before a single step of filtering is complete), using the read-only 3D textures available to CUDA is a natural choice. 3D texturing has hardware support for 3D cache which accelerates any texture reads in succession. To load a 3D data (a small subset of the volume) from the global memory into the shared memory could be a little tricky and might not result in the same performance as provided by the specialized hardware for 3D texture cache. In our application, data writes are made to the global memory.

Signed distance transform is a global operation and cannot be implemented in a straightforward manner. We compute a local approximation of the Euclidean distance transform using the Chamfer distance. A narrow band distance transform is computed layer by layer using, what we call a d-pass algorithm. Every pass of the method adds a layer of distance values on the existing distance transform. The distance values are local distance increments computed in a $3 \times 3 \times 3$ neighborhood. Therefore, every single pass needs only local information to compute the distance values except at the border of the sub-volume. We therefore support every sub-volume with a one voxel cover from other adjoining sub-volumes, thereby reducing the computational domain to a volume of size $126 \times 126 \times 126$. The CPU scheduler takes care of the voxel cover. At the beginning, the interface (zero level) is initialized to a used specified bounding cuboid or a super-ellipsoid.

Computing average intensities (c_1 and c_2) is an operation that cannot be easily computed in a parallel fashion, and a reduction like method is required for the same. We employ a slightly different scheme to compute averages by using three accumulator sub-volumes on the GPU. These accumulators are essentially 3D sub-volumes of the same dimensions as of the textures. Every voxel in the accumulators accumulates (adds up), the values for H, $I \cdot H$, and $I \cdot (1 - H)$ for all the sub-volumes in the CPU volume(s). We then sum up the small sub-volume on the CPU to get the final sum and compute c_1 and c_2 values from it. Using a mixed mode CPU-GPU computation not only reduces the complexity of an inherently non-parallel operation, but also performs better by moving less expansive parts of the computation to the CPU.

Computing median on the GPU is not very straightforward since it is an order statistic and requires that the data be sorted. Therefore the computation of weighted median is very different than the one for average intensity value. Since sorting values of order of millions in every iteration of the solver is not a computationally good solution, we resort to the alternative definition of the median. A median is a value that divides the data-set into two sets of equal cardinalities. This definition is generalized for a weighted median. Therefore, for a data-set V with weights W associated with each value in the set, the median value V_k is the value for which the following holds:

$$\sum_{i=0}^{k} W_i = \sum_{i=k+1}^{n} W_i$$

This equation can only be solved iteratively, starting with a guess index value k_0. In our CUDA implementation, we start with V_{k_0} to be the mean value c_1 and iteratively reach the weighted median. In every iteration, the increment $\triangle i$ for the index k_0 is computed as:

$$\triangle i = \begin{cases} \dfrac{\sum\limits_{i=0}^{k} W_i - \sum\limits_{i=k+1}^{n} W_i}{\sum\limits_{i=0}^{k} W_i}, & \text{if } \sum_{i=0}^{k} W_i > \sum_{i=k+1}^{n} W_i \\ \dfrac{\sum\limits_{i=k+1}^{n} W_i - \sum\limits_{i=0}^{k} W_i}{\sum\limits_{i=k+1}^{n} W_i}, & \text{if } \sum_{i=k+1}^{n} W_i > \sum_{i=0}^{k} W_i \end{cases}$$

The increment $\triangle i$ can be adaptively controlled to give results as precise as desired.

4) Solver update: A PDE update in the level set method comprises of computing the curvature energy and the external energy. In order to compute the curvature term (involving double derivatives) for a voxel in a sub-volume by centered differencing, we need information from a $5 \times 5 \times 5$ neighborhood with the current voxel at its center. Therefore, the sub-volume size needs a cover of two voxels on all sides, thus reducing the computational domain further down to $124 \times 124 \times 124$. The memory schedular performs additional computations to effectively cover the whole volume with the new setup. Once the energy terms are computed, the PDE solver kernel updates ϕ_{GPU} and uses c_1 to update I_{GPU}. These sub-volumes are then updated to the CPU main volume. It is often convenient to perform anisotropic diffusion on the input image so that the evolution of the level curve is smooth and ϕ is well behaved. Finally, the zero level surface is extracted from the evolved ϕ using the Marching Cubes method.

III. Results

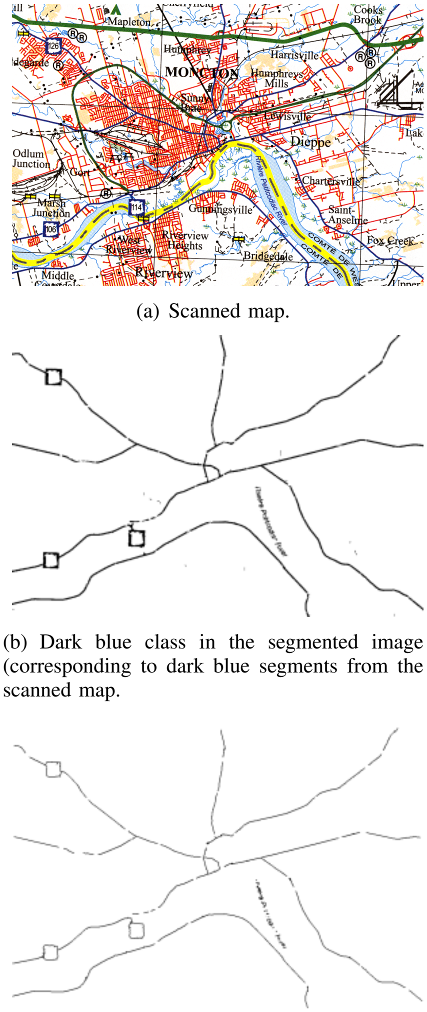
We observe that using the GPU for the parallelization of the processing of the imagery induces an overall speed-up that ranges between 10 and 20 times over CPU algorithms.

A. High signal to noise ratio imagery

In this section, we show results of our new combined GPU-based methodology on high signal to noise ratio imagery (scanned maps and satellite imagery).

The extraction of a road network from a scanned map is shown in Figure 6.

In the following example, the coastline is extracted as the boundary of the selected object. The accuracy of the coastline rendition depends on the spatial resolution of the imagery. The beach of Seychelles shown in Figure 7(a) is mainly sandy and shows a wide variation in the ocean color. The color variation is primarily due to the depth of water. Figure 7(b) shows the result of the segmentation using mean shift and Voronoi skeletonization on Figure 7(a). Regions in Figure 7(b) that form the ocean are combined to extract the ocean boundary (see Figure 7(c)). The extended coastline representing the boundary of the sand beach is shown in Figure 7(d). The extended coastline shows the presence of a number of small polygons, since the roads connecting the beach have the same color value in the image, and are included in the selection.



(a) Scanned map.

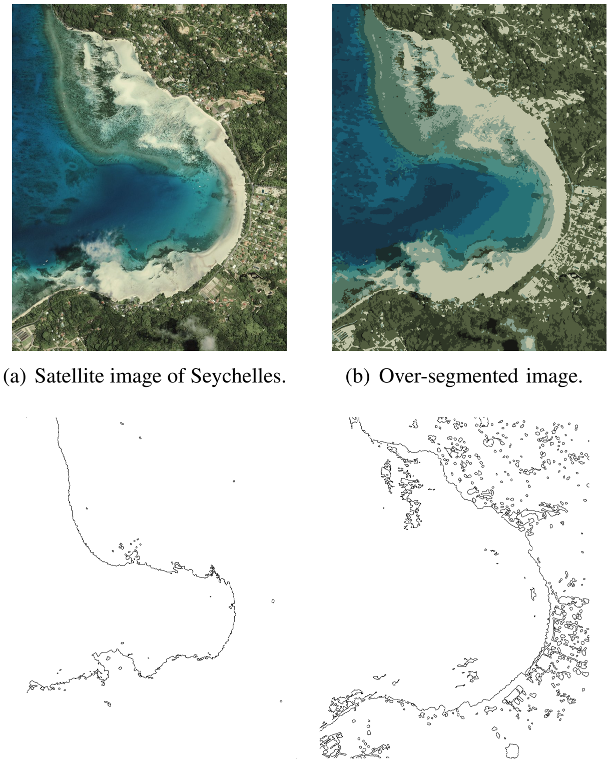
(b) Dark blue class in the segmented image (corresponding to dark blue segments from the scanned map.

(c) Medial axis of the polygons represented in (c).

Fig. 6. City Map of Moncton, New Brunswick, Canada.

B. Low signal to noise ratio imagery

We present experimental results on 2D multi-beam echo sounder (acoustic) images to show that the suppression scheme works well on such images. Figure 8 shows evolution of the level set. The parameters for this evolution were chosen to be: $\mu = 0.0005$, $\nu = 0$, $\lambda_1 = \lambda_2 = 1$, and $\epsilon = 2.5$. It can be seen that the original image suffers from speckle noise and that the final zero level contour approximates the fish boundaries very

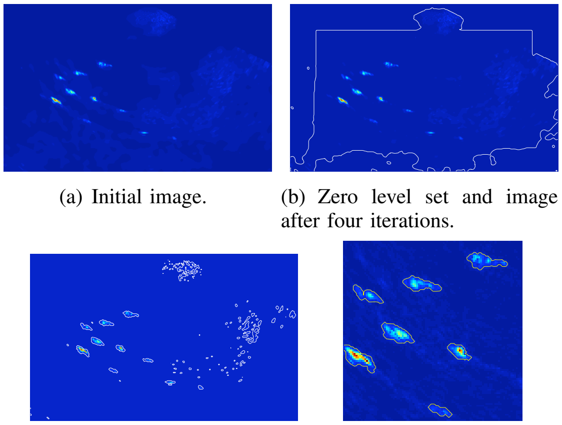


(a) Satellite image of Seychelles. (b) Over-segmented image.

(c) Extracted ocean boundary. (d) Extracted beach boundary.

Fig. 7. Feature polygon extraction from the satellite image of Seychelles.

well.

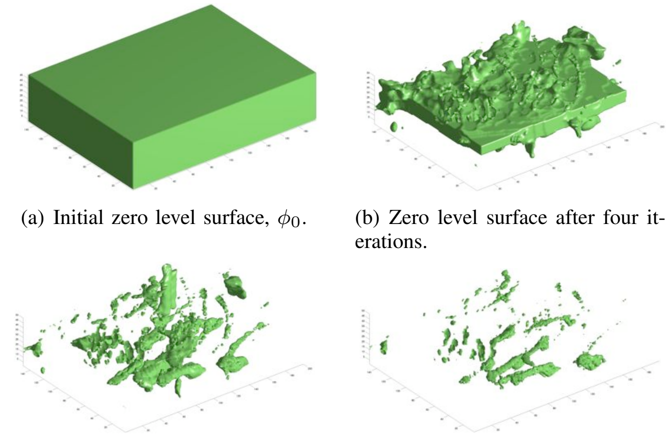


(a) Initial image. (b) Zero level set and image after four iterations.

(c) End of evolution after 16 iterations. (d) Final contour shown on part of the original image.

Fig. 8. Level set evolution on sample image, $\epsilon = 2.5$.

We next show results of application of the level set equation and the noise suppression scheme on a small 3D multi-beam echo sounder (acoustic) volume of size $150 \times 100 \times 50$. Fish intensities can be identified in dark green against a noisy background. The level set equation was initialized with the zero level set of ϕ_0 as the bounding box of the volume. The level set is then allowed to evolve with parameters, $\mu = 0.0005$, $\nu = 0$, $\lambda_1 = \lambda_2 = 1$, and $\epsilon = 1.0$. Figure 9 shows the evolution at different time steps and the final level surface.



(a) Initial zero level surface, ϕ_0. (b) Zero level surface after four iterations.

(c) Zero level surface after six iterations. (d) End of evolution after nine iterations.

Fig. 9. Level set evolution on sample volume, $\epsilon = 1.0$.

We test the CUDA solver on a larger volume of size $686 \times 1234 \times 100$. This volume uses about 470 MB of CPU memory along with the same amount of memory consumed by the signed distance field. We test our implementation with the mobile GPU, GeForce 8600M GT (NVDIA CUDA compute capability of 1.1) with 256 MB of memory on a Mac OS X notebook with 2 GB of host memory. The total number of iterations required until convergence were 29, with a compute time of about 52 seconds per iteration (32 seconds without median computation). The signed distance field was reconstructed in a narrow band of width 20 voxels in every iteration.

In order to compare the 2D and 3D reconstructions, we show an overlay of 2D curves over the extracted 3D surface. This is shown in Figure 10. The results agree very well when the 2D image contains high intensity objects. The acoustic images were taken by scanning fishes in an aquarium and the images corresponding to the bottom of the aquarium (time slices with higher depth, 30 to 50 in Figure 10) contain almost no fishes. Therefore, these images contain very little useful information. The 2D level set evolution fails to detect fishes in these images. Therefore, the 3D results should be trusted since the 2D reconstruction does not consider information present in other image planes. We would like to comment that a ground truth segmentation is not practically possible for open sea. Evaluation of the extracted fish trails/schools by domain experts is under process because of marine surveys.

IV. CONCLUSIONS

This research work succeeds in achieving its primary goals by designing an effective GPU-based highly parallel methodology for automated vectorization of imagery. Based on the methodology, an interactive software application has been developed. It incorporates object extraction from input images

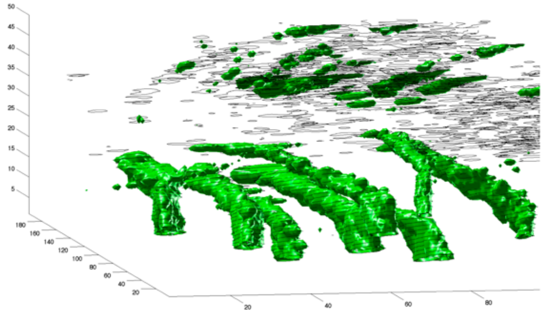


Fig. 10. Comparison of zero level 2D curves with the zero level 3D surface.

using color image segmentation or level sets evolution. The application allows either automated extraction of all features or manual selection of multiple objects for semi-automated extraction of targeted feature classes. It is worth noticing here that the extracted feature might not be a complete map object in the scanned image. This is due to labels and other features drawn over the feature of interest on a paper map.

Applicability of the Voronoi-based methodology to satellite imagery has been shown by extracting natural features like coastlines. Experiments done with satellite imagery show acceptable results too. Coastline delineation, snow cover mapping, cloud detection, and dense forest mapping are a few areas where satisfactory results can be obtained. Applicability of the level set evolution methodology has been shown on underwater acoustic images. Current work focuses on level set evolution based Digital Terrain Model (DTM) generation from Light Detection And Ranging (LIDAR) data sets. Any imagery can be processed using the above methodology: not only raster images, but also 2D and 3B beam-formed datasets like Computer Tomography or multi-beam echo sounder data.

REFERENCES

[1] F. Anton, D. Mioc, and A. Fournier, “2D Image Reconstruction using Natural Neighbour Interpolation,” *The Visual Computer*, vol. 17, no. 3, pp. 134–146, 2001.

[2] S. Bagli and P. Soille, “Morphological automatic extraction of coastline from pan-European Landsat TM images,” in *Proceedings of the Fifth International Symposium on GIS and Computer Cartography for Coastal Zone Management*, vol. 3, Genova, 2003, pp. 58–59.

[3] G. Bo, S. Delleplane, and R. D. Laurentiis, “Coastline extraction in remotely sensed images by means of texture features analysis,” in *Geoscience and Remote Sensing Symposium, IGARSS ’01*, vol. 3, Sydney, NSW, Australia, 2001, pp. 1493–1495.

[4] T. Chan and L. Vese, “A level set algorithm for minimizing the Mumford-Shah functional in image processing,” *IEEE/Computer Society Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision*, pp. 161–168, 2001.

[5] ——, “Active Contours Without Edges,” *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 10, no. 2, 2001.

[6] Y. Cheng, “Mean shift, mode seeking, and clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[7] D. Comaniciu and P. Meer, “Robust analysis of feature spaces: color image segmentation,” in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR ’97)*. Washington, DC, USA: IEEE Computer Society, 1997, pp. 750–755.

[8] ——, “Mean Shift: A Robust Approach Toward Feature Space Analysis,” *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[9] K. Di, J. Wang, R. Ma, and R. Li, “Automatic Shoreline Extraction from High-Resolution Ikonos Satellite Imagery,” in *Proceeding of ASPRS 2003 Annual Conference*, vol. 3, Anchorage, Alaska, 2003.

[10] K. R. Gabriel and R. R. Sokal, “A new statistical approach to geographic variation analysis,” *Systematic Zoology*, vol. 18, no. 3, pp. 259–278, 1969.

[11] C. M. Gold, “Crust and anti-crust: A one-step boundary and skeleton extraction algorithm,” in *Symposium on Computational Geometry*. New York, NY, USA: ACM Press, 1999, pp. 189–196.

[12] C. M. Gold and J. Snoeyink, “A one-step crust and skeleton extraction algorithm,” *Algorithmica*, vol. 30, no. 2, pp. 144–163, Jun 2001.

[13] C. M. Gold and D. Thibault, “Map generalization by skeleton retraction,” in *Proceedings of the 20th International Cartographic Conference (ICC)*, Beijing, China, August 2001, pp. 2072–2081.

[14] N. K. Govindaraju, B. Lloyd, W. Wang, M. Lin, and D. Manocha, “Fast computation of database operations using graphics processors,” in *SIGGRAPH ’05: ACM SIGGRAPH 2005 Courses*. New York, NY, USA: ACM, 2005, p. 206.

[15] L. Guibas and J. Stolfi, “Primitives for the manipulation of general subdivisions and the computation of Voronoi Diagrams,” *ACM Transactions on Graphics*, vol. 4, no. 2, pp. 74–123, 1985.

[16] M. Lianantonakis and Y. Petillot, “Sidescan sonar segmentation using active contours and level set methods,” *Oceans 2005-Europe*, vol. 1, 2005.

[17] H. Liu and K. C. Jezek, “A Complete High-Resolution Coastline of Antarctica Extracted from Orthorectified Radarsat SAR Imagery,” *Photogrammetric Engineering and Remote Sensing*, vol. 70, no. 5, pp. 605–616, 2004.

[18] R. Malladi and J. Sethian, “Image Processing Via Level Set Curvature Flow,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 15, pp. 7046–7050, 1995.

[19] D. Mumford and J. Shah, “Optimal approximations by piecewise smooth functions and associated variational problems,” *Commun. Pure Appl. Math.*, vol. 42, no. 5, pp. 577–685, 1989.

[20] R. L. Ogniewicz, “Skeleton-space: A multiscale shape description combining region and boundary information,” in *Proceedings of Computer Vision and Pattern Recognition, 1994*, 1994, pp. 746–751.

[21] S. Osher and R. Fedkiw, *Level sets and dynamic implicit surfaces*. Springer New York, 2003.

[22] S. Osher and J. Sethian, “Fronts propagation with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations,” *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.

[23] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[24] J. Sethian, “Theory, algorithms, and applications of level set methods for propagating interfaces,” *Acta Numerica 1996*, pp. 309–395, 1996.

[25] O. Sharma and F. Anton, “CUDA based Level Set Method for 3D Reconstruction of Fishes from Large Acoustic Data,” in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision ; 17*, 2009.

[26] O. Sharma, D. Mioc, and F. Anton, “Feature extraction and simplification from colour images based on colour image segmentation and skeletonization using the quad-edge data structure,” in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision WSCG 2007*. University of West Bohemia, Plzen, Czech Republic, 2007, pp. 225–232.

[27] O. Sharma, D. Mioc, and A. Habib, “Road extraction from satellite imagery using fractals and morphological image processing,” in *Proceedings of the 13th International Conference on Geoinformatics*, Toronto, Canada, 2005.

[28] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. PWS publishing, 1999.

[29] L. Yin, R. Yang, M. Gabbouj, and Y. Neuvo, “Weighted median filters: a tutorial,” *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on [see also Circuits and Systems II: Express Briefs, IEEE Transactions on]*, vol. 43, no. 3, pp. 157–192, 1996.

Spatial Characterization of Extreme Precipitation in Madeira Island Using Geostatistical Procedures and a 3D SOM

Jorge Gorricha, Victor Lobo
CINAV-Escola Naval
Portuguese Naval Academy
Almada, Portugal
lourenco.gorricha@marinha.pt, vlobo@isegi.unl.pt

Ana Cristina Costa
ISEGI-UNL
Universidade Nova de Lisboa
Lisbon, Portugal
ccosta@isegi.unl.pt

Abstract— **Extreme precipitation events can be analyzed from multiple perspectives. Precipitation indices, estimated from the empirical distribution of the daily observations, are increasingly being used not only to investigate trends in observed precipitation records, but also to examine scenarios of future climate change. In this study, we propose a methodology for characterizing the spatial patterns of extreme precipitation in Madeira Island that is based on two types of approaches. The first one uses linear models, such as Ordinary Kriging and Ordinary Cokriging, to produce continuous surfaces of five extreme precipitation indices. The second one uses a 3D Self-Organizing Map (SOM) to visualize the phenomenon from a global perspective, allowing identifying and characterizing homogenous areas in a geo-spatial perspective. The methodology was applied to a set of precipitation indices, which were computed using daily precipitation data from 1998 to 2000 measured at 19 meteorological stations located in Madeira Island. Results show that the island has distinct climatic areas in relation to extreme precipitation events. The northern part of the island and the higher locations are characterized by heavy precipitation events, whereas the south and northwest of the island exhibit low values in all indices. The promising results from this study indicate the proposed methodology, which combines linear and nonlinear approaches, as a valuable tool to deepen the knowledge on the local spatial patterns of extreme precipitation.**

Keywords-Geostatistics; Kriging; Precipitation patterns; Self-Organizing Map.

I. INTRODUCTION

The occurrence of extreme weather events, such as the extreme precipitation, is often associated to climate change and constitutes an enormous challenge to society. In fact, the monitoring of risk associated with such phenomena is a key element in ensuring the sustainability of economic development and living conditions of populations. It is in this context that we have been witnessing an increase in information on this type of extreme weather [1].

Extreme precipitation events can be characterized using several approaches. To gain a uniform perspective on observed changes in precipitation extremes, a core set of standardized indices was defined by the joint working group CCI/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI).

Numerous studies of changes in extreme weather events focus on linear trends in the indices, aiming to determine whether there has been a statistically significant shift in such indices of extremes [2-5], but only a few focus on their local spatial patterns [6].

Madeira is a Portuguese subtropical island located in the North Atlantic. It is considered a Mediterranean biodiversity 'hot-spot' and is especially vulnerable to climate change [7]. During the winter season, eastward moving Atlantic low-pressure systems bring precipitation to the island and stationary depressions can cause extreme precipitation events [7]. The characterization of precipitation in Portuguese islands has been less studied than in mainland Portugal [3].

The work reported herein investigates the spatial patterns of extreme precipitation in Madeira Island during three hydrological years (1998-2000). Among the eleven precipitation indices proposed by the ETCCDI, five indices were selected (R1, R1d, CWD, SDII and Rx5d), hoping to achieve a global characterization of the phenomenon in its different perspectives. The selected indices capture not only the precipitation intensity, but also the frequency and length of heavy precipitation events. Although the period chosen is not significant for a robust characterization of extreme precipitation events in Madeira Island, it is sufficient to test the proposed methodology and provide an exploratory analysis of the phenomenon.

First, and for spatial interpolation purposes, the spatial continuity models of the five precipitation indices will be computed using geostatistical procedures, such as Ordinary Kriging (OK) and Ordinary Cokriging (OCK). Finally, the estimated surfaces of all the precipitation indices will be analyzed using a clustering tool especially adapted for visualizing multidimensional data: the SOM [8-10].

This paper is organized into five Sections as indicated: Section 2 presents the study area and the main data characteristics; Section 3 provides a description of the methodology; Section 4 reports the results obtained; and, finally, some concluding remarks are made in Section 5.

II. STUDY REGION AND DATA

This Section provides a description of the study region and of the data used to characterize the extreme precipitation patterns in Madeira Island.

ISBN: 978-1-61208-178-6

A. Madeira Island

The study area corresponds to Madeira Island, which is located in the Atlantic Ocean between latitudes 32° 30' N – 33° 30' N and longitudes 16° 30' W – 17° 30' W. The island has an area of approximately 737 km2 distributed over a mountain range of 58 km oriented in the direction WNW-ESE (Fig. 1).

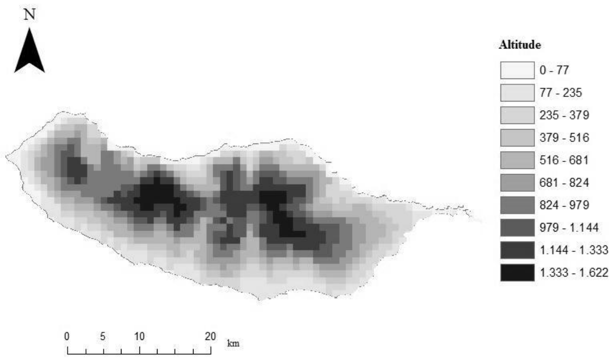


Figure 1. Madeira's Island Elevation Model.

The climate of the island is extremely affected by the Atlantic Azores anticyclone and also by its own characteristics of altitude and relief direction [11]. In fact, the Island topography orientation causes a barrier, almost perpendicular to the most frequent wind direction (northeast). As a result of this natural barrier, there is a continuous ascent of moist air masses from the Atlantic, causing frequent precipitation in the northern part of the island [11].

Despite the small size of the island, there are significant differences in the climate of its two halves [12]: the northern part of the island is colder and wetter, and the southern part is warmer and drier. Also, and as expected, the precipitation on the island increases with altitude but presents significant differences between those two halves.

The highest annual precipitation occurs in the highest parts of the island and the lower rainfall amounts are observed in lowland areas, such as *Funchal* and *Ponta do Sol* [13].

B. Precipitation indices

The daily precipitation data used to compute the indices were observed at 19 meteorological stations of the National Information System of Hydric Resources (NISHR) in the period 1998–2000 (Fig. 2), and downloaded from the NISHR database (http://snirh.pt). In the present study, only annually specified indices are considered. A wet day is defined as a day with an accumulated precipitation of at least 1.0 mm. The precipitation indices computed on an annual basis can be described as follows:

- R1 is the number of wet days (in days);
- Rx1d is the maximum 1-day precipitation (in mm);
- CWD is the maximum number of consecutive wet days (in days);
- SDII is named simple daily intensity index, and is equal to the ratio between the total rain on wet days and the number of wet days (in mm);
- Rx5d is the highest consecutive 5–day precipitation total.

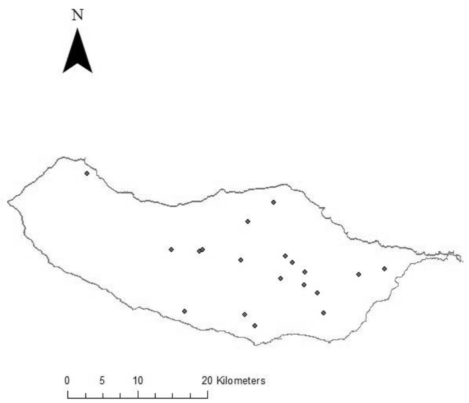


Figure 2. Distribution of meteorological stations over the island (NISHR network).

The precipitation data used in the subsequent analysis corresponds to the simple annual average of each index from October 1998 to September 2000, at each station location. Summary statistics of these data are presented in Table I. The combined analysis of the 5 indices allows characterizing extreme precipitation situations under different perspectives, namely considering the intensity, length and frequency of the precipitation events.

TABLE I. SUMMARY STATISTICS OF THE PRECIPITATION INDICES VALUES AVERAGED IN THE PERIOD 1998–2000

Variable	CWD	R1	Rx1d	SDII	Rx5d
Min	5	52	50	8	64
Median	9	104	114	15,00	216
Max	15	141	169	26	390
Mean	9,53	94,95	114,74	15,48	218,2
Standard-deviation	3,1	27,47	35,0	4,26	92,9
Skewness	0,44	-0,25	-0,06	1,11	0,18
Kurtosis	-0,77	-1,22	-1,22	2,15	-0,63

The data and ancillary information used in this study, particularly the island map and its Terrestrial Digital Elevation Model (Fig. 1) were downloaded from the Portuguese Hydrographic Institute website and from the GeoCommunity™ portal, respectively.

III. METHODOLOGY

The methodology used in this study integrates two main steps: first, the values of each variable at unsampled locations are estimated using geostatistical procedures; second, the variables are visualized using the SOM.

A. *Geostatistical modeling of precipitation indices*

As the ultimate goal is to get an insight of the spatial patterns of extreme precipitation over the island, the first step corresponds to the spatial interpolation of each averaged index, i.e., to estimate the values of each primary variable at unsampled locations.

Deterministic interpolation methods, such as Inverse Distance Weighting (IDW), were not considered because this methods produce inaccurate results when applied to clustered data [14]. In fact, not only the number of stations is small, but also the stations are not distributed equally over the island.

Geostatistical methods, known as Kriging, are usually preferred to estimate unknown values at unsampled locations because they account for the attribute spatial continuity. In this study, we will focus on two particular cases of this group of linear regression estimators: the OK and the OCK. The main difference between these two Kriging variants is that OCK explicitly accounts for the spatial cross-correlation between the primary variable and secondary variables [15]. The elevation model of the Madeira Island will be used as secondary information as some primary variables are strongly correlated with elevation.

A key step of Kriging interpolation is the spatial continuity modeling, which corresponds to fit an authorized semivariogram model (e.g., exponential, spherical, Gaussian, etc.) to the experimental semivariogram cloud of points [15]. This procedure is extremely important for structural analysis and is essential to get the Kriging parameters [16]. The modeling results of this stage will be detailed in the next Section. The methodology used to model the spatial continuity of each index can be summarized as follows:

- Determine the experimental semivariogram for the two main directions of the island relief orientation (if there is significant evidence of geometric anisotropy). Isotropy can be assumed only if the semivariogram is not dependent on direction [17];
- In the remaining cases assume isotropy;
- If there is evidence of strong correlation and linear relationship between some primary variable and the existing secondary information (i.e., elevation), the model of co-regionalized variables is considered in the semivariogram modeling phase;
- After modeling the experimental semivariograms, the OK/OCK methods are applied. The interpolation model selected to describe each index will be chosen based on the Mean Error (ME) of the cross-validation (or "leave-one-out" cross-validation) results. This criterion is especially appropriate for determining the degree of bias in the estimates [14], but it tends to be lower than the real error [18]. Therefore, the final decision will also consider the Root Mean Square Error (RMSE) of the cross-validation results, which is an error statistic commonly used to check the accuracy of the interpolation method.

B. *Using the SOM to Visualize the Precipitation Indices*

After producing the spatial surface of each averaged precipitation index, the main goal is to visualize this set of indicators in order to identify areas with similar patterns of occurrence of extreme precipitation. To achieve this, we propose the use of the SOM, a data visualization tool that has been proposed for visualizing spatial data [19, 20].

The SOM is an artificial neural network based on an unsupervised learning process that performs a gradual and nonlinear mapping of high dimensional input data onto an ordered and structured array of nodes, generally of lower dimension [10]. As a result of this process, and by combining the properties of an algorithm for vector quantization and vector projection, the SOM compresses information and reduces dimensionality [21].

Because the SOM converts the nonlinear statistical relationships that exist in data into geometric relationships, able to be represented visually [9, 10], it can be considered as a visualization method for multidimensional data especially adapted to display the clustering structure [22, 23], or in other words, as a diagram of clusters [9]. When compared with other clustering tools, the SOM is characterized mainly by the fact that, during the learning process, the algorithm tries to guarantee the topological order of its units, thus allowing an analysis of proximity between the clusters and the visualization of their structure [24].

Typically, a clustering tool must ensure the representation of the existing patterns in data, the definition of proximity between these patterns, the characterization of clusters and the final evaluation of output [25]. In the case of spatial data, the clustering tool should also ensure that the groups are made in line with the geographical closeness [24]. The geo-spatial perspective is, in fact, a crucial point that makes the difference between spatial clustering and clustering in common data. Recognizing this, there are several approaches, including some variants to the SOM algorithm [26], proposed to visualize the SOM in order to deal with geo-spatial features.

In this context, an alternative way to visualize the SOM taking advantage of the very nature of geo-referenced data can be reached by coloring the geographic map with label colors obtained from the SOM units [24]. One such approach is the “Prototypically Exploratory Geovisualization Environment” [27] developed in MATLAB®. This prototype incorporates the possibility of linking SOM to the geographic representation by color, allowing dealing with data in a geo-spatial perspective.

In this study, we propose to use a clustering method for spatial data based on the visualization of the output space of a 3D SOM [28]. This approach is based on the association of each of the three orthogonal axes (x, y and z) that define the SOM grid to one of the three primary colors: red, green and blue (RGB scheme). As a result, each of the three dimensions of the 3D SOM will be expressed by a change in tone of one particular primary color (RGB), and each SOM unit will have a distinct color label. Therefore, each geo-referenced element can be painted with the color of its Best Matching Unit (BMU), i.e., the SOM unit where each geo-referenced element is mapped.

Fig. 3 represents schematically a SOM with 27 units (3x3x3) in the RGB space followed by the geographical representation of several geo-referenced elements painted with the color labels of their BMU's.

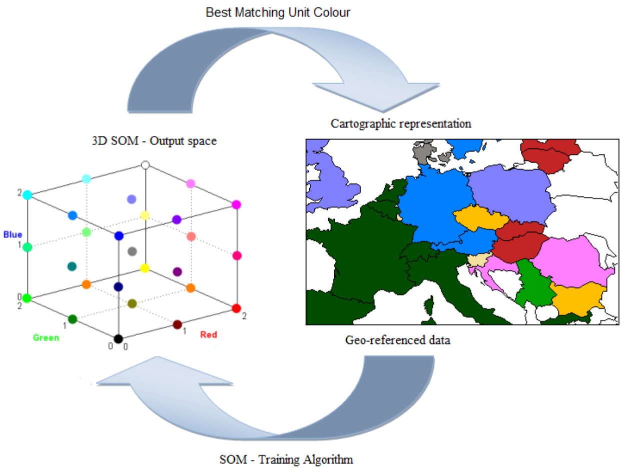


Figure 3. Linking SOM's knowledge to cartographic representation. A color is assigned to each SOM unit (following the topological order). Then the geo-referenced elements are painted with the color of their BMU's.

IV. Results

In this Section we present the spatial interpolation of the precipitation indices and the spatial patterns of extreme precipitation obtained using the methodology proposed in the previous Section.

A. *Spatial interpolation of precipitation indices*

The semivariogram modeling was conducted using the GeoMS® software and the spatial prediction models were obtained using ARCGIS®. The final visualization of the extreme precipitation was produced through routines and functions implemented in MATLAB®.

Not surprisingly, the most correlated indices are Rx1d and Rx5d (R2=0.804). The remaining indices are moderately or weakly correlated, which indicates their suitability to characterize different features of the precipitation regime in the Madeira Island. Moreover, Rx5d and CWD are moderately correlated with elevation (Table II).

TABLE II. Correlation Matrix Between Indices and Elevation (Elev.)

Variables	Elev.	CWD	R1	Rx1d	SDII	Rx5d
Elev.	1					
CWD	0,768	1				
R1	0,424	0,684	1			
Rx1d	0,393	0,242	0,489	1		
SDII	0,308	-0,134	-0,098	0,627	1	
Rx5d	0,616	0,440	0,542	0,804	0,62	1

Taking into account the results obtained in the exploratory analysis (IDW models not shown), several modeling strategies were compared taking into account the spatial continuity behavior assumed for each index and its correlation with elevation (Table III). Although the relief of the island is in direction WNW-ESE, the analysis of the estimated surfaces obtained with IDW (not shown) shows no evidence of anisotropy, except for variable Rx5d. This means that the spatial variability of all other indices was assumed identical in all directions (i.e., isotropic).

Table IV summarizes the semivariogram parameters estimated for the models indicated in Table III.

TABLE III. Experimental Semivariogram Modeling Strategies

Index model number	Semivariogram	Spatial behavior assumed
CWD-1	Omnidirectional	Isotropic
CWD-2	Linear model of co-regionalization with elevation	Isotropic
R1	Omnidirectional	Isotropic
Rx1d	Omnidirectional	Isotropic
SDII	Omnidirectional	Isotropic
Rx5d-1	Omnidirectional	Isotropic
Rx5d-2	Semivariogram models for the azimuth directions 100º and 10º	Anisotropic
Rx5d-3	Linear model of co-regionalization with elevation	Isotropic

TABLE IV. Semivariogram Parameters Estimated for the Models Indicated in Table III

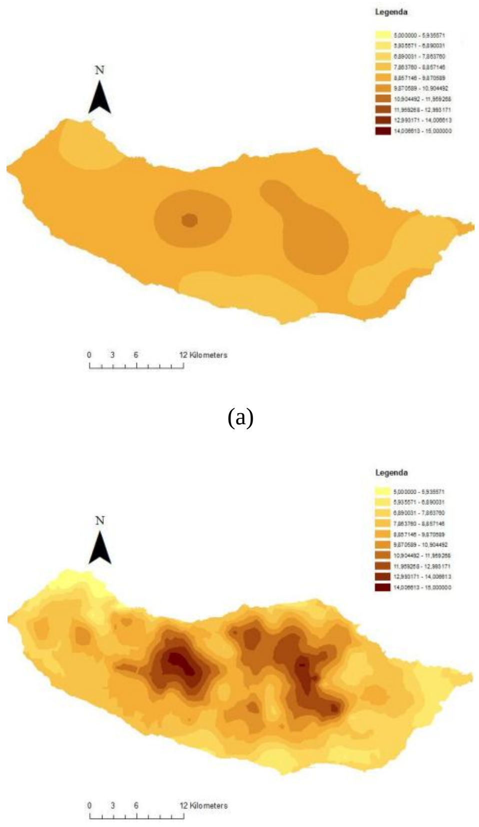
Index model number	Model type	Nugget	Partial sill	Spatial range (Km)
CWD-1	Spherical	6	3	11.7
CWD-2	Exponential (Exp.)	0	9 (CWD) 940 (CWD-Elevation) 166272 (Elevation)	13.4
R1	Exp.	0	714	12.6
Rx1d	Exp.	0	1157	8.2
SDII	Exp.	0	17	5.3
Rx5d-1	Gaussian	1165	6992	12.7
Rx5d-2	Gaussian	1371	6794	14.3 (major) 8.2 (minor)
Rx5d-3	Spherical	0	16440 (Rx5d) 23891 (Rx5d-Elevation) 166380 (Elevation)	12.6

OCK with elevation was used in the spatial interpolation of the averaged Rx5d and CWD, whereas all other variables were interpolated through OK (Fig. 4-8).

The final interpolation model selected to describe the spatial distribution of Rx5d and CWD depends on the error statistics of the cross-validation (Table V). ME values close to zero indicate a small bias in the estimation. Hence, the best interpolation strategy for both variables is OCK with the semivariogram models Rx5d-3 and CWD-2, respectively.

TABLE V. CROSS-VALIDATION ERROR STATISTICS OBTAINED IN THE VARIOUS SPATIAL INTERPOLATION STRATEGIES (SELECTED MODELS ARE IN ITALICS)

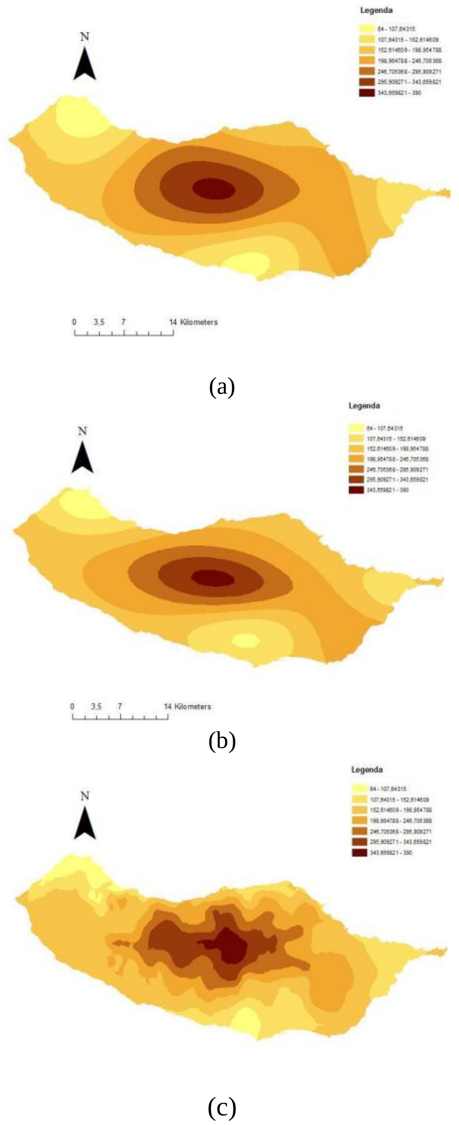
Indices	Spatial interpolation model	ME	RMSE
CWD	OK with the semivariogram model CWD-1	0,045	3,13
	OCK with the semivariogram model CWD-2	-0,02	3,214
R1	OK	0,529	20,77
Rx1d	OK	2,68	31,67
SDII	OK	-0,01	5,012
Rx5d	OK with the semivariogram model Rx5d-1	5,647	59,52
	OK with the semivariogram model Rx5d-2	4,493	56,5
	OCK with the semivariogram model Rx5d-3	-0,853	69,04



(a)

(b)

Figure 4. Interpolation of the averaged CWD index using: (a) OK and the semivariogram model CWD-1; (b) OCK and the semivariogram model CWD-2.



(a)

(b)

(c)

Figure 5. Interpolation of the averaged Rx5d index using: (a) OK and the semivariogram model Rx5d-1; (b) OK and the semivariogram model Rx5d-2. (c) OCK and the semivariogram model Rx5d-3.

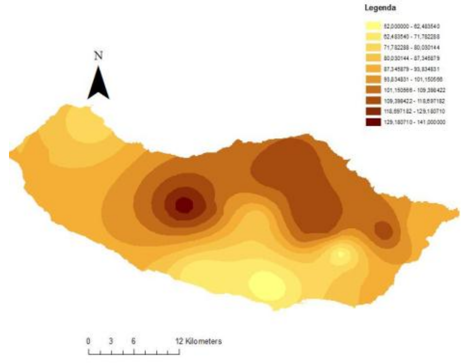


Figure 6. OK interpolation of the averaged R1 index.

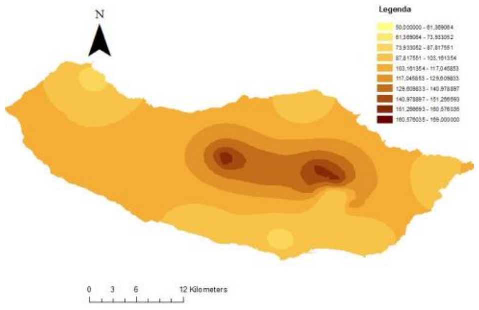


Figure 7. OK interpolation of the averaged Rx1d.

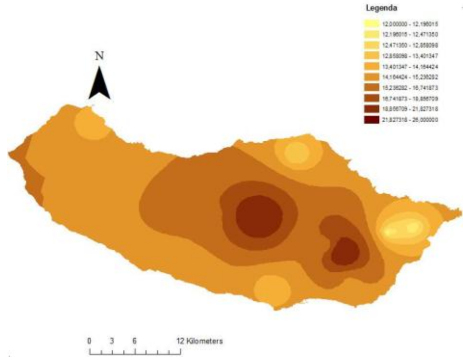


Figure 8. OK interpolation of the averaged SDII index.

B. *Spatial patterns of extreme precipitation*

In order to visualize the spatial patterns of extreme precipitation from a global perspective, a 3D SOM was applied to the indices surfaces obtained through Kriging. First, the selected models (Table V), obtained in raster format, were converted back to point data, sampled at regular intervals. Afterwards, the indices values were normalized to ensure equal variance in all variables and the SOM was parameterized as follows:

- The output space was set with 3 dimensions [4 × 4 × 4], which corresponds to 64 units in total;
- The neighborhood function selected was Gaussian;
- The length of the training was set to "long" (8 epochs);
- Random initialization.

As the final results depend on the initialization of the SOM, 100 models were obtained and the best model was chosen according to the criterion of best fit, i.e., the lowest quantization error (Table VI).

TABLE VI. 3D SOM RESULTS (100 MODELS)

	Quantization Error	Topological Error
Selected Model	0,74747	0,042616
Average Model	0,78156	0,041578

To each unit of the SOM (output space of the network) was then assigned a RGB color according to its output space coordinates. In turn, each raster cell was represented cartographically with the color assigned to the unit of the SOM where that cell is mapped, i.e., its BMU (Fig. 9). This means that each color corresponds to a homogeneous zone in terms of the various indices values.

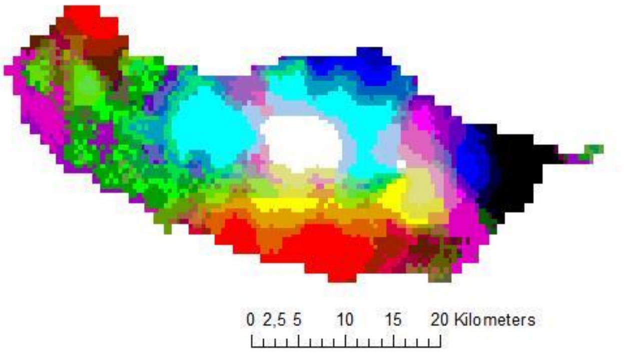


Figure 9. Visualization of the five precipitation indices using the output of the SOM mapped to a 3D RGB space. Areas with similar colors have similar characteristics.

Table VII summarizes the characteristics of each area identified in Fig. 9. There are significant differences between the different areas (colors). Table VII allows comparing the predicted mean values for the whole island.

TABLE VII. SUMMARY OF THE AVERAGE VALUES FOR EACH AREA

Color\Index	CWD	R1	Rx1d	SDII	Rx5d
Black	6,58	91	100,2	13,73	124,89
White	10,12	92,60	138	19,87	336,1
Yellow	9,57	75,86	101	16,99	206,24
Light Blue	12,88	116,46	115,4	15,62	301,5
Blue	9,465	109,81	106,3	14,14	198,82
Green	8,91	92,3	105	15,09	188,0
Red	7,52	72,46	95	14,60	132,08
Violet	7,87	94,71	110,9	15,51	184,6

Despite its small size, Madeira Island has distinct zones in relation to extreme precipitation events. The white area corresponds to the higher regions of the island characterized by higher values in all indices, whereas the darkest area (black in the far east of the island), is characterized by the lowest values in all indices. The north of the island, which is colored dark blue and light blue, corresponds to high values in all indices (although much smaller than in the white colored area), with particularly high R1 index values. Finally, the area colored in red is characterized by low values in all indices. The green area is very close to the average values (a phenomenon that is partly explained by the lack of information in the area). There are no significant differences between the green and violet zone (analysis of Euclidean distance).

ISBN: 978-1-61208-178-6

V. Conclusion

In this paper, we propose a methodology for characterizing the spatial patterns of extreme precipitation in Madeira Island. This methodology combines two different approaches: the first one is based on geostatistical procedures, and the second one is based on the 3D SOM. The first approach is used to estimate spatial surfaces of extreme precipitation indices, and the second one allows visualizing the phenomenon from a global perspective, thus enabling the identification of homogeneous areas in relation to extreme precipitation events.

The spatial and temporal resolution of the data set considered is too small to thoroughly characterize the extreme precipitation phenomenon in Madeira Island. Nevertheless, the results indicate the proposed methodology as a valuable tool to provide a set of maps that can effectively assist the spatial analysis of a phenomenon. It can have multiple perspectives and deal with high dimensional data, which requires a global view. The results of this particular application open perspectives for new applications not only in the climate context, but also in other domains.

References

[1] A.M.G.K. Tank, F.W. Zwiers and X. Zhang, Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation, WMO-TD, WMO, 2009.

[2] A.C. Costa and A. Soares, “Trends in extreme precipitation indices derived from a daily rainfall database for the South of Portugal,” International Journal of Climatology, vol. 29, no. 13, 2009, pp. 1956-1975; DOI 10.1002/joc.1834.

[3] M.I.P. de Lima, S.C.P. Carvalho and J.L.M.P. de Lima, “Investigating annual and monthly trends in precipitation structure: an overview across Portugal,” Nat. Hazards Earth Syst. Sci., vol. 10, no. 11, 2010, pp. 2429-2440; DOI 10.5194/nhess-10-2429-2010.

[4] G.M. Griffiths, M.J. Salinger and I. Leleu, “Trends in extreme daily rainfall across the South Pacific and relationship to the South Pacific Convergence Zone,” International Journal of Climatology, vol. 23, no. 8, 2003, pp. 847-869; DOI 10.1002/joc.923.

[5] M. Haylock and N. Nicholls, “Trends in extreme rainfall indices for an updated high quality data set for Australia, 1910–1998,” International Journal of Climatology, vol. 20, no. 13, 2000, pp. 1533-1541; DOI 10.1002/1097-0088(20001115)20:13<1533::aid-joc586>3.0.co;2-j.

[6] A.C. Costa, R. Durão, M.J. Pereira and A. Soares, “Using stochastic space-time models to map extreme precipitation in southern Portugal,” Nat. Hazards Earth Syst. Sci., vol. 8, no. 4, 2008, pp. 763-773; DOI 10.5194/nhess-8-763-2008.

[7] M.J. Cruz, R. Aguiar, A. Correia, T. Tavares, J.S. Pereira and F.D. Santos, “Impacts of climate change on the terrestrial ecosystems of Madeira,” International Journal of Design and Nature and Ecodynamics, vol. 4, no. 4, 2009, pp. 413-422.

[8] T. Kohonen, “The self-organizing map,” Proceedings of the IEEE, vol. 78, no. 9, 1990, pp. 1464 -1480.

[9] T. Kohonen, “The self-organizing map,” Neurocomputing, vol. 21 no. 1-3, 1998, pp. 1-6.

[10] T. Kohonen, Self-organizing Maps, Springer, 2001.

[11] S. Prada, M. Menezes de Sequeira, C. Figueira and M.O. da Silva, “Fog precipitation and rainfall interception in the natural forests of Madeira Island (Portugal),” Agricultural and Forest Meteorology, vol. 149, no. 6-7, 2009, pp. 1179-1187.

[12] J.J.M. Loureiro, “Monografia hidrológica da ilha da Madeira,” Revista Recursos Hídricos, vol. 5, 1984, pp. 53-71.

[13] S. Prada, “Geologia e Recursos Hídricos Subterrâneos da Ilha da Madeira,” Universidade da Madeira, 2000.

[14] E.H. Isaaks and R.M. Srivastava, An introduction to applied geostatistics, Oxford University Press, 1989.

[15] P. Goovaerts, Geostatistics for natural resources evaluation, Oxford University Press, 1997.

[16] P.A. Burrough and R.A. McDonnell, Principles of Geographical Information Systems, Oxford University Press, 1998.

[17] A.D. Hartkamp, K.D. Beurs, A. Stein and J.W. White, Interpolation Techniques for Climate Variables, CIMMYT, 1999.

[18] I.A. Nalder and R.W. Wein, “Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest,” Agricultural and Forest Meteorology, vol. 92, no. 4, 1998, pp. 211-225.

[19] E.L. Koua, “Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets,” Proc. Proceedings of 21st International Cartographic Renaissance (ICC), International Cartographic Association, 2003, pp. 1694-1702.

[20] F. Bação, V. Lobo and M. Painho, “Applications of Different Self-Organizing Map Variants to Geographical Information Science Problems,” Self-Organising Maps: applications in geographic information science, A. Skupin and P. Agarwal, eds., John Wiley & Sons, 2008, pp. 22-44.

[21] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, SOM Toolbox for Matlab 5, Helsinki University of Techology, 2000.

[22] J. Himberg, “A SOM based cluster visualization and its application for false coloring,” Proc. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000, pp. 587- 592.

[23] S. Kaski, J. Venna and T. Kohonen, “Coloring that reveals high-dimensional structures in data,” Proc. Proceedings of 6th International Conference on Neural Information Processing, IEEE, 1999, pp. 729-734.

[24] A. Skupin and P. Agarwal, “What is a Self-organizing Map?,” Self-Organising Maps: applications in geographic information science, P. Agarwal and A. Skupin, eds., John Wiley & Sons, 2008, pp. 1-20.

[25] A.K. Jain, M.N. Murty and P.J. Flynn, “Data Clustering: A Review,” ACM Computing Surveys, vol. 31, no. 3, 1999, pp. 264-323.

[26] F. Bação, V. Lobo and M. Painho, “The self-organizing map, the Geo-SOM, and relevant variants for geosciences,” Computers & Geosciences, vol. 31, no. 2, 2005, pp. 155-163.

[27] E.L. Koua and M. Kraak, “An Integrated Exploratory Geovisualization Environment Based on Self-Organizing Map,” Self-Organising Maps: applications in geographic information science, P. Agarwal and A. Skupin, eds., John Wiley & Sons, 2008, pp. 45-86.

[28] J. Gorricha and V. Lobo, “On the Use of Three-Dimensional Self-Organizing Maps for Visualizing Clusters in Georeferenced Data ” Information Fusion and Geographic Information Systems, Lecture Notes in Geoinformation and Cartography 5, V. V. Popovich, et al., eds., Springer Berlin Heidelberg, 2011, pp. 61-75.

Fast and Accurate Visibility Computation in a 3D Urban Environment

Oren Gal
Mapping and Geo-information Engineering
Technion - Israel Institute of Technology
Haifa, Israel
e-mail: orengal@technion.ac.il

Yerach Doytsher
Mapping and Geo-information Engineering
Technion - Israel Institute of Technology
Haifa, Israel
e-mail: doytsher@technion.ac.il

***Abstract*—This paper presents a unique solution to the visibility problem in 3D urban environments. We shall introduce a visibility algorithm for a 3D urban environment, based on an analytic solution for basic building structures. A building structure is presented as a continuous parameterization approximating of the building's corners. The algorithm quickly generates the visible surfaces' boundary of a single building. Using simple geometric operations of projections and intersections between visible pyramid volumes, hidden surfaces between buildings are rapidly computed. The algorithm, demonstrated with a schematic structure of an urban environment and compared to the Line of Sight (LOS) method, demonstrates the computation time efficiency. The basic building structure can be modified to complex urban structures by merging together a number of basic structures.**

Keywords-Visibility; 3D; Urban environment; Spatial analysis

I. INTRODUCTION

In the last few years, the 3D GIS domain has developed rapidly, and has become increasingly accessible to different disciplines. Spatial analysis in a 3D environment seems to be one of the most challenging topics in the communities currently dealing with spatial data. One of the most basic problems in spatial analysis is related to visibility computation in such an environment. Visibility calculation methods aim to identify the parts visible from a single point, or multiple points, of objects in the environment.

The visibility problem has been extensively studied over the last twenty years, due to the importance of visibility in GIS, computer graphics, computer vision and robotics. Most previous works approximate the visible parts to find a fast solution in open terrains, and do not challenge or suggest solutions for a dense urban environment. The exact visibility methods are highly complex, and cannot be used for fast applications due to the long computation time. Other fast algorithms are based on the conservative Potentially Visible Set (PVS) [8]. These methods are not always completely accurate, as they may include hidden objects' parts as visible due to various simplifications and heuristics.

In this paper, we introduce a new fast and exact solution to the 3D visibility problem from a viewpoint in urban environment, which does not suffer from approximations. We consider a 3D urban environment building modeled as a cube (3D box) and present analytic solution to the visibility problem. The algorithm computes the exact visible and hidden parts from a viewpoint in urban environment, using an analytic solution, without the expensive computational process of scanning all objects' points. The algorithm is demonstrated by a schematic structure of an urban environment, which can also be modified for other complicated urban environments, with simple topological geometric operators. In such cases, computation time grows linearly.

Our method uses simple geometric relations between the objects and the lines connecting the viewpoint and the objects' boundaries by extending the visibility boundary calculation from 2D to a 3D environment by using approximated singular points [9]. The spatial relationship between the different objects is computed by using fast visible pyramid volumes from the viewpoint, projected to the occluded buildings.

The current research tackles the basic case of a single viewpoint in an urban environment, which consists of buildings that are modeled as cubes. More complex urban environments can be defined as a union between the basic structures of several cubes. Further research will focus on modeling more complex urban environments, and facing multiple viewpoints for optimal visibility computation in such environments.

II. PROBLEM STATEMENT

We consider the basic visibility problem in a 3D urban environment, consisting of 3D buildings modeled as 3D cubic parameterization $\sum_{i=1}^{N} C_i(x, y, z = {}^{h_{\max}}_{h_{\min}})$, and viewpoint $V(x_0, y_0, z_0)$.

Given:

- A viewpoint $V(x_0, y_0, z_0)$ in 3D coordinates
- Parameterizations of N objects $\sum_{i=1}^{N} C_i(x, y, z = {}^{h_{\max}}_{h_{\min}})$ describing a 3D urban environment model

ISBN: 978-1-61208-178-6

Computes:

- Set of all visible points in $\sum_{i=1}^{N} C_i(x, y, z =_{h_{\min}}^{h_{\max}})$ from V(x$_0$, y$_0$, z$_0$).

This problem seems to be solved by conventional geometric methods, but as mentioned before, it demands a long computation time. We introduce a fast and efficient computation solution for a schematic structure of an urban environment that demonstrates our method.

III. Analytic Visibility Computation

A. Analytic Solution for a Single Object

In this section, we first introduce the visibility solution from a single point to a single 3D object. This solution is based on an analytic expression, which significantly improves time computation by generating the visibility boundary of the object without the need to scan the entire object's points.

Our analytic solution for a 3D building model is an extension of the visibility chart in 2D introduced by Elber et al. [9] for continuous curves. For such a curve, the silhouette points, i.e. the visibility boundary of the object, can be seen in Figure 1:

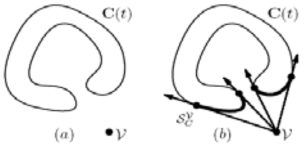


Figure 1. Visible Silhouette Points $S_C{}^V$ from viewpoint V to curve *C(t)* (source: [9]).

The visibility chart solution was originally developed for dealing with the Art Gallery Problem for infinite viewpoint; it is limited to 2D continuous curves using multivariate solver [9], and cannot be used for on-line application in a 3D environment.

Based on this concept, we define the visibility problem in a 3D environment for more complex objects as:

$$C'(x, y)_{z_{const}} \times (C(x, y)_{z_{const}} - V(x_0, y_0, z_0)) = 0 \quad (1)$$

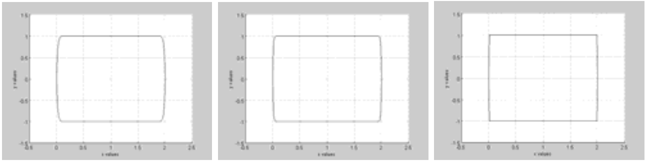
where 3D model parameterization is $C(x, y)_{z_{const}}$, and the viewpoint is given as $V(x_0, y_0, z_0)$. Solutions to equation (1) generate a visibility boundary from the viewpoint to an object, based on basic relations between viewing directions from V to $C(x, y)_{z_{const}}$ using cross-product characters.

A three-dimensional urban environment consists mainly of rectangular buildings, which can hardly be modeled as continuous curves. Moreover, an analytic solution for a single 3D model becomes more complicated due to the higher dimension of the problem, and is not always possible. Object parameterization is therefore a critical issue, allowing us to find an analytic solution and, using that, to generate the visibility boundary very fast.

1) 3D Building Model: Most of the common 3D City Models are based on object-oriented topologies, such as 3D Formal Data Structure (3D FDS), Simplified Spatial Model (SSS) and Urban Data Model (UDM) [24]. These models are very efficient for web-oriented applications. However, the fact that a building consists of several different basic features makes it almost impossible to generate analytic representation. A three-dimensional building model should be, on the one hand, simple enabling analytic solution, and on the other hand, as accurate as possible. We examined several building object parameterizations, and the preferred candidate was an extended n order sphere coordinates parameterization, even though such a model is a very complex, and will necessitate a special analytic solution. We introduce a model that can be used for analytic solution of the current problem. The basic building model can be described as:

$$x = t, y = \begin{pmatrix} x^n - 1 \\ 1 - x^n \end{pmatrix}, z = c \quad (2)$$
$$-1 \le t \le 1, n = 350, c = c + 1$$

This mathematical model approximates building corners, not as singular points, but as continuous curves. This building model is described by equation (2), with the lower order badly approximating the building corners, as depicted in Figure 2. Corner approximation becomes more accurate using *n=350* or higher. This approximation enables us to define an analytic solution to the problem.



(a) (b) (c)

Figure 2. Topside view of the building model using equation (2) - (a) n=50; (b) *n=200*; (c) *n=350*.

We introduce the basic building structure that can be rotated and extracted using simple matrix operators (Figure 3). Using a rotation matrix does not affect our visibility algorithm, and for simple demonstration of our method we present samples of parallel buildings.

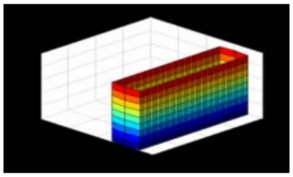


Figure 3. A Three-dimension Analytic Building Model with Equation (2), where $z_{h_{\min}=0}^{h_{\max}=9}$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-178-6

2) Analytic Solution for a Single Building: In this part we demonstrate the analytic solution for a single 3D building model. As mentioned above, we should integrate building model parameterization to the visibility statement. After integrating eq. (1) and (2):

$$\begin{aligned} &C'(x,y)_{z_{const}} \times (C(x,y)_{z_{const}} - V(x_0,y_0,z_0)) = 0 \rightarrow \\ &x^n - V_{y_0} - n\cdot x^{n-1}(x - V_{x_0}) - 1 = 0 \\ &x^n + V_{y_0} - n\cdot x^{n-1}(x - V_{x_0}) - 1 = 0 \\ &n = 350, -1 \le x \le 1 \end{aligned} \tag{3}$$

where the visibility boundary is the solution for these coupled equations. As can be noticed, these equations are not related to Z axis, and the visibility boundary points are the same ones for each x-y surface due to the model's characteristics. Later on, we treat the relations between a building's roof and visibility height in our visibility algorithm, as part of the visibility computation.

The visibility statement leads to two polynomial *N* order equations, which appear to be a complex computational task. The real roots of these polynomial equations are the solution to the visibility boundary. These equations can be solved efficiently by finding where the polynomial equation changes its sign and cross zero value; generating the real roots in a very short time computation (these functions are available in Matlab, Maple and other mathematical programs languages). Based on the polynomial cross zero solution, we can compute a fast and exact analytic solution for the visibility problem from a viewpoint to a 3D building model. This solution allows us to easily define the Visible Boundary Points.

Visible Boundary Points (VBP) - we define VBP of the object *i* as a set of boundary points $j=1..N_{bound}$ of the visible surfaces of the object, from viewpoint $V(x_0, y_0, z_0)$.

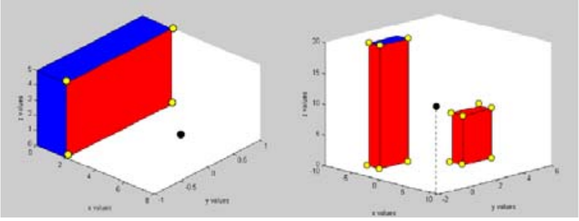
$$VBP_{i=1}^{j=1..N_{bound}}(x_0,y_0,z_0) = \begin{bmatrix} x_1, y_1, z_1 \\ x_2, y_2, z_2 \\ .. \\ x_{N_{bound}}, y_{N_{bound}}, z_{N_{bound}} \end{bmatrix} \tag{4}$$

Roof Visibility – The analytic solution in equation (3) does not treat the roof visibility of a building. We simply check if viewpoint height $V(z_0)$ is lower or higher than the building height $h_{\max_{C_i}}$ and use this to decide if the roof is visible or not:

$$V_{z_0} \ge Z = h_{\max_{C_i}} \tag{5}$$

If the roof is visible, roof surface boundary points are added to VBP. Roof visibility is an integral part of VBP computation for each building.

Two simple cases using the analytic solution from a visibility point to a building can be seen in Figure 4. The visibility point is marked in black, the visible parts colored in red, and the invisible parts colored in blue. The visible volumes are computed immediately with very low computation effort, without scanning all the model's points, as is necessary in LOS-based methods for such a case.



(a) (b)

Figure 4. Visibility Volume computed with the Analytic Solution. Viewpoint is marked in black, visible parts colored in red, and invisible parts colored in blue. VBP marked with yellow circles - (a) single building; (b) two non-overlapping buildings.

B. *Visibility Computation in Urban Environments*

In the previous sections, we treated a single building case, without considering hidden surfaces between buildings, i.e. building surface occluded by other buildings, which directly affect the visibility volumes solution. In this section, we introduce our concept for dealing with these spatial relations between buildings, based on our ability to rapidly compute visibility volume for a single building generating VBP set.

Hidden surfaces between buildings are simply computed based on intersections of the visible volumes for each object. The visible volumes are defined easily using VBP, and are defined, in our case, as Visible Pyramids. The invisible components of the far building are computed by intersecting the projection of the closer buildings' VP base to the far building's VP base as described in 4.2.2.

1) The Visible Pyramid (VP): we define $VP_i^{j=1..Nsurf}(x_0, y_0, z_0)$ of the object *i* as a 3D pyramid generated by connecting VBP of specific surface *j* to a viewpoint $V(x_0, y_0, z_0)$. Maximum number of N_{surf} for a single object is three. VP boundary, colored with green arrows, can be seen in Figure 5. The intersection of VPs allows us to efficiently compute the hidden surfaces in urban environments, as can be seen in the next sub-section.

2) Hidden Surfaces between Buildings: As we mentioned earlier, invisible parts of the far buildings are computed by intersecting the projection of the closer buildings' VP to the far buildings' VP base.

For simplicity, we demonstrate the method with two buildings from a viewpoint $V(x_0, y_0, z_0)$ one (denoted as the first one) of which hides, fully or partially, the other (the second one).

As can be seen in Figure 6, in this case, we first compute VBP for each building separately, $VBP_1^{1..4}$, $VBP_2^{1..4}$, based on these VBPs, we generate VPs for each building, VP_1^1, VP_2^1. After that, we project VP_1^1 base to VP_2^1 base plane, as seen in Figure 7, if existing. At this point, we intersect the projected surface in VP_2^1 base plane and update $VBP_2^{1..4}$ and VP_2^1 (decreasing the intersected part).

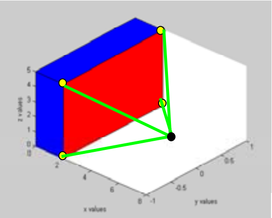
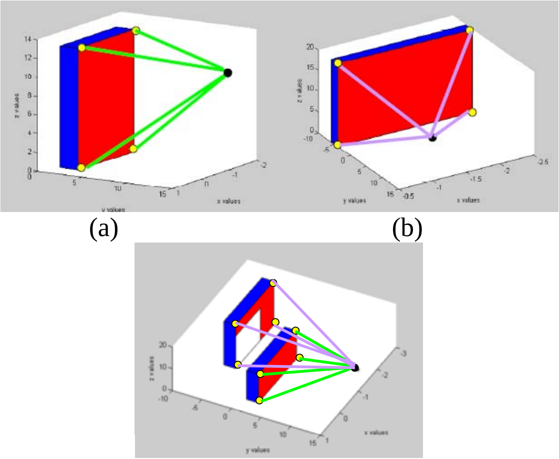


Figure 5. A Visible Pyramid from a viewpoint (marked as a black point) to VBP of a specific surface



(a) (b)

(c)

Figure 6. Generating VP - (a) VP_1^1 boundary colored in green arrows; (b) VP_2^1 boundary colored in purple lines; (c) the two buildings - VP_1^1in green and VP_2^1 in purple, from the viewpoint.

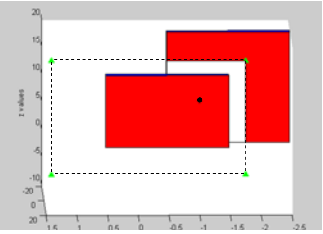


Figure 7. Projection of VP_1^1 *to* VP_2^1base plane marked with dotted lines.

The intersected part is the invisible part of the second building from viewpoint *V(x*$_0$*, y*$_0$*, z*$_0$*)* hidden by the first building, which is marked in white in Figure 8. In the case of a third building, in addition to the buildings introduced in Figure 8, the projected VP will only be the visible ones, and the VBP and VP of the second building will be updated accordingly (as is described in the next sub-section - stage 2.3.4.3).

We demonstrated a simple case of an occluded building. A general algorithm for more a complex scenario, which contains the same actions between all the combinations of VP between the objects, is detailed in the next sub-section. Projection and intersection of 3D pyramids can be done with simple computational geometry elements, which demand a very low computation effort.

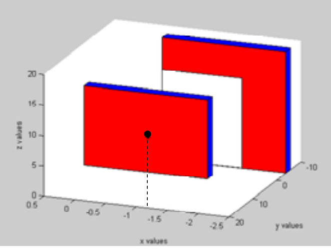


Figure 8. Computing Hidden Surfaces between Buildings by using the Visible Pyramid Colored in White on VP_2^1 Base Plane.

C. *Visibility Algorithm Pseudo - Code*

```
1. Given viewpoint $V(x_0, y_0, z_0)$
2. For $i=1:1:N_{models}$ building model
    2.1. Calculate Azimuth $\theta_i$ and Distance $D_i$ from viewpoint to object
    2.2. Set and Sort Buildings Azimuth Array $\theta[i]$
    2.3. IF Azimuth Objects $(i, 1..i-1)$ Intersect
        2.3.1. Sort Intersected Objects $j=1:1:N_{insect}$ by Distance
        2.3.2. Compute VBP for each intersected building, $VBP_{j=1..N_{\text{int sec}}}^{1..N_{bound}}$ .
        2.3.3. Generate VP for each intersected building, $VP_{j=1..N_{\text{int sec}}}^{1..N_{surf}}$
        2.3.4. For $j=1:1:N_{insect}-1$
            2.3.4.1. Project $VP_j^{1..N_{surf}}$ base to $VP_{j+1}^{1..N_{surf}}$ base plane, if exist.
            2.3.4.2. Intersect projected surfaces in $VP_{j+1}^{1..N_{surf}}$ base plane.
            2.3.4.3. Update $VBP_{j+1}^{1..N_{bound}}$ and $VP_{j+1}^{1..N_{surf}}$ .
            End
    Else
            Locate Building in Urban Environment
    End
End
```

D. *Visibility Algorithm – Complexity Analysis*

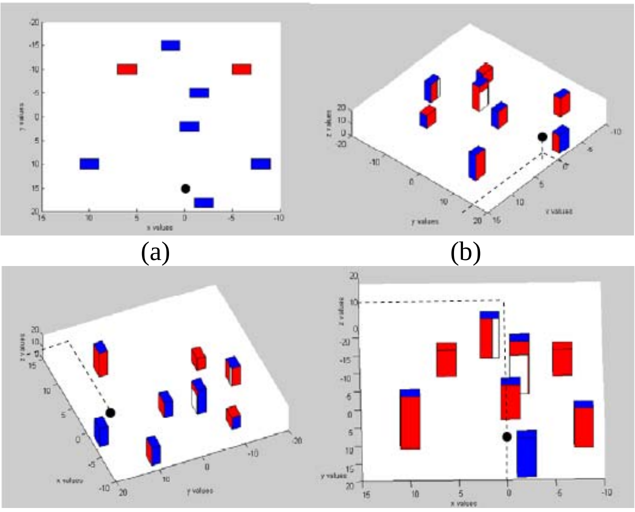
We analyze our algorithm complexity based on the pseudo code presented in the previous section, where *n* represents the number of buildings. In the worst case, *n* buildings hide each other. Visibility complexity consists of generating VBP and VP for *n* buildings, $n \cdot O(1)$ complexity. Projection and intersection are also $n \cdot O(1)$ complexity.

The complexity of our algorithm, without considering data structure managing for urban environments, is $n \cdot O(n)$.

We analyze the visibility algorithm complexity of the LOS methods, where *n* represents the number of buildings and *k* represents the resolution of the object. The exact visibility computation requires scanning each object and each object's points, *O(nk)* where usually *k>>n*.

IV. RESULTS

We have implemented the presented algorithm and tested some urban environments on a 1.8GHz Intel Core CPU with Matlab. First, we analyze the versatility of our algorithm on two different test scenes with different occluded elements. After that, we compare our algorithm to the basic LOS visibility computation, to prove accuracy and computational efficiency. Test scenes can be seen in FigureT 9.



(a) (b)

(c) (d)

Figure 9. Scene number 1: Eight buildings in an Urban Environment, $V(x_0, y_0, z_0)$= (0,15,10) - (a) Topside view; (b)-(d) Different views demonstrating the visibility computation using our algorithm. CPU time was 0.15 sec.

A. Computation Time and Comparison to LOS

The main contribution of this research focuses on a fast and accurate visibility computation in urban environments. We compare our algorithm time computation with common LOS visibility computation demonstrating algorithm's computational efficiency.

1) Visibility Computation Using LOS: The common LOS visibility methods require scanning all of the object’s points. For each point, we check if there is a line connecting the viewpoint to that point which does not cross other objects. We used LOS2 Matlab function, which computes the mutual visibility between two points on a Digital Elevation Model (DEM) model. We converted our first test scene with one to eight buildings to DEM, operated LOS2 function, and measured CPU time after model conversion. Each building with DEM was modeled homogonously by 50 points. The visible parts using the LOS method were the exact parts computed by our algorithm. Obviously, computation time of LOS method was about 500 times longer than our algorithm using analytic solution. CPU time of our analytic solution and LOS method are introduced in Figure 10.

Over the last years, efficient LOS-based visibility methods for DEM models, such as *Xdraw*, have been introduced in order to generate approximate solutions [12]. However, the computation time of these methods is at least *O(n(n-1))*, and, above all, the solution is an approximate one.

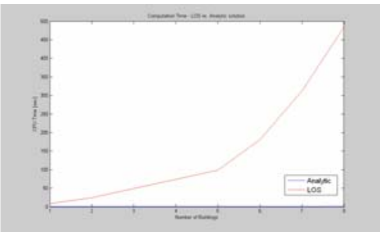


Figure 10. CPU Computation Time of LOS and our algorithm. CPU was measured in the first scene with an increasing number of buildings from one to eight. LOS method was 500 times longer than our algorithm.

V. RELATED WORK

Accurate visibility computation in 3D environments is a very complicated task demanding a high computational effort, which can hardly been done in a very short time using traditional well-known visibility methods [1], [16]. Previous research in visibility computation has been devoted to open environments using DEM models, representing raster data in 2.5D (Polyhedral model). Most of these works have focused on approximate visibility computation, enabling fast results using interpolations of visibility values between points, calculating point visibility with the LOS method [6], [12].

A vast number of algorithms have been suggested for speeding up the process and reducing the computation time [15]. Franklin [11] evaluates and approximates visibility for each cell in a DEM model based on greedy algorithms. An application for siting multiple observers on terrain for optimal visibility cover was introduced in [13]. Wang et al. [21] introduced a Grid-based DEM method using viewshed horizon, saving computation time based on relations between surfaces and Line Of Sight (LOS), using a similar concept of Dead-Zones visibility [3]. Later on, an extended method for viewshed computation was presented, using reference planes rather than sightlines [22].

One of the most efficient methods for DEM visibility computation is based on shadow-casting routine. The routine cast shadowed volumes in the DEM, like a light bubble [17]. Other methods related to urban design environment and open space impact treat abstract visibility analysis in urban environments using DEM, focusing on local areas and approximate openness [10], [23]. Extensive research treated Digital Terrain Models (DTM) in open terrains, mainly Triangulated Irregular Network (TIN) and Regular Square Grid (RSG) structures. Visibility analysis on terrain was classified into point, line and region visibility, and several algorithms were introduced based on horizon computation describing visibility boundary [4], [5].

Only a few works have treated visibility analysis in urban environments. A mathematical model of an urban scene, calculating probabilistic visibility for a given object from a specific viewcell in the scene, has been presented by [14]. This is a very interesting concept, which extends the traditional deterministic visibility concept. Nevertheless, the buildings are modeled as circles, and the main challenges of spatial analysis and building model were not tackled.

Other methods were developed, subject to computer graphics and vision fields, dealing with exact visibility in 3D scenes, without considering environmental constraints. Plantinga and Dyer [16] used the *aspect graph* – a graph with all the different views of an object. Shadow boundaries computation is a very popular method, studied by [20], [7], [19]. All of these works are not applicable to a large scene, due to computational complexity.

As mentioned, online visibility analysis is a very complicated task. Recently, off-line visibility analysis, based on preprocessing, was introduced. Cohen-Or et al. [2] used a ray-shooting sample to identify occluded parts. Schaufler et al. [18] use blocker extensions to handle occlusion.

VI. Conclusions and Future Work

We have presented an efficient algorithm for visibility computation in an urban environment, modeling basic building structure with mathematical approximating for presentation of buildings' corners. Our algorithm is based on a fast visibility boundary computation for a single object, and on computing the hidden surfaces between buildings by using projected surfaces and intersections of the visible pyramids. Complexity analysis of our algorithm has been presented, as well as the computational running time as compared to LOS visibility computation showing significant improvement of time performance.

The main contribution of the method presented in this paper is that it does not require special hardware, and is suitable for on-line computations based on the algorithms' performances, as was presented above. The visibility solution is exact, defining a simple problem that can be a basic form of other complicated environments. Further research will focus on modeling more complex urban environments and facing multi viewpoints for optimal visibility computation in such environments, generalizing the presented building model for more complex ones.

VII. References

[1] Chrysanthou Y., "Shadow Computation for 3D Interactive and Animation," Ph.D. Dissertation, Department of Computer Science, College University of London, UK, 1996.

[2] Cohen-Or D., Fibich G., Halperin D., and Zadicario E., "Conservative Visibility and Strong Occlusion for Viewspace Partitioning of Densely Occluded Scenes," In EUROGRAPHICS'98, 1998.

[3] Cohen-Or D. and Shaked A., "Visibility and Dead- Zones in Digital Terrain Maps," Eurographics, vol. 14(3), pp. 171- 180, 1995.

[4] De Floriani L. and Magillo P., "Visibility Algorithms on Triangulated Terrain Models," International Journal of Geographic Information Systems, vol. 8(1), pp. 13-41, 1994.

[5] De Floriani L. and Magillo P., "Intervisibility on Terrains," In P.A. Longley, M.F. Goodchild, D.J. Maguire & D.W. Rhind (Eds.), Geographic Information Systems: Principles, Techniques, Management and Applications,1999, pp. 543-556. John Wiley & Sons.

[6] Doytsher Y. and Shmutter B., "Digital Elevation Model of Dead Ground," Symposium on Mapping and Geographic Information Systems (Commission IV of the International Society for Photogrammetry and Remote Sensing), Athens, Georgia, USA, 1994.

[7] Drettakis G. and Fiume E., "A Fast Shadow Algorithm for Area Light Sources Using Backprojection," In Computer Graphics (Proc. of SIGGRAPH '94), 1994, pages 223–230.

[8] Durand F., "3D Visibility: Analytical Study and Applications," PhD thesis, Universite Joseph Fourier, Grenoble, France, 1999.

[9] Elber G., Sayegh R., Barequet G. and Martin R., "Two-Dimensional Visibility Charts for Continuous Curves," Shape Modeling International 05, MIT, Boston, USA, 2005, pp. 206-215.

[10] Fisher-Gewirtzman D. and Wagner I.A., "Spatial Openness as a Practical Metric for Evaluating Built-up Environments," Environment and Planning B: Planning and Design vol. 30(1), pp. 37-49, 2003.

[11] Franklin W.R., "Siting Observers on Terrain," in D. Richardson and P. van Oosterom, eds, Advances in Spatial Data Handling: 10th International Symposium on Spatial Data Handling. Springer-Verlag, 2002, pp. 109–120

[12] Franklin W.R. and Ray C., " Higher isn't Necessarily Better: Visibility Algorithms and Experiments," In T. C. Waugh & R. G. Healey (Eds.), Advances in GIS Research: Sixth International Symposium on Spatial Data Handling, 1994, pp. 751–770. Taylor & Francis, Edinburgh.

[13] Franklin W.R. and Vogt C., "Multiple Observer Siting on Terrain with Intervisibility or Lores Data," in XXth Congress, International Society for Photogrammetry and Remote Sensing. Istanbul, 2004.

[14] Nadler B., Fibich G., Lev-Yehudi S. and Cohen-Or D.,"A Qualitative and Quantitative Visibility Analysis in Urban Scenes," Computers & Graphics, 1999, pp. 655-666.

[15] Nagy G., "Terrain Visibility," Technical report, Computational Geometry Lab, ECSE Dept., Rensselaer Polytechnic Institute, 1994

[16] Plantinga H. and Dyer R., "Visibility, Occlusion, and Aspect Graph," The Int. Journal of Computer Vision, vol. 5(2), pp.137-160, 1990.

[17] Ratti C, "The Lineage of Line: Space Syntax Parameters from the Analysis of Urban DEMs'," Environment and Planning B: Planning and Design, vol. 32, pp. 547-566, 2005.

[18] Schaufler G., Dorsey J., Decoret X. and Sillion F.X., "Conservative Volumetric Visibility with Occluder Fusion," In Computer Graphics, Proc. of SIGGRAPH 2000, pp. 229-238.

[19] Stewart J. and Ghali S., "Fast Computation of Shadow Boundaries Using Spatial Coherence and Backprojections," In Computer Graphics, Proc. of SIGGRAPH 1994, pp. 231-238.

[20] Teller S. J., "Computing the Antipenumbra of an Area Light Source," Computer Graphics, vol. 26(2), pp.139-148, 1992.

[21] Wang, J., G.J. Robinson, and K. White, "A Fast Solution to Local Viewshed Computation Using Grid-based Digital Elevation Models," Photogrammetric Engineering & Remote Sensing, vol. 62, pp.1157-1164, 1996.

[22] Wang, J., G.J. Robinson, and White K., "Generating Viewsheds without Using Sightlines," Photogrammetric Engineering & Remote Sensing, vol. 66, pp. 87-90, 2000.

[23] Yang, P.P.J., Putra, S.Y. and Li, W., "Viewsphere: a GIS-based 3D Visibility Analysis for Urban Design Evaluation," Environment and Planning B: Planning and Design, vol. 43, pp.971-992, 2007.

[24] Zlatanova S., Rahman A., and Wenzhong S., "Topology for 3D Spatial Objects," Int. Sym. and on Geoinformation, 2002.

Combining Territorial Data With Thermal Simulations to Improve Energy Management of Suburban Areas

An application to the Walloon region of Belgium

Anne-Françoise Marique and Sigrid Reiter
Local Environment: Management and Analysis
University of Liège
Liège, Belgium
afmarique@ulg.ac.be, Sigrid.Reiter@ulg.ac.be

Asma Hamdi
Energétique des Bâtiments et Systèmes Solaires
Ecole Nationale d'Ingénieurs de Tunis
Tunis, Tunisie
Asma.hamdi86@gmail.com

Maud Pétel
Ecole supérieure du Génie Urbain
Ecole des Ingénieurs de la Ville de Paris
Paris, France
Maud.Petel@eivp-paris.fr

***Abstract*—Urban sprawl has been identified as a major issue for sustainable development. Energy consumption in suburban buildings, in particular, is a widespread issue because detached types of houses require significantly more energy to be heated than more compact urban forms. Energy efficiency is often presented as a viable approach to the mitigation of climate change, but research and studies mainly contend with individual buildings and do not address this issue at larger territorial scales or for a whole building stock. In this respect, this paper first presents a morphological definition of urban sprawl. This definition uses territorial and cadastral data available for the Walloon region of Belgium. Using this definition, a suburban type classification adapted to thermal studies is drawn up. A representative block of each type is selected to model energy use and to determine the total energy consumption of the whole suburban building stock. An application is then presented concerning a comparison of potential energy savings associated with several renovation strategies. The results of this exercise are presented and highlight the benefits of combining Geographic Information Systems (GIS) tools, territorial data and thermal simulations for the efficient energy management of suburban areas at the scale of the whole building stock.**

Keywords-urban sprawl; territorial data; urban GIS; energy consumption.

I. Introduction

The expansion of urban areas, commonly referred to as urban sprawl, has been identified as a major issue for sustainable development [1]. Although opponents of sprawl argue that more compact urban forms would significantly reduce energy consumption both in the building and transportation sectors [2][3][4][5], low-density residential suburban districts are a reality in our territories and continue to grow. Such patterns of development are found in both developed and developing countries [6][7][8].

Much research has focused on urban sprawl and has in particular identified energy consumption in suburban houses as a major issue because detached houses consume significantly more energy than compact urban forms [1][9][10][11]. In the actual context of increasing environmental awareness, energy efficiency in buildings is in fact a topic widely studied in the literature and often presented as a viable approach to the mitigation of climate change. It has also become a central policy target in the European Union at both the national and local levels [12]. A well-known example of this dynamic is the adoption in 2002 and the progressive integration into local laws of the European Energy Performance of Buildings Directive (EPBD). This directive requires all European countries to strongly enhance their building regulations and aims primarily to establish minimum standards for the energy performance of new buildings and existing buildings larger than 1000 m² that are subject to major renovation [13].

Although this is a good step towards more sustainability in the building sector, two objections can be made to this directive and to much of the existing research and models. First of all, they adopt the perspective of the individual building as an autonomous entity and neglect the importance of phenomena linked to larger scales, although decisions made at the neighborhood and regional levels have important consequences for the performance of individual buildings and the transportation habits of the occupants [11][14][15]. Moreover, this approach is difficult to generalize to address the sustainability of a whole territory. Secondly, the EPDB Directive mainly applies to new buildings, whereas the existing building stock is huge, often poorly or non-insulated and takes a very long time to be renovated.

To effectively address the issue of energy efficiency in the whole suburban building stock and to help reach the climate change targets adopted in the scope of international agreements, it is essential to surpass this "single-new-building" approach. A large amount of territorial data is available, but it is rarely used where the sustainability of the territory, in particular the issue of energy consumption in buildings, is concerned. We propose to exploit these data, in combination with thermal simulations and energy performance indicators, to draw a regional cadastre of energy consumption in suburban areas and to test the impact of regional strategies applied to the whole suburban building stock. Note that the definition and the building type classification presented in this paper can also be adapted to urban and rural areas to cover the whole regional territory.

In Section 2, the paper presents a morphological definition of urban sprawl adapted to thermal studies. The suburban building stock referred to by this definition is then classified into representative categories of buildings and neighborhoods (Section 3). This typological approach has already been used in the literature and proved to be of interest [2][11][16][17]. Two applications are proposed in Section 4 to highlight the usefulness of this approach: a calculation of the energy consumption of the whole suburban area of the Walloon region and a comparison of three renovation strategies at a regional scale. Our main findings and the reproducibility of this approach are discussed in Section 5.

II. A Morphological Definition of Urban Sprawl

In this section, the existing classifications commonly used in Belgium are presented together with their limitations as far as morphological studies are concerned. Developed on this basis, our classification of urban types is presented and compared with existing ones and with an empirical survey.

A. The existing classifications

Based on qualitative and quantitative data, Van der Haegen and Van Hecke's urban type classification [18][19] (Figure 1) assigns the 262 Walloon municipalities (589 in Belgium) to four categories based on the level of facilities provided and on residents' locations with respect to work, shopping and services. This classification follows the same philosophy as the UK's OPCP and the ECOTEC's urban categories. The "operational agglomeration" is based on the morphological agglomeration. The "suburbs" are the first suburban area of a city. The density of population remains less than 500 inhabitants per square kilometer. Areas located further from the city, while keeping strong relationships with it (namely through home-to-work commutes), constitute the "alternating migrants area," whereas the remaining areas are grouped under the "other areas" term. This category thus comprises not only rural and suburban districts but also secondary cities and municipalities located outside the influence of the main agglomerations. Although urban sprawl is particularly linked to the "suburbs" and the "alternating migrants areas" [20], low-density suburban neighborhoods are found in the four urban categories because the boundaries of the urban types are adapted to administrative borders, as observed in Figure 2.

The Belgian urban settlements zoning (Figure 3) is a finer representation [21]. The size of an urban settlement (defined as groups of population living in neighborhood buildings) varies from a census block (i.e., a neighborhood in urban areas and a village in rural areas with more than 150 inhabitants) to an aggregation of several census blocks separated by less than 100 meters. Updated in 2001 [22], this classification more adequately fulfills our purpose as it embraces morphologically contiguous urbanized areas and crosses over municipalities' boundaries. However, rural and suburban areas are not differentiated.

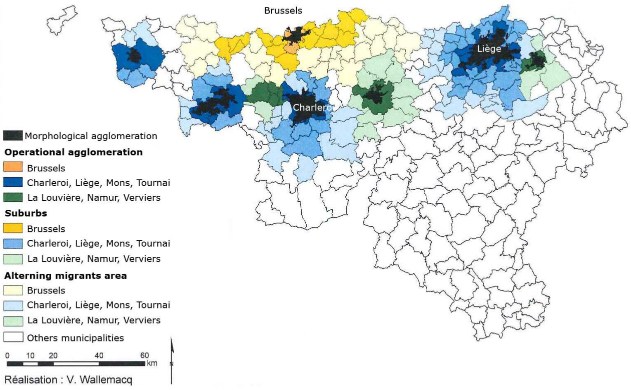

Brussels
Liège
Charleroi
Morphological agglomeration
Operational agglomeration
Brussels
Charleroi, Liège, Mons, Tournai
La Louvière, Namur, Verviers
Suburbs
Brussels
Charleroi, Liège, Mons, Tournai
La Louvière, Namur, Verviers
Alterning migrants area
Brussels
Charleroi, Liège, Mons, Tournai
La Louvière, Namur, Verviers
Others municipalities
0 10 20 40 60 km
Réalisation : V. Wallemacq

Figure 1. Van der Haegen and Van Hecke's urban type classification. "Operational agglomerations" are black; "suburbs" are dark blue, dark yellow and dark green; "alternating migrants areas" are light blue and light green and "other areas" are white.

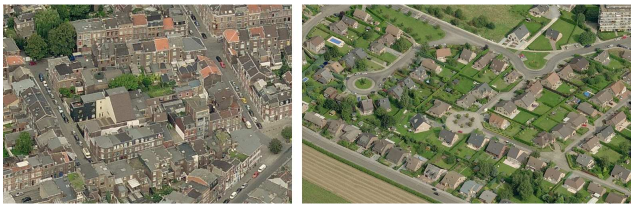


Figure 2. Very different types of districts are found inside each urban type highlighted in Van der Haegen and Van Hecke's classification. Example of a very dense district and a suburban district located in the "operational agglomeration" (municipality of Liège).

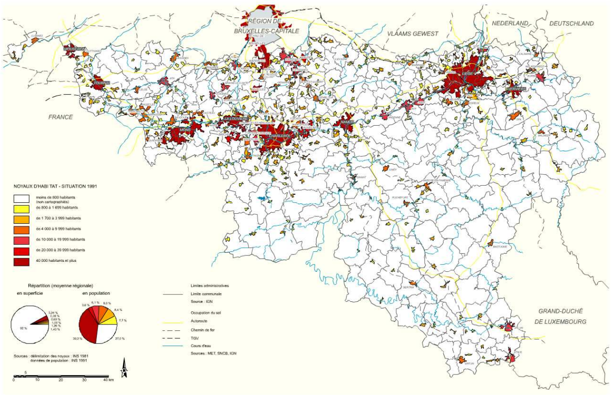

RÉGION DE BRUXELLES-CAPITALE
VLAAMS GEWEST
NEDERLAND
DEUTSCHLAND
FRANCE
GRAND-DUCHÉ DE LUXEMBOURG
NOYAUX D'HABITAT - SITUATION 1991
Répartition (moyenne régionale)
en superficie
en population

Figure 3. The Belgian urban settlements zoning (in red).

B. *Drawing a morphological urban type classification*

Although the two previous classifications are often used in national and regional research dealing with population and socio-economic issues, they do not seem adapted to morphological studies and do not allow the clear identification of suburban areas. To propose a better definition, an extensive review of the literature dedicated to urban sprawl was performed and three main characteristics of this phenomenon were highlighted: (1) low density, (2) mono-functionality and (3) discontinuity with traditional urban cores. The first parameter in particular is closely linked to the morphology of buildings.

To determine our morphological urban type classification, we used the following territorial data set at a disaggregated scale:

- Cartography (1/10,000) of the buildings and plots of the Walloon region drawn by the regional administration in charge of cartography;
- Cadastral database: buildings' date of construction, type of buildings (i.e., housing, commercial).

We calculated the density of dwellings (shops, schools and others buildings were eliminated) in each of the 9,730 census blocks of the Walloon region of Belgium. These census blocks were then classified according to the value of this index. The frequency distribution was divided into four parts, each containing a quarter of the population (Figure 4). The two central intervals (density in the range of 5 to 12 dwellings per hectare) are identified as the suburban territory, as 52% of the building stock of Belgium is composed of detached and semi-detached houses. Census blocks presenting a dwelling density higher than 12 dwellings per hectare are identified as "urban districts", and those with a density lower than 5 dwellings per hectare are considered to be "rural districts". Figures 5, 6 and 7 present the census blocks associated with each type. This approach crosses over municipalities' boundaries and distinguishes three urban types based on morphological criteria. Note that in the rest of the paper, we will only consider the suburban territory (Figure 6). However, the developments presented below are also easily applicable to urban and rural areas.

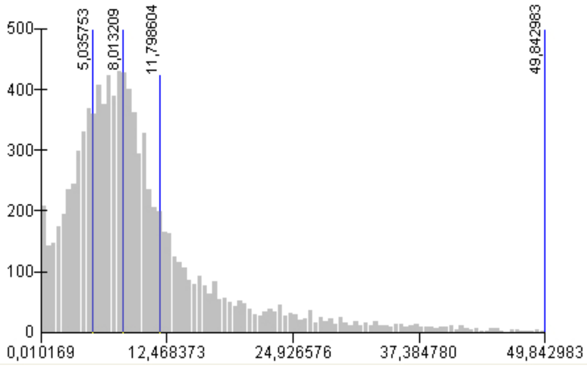


Figure 4. The frequency distribution of the density of the 9,730 census blocks. Quartiles are used to determine the urban area (density > 12 dwellings per hectare), the suburban area (5 dw/ha < density < 12 dw/ha) and the rural area (density < 5 dw/ha).

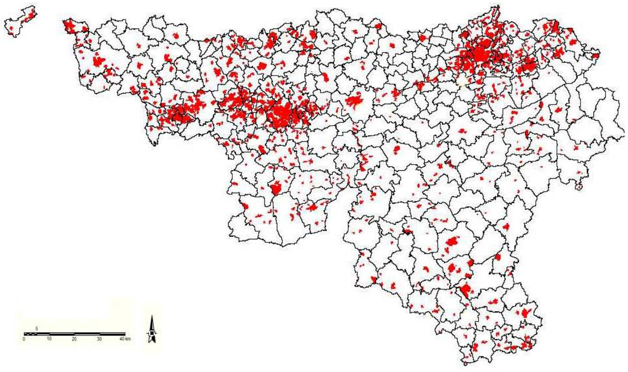


Figure 5. The urban area: districts presenting a built density higher than 12 dwellings per hectare.

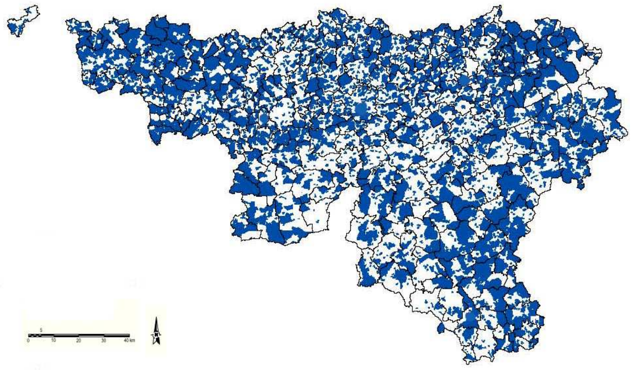


Figure 6. The suburban area: districts presenting a built density in the range of 5 to 12 dwellings per hectare.

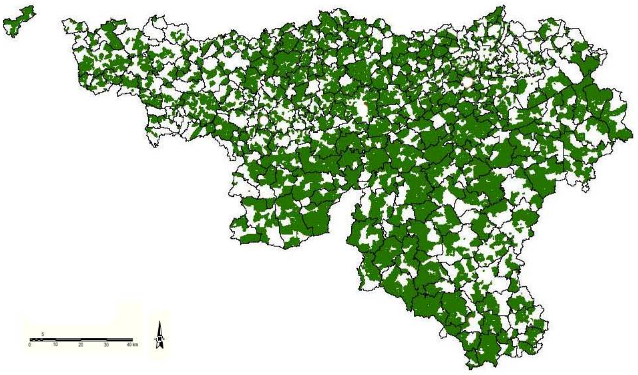


Figure 7. The rural area: districts presenting a built density lower than 5 dwellings per hectare (including unbuilt census blocks).

C. *Comparison of our classification with existing ones*

The main assumption that guided the definition of our classification is that suburban districts are spread out over the whole territory, as observed in Figure 6. This definition thus resolves the issue highlighted in Section 2 and overcomes the two disadvantages of the existing classifications as far as morphologic research is concerned. Note that our definition of the urban areas corresponds fairly well to the urban settlements zoning. In addition, we are able to differentiate suburban and rural areas.

The number of suburban districts in each municipality was then calculated and is presented in Figure 8 to highlight the parts of the territory where urban sprawl is common. This confirms that central municipalities (in particular the municipality of Namur in the center of the regional territory) may also contain a huge number of suburban low-density districts and that classification based on municipalities' boundaries is not adapted to research dealing with the morphology of the urban form. The north part of the Walloon region (the Walloon Brabant, in the influence area of the metropolitan area of Brussels) is particularly concerned with urban sprawl. To a lesser extent, suburban districts are also located in the southern, less densely populated part of the region, and particularly in the extreme south under the influence of the metropolitan area of Luxembourg.

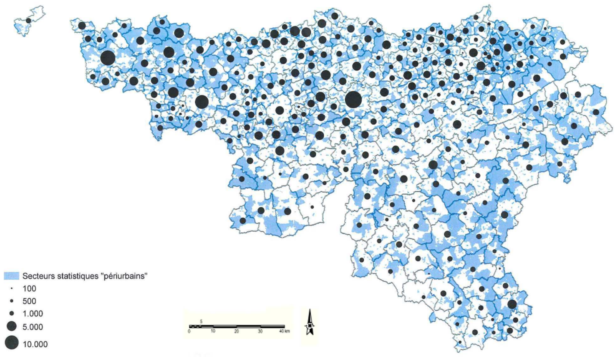


Figure 8. The number of suburban districts (built density in the range of 5 to 12 dwellings per hectare) per municipality.

D. *Comparison of our classification with a sample of inhabitants*

To test the relevance of the definition of urban sprawl in our urban type classification, a survey was conducted amongst a sample of 480 people working at the University of Liège [23]. These people were asked, amongst other questions dealing with urban sprawl and their quality of life, to give their address and to specify in which area they live (with a choice available between urban, suburban or rural districts) without any former indications. Their responses were encoded in a GIS and compared with our classification based on the built density. The results are quite good, as the majority of the answers given by the sample of people corresponded to our classification. Only 2.7% of respondents chose "suburban area" instead of "urban area", 1.9% answered "suburban area" instead of "rural area" and 6.0% "rural area" instead of "suburban area". This last result can be explained by the large dispersion of suburban districts within the territory; several people chose "rural area" because of a huge distance to the city center. Amongst the criteria given by the respondents to define urban sprawl and justify their choice, low-density, a large detached house with a garden located in a green environment, a quiet environment and the distance to city centers were most often cited.

III. The Typological Classification

Finally, we studied a random selection of 300 houses amongst the 702,000 suburban buildings identified by our definition (Figure 6). We extracted, for each one, the following data: the type of neighborhood in which the house is located, the dimensions, the number of levels, the type of house (detached, semi-detached and terraced) and the dimensions of the plot on which the house is built.

The neighborhood classification only takes into account the shape of the neighborhood, as our study deals with morphological criteria. The neighborhood classification assigns the 300 samples to four main categories (Figure 9). The linear district is composed of houses located on both sides of a road linking former villages or towns. This type represents 21% of the suburban building stock. The semi-detached district (8%) consists of detached and semi-detached houses. The "plot" district (20%) comes from the division by a private developer of a large site into individual plots and internal roads. The mixed district (30%) is more heterogeneous and is made up of individual houses together with older types of buildings (farms, old houses, etc.). A "composite" type is added to classify suburban blocks consisting of two different types of structures (9%). 12% of our sample does not correspond to any of these types. We considered these biases as acceptable in a study dealing with the whole suburban building stock of a region.

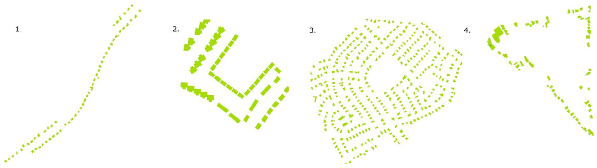


Figure 9. Illustrations of the four suburban blocks: 1. The linear block; 2. The semi-detached block; 3. The plot block; 4. The mixed block.

Analyses were then performed to identify the main characteristics of the suburban building stock. A large amount of the suburban building stock is composed of houses with a surface area in the range of 101 to 150 m². The mean surface area of the suburban building stock is 120 m² (standard deviation=40). The linear type is particularly made up of houses with a surface area in the range of 101 to 150 m². Large houses are also mainly found in plot districts, whereas semi-detached districts are mainly made up of buildings with a surface area less than 100 m². The mixed type is characterized by a wider distribution among the four classes, confirming the definition of this type, which is based on the diversity of the type and surfaces of buildings encountered. Buildings larger than 200 m^2 are exclusively found in the plot (very large houses on very large plots built by rich households) and mixed (old farms) types of districts. The size of the plots is not linked to the size of the houses as a R^2 of only 0.2 has been calculated.

The "before 1930" and "between 1961 and 1980" classes of ages are well represented, which highlights different phenomena. First of all, the existing suburban building stock is old, and the renovation of the building stock is particularly low. Secondly, urban sprawl was particularly popular from the "golden sixties" until the eighties. After that, the phenomenon was still present but slowed down. Buildings built before 1930 are mainly located in mixed districts (64.8%), whereas those built after 1960 are mainly found in plot and, to a lesser extent, linear districts. Semi-detached districts are represented in each class of age of construction, which tends to prove that these kinds of suburban forms (mainly social housing built by public developers) are developed in all time periods.

TABLE I. SURFACE AREA OF THE SUBURBAN BUILDING STOCK

Partition by age class			
50-100 m^2	***101-150 m^2***	***150-200 m^2***	***> 200 m^2***
31.5%	4.5%	15.0%	6.0%

TABLE II. AGE OF THE SUBURBAN BUILDING STOCK

Partition by age class				
Before 1930	***1931-1960***	***1961-1980***	***1981-1996***	***After 1996***
38.3%	14.3%	30.0%	10.4%	7.0%

We finally combined the surface area of the house, the age of construction and the type of district to highlight the most common combinations:

- 101-150 m^2 houses built before 1930 in a mixed district (11.9% of the suburban building stock).
- 101-150 m^2 houses built between 1961 and 1980 in a linear district (6.9%).
- 50-100 m^2 houses built before 1930 in a mixed district (5.1%).
- 101-150 m^2 houses built between 1961 and 1980 in a plot district (5.0%).
- 101-150 m^2 houses built between 1961 and 1980 in a plot district (5.0%).

IV. APPLICATION: ENERGY MODELING

In this section, the energy modeling of existing building stock based on the previous classification of urban types is first presented. The suburban type classification is then used to compare three renovation strategies at a regional scale.

A. *Energy modeling of the existing building stock*

A representative block and a representative building of each type highlighted in the classifications were selected to model the energy use of the whole suburban territory of the Walloon region of Belgium. Based on the evolution of regional policies concerning building energy performance and the evolution of construction techniques, the five age categories (pre-1930, 1931-1960, 1961-1980, 1981-1996 and 1996-2008) were used to approximate a mean thermal conductivity of external façades from a "standard" composition of façades and from glazing attributes for the building envelope in each category. Detailed values (glazing and wall heat transfer coefficients and composition) are available in [11]. Dynamic thermal simulation software was then used to model each type of building and to calculate their energy requirements for heating (in kWh/m^2.year).

The annual energy requirement for heating the whole suburban building stock was calculated according to the partition of each type of building in the whole suburban area of the Walloon region. The total annual energy requirement is equal to 19,914 GWh. The mean consumption for heating is equal to 232.8 kWh/m^2.year.

A clear difference is observed between the heating energy requirements of houses and neighborhoods built before and after the first thermal regulations adopted in the Walloon region. Houses built after the first regulations consume 130 kWh/m^2 or less annually, whereas those built before 1980, especially dispersed houses, consume from 235 to 401 kWh/m^2 annually. For semi-detached and terraced houses, the annual energy consumption falls between 84 and 319 kWh/m^2 depending on the age of the building. For buildings of the same age, semi-detached and terraced houses consume 14.6% to 23.6% less energy than detached houses, highlighting the effect of connectivity on the energy performance of buildings.

B. *The impact of renovation strategies*

Figure 10 presents the potential energy savings associated with three renovation strategies dealing with energy efficiency in comparison with the existing situation (EX as calculated in the previous section). We determined the potential energy savings associated with a light upgrade (insulation in the roof and new high-performance glazing) of 50% of the pre-1981 building stock (SC 1). This policy could reduce energy consumption of the whole suburban building stock by 10.8%. Adopting a more ambitious policy (improvement of the insulation of the whole building envelope) targeted at the 1961-1980 building stock (SC 2) is more interesting as far as energy savings are concerned.

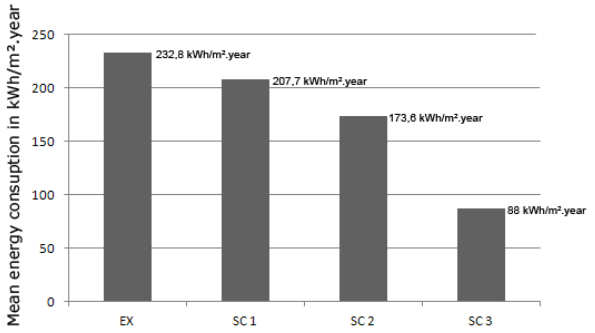


Figure 10. Potential energy savings (in kWh/m^2.year) related to three renovation strategies dealing with the improvement of energy efficiency in the existing suburban building stock.

In fact, as cavity walls became widely used after 1960, these houses are particularly well adapted to new insulation techniques used to retrofit existing walls (insulation is blown into the cavity), and insulating the roofs, the slabs and replacing glazing is fairly easy to achieve. This approach could reduce energy consumption of the whole suburban building stock by 26.5%. In the last scenario (SC 3), we assumed that the whole suburban building stock was retrofitted to fit actual standards for new buildings. The resulting energy savings are huge (62.6%). However, retrofitting the whole building stock is a very difficult target to reach even if households could significantly reduce their energy consumption.

V. Conclusions and Future Work

Based on cadastral and territorial data, we developed a morphological definition of the Walloon suburban territory and a typological classification of suburban districts and residential buildings. The proposed definition allowed us to question the relevance of the existing classification of urban types and to prove the validity of a finer representation, especially as far as morphological studies are concerned. Thermal simulations were thus performed on a representative block / building of each type highlighted in the classification to estimate the energy consumption of the whole suburban building stock at a regional scale. The classification was then used to assess the impact of two renovation strategies and to compare their value as far as energy savings are concerned. The paper has thus highlighted the benefits of using GIS tools for territorial management and thermal topics. The combination of thermal simulations with territorial data, in particular, is relevant and useful to effectively address the issues related to energy efficiency in the building sector at a territorial scale. The same exercise will now be used for the urban and rural parts of the territory to address the whole regional building stock and to highlight the urban blocks that are a priority for retrofitting or for increased density. Finally, the method is sufficiently general, and the data used in Belgium are widely available in other regions, which makes the definition and the suburban type classification easily reproducible in other territories.

Acknowledgment

This research is funded by the Walloon region of Belgium (DG04, Department of Energy and Sustainable Building) in the framework of the “Suburban Areas Favoring Energy efficiency” project (SAFE).

The research team thanks Charlotte Pierson, student at the University of Liège, and Véronique Wallemacq, researcher, for their contributions to this paper.

References

[1] EEA, “Urban sprawl in Europe. The ignored challenge,” Report EEA l 0/2006, European Environment Agency, 2006.

[2] M. Maïzia, C. Sèze, S. Berge, J. Teller, S. Reiter, and R. Ménard, “Energy requirements of characteristic urban blocks,” Proc. CISBAT 2009 Int. Scientific Conf. on Renewables in a Changing Climate: From Nano to Urban scale, 2009, pp. 439-44.

[3] P. Newman and J.R. Kenworthy, “Cities and Automobile Dependence: A sourcebook,” Aldershot: Gower Publishing Co, 1989.

[4] P. Newman and J.R. Kenworthy, “Sustainability and Cities: overcoming automobile dependence,” Washington DC: Island Press, 1999.

[5] K. Steemers, “Energy and the city: density, buildings and transport,”. Energy and Buildings, vol. 35(1), 2003, pp. 3-14.

[6] K. S. Nesamani, “Estimation of automobile emissions and control strategies in India," Science of the Total Environment, vol. 408, 2010, pp. 1800-11.

[7] A. N. R. da Silva, G.C.F. Costa, and N.C.M. Brondino, “Urban sprawl and energy use for transportation in the largest Brazilian cities,” Energy for Sustainable Development, vol.11(3), 2007, pp. 44-50.

[8] W. Yaping and Z. Min, “Urban spill over vs. local urban sprawl: Entangling land-use regulations in the urban growth of China’s megacities,” Land Use Policy, vol. 26, 2009, pp. 1031-45.

[9] A. F. Marique and S. Reiter, “A method to assess global energy requirements of suburban areas at the neighborhood scale,” Proc. 7th International Conference on Indoor Air Quality, Ventilation and Energy Conservation in buildings (IAQVEC 2010), 2010.

[10] A. F. Marique and S. Reiter, “A Method to Evaluate the Energy Consumption of Suburban Neighbourhoods,” HVAC&R Research, in press.

[11] G. Verbeek and H. Hens, “Energy savings in retrofitted dwellings: economically viable?,” Energy and Buildings, vol. 37, 2005, pp. 747-754.

[12] CEC, “Green Paper on Energy Efficiency or Doing More With Less,” Report CEC COM(2005) 265, Commission of the European Communities, Belgium, 2005.

[13] EP, “Directive 2001/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings,” Official Journal of the European Community, European Parliament and the Council, 2002.

[14] E. Popovici and B. Peuportier, “Using life cycle assessment as decision support in the design of settlements,” Proc. 21th Conference on Passive and Low Energy Architecture (PLEA 2004), 2004, pp. 1-6.

[15] C. Ratti, N. Bakker, and K. Steemers. “Energy consumption and urban texture,” Energy and Buildings, vol.37(7), 2005, pp. 762-76.

[16] P. J. Jones, S. Lannon, and J. Williams, “Modeling building energy use at urban scale,” Proc. of the 7th International IBSPA Conference, 2001, pp. 175-80.

[17] E. Popovici, “Contribution to the Life Cycle Assessment of settlements,”. Ph.D. Thesis, Ecole des Mines de Paris, 2006.

[18] H. Van der Haegen, E. Van Hecke, and G. Juchtmans, “Les regions urbaines belges en 1991,” Etudes statistiques de l’INS, vol. 104, 1996.

[19] J. A. Sporck, H. Vand der Haegen, and M. Pattyns, “L’organisation spatiale de l’espace urbain,” La cité belge d’aujourd’hui, quel devenir?, 1985.

[20] L. Brück, “La périurbanisation en Belgique,” SEGEFA, 2002.

[21] H. Van der Haegen, M. Pattyn, and S. Rousseau, “Dispersion et relations de niveau élémentaire des noyaux d’habitat en Belgique: Situation en 1980,” Bulletin de Statistique, vol.67, 1991.

[22] E. Van Hecke, J. M. Halleux, J. M. Decroly, and B. Merenne-Schoumaker, “Noyaux d’habitat et régions urbaines dans une Belgique urbanisée,” Monographie n°9. SPF Economie, P.M.E., Classes moyennes et Energie, 2009.

[23] C. Pierson, “Approche sociologique de l’habitat périurbain,” Master thesis, University of Liège, unpublished.

Vehicle Position Determination — Using Markers and Speed Reports

Bruce Beyeler and David C. Pheanis
Computer Science and Engineering
Ira A. Fulton School of Engineering
Arizona State University
Tempe, Arizona 85287-8809 USA
Bruce.Beyeler@ASU.edu *or* David.Pheanis@ASU.edu

***Abstract*—The use of cell phones as data-collection devices for obtaining automotive traffic-flow information provides the potential for instrumenting large numbers of vehicles at a minimal cost. Effectively incorporating cell phones as sensors in traffic-flow collection systems requires a clear understanding of the accuracy of the data produced by each cell phone. Previous experiments and field trials have typically measured the accuracy of cell-phone data at large — comparing all of the collected cell-phone reports across a given segment of the road against data obtained with traditional techniques such as loop detectors. The approach that we take in this research differs by comparing each individual cell-phone report with the known position of the vehicle at the time of the report. This paper describes the technique we used for accurately determining the actual speed and position of a vehicle at any given point in time during a test trip by using published map data, speed reports from the vehicle itself, position reports from a hand-held GPS unit with an external antenna, and operator inputs.**

***Index Terms*—GPS, traffic flow, cell phones, data collection.**

I. Introduction

Rapid advances in cell-phone technology coupled with the penetration of cell phones into the general population provide an opportunity to utilize cell phones as mobile probes that provide real-time traffic-flow information. Multiple studies and experiments show the feasibility of using cell phones to collect both basic traffic-flow information such as vehicle position and speed [1][2] and additional types of information including road conditions, bumps, braking conditions, and the presence of honking [3].

Incorporating cell phones into traffic-flow data-collection techniques involves some challenges that we must address, most notably, determining the accuracy of the reports that cell phones generate [4][5][1]. We can determine the accuracy of data reports by comparing the reports with the known ground truth. In most cases, the ground truth for individual vehicles is not easy to determine. Most field tests utilize passive collection techniques, such as inductive loops, to provide the ground truth for comparison purposes.

Rather than compare collected data in the traditional sense, this research focuses on individual cell-phone reports and comparing each report with the ground truth for that cell phone at the time of the report generation. In order to carry out this comparison, we must determine the ground truth for a given vehicle at any point in time. With our approach there is no need for physical roadway support such as loop detectors, and our approach works over the entire length of any desired roadway, not just where sensors are available. This paper describes the technique of combining various inputs in order to determine the vehicle's precise location and speed at any point in time.

The next section of this paper discusses related work, and Section III highlights some of the challenges that we face while also providing an overview of our approach. In subsequent sections we explain how we use map data, vehicle speed data, GPS data, and operator inputs. Next we detail the calculations that we use for determining the actual position of a test vehicle at any given time, and then we itemize the steps for finding the vehicle position. Finally, we summarize the results of our work and mention what our next steps will be going forward.

II. Related Work

The growing interest in using cell phones in probe vehicles to detect and report traffic-flow conditions has generated a number of research projects and experiments. Some of these experiments exploit the characteristics of the cellular network to calculate a vehicle's location and speed [6][7] while other experiments utilize the GPS capabilities available in most cell phones today [8][1].

While assessing the accuracy of the cell phone data, most of these experiments compare the data collected by cell phones with that collected by traditional means — most often inductive loops embedded in roadways. Loop sensors or other types of point sensors count the number of vehicles that pass by the sensor and provide data for calculating the average speeds and time of traversal for the segments of the road between adjacent sensors [9]. In order to compare the accuracy of cell phones against this traditional data, some of these experiments average cell-phone reports over the same segments of the road and compare them against the data that the inductive loops provide. Other tests include additional data such as the speed of the vehicle obtained directly from the vehicle [10]. The resulting comparisons provide insight into the overall accuracy of the cell-phone reports but do not characterize the accuracy of individual cell phones. Furthermore, this technique provides no insight into the particular factors that directly affect a cell phone's accuracy.

Our overall research focuses on optimizing data-collection strategies by incorporating knowledge about the accuracy of individual cell phones. Minh and Kamioka suggest a similar technique in their research with their Pinpoint approach [11]. Their concept of sending the "right" data at the "right" time aligns closely with our research approach. A key distinction of our research involves incorporating the reported accuracy of individual cell phones into the reporting algorithms.

In order to characterize the accuracy of individual cell phones under various conditions, we take a different approach from the works cited above. Our field tests constrain the probe vehicles to a known path and the phones to various positions in the vehicle. In addition, we incorporate outside information including map data, speed information from the vehicle itself, and operator input. This paper describes our techniques for determining the ground truth, exactly where the probe vehicle is at any given point in time during the experiment, for each cell-phone report — something typically reserved for simulations. Establishing the ground-truth data for each individual report allows us to analyze the accuracy of each individual cell phone under a variety of conditions.

III. Challenge and Approach

Measuring the accuracy of a position report from a static (i.e., not moving) cell phone is easy. Simply put the cell phone at a known location, and compare the position report from the cell phone to the actual location of the cell phone. However, to use cell phones for collecting vehicular traffic-flow data, we must get reports from cell phones that are inside moving vehicles. Multiple factors may affect the accuracy of a cell phone's speed-and-position reports when the cell phone is in a moving vehicle. For example, the cell phone may not have a consistently clear view of the sky because of terrain, buildings, or the structure of the vehicle itself. Also, the cell phone may not be able to produce accurate reports due to the simple fact that the cell phone is moving, especially when the vehicle is changing direction.

Before we can assess the accuracy of speed-and-position reports from a cell phone in a moving vehicle, we must somehow determine the *actual* speed and position of the vehicle. Our goal in this paper is to determine the actual speed and location of a test vehicle at any given time in a test trip so we can measure the accuracy of a cell-phone report whenever the cell phone makes a report. We can determine the actual speed directly from the vehicle itself by using the vehicle's diagnostic interface. However, determining our exact position on the road is a bit more of a challenge. By combining static data published by the Arizona Department of Transportation (ADOT) with dynamic data that we collect during the test run, we can determine the vehicle's actual location at any point in time.

Our technique utilizes map data published by the Arizona Department of Transportation (ADOT), speed data that we collect from the vehicle, GPS data that we collect from a hand-held GPS unit with an external antenna, and operator input that we collect via a laptop PC in the vehicle under test. Our data-collection procedures [12], [13], [14] dictate that we follow the same center-line path that the ADOT data-collection vehicles used, so we have a known path. In addition, the test vehicle will be passing precisely located milepost markers, and the operator will use the laptop PC to flag the milepost markers in the data log, thus giving us a set of time-correlated positions. We can determine the appropriate milepost marker corresponding to each item in the operator log by using the GPS reports at or near the time of the operator's mark.

IV. Map Information

ADOT provides two relevant sets of map information — data for the centerline track and data for the milepost markers. The data values for the centerline track provide points in the middle of our lane along the path that we will follow, mirroring the ADOT collection vehicles as part of our collection strategy. The data values for the milepost markers identify the locations (on the centerline track) of the mile markers along the path. By marking the time when we pass each milepost, we will be creating a set of known positions at specific points in time.

Note that the data values for a milepost marker conveniently specify the geographical coordinates of the centerline point corresponding to the mile marker, not the physical location of the milepost marker itself. We are interested in the position of our vehicle as it passes the milepost marker, so the data values for the milepost markers in the ADOT data are exactly what we need. Figure 1 shows how the centerline points and the milepost points relate to each other.

The centerline points are not regularly spaced, but vary according to the curvature of the road. Straight segments of the road require relatively few data points while curved segments of the road require multiple points. Milepost markers, on the other hand, define regular intervals of one mile and are evenly spaced one mile from each other.

V. Vehicle Speed Data and GPS Data

We can retrieve the vehicle's speed directly from the vehicle. By tapping into the vehicle's diagnostic port and requesting speed data, we can collect the actual speed of the vehicle throughout the test trip. All vehicles supporting the onboard diagnostics (ODB-II) interface provide a basic set of powertrain parameters — speed being one of those parameters. The ODB-II protocol allows us to query the data once per second, which will be frequent enough for our use since our test vehicle will not change speed dramatically from one second to the next.

In order to request the speed information from the vehicle, we are using a vehicle interface adapter from Multiplex Engineering. This adapter converts the laptop interface (RS-232 and command packets) to SAE-J1708 diagnostic requests (ISO-9141 signals and SAE-J1708 packet data). The data-collection program in the laptop PC contains a thread to do the following:

- Establish communications with the test vehicle.
- Periodically (once per second) send a request for the speed of the vehicle.
- Read the response from the vehicle.
- Convert the response to vehicle-report format [14].
- Send the report to the logging system.

The logging software timestamps each vehicle speed report. These reports give us a set of speed reports of the form (v, t) where v represents the vehicle's speed at the time t.

Figure 1 illustrates the vehicle data collections in relation to the map information. Notice that we collect the vehicle data at regular intervals with a period of approximately one second.

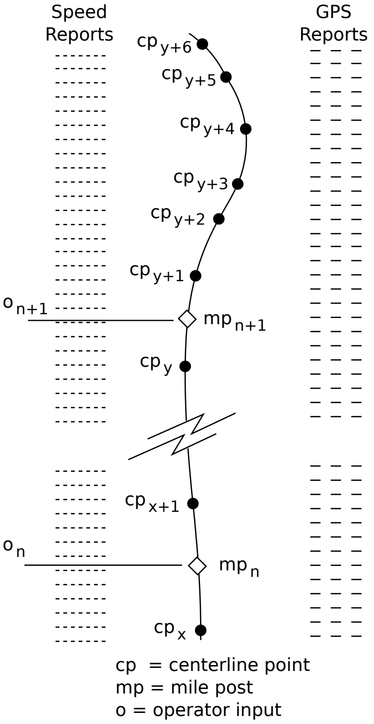


Fig. 1. Map Info with Collected Data

We collect GPS data from a hand-held GPS unit with an external roof-mounted antenna in a similar fashion. The GPS unit reports its data with a period of one second. The data-collection system receives these reports and logs them. Even though the GPS unit reports data with the same frequency as the vehicle speed data, the reports are not synchronized in any manner. As the next section explains, the GPS reports provide the location of the vehicle with more than enough accuracy to determine exactly which milepost marker the vehicle has just passed when the operator makes an entry to mark the passing of a milepost. We thereby fix our vehicle position at a specific location and at a specific point in time. Figure 1 shows how the GPS reports relate to the map information and the speed reports.

VI. Operator Input

Operator inputs provide the key component to determining the exact vehicle position at any point during the test run. By creating a marker in the logs whenever the vehicle passes a milepost, the operator establishes a set of known positions, at each milepost marker, at specific times during the test run [14]. The software records the operator input in the logs as a simple landmark notation that includes the timestamp but does not include any positional data.

We collect GPS reports from a hand-held GPS unit at one-second intervals throughout the test trip. We use the GPS reports only to determine the closest milepost at each operator's mark. Once we have determined which milepost the operator was marking, we have established a known position at a specific moment in time. Since the mileposts are one mile apart, we need only enough accuracy from the GPS reports to pick between two positions that are one mile apart. This modest accuracy requirement is well within the range of the hand-held GPS with an external antenna. Figure 1 illustrates the operator inputs at each milepost marker.

VII. Distance Calculations

After we have collected the data, we have all of the the information necessary to determine the vehicle position at any point in time. We have a fixed path, several known milepost positions at specific points in time, and known vehicle speed at numerous points between our known milepost positions. The approach that we take is similar to the one that people use in computer animation for moving an object along a curved path [15]. We calculate the arc length for each segment, sum those distances to obtain the total distance traveled between known positions, determine the scaling factors required to project the segments onto the known path, and, finally, use the adjusted formulas to determine the absolute position of the vehicle for any given time t.

A. Calculating Segment Distances

Figure 2 illustrates the speed-time relation, where each point represents one speed report — a specific speed at a given point in time. The shaded areas below the speed line represent the distance covered between two speed reports.

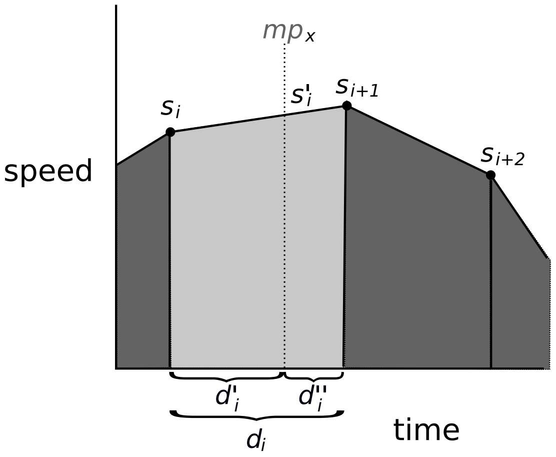


Fig. 2. Speed-Time Relation

Each speed report s_i has two components and occurs in the data as the pair (v, t) where v represents the vehicle's velocity at time t. Individually, we will refer to these components as v_i and t_i. The distance covered between speed reports s_i and s_{i+1} corresponds to the shaded area below those two reports — d_i. When we need to determine a speed at a time between t_i and t_{i+1} we will generate an interpolated speed report s'_i. We denote the distance between s_i and s'_i as d'_i, and we denote the distance between s'_i and s_{i+1} as d''_i.

We can calculate the distance covered between s_i and s_{i+1} as:

$$d_i = \int_{t_i}^{t_{i+1}} f(t)dt \tag{1}$$

where $f(t)$ represents the velocity of the vehicle over time.

However, the only data values that we have are the known speeds at t_i and t_{i+1}. Given that the speed reports occur at a rate of approximately one per second and based on the physics of vehicles in motion, we can safely assume a constant acceleration/deceleration between t_i and t_{i+1}. We can therefore simplify Equation (1) to the following:

$$d_i = v_{avg}(t_{i+1} - t_i) \tag{2}$$

which, when given that

$$v_{avg} = \frac{1}{2}(v_i + v_{i+1}) \tag{3}$$

yields:

$$d_i = \frac{1}{2}(v_i + v_{i+1})(t_{i+1} - t_i) \tag{4}$$

If the milepost reports were synchronized with the speed reports, then we could calculate the total distance between two mileposts by summing the distances covered by each speed report. For example, if s_i occurred at mp_x and s_{i+n} occurred at mp_{x+1}, then we could calculate the total distance, D_x, between the two mileposts as:

$$D_x = \sum_{j=i}^{i+n-1} d_j \tag{5}$$

Since the milepost reports typically fall between two speed reports rather than being synchronized with a speed report, we can generate an interpolated speed report to coincide with the time of the milepost report (t'). If s_i is the latest speed report before the milepost report, then we can calculate the velocity for the generated speed report s'_i at t' as follows:

$$tan(\Theta) = \frac{v_{i+1} - v_i}{t_{i+1} - t_i} \tag{6}$$

$$v' = v_i + (tan(\Theta)) * (t' - t_i) \tag{7}$$

$$v' = v_i + \left(\frac{v_{i+1} - v_i}{t_{i+1} - t_i}\right)(t' - t_i) \tag{8}$$

The generated speed report s'_i would be (t', v'). We can generate a similar speed report s'_{i+n} for the ending milepost by using the speed report s_{i+n} that was the latest speed report before the mp_{x+1} report for the next milepost marker. For each speed report that falls just before a milepost report, we can separate the distance into two areas, d'_i and d''_i as shown in Figure 2. Note that the milepost report can be anywhere between the two speed reports. With these definitions, we can rewrite Equation (5) for calculating the distance between the two mileposts mp_x and mp_{x+1} as:

$$D_x = d''_i + \sum_{j=i+1}^{i+n-1} d_j + d'_{i+n} \tag{9}$$

B. Aligning Calculations with Map Data

We can calculate the total distance between operator inputs (two known positions at marked times) by using Equation (9) of Section VII-A. This calculated value will most likely be different from the actual value (from the ADOT-provided data). A variety of factors can account for this difference: inexactness of the operator's marks, speedometer error, inaccuracy of the ADOT-provided data, etc.

Based on Equation (9), each d_i accounts for a given portion of the entire length of D_x. Specifically, each d_i represents d_i/D_x of the total distance D_x. If k is the scale factor such that:

$$kD_x = L_x \tag{10}$$

where L_x is the ADOT-provided distance over the same two milepost markers, then

$$k = \frac{L_x}{D_x} \tag{11}$$

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-178-6

and we can rewrite Equation (9) to reflect the distances scaled to yield L_x as the result:

$$L_x = kd''_i + \sum_{j=i+1}^{i+n-1} kd_j + kd'_{i+n} \qquad (12)$$

VIII. Determining Vehicle Position

We can now determine the position of the test vehicle at any given time t (corresponding to a report from a cell phone whose accuracy we are trying to measure) by performing the following steps:

- Determine the milepost reports that surround time t (M_x, M_{x+1}).
- Retrieve all speed reports between and encompassing M_x and M_{x+1}.
- Calculate D_x from Equation (9).
- Calculate L_x from ADOT-provided data, and determine k from Equation (11).
- Sum the adjusted distances between M_x and the speed report just before time t.
- Add the partial distance covered in the segment containing t by generating an artificial speed report as in Equation (8).
- Travel the known path from M_x toward M_{x+1} for the calculated covered distance.

IX. Conclusion

We now have a method for determining the actual position of a test vehicle at any given point in time, so we can process the cell-phone GPS reports and calculate the error value for each report. In a batch-processing mode following the conclusion of a test run, we can optimize the steps listed above by storing the calculated distances for each individual segment as well as the scaling values and distances between milepost reports.

The procedures we have described in this paper provide the foundation for evaluating the accuracy of the reports from a cell phone in a moving vehicle. Our inputs include ADOT data to specify a known path of travel, operator inputs to place the vehicle at known positions at specific times, and the speed data that the vehicle itself reports.

X. Future Work

With the ability to determine the accurate position of a test vehicle along a traveled path at any given point in time, we plan to take cell-phone GPS reports across a number of test runs and begin creating accuracy models for different types of phones under various conditions. These models will help determine optimal methods for collecting data and processing that data both for real-time traffic-flow information and for data-mining applications.

References

[1] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568–583, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X09001430

[2] Hillel and Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380–391, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X07000393

[3] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*, ser. SenSys '08. New York, NY, USA: ACM, 2008, pp. 323–336. [Online]. Available: http://doi.acm.org/10.1145/1460412.1460444

[4] D. Valerio, A. D'Alconzo, F. Ricciato, and W. Wiedermann, "Exploiting cellular networks for road traffic estimation: A survey and a research roadmap," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, April 2009, pp. 1–5.

[5] Y. Yuan and W. Guan, "Using cellular network for inner-suburban freeway traffic monitoring," in *Logistics Engineering and Intelligent Transportation Systems (LEITS), 2010 International Conference on*, November 2010, pp. 1–4.

[6] B. L. Smith and S. T. Laboratory, *Cellphone probes as an ATMS tool.* Charlottesville, VA: The Laboratory, 2003.

[7] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 141–151, March 2011.

[8] S. Amin, S. Andrews, S. Apte, J. Arnold, J. Ban, and M. B. et. al., "Mobile century using gps mobile phones as traffic sensors: A field experiment," in *The 15th World Congress on Intelligent Transportation Systems*, November 2008. [Online]. Available: http://www.ce.berkeley.edu/~bayen/conferences/its08.pdf

[9] L. A. Klein, M. K. Mills, D. R. P. Gibson, U. States, and T.-F. H. R. Center, *Traffic detector handbook*, 3rd ed. McLean, VA: U.S. Dept. of Transportation, Federal Highway Administration, Research, Development, and Technology, Turner-Fairbank Highway Research Center, 2006, vol. 1. [Online]. Available: http://purl.access.gpo.gov/GPO/LPS91983

[10] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson, "Virtual trip lines for distributed privacy-preserving traffic monitoring," in *Proceeding of the 6th international conference on Mobile systems, applications, and services*, ser. MobiSys '08. New York, NY, USA: ACM, 2008, pp. 15–28. [Online]. Available: http://doi.acm.org/10.1145/1378600.1378604

[11] Q. T. Minh and E. Kamioka, "Pinpoint: An efficient approach to traffic state estimation system using mobile probes," in *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, September 2010, pp. 1–5.

[12] B. Beyeler and D. C. Pheanis, "Using gps-enabled cell phones for traffic-flow data collection," in *Computers and Their Applications*, G. Hu, Ed. ISCA, 2005, pp. 253–258.

[13] ——, "Measuring cell-phone gps accuracy," in *Computers and Their Applications*, D. J. Jackson, Ed. ISCA, 2006, pp. 401–406.

[14] ——, "Cell-phone accuracy tests: Data-collection techniques," in *Computers and Their Applications*, B. Gupta, Ed. ISCA, 2007, pp. 271–276.

[15] B. Guenter and R. Parent, "Computing the arc length of parametric curves," *Computer Graphics and Applications, IEEE*, vol. 10, no. 3, pp. 72–78, May 1990.

A GIS-Based Approach for Representation of Air Pollution
A Case Study: Tabriz City

Davood Parvinnezhad Hokmabadi
Marand Faculty of Engineering
University of Tabriz
Tabriz, Iran
dparvinnezhad@tabrizu.ac.ir

Arash Rahmanizadeh
Marand Faculty of Engineering
University of Tabriz
Tabriz, Iran
arahmanizadeh@tabrizu.ac.ir

Shahla Taghizadeh Sufiani
Faculty of Environment and Energy
Islamic Azad University, Science and Research Branch
Tehran, Iran
sh.taghizade@gmail.com

***Abstract*— Pollution from urban transport has a big impact on community health. Environmental organizations in different countries of the world have installed sensor devices of air pollutants in different parts of their cities for informing the air pollution and its necessary and timely warnings and these devices record pollutants data in 24 hours of everyday. Today, these data are presented on large display boards installed in the important places of big cities. In this research, the results of pollutant data are presented visually on Tabriz map by GIS that it is better than statistical presentation.**

Keywords- Air Pollutants; Air Particles; Air Quality Index; GIS.

I. INTRODUCTION

Transportation is one of the vital components in modern human daily life. However, it has both productive effects on human developments and detrimental effects on public health. The number of motor vehicles is estimated to be over 800 million worldwide and is increasing almost everywhere at higher rates than human population, and road traffic may be growing even more rapidly. The number of private cars worldwide rose to 500 million in 1990 from 50 million in 1950. Road traffic is related to undesirable health effects caused by air pollution, noise and accidents. This wide range of negative health effects includes increased mortality, cardiovascular, respiratory and stress-related diseases, cancer and physically injuries. The negative effects are felt not only by transport users but also by the whole population especially in the vulnerable group of children and elderly people, pedestrians and cyclists. The effect of air pollution on public health depends on factors such as: the chemical composition of a particular pollutant, the level of concentration; the presence of other pollutants; the existing health of individuals; and periods of exposure [3].

Environmental organizations in different countries of the world installed sensor devices of air pollutants in different parts of their cities for informing the air pollution and its necessary and timely warnings and these devices record pollutants data in 24 hours of day. In our country, Iran, the Environmental Organization has installed 13 pollutant stations in Tehran and a few stations in other big cities. In Tabriz city 5 stations have been installed and we used all of them. More stations have best results and we can do best and complete analyses by GIS and then authorities can take decisions more effectively. Unfortunately due to the low monitoring stations in Tabriz, it is not possible we can perform GIS famous analyses such as Geostatistical analysis. Therefore, another GIS tools are used for visual presentation the results instead of statistical presentation. In this paper, a program has been developed in ARCObjects^{TM} [2] that it first calculates AQI values for whole of pollutant data and selects the maximum of them, then merges these values properly to attribute information of pollutant monitoring stations. Finally, based on these attribute information, the pollution status on the relevant maps or satellite images is represented as daily, monthly and yearly with various tools.

II. AIR POLLUTANTS AND THEIR ROLE IN HUMAN HEALTH

The main air pollutants are ozone, particles, CO, Nitrogen oxides, Sulfur dioxide and lead. In Table I, one can see these pollutants and their effects on society health [4].

TABLE I. THE MAIN AIR POLLUTANTS AND THEIR EFFECTS [4]

Pollutant	Impact
Ozone	Burning nose and watering eyes; Tightening of the chest Coughing, wheezing and throat irritation; Rapid, shallow, painful breathing; Susceptibility to respiratory infections; Inflammation and damage to the lining of the lungs; Aggravation of asthma; Fatigue; Cancer

ISBN: 978-1-61208-178-6

Particles	Stuffy noses, sinusitis; Sore throats; Wet cough, dry cough, phlegm; Head colds; Burning eyes; Wheezing; shortness of breath; Lung disease; Chest discomfort or pain
CO	Toxicity of the central nervous system and heart; Headaches, dizziness, nausea and unconsciousness; Loss of vision; Decreased muscular coordination; Abdominal pain; Severe effects on the baby of a pregnant woman; Impaired performance on simple psychological tests and arithmetic; loss of judgment of time;
NOx	Increased incidence of respiratory illness; Increased airway resistance; Damage to lung tissue; Chronic obstructive pulmonary disease, or COPD (narrowing of the airways); Emphysema (as part of COPD); Pulmonary edema; Infant and cardiovascular death
SO2	Irritation of eyes, nose, throat; Damage to lungs when inhaled; Acute and chronic asthma; Bronchitis and emphysema; Lung cancer
Lead	Mortality; Hypertension nonfatal coronary heart disease;

III. Study Area

Tabriz is located in the north-west of Iran and its population is about 1.5 millions. Tabriz is the fourth most populous city in Iran after Tehran, Mashhad, and Esfahan, and is also a major Iranian heavy industrial and manufacturing center. Some of these industries include automobile, machine tools, oil and petrochemical and cement production. Therefore, we used pollution data of it that were observed from 5 ground stations in 2011. These stations geographically located in Rahahan, Rastekouche, Baghshomal, Abrasan, and Hakimnezami zones (Fig. 1). Ground pollution monitoring stations collect daily amounts of pollutants in different parts of the city. In this work, all of pollutant data for determining the emergency situation are used. For example in Table II there is a sample data of air pollutants for one of pollutant stations during 5 hours a day. Based on pollution data, the Air Quality Index is calculated from Equation I and then emergency situation is warned.

TABLE II. A Sample Data of Air Pollutants for One of Pollutant Stations

Date	Time	PM_{10} [μg/m^3]	SO_2 [ppb]	NO_2 [ppb]	CO [ppb]	O_3 [ppb]
2011/07/23	00:00	40.5	5.4	10.1	1.2	5.4
2011/07/23	01:00	247.9	6.1	12.2	1.3	6.1
2011/07/23	02:00	45.8	6.3	12.1	1.3	6.3
2011/07/23	03:00	104.3	4.7	6.0	0.7	4.7
2011/07/23	04:00	478.4	5.6	12.4	1.4	5.6

IV. Air Quality Index

Air quality indices (AQI) are numbers used by government agencies to characterize the quality of the air at a given location. As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects [1].

To compute the AQI everyone requires an air pollutant concentration from a monitor or model. The function used to convert from air pollutant concentration to AQI varies by pollutant, and is different in different countries. Air quality index values are divided into ranges, and each range is assigned a descriptor and a color code. Standardized public health advisories are associated with each AQI range. An agency might also encourage members of the public to take public transportation or work from home when AQI levels are high [1].

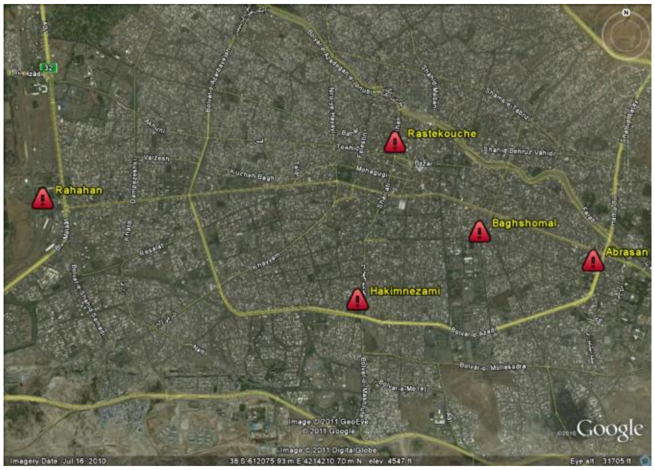


Figure 1. The pollutant stations of Tabriz city

The air quality index is a piecewise linear function of the pollutant concentration. At the boundary between AQI categories, there is a discontinuous jump of one AQI unit. To convert from concentration to AQI the equation I is used [1]:

$$AQI = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low} \qquad (1)$$

I = the (Air Quality) index,
C = the pollutant concentration,
C_{low}= the concentration breakpoint that is ≤ C,
C_{high}= the concentration breakpoint that is ≥ C,
I_{low}= the index breakpoint corresponding to C_{low},
I_{high}= the index breakpoint corresponding to C_{high}.

For example, suppose a monitor records a 24-hour average fine particle concentration of PM10=85.3 micrograms per cubic meter. Based on Table III this value is in the 55-154 intervals, then:

$$C_{low} = 55,\ C_{high} = 154 \Rightarrow I_{low} = 51,\ I_{high} = 100$$

Now, we can calculate the AQI:

$$AQI = \frac{100-51}{154-55}(85.3-55)+51=66$$

Based on calculated value for AQI above, the emergency status is *moderate*. (Table III)

TABLE III. AQI INTERVALS CORRES PONDING TO AIR POLLUTANTS [1]

Emergency Status	AQI	Air Pollutant						
		NO_2 (ppm)	SO_2 (ppm)	CO (ppm)	PM2.5 (µg/m^3)	PM10 (µg/m^3)	O_3 (ppm) 1-Hour	O_3 (ppm) 8-Hour
Good	0-50	-	0-0.034	0-4.4	0-15.4	0-54	-	0-0.0059
Moderate	51-100	-	0.035-0.144	4.5-9.4	15.5-40.4	55-154	-	0.0060-0.0075
Unhealthy for Sensitive Groups	101-150	-	0.145-0.224	9.5-12.4	40.5-65.4	155-254	0.125-0.164	0.0076-0.0095
Unhealthy	151-200	-	0.225-0.304	12.5-15.4	65.5-150.4	255-354	0.165-0.204	0.0096-0.0115
Very unhealthy	201-300	0.65-1.24	0.305-0.604	15.5-30.4	150.5-250.4	355-424	0.205-0.404	0.0116-0.0374
Hazardous	301-400	1.25-1.64	0.605-0.804	30.5-40.4	250.5-350.4	425-504	0.405-0.504	-
Hazardous	401-500	1.65-2.04	0.805-1.004	40.5-50.4	350.5-500.4	505-604	0.505-0.604	-

TABLE IV. A CALCULATED AQI OF AIR POLLUTANTS FOR ONE OF POLLUTANT STATIONS AND MAXIMUM OF THEM

Date	PM_{10} AQI	SO_2 AQI	NO_2 AQI	CO AQI	O_3 AQI	Maximum AQI
2011/07/23	37	7	0	14	0	37
2011/07/23	147	9	0	15	0	147
2011/07/23	43	9	0	15	0	43
2011/07/23	75	7	0	8	0	75
2011/07/23	367	9	0	16	0	367

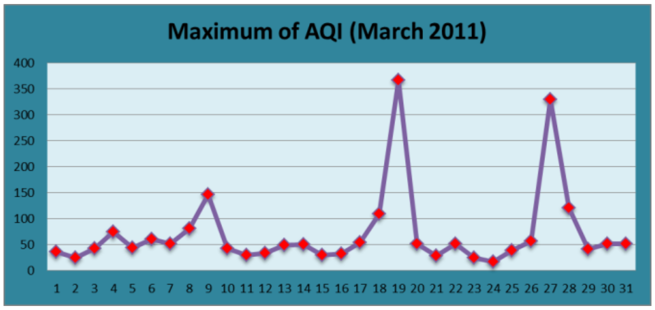


Figure 2. Maximum of AQI for Baghshomal monitoring station in March 2011.

In Table IV, for example, the calculated AQI and their maximum for concentrations related to Table II are presented. The Maximum of AQI for Baghshomal monitoring station in March 2011 is shown in Fig. 2. It is determined from Fig. 2 there are about 5 days in March that air pollution is very much and the emergency status is *Hazardous*.

V. IMPLEMENTATION

In this paper, a program has been developed in ARCObjects^{TM} [2] that it first calculates AQI values for whole of pollutant data and selects the maximum of them then merges these values properly to attribute information of pollutant monitoring stations. Finally, based on these attribute information, the pollution status on the relevant maps or satellite images is represented as daily, monthly and yearly with various statistical tools. For example Fig. 4 and Fig. 6 represent the maximum AQI of pollutants in March and July 2011 with variable slice sizes and Fig. 5 and Fig. 7 represent the same maximum AQI with identical slice sizes. In these figures the sequence of days is based on Fig. 3.

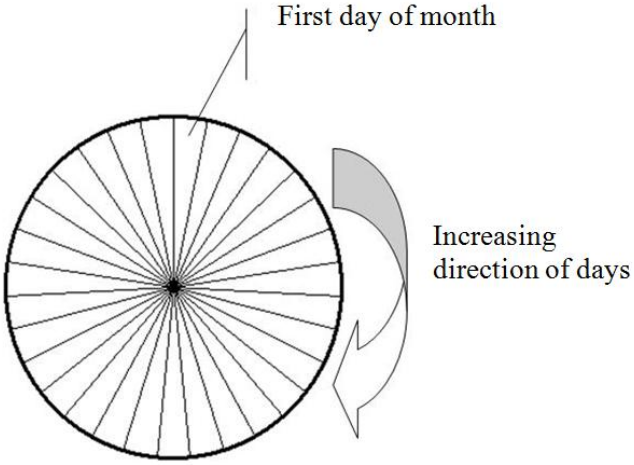


Figure 3. The sequence of days in one month

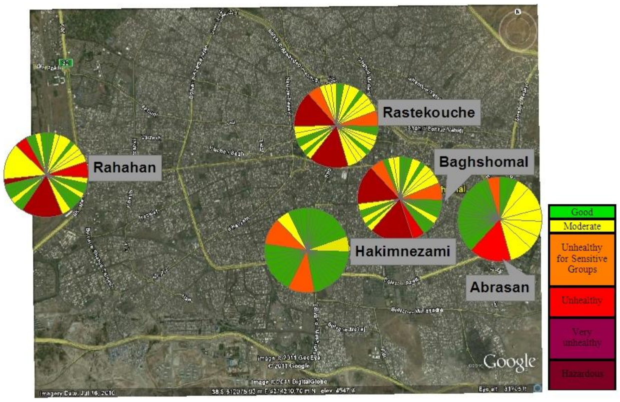


Figure 4. Maximum of AQI corresponding to pollutants of Tabriz in March 2011 with variable slices

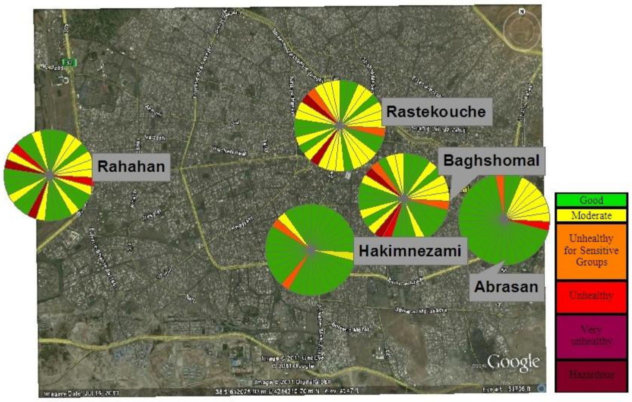


Figure 5. Maximum of AQI corresponding to pollutants of Tabriz in March 2011 with identical slices

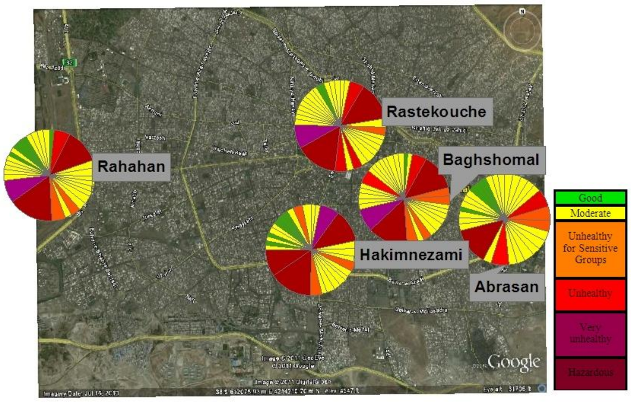


Figure 6. Maximum of AQI corresponding to pollutants of Tabriz in July 2011 with variable slices

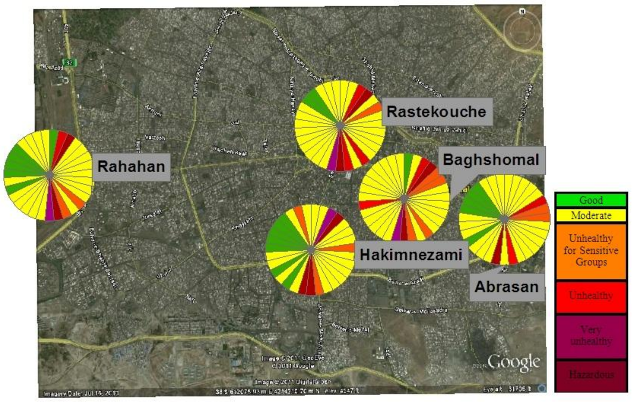


Figure 7. Maximum of AQI corresponding to pollutants of Tabriz in July 2011 with identical slices

Also, the maximum value of AQI corresponding to the pollutant data during the selected 6 months, March to August, of 2011 was calculated (Fig. 9) In these figures, the sequence of months is based on Fig. 8.

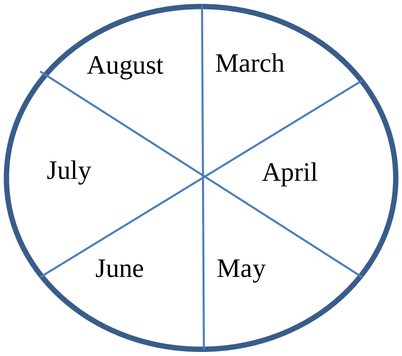


Figure 8. The sequence of the selected 6 months of 2011

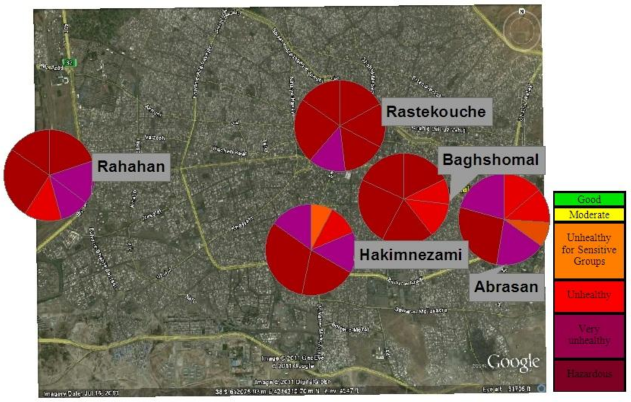


Figure 9. Maximum of AQI corresponding to pollutants of Tabriz in six months of 2011

It is seen from Fig. 9 that Tabriz city has air pollution in majority months of the year.

By more developing of the written program and customizing it, some radio buttons and check boxes can develop for best interface and GUI that every user can select daily, monthly or yearly and so select other options to best visualization and interpretation of results. This program can run on the Web for online visualization the results.

VI. CONCLUSION

By using GIS with various representations everyone can utilize visual, not only statistical, analyses based on different data especially pollution data. If the AQI is represented with identical slices, everyone can rapidly conclude that in which interval of month, 1st week or 2nd week and so on, the air pollution has maximum value and if the AQI is represented with variable slices, everyone rapidly conclude that what days have the maximum pollution.

Based on the data, it is obvious that PM_{10} is the major air pollutant in Tabriz city and other pollutants have a minor role in the pollution of its air.

If the number of pollutant monitoring stations in our city increase and the resulted data are available in the hands of analysts in every time and everywhere, they can represent the various results of these data in the different places of the cities for public, and also they can use GIS complex analyses such as Geostatistical Analysis for better analyzing the results and estimating the quantity of pollution in other places of city. It is expected that by this article and articles like it, such matters to be provided in the near future.

VII. ACKNOWLEDGMENTS

I would like to express my special gratitude and thanks to Mr. Masoud Shakeri, an employee of Department of Environment of Tabriz, for providing the ground station data and Mr. Fereydoon Naghibi, GIS doctorate student in Tehran University, for assisting us in developing program in ARCObjects^{TM}.

REFERENCES

[1] Johnson, D.L., S.H. Ambrose, T.J. Bassett, M.L. Bowen, D.E. Crummey, J.S. Isaacson, D.N. Johnson, P. Lamb, M. Saul, and A.E. Winter-Nelson. 1997. Meanings of environmental terms. Journal of Environmental Quality 26: 581-589.

[2] Kang Tsung Chang, Programming ArcObjects with VBA-A Task-Oriented Approach-Second Edition, 2008-CRC Press - Taylor & Francis Group.

[3] Khaled Ahmad Ali Abdulla Al Koas- 2010- GIS-Based Mapping and Statistical Analysis of Air Pollution and Mortality in Brisbane, Australia- School of Built Environment and Engineering – Queensland University of Technology.

[4] Technical Assistance Document for the Reporting of Daily Air Quality - the Air Quality Index (AQI) - February 2011- U.S. Environmental Protection Agency Office of Air Quality Planning and Standards -Research Triangle Park- North Carolina 27711- EPA-454/B-09-001.

Study on the Composition of the Residential Environment and Environmental Cognition in Collective Housing

Hirotomo Ohuchi
Dept. of Architecture College of Industrial Technology
Nihon University
Chiba, Japan
oouchi.hirotomo@nihon-u.ac.jp

Setsuko Ouchi
Ohuchi Environmental Design Lab.
Chiba, Japan
setsukoouchi@gmail.com

Katsuhito Chiba
Graduate School of Industrial Technology
Nihon University
Chiba, Japan
k.chiba0730@gmail.com

Yuta Takano
Graduate School of Industrial Technology
Nihon University
Chiba, Japan
yuta_t860709@yahoo.co.jp

***Abstract*—This paper investigates, analyzes, and considers the living environment, the characteristics of environmental perception, and their relationships, based on a survey of the residents of the housing complex in Makuhari Baytown, which was built using a new planning technique. In this study, it was found that the average size of the cognitive domain of the residents that has a courtyard which is available for a non-resident tends to increase.**

Keywords-Environmental recognition; Cognitive area; Collective housing; Makuhari Baytown; Living environment

I. INTRODUCTION

The plan for the center of the complex was to make it multistory, and to use standardization, which is one of the modern city theories followed in the design of a housing complex on an urban scale, with the aim of alleviating the shortage of houses. However, a uniform plan that does not consider the surrounding environment would result in a housing complex that would deteriorate after a time, and would eventually be destroyed, be rebuilt, or become antiquated, fall into ruin, and transform the district, making it quite different from the original plan. It is necessary to consider the environmental conditions of existing buildings and the original city plan and building codes, along with the surrounding environment, in order to ensure the plan will result in a sustainable housing complex. This consideration changes the existing problem, which is quantitative, to a qualitative problem.

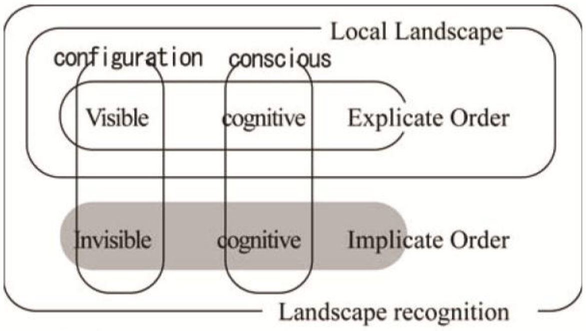
In the previous study, [6] Sketched maps drawn by children have analysed and considered the relationships between the actual physical environment of the town and the images that children drew of their environment, as well as the mutual relationships between the environmental changes that affect the children's spatial cognition.

Landscape recognition is the process by which an inhabitant recognizes the regional landscape. Regional symbiosis, which has been increasingly important in the recent years, is based on the premise that the resident feels that the surrounding environment is common property. Therefore, the shared recognition of the local landscape, which is valuable common property, is important.

The local landscape is defined by the correlation between morphological character, a physical-environment characteristic of space, and cognition based on spatial vision. Moreover, the visual information a person receives in daily life defines the cognitive region, which is a composite image of visible things and invisible things in specific positions. Landscape recognition is formed by relationship between the explicate order (visible cognitive region of consideration) and the implicate order (subconscious invisible cognitive region). The explicate and implicate orders are inseparable reciprocal processes (Figure 1).

Therefore, environmental recognition is analyzed based on landscape recognition.

Figure 1. Conception of Landscape recognition.



II. Investigation and Outline of Analysis

A. Region for research investigation

The investigation covered Makuhari Baytown, a residential housing area in the new center of the city of Makuhari, and constructed by a Chiba prefecture corporate agency in the Tokyo Bay shore area. It was possible to house 26,000 people in the 84-ha area, and 8,900 households were planned (Figure 2) .The route enclosing the housing area was constructed as the first part of the residential quarter plan (Figure 3).

The design was based on the guidelines, and an inside layer, a multistory section, and a high rise area are allocated in every city block. The layout of the plan included the development of a route enclosing the residential housing, and another enclosing the entire residential quarter (Figure 3).

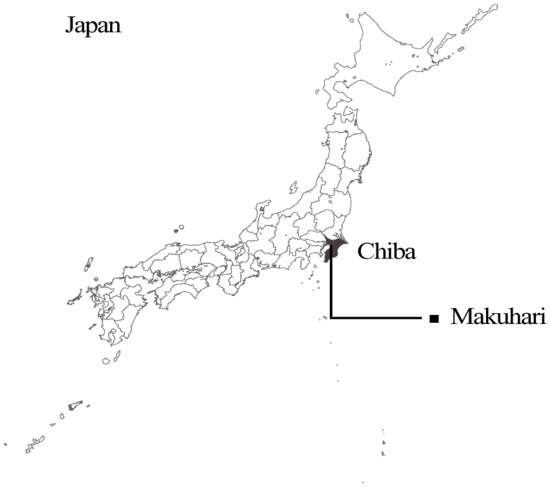


Figure 2. Location map of Japan in Makuhari

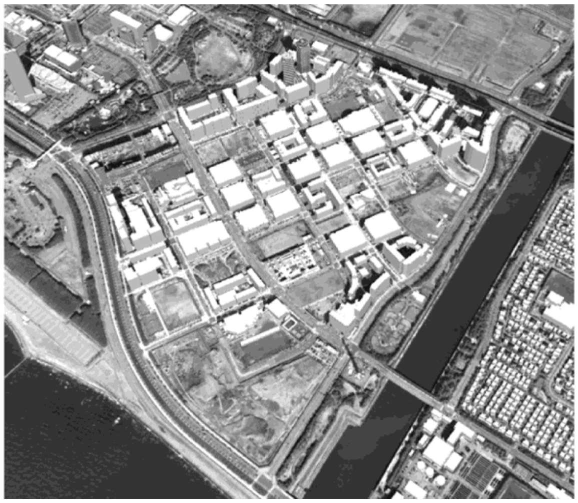


Figure 3. Investigation aerial photograph (2006)

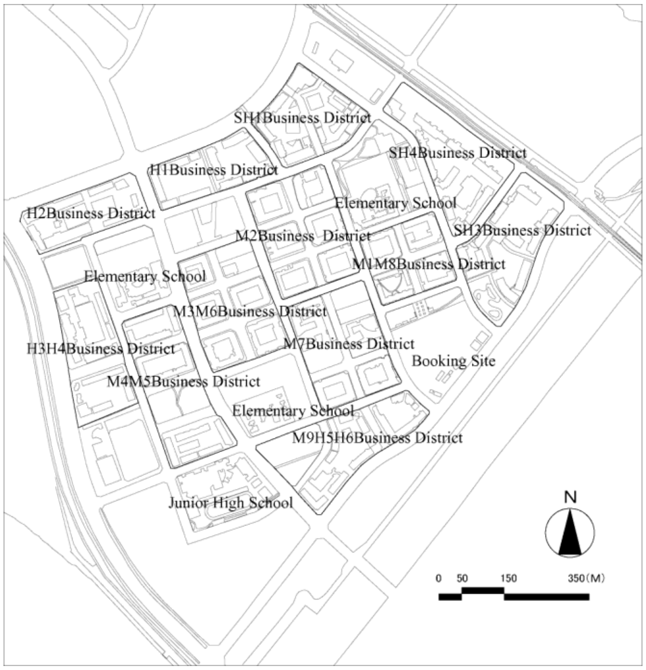


Figure 4. Arrangement plan

In many commercial and residential buildings, the lower floors house commercial facilities, which led to the invention of “The city is made in the house” development concept. The population, the number of houses, and the number of households, have all expanded during the passage of 15 years from the town’s opening. As of August 2009, around 40 housing complexes have been built so far, and the population now exceeds 23,000.

B. Methods of analysis

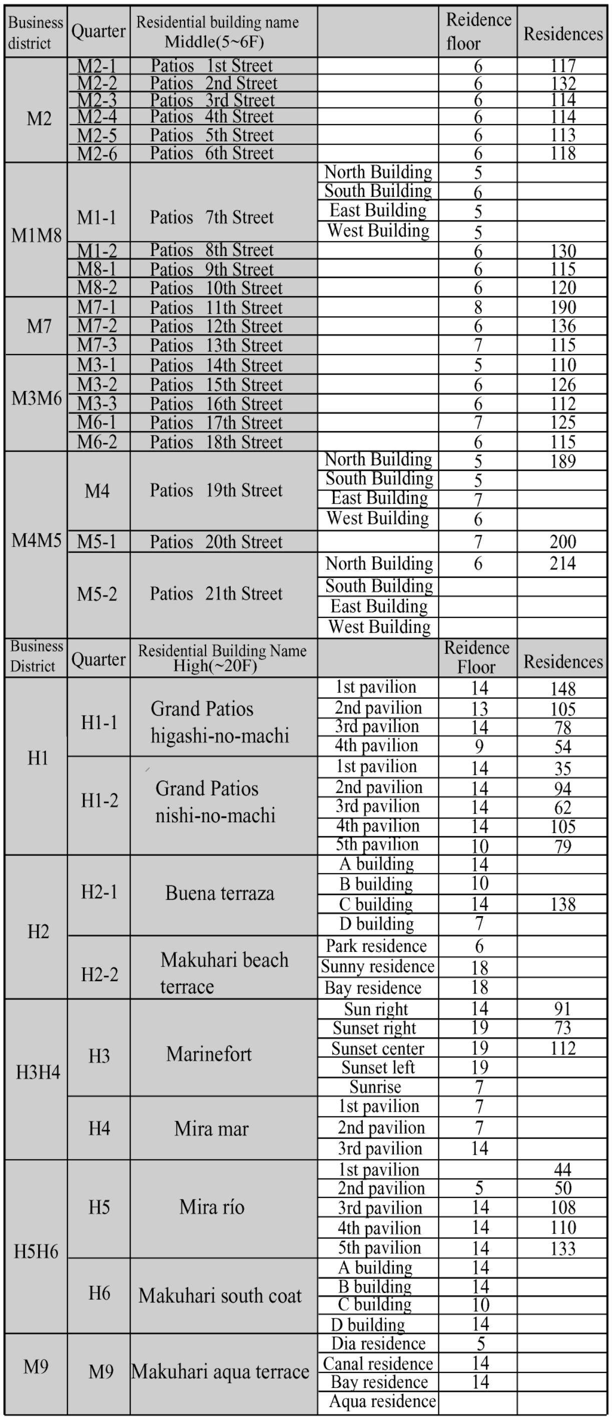
In the present study, residents of apartments and condominium complexes in the town were asked to complete a perceived area questionnaire survey. Next, a multivariate analysis was done to analyze the address space configuration of the research zone and the residents’ perceptions of the space surrounding them, which were quantified. The individual data obtained from the questionnaire was analyzed and a factor axis was extracted. The patterns were analyzed, and the features of the perceptions of the residents in the apartment and condominium complexes in Makuhari Baytown were obtained from a series of analyses.

C. Outline of the investigation

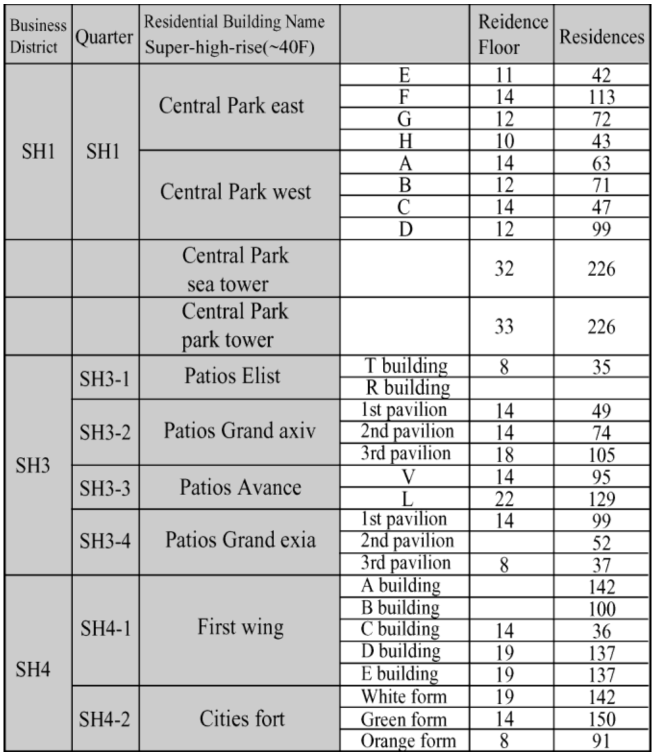
Investigation period: 25, 26 July, 1, 2, 7, and 31 August, and 8 September, 2010.

The search procedure: Residents of 43 residential buildings in the housing complex were given the questionnaire survey to complete.

TABLE I . OUTLINE



Business district	Quarter	Residential building name Middle(5~6F)		Reidence floor	Residences
M2	M2-1	Patios 1st Street		6	117
	M2-2	Patios 2nd Street		6	132
	M2-3	Patios 3rd Street		6	114
	M2-4	Patios 4th Street		6	114
	M2-5	Patios 5th Street		6	113
	M2-6	Patios 6th Street		6	118
M1M8	M1-1	Patios 7th Street	North Building	5	
			South Building	6	
			East Building	5	
			West Building	5	
	M1-2	Patios 8th Street		6	130
	M8-1	Patios 9th Street		6	115
	M8-2	Patios 10th Street		6	120
M7	M7-1	Patios 11th Street		8	190
	M7-2	Patios 12th Street		6	136
	M7-3	Patios 13th Street		7	115
M3M6	M3-1	Patios 14th Street		5	110
	M3-2	Patios 15th Street		6	126
	M3-3	Patios 16th Street		6	112
	M6-1	Patios 17th Street		7	125
	M6-2	Patios 18th Street		6	115
M4M5	M4	Patios 19th Street	North Building	5	189
			South Building	5	
			East Building	7	
			West Building	6	
	M5-1	Patios 20th Street		7	200
	M5-2	Patios 21th Street	North Building	6	214
			South Building		
			East Building		
			West Building		
Business District	Quarter	Residential Building Name High(~20F)		Reidence Floor	Residences
H1	H1-1	Grand Patios higashi-no-machi	1st pavilion	14	148
			2nd pavilion	13	105
			3rd pavilion	14	78
			4th pavilion	9	54
	H1-2	Grand Patios nishi-no-machi	1st pavilion	14	35
			2nd pavilion	14	94
			3rd pavilion	14	62
			4th pavilion	14	105
			5th pavilion	10	79
H2	H2-1	Buena terraza	A building	14	
			B building	10	
			C building	14	138
			D building	7	
	H2-2	Makuhari beach terrace	Park residence	6	
			Sunny residence	18	
			Bay residence	18	
H3H4	H3	Marinefort	Sun right	14	91
			Sunset right	19	73
			Sunset center	19	112
			Sunset left	19	
			Sunrise	7	
	H4	Mira mar	1st pavilion	7	
			2nd pavilion	7	
			3rd pavilion	14	
H5H6	H5	Mira río	1st pavilion		44
			2nd pavilion	5	50
			3rd pavilion	14	108
			4th pavilion	14	110
			5th pavilion	14	133
	H6	Makuhari south coat	A building	14	
			B building	14	
			C building	10	
			D building	14	
M9	M9	Makuhari aqua terrace	Dia residence	5	
			Canal residence	14	
			Bay residence	14	
			Aqua residence		



Business District	Quarter	Residential Building Name Super-high-rise(~40F)		Reidence Floor	Residences
SH1	SH1	Central Park east	E	11	42
			F	14	113
			G	12	72
			H	10	43
		Central Park west	A	14	63
			B	12	71
			C	14	47
			D	12	99
		Central Park sea tower		32	226
		Central Park park tower		33	226
SH3	SH3-1	Patios Elist	T building	8	35
			R building		
	SH3-2	Patios Grand axiv	1st pavilion	14	49
			2nd pavilion	14	74
			3rd pavilion	18	105
	SH3-3	Patios Avance	V	14	95
			L	22	129
	SH3-4	Patios Grand exia	1st pavilion	14	99
			2nd pavilion		52
			3rd pavilion	8	37
SH4	SH4-1	First wing	A building		142
			B building		100
			C building	14	36
			D building	19	137
			E building	19	137
	SH4-2	Cities fort	White form	19	142
			Green form	14	150
			Orange form	8	91

To clarify each resident's cognitive domain, the questionnaire survey was done by the by Sphere graphic method[*]. Residents had to have lived in the housing complex three years or more, and had to be at least 10 years old. To ensure there was no bias, the questionnaire survey covered each region of "Makuhari Baytown".

The survey content. The following items were investigated:

①Attribute investigation

②Consideration range survey recognized local resident

③Rote investigation in daily life

④Perceived area survey in range of action

⑤Perceived area component investigation in range of action

⑥Perceived area survey of responses to phrases such as familiar waterside, familiar green space, and bustle

⑦Perceived area component investigation

⑧Landmark investigation

⑨Visible consideration investigation of component of ②, ⑤, ⑦ and ⑧.

⑩Comparative study of the current housing and previous type of residence

The survey was conducted according to the overview above, and we obtained 164 valid responses, which are summarized in TABLE I and TABLE II. Cognitive domain data were obtained, and we considered the data analysis as a multivariate analysis.

ISBN: 978-1-61208-178-6

TABLE II． DATA ON THE RESPONDENTS

Age	Teens	24	Apartment name	Patios 1st Street	3
	Twenties	9		Patios 2nd Street	5
	Thirties	17		Patios 3rd Street	5
	Forties	69		Patios 4th Street	6
	Fifties	22		Patios 5th Street	5
	Sixties	16		Patios 6th Street	4
	Seventies	6		Patios 7th Street	4
	Eighties	1		Patios 8th Street	4
Sex	A man	75		Patios 9th Street	3
	Woman	88		Patios 10th Street	6
Residence Year	3~6years	56		Patios 11th Street	6
	7~10years	52		Patios 12th Street	3
	11~13years	28		Patios 13th Street	4
	14~17years	27		Patios 14th Street	3
	18years~	1		Patios 15th Street	4
Residence Floor	1~5	100		Patios 16th Street	4
	6~10	46		Patios 17th Street	4
	11~15	11		Patios 18th Street	3
	16~20	1		Patios 19th Street	
	21~24	1		Patios 20th Street	3
	25~30	3		Patios 21th Street	3
	31~	2		Patios 22th Street	3
The Direction of the room	North	13		Patios Avance	5
	South	59		Patios Elist	3
	East	14		Patios Grand axiv	4
	West	22		Patios Grand exia	4
	Northeast	1		Central Park east	6
	Northwest	3		Central Park west	8
	Southwest	20		Central Park sea tower	6
	Southeast	25		Central Park park tower	3
Employment	Company emploee	57		Grand Patios higashi-no-machi	6
	Civil servant	4		Grand Patios nishi-no-machi	4
	Independent enterprise	6		Buena terraza	3
	Profession	2		Makuhari beach terrace	3
	University student	27		Marinefort	4
	High shcool student	19		Mira mar	3
	Junior high shcool student	37		Mira río	4
	Part-time job	3		Makuhari southcoat	4
	Full time housewife	4		First wing	4
	The unemployed	5		Cities fort	5
Past Resident Status	Detached house	25	The effective number of answers：164samples		
	Lodgings	3			
	Private apartment	80			
	Company residence	23			
	Public corporation	25			
	Others	4			

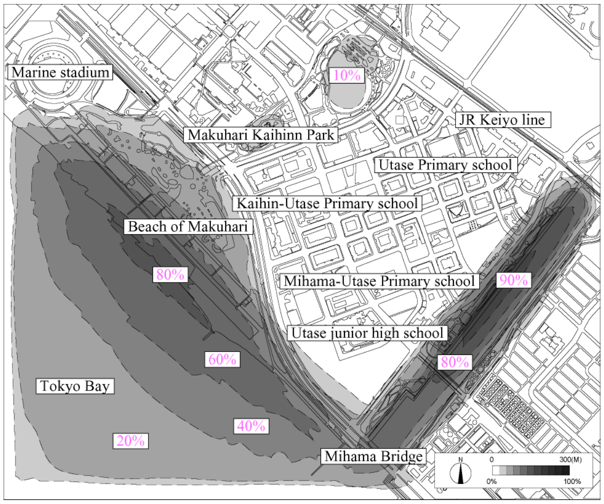


Figure 5. Cognitive Region Map “Familiar Waterside” (All)

*Sphere graphic method
This method is effective when focused on a subject who has adequate recognition of the area. It is suitable for studying elatively limited spaces in small areas, such as the area surrounding a personal dwelling. The subject’s cognitive area is obtained by indirectly exploring the structure through a spread, a spatial break, etc.

III. CONSIDERATION OF RESIDENTS’ ENVIRONMENTAL PERCEPTION

Subjects from the study demonstrated their cognitive domains by sphere graphic method such as “My Town,” “familiar waterside,” “familiar green,” “buzz,” “action range,” and “neighbors.” From their responses, cognitive domain diagrams were created (Figure 4).

The perceived area chart, in which all residential buildings in Makuhari Baytown had been summarized, was analyzed. The division perceived area chart was made for an inside layer, multistory, and high rise residential buildings to compare the analyses of the cognitive domains of each residential building type by height. This comparison analysis was carried out.

IV. CONSIDERATION OF COGNITIVE CONSTRUCT THAT USES MULTIVARIATE ANALYSIS

This section considers the psychological impact of the Makuhari Baytown design on its residents. An important part in the plan was the “Enclosed roadside” concept to target residents of residential mid-rise buildings (Figure 6).

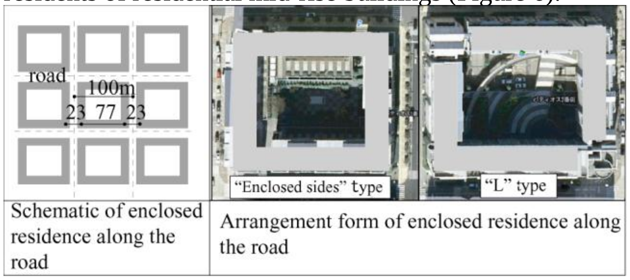

road
100m
23 77 23
“Enclosed sides” type
“L” type
Schematic of enclosed residence along the road
Arrangement form of enclosed residence along the road

Figure 6. Characteristic of enclosed residence along the road

TABLE III. ITEM CATEGORY TABLE

IN(Item Number),CN(Category Number),PN(Plot Number),FRE(Frequency)

IN	Item	CN	PN	FRE
1	Sex	1	A man	39
		2	Women	43
2	Age	1	10~29	18
		2	30~45	27
		3	46~59	26
		4	60~79	11
3	Residence Year	1	3~5years	17
		2	6~8years	10
		3	9~11years	11
		4	12~14years	24
		5	15~18years	20
4	Residence Floor	1	1F	5
		2	2F	16
		3	3F	20
		4	4F	21
		5	5F	9
		6	6F~	11
5	Perceived neighborhood (horizon)	1	0~10ha	36
		2	10~50ha	21
		3	50~90ha	14
		4	90ha~	11
6	Attribute	1	Point	46
		2	Line	1
		3	Respect	29
		4	Time	6
7	Number of components	1	1	61
		2	2	11
		3	3~	10
8	Visibility neighborhood	1	visibile	34
		2	unvisibile	29
		3	no cognition	19
9	Liberating level of courtyard	1	possible	31
		2	impossible	51
10	Living building arrangement	1	Type “L”	35
		2	Type “Enclosed sides”	47

Examples of the shaft to extract the common factors by quantification III[*] by using multivariate data obtained from surveys of previous chapters, we discuss factors that are critical to the recognition of residents (TABLE III). Quantification III was employed for 82 samples and residential buildings in which residents lived in the middle floors.

The first axis “Correlation coefficient:” 0.492140
The number of years in residence increased more than the item category plot chart, and the thing that decreases is understood (Figure 7). Moreover, it ranks highest in an item range high-ranking table (TABLE IV).The first axis is interpreted from the above-mentioned factor as a settled axis.

The second axis “Correlation coefficient:” 0.450089
The attribute becomes a quick fact from the item category plot chart in the positive direction, and the thing that is a physical element is understood in the negative direction (Figure 6). It ranks highest in an item range high-ranking table (TABLE IV). The second axis is interpreted from the above-mentioned factor as the axis of the formation of the local resident perceived area.

The third axis “Correlation coefficient:” 0.418563
It becomes impossible to use the degree of openness to the courtyard from the item category plot chart in the positive direction, and the thing that can be used in the negative direction is understood. The arrangement form understands similarly and the thing that there is all sides enclosing type in the positive direction, and L type is located in the negative direction is understood . The third axis is interpreted from the above-mentioned factor as the axis of the degree of freedom of access to the courtyard. It has been understood that “Liberating level of the courtyard” is a factor that alters the perception characteristics of the residents.

TABLE IV. TOP TABLE ITEM RANGE

The 1st axis			2nd axis			3rd axis		
IN	Item	Range	IN	Item	Range	IN	Item	Range
3	Residence Year	3.733166	6	Attribute	9.777477	6	Attribute	6.880096
5	Cognitive region area	3.438209	4	Residence floor	5.886572	3	Residence year	4.172621
2	Age	2.694054	3	Residence year	4.866726	4	Residence floor	3.478414
7	Number of components	2.61696	7	Number of components	4.708679	9	Liberating level of courtyard	2.280037
4	Residence floor	2.453967	2	Age	3.732348	10	Arrangement form	1.986933
9	Liberating level of courtyard	2.452993	5	Cognitive region area	3.602709	2	Age	1.856149
6	Attribute	1.909773	8	Visibility	2.260094	7	Number of components	1.666257
8	Visibility	1.798911	10	Arrangement form	0.935666	1	Sex	1.645543
1	Sex	0.749161	9	Liberating level of courtyard	0.336151	8	Visibility	1.537859
10	Arrangement form	0.066751	1	Sex	0.046749	5	Cognitive region area	1.120360

V. CONSIDERATION OF STRUCTURAL CHANGES IN THE RESIDENT FACTOR ANALYSIS AND TYPOLOGY

This study used samples obtained from the results in a score quantification # III, cluster analysis (Ward method), and revealed the characteristic features of each type of pattern recognition by each resident that typify the be. This paper identified at least 40 patterns in types I-IV. Euclidean distance cluster analysis helped obtain a dendrogram (Figure 6, Figure 7).

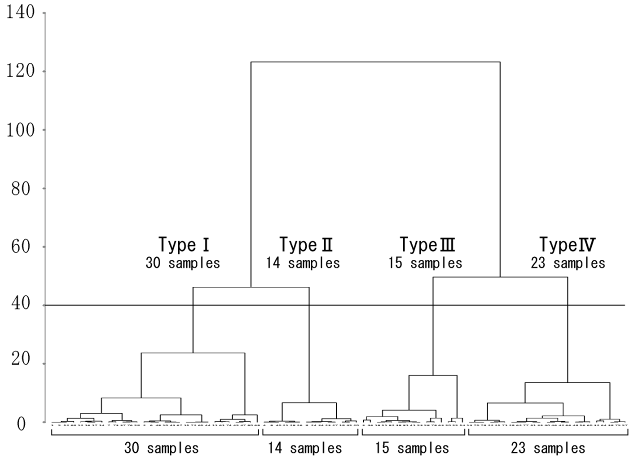


Figure 7. Cluster analysis dendrogram (Ward method)

Type I (30 samples)

Type I represents large residential buildings with open entry to a courtyard that is open to residents only. The average length of residency was 11.3 years. Many factors are involved, but they do not include a time element.

Type II (14 samples)

This type consists of a residential building with an open courtyard that is accessible to residents only. The average length of residency was 15 years.

Type III (15 samples)

In regard to the degree of openness of the courtyard entry, this type consists of large residential buildings with an open entry to the courtyard, which is also available to non-residents. Attributes of the neighborhood that form the cognitive domain include a time element.

Type IV (23 samples)

Type IV comprises large residential buildings with an open courtyard also available only to non-residents. The average length of residency was 7.5 years. The cluster analysis was made based on the four patterns of local resident perceived area charts (Figure 8).

TypeI: Perceived area mean value: 11.61 ha
TypeII: Perceived area mean value: 11.26 ha
TypeIII: Perceived area mean value: 95.94 ha
TypeIV: Perceived area mean value: 42.81 ha

[*] Quantification#III
The purpose of this analysis is to classify samples from relationship between categories (characteristic items) and the samples. The result is shows as scatter diagrams.
The procedure of the analysis is
1) the relationship between categories are analysed,
2) from the result, reveal latent common factor showed as axes of scatter diagrams (Item category plotting fig.) .
3) By the possession of samples on these scatter diagrams (Sample plotting fig.), they are classified, and their characteristics are grasped.

ISBN: 978-1-61208-178-6

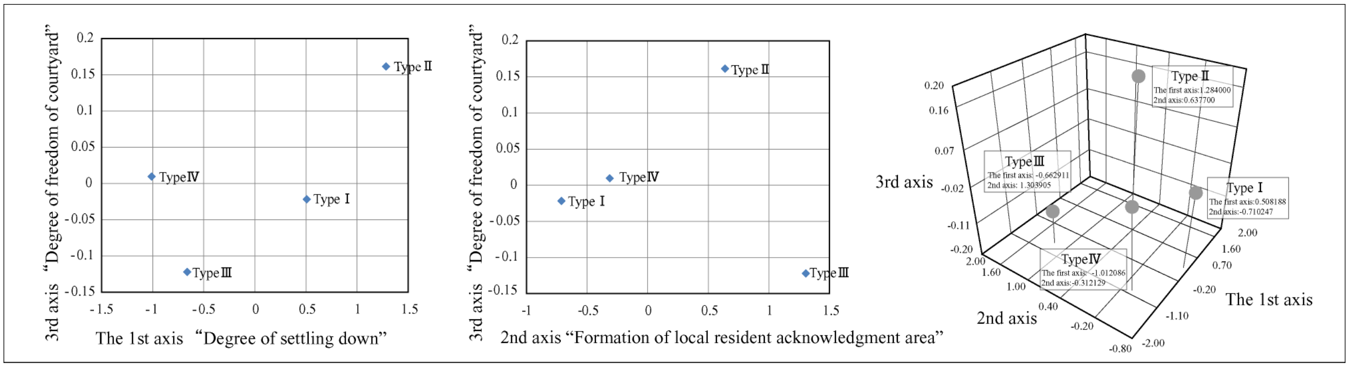


Figure 8. Pattern sample plot chart

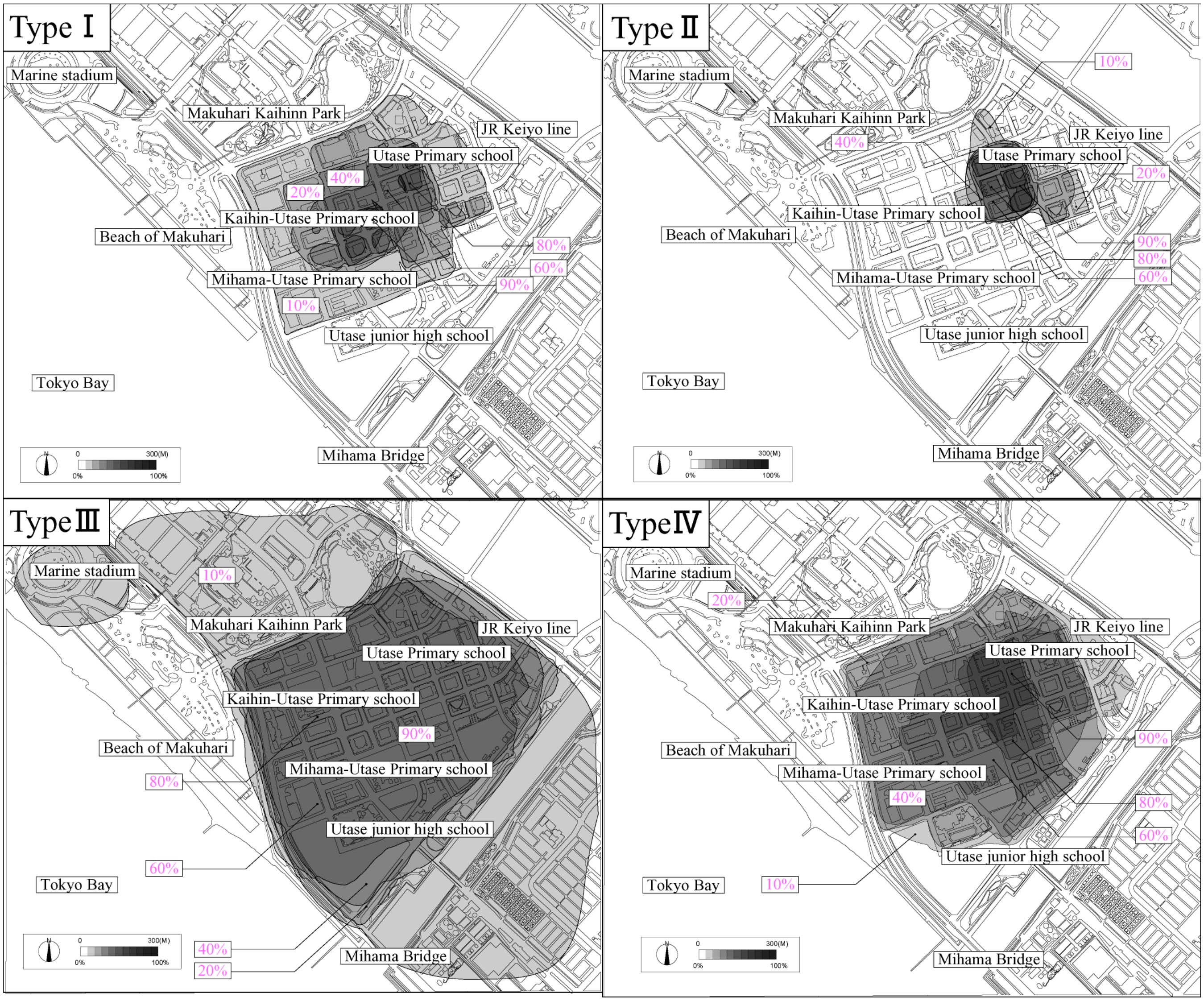


Figure 9. Cognitive region map "neighborhood"

VI. SUMMARY

The results of the above-mentioned analysis show the relationship between the characteristics of the living environment in Makuhari Baytown and the perception of the environment by the residents.

1. When the patternizing of the residents in the courtyard type route house is done according to the local resident perception characteristic, the factor of “Settle down”, “Formation of the local resident perceived area,” and “Degree of freedom of access to the courtyard” are the key indicators.

2. As the pattern characteristic, in pattern 1, there are a lot of ratios of the residential building with a courtyard that only the resident can use, and the length of residency was the longest in Type I. In Type Ⅱ, the residential building has a courtyard that only the resident can use, and the average length of residency is the longest in this type. The perceived area mean value for this type is the lowest. Type III is a pattern that there are a lot of ratios of the residential building with a courtyard that can be used exclusively by the resident, and the residency years are short. The perceived area mean value is the maximum. Type IV is a pattern that there are a lot of ratios of the residential building with a courtyard that can be used exclusively by the resident, and in this type the average length of residence is the shortest. The design characteristics of the courtyard type and the composition of the cognitive area were able to be considered by analyzing the living environment of the type of courtyard and “Local resident” perception characteristics. It is thought that the results achieved in this master’s thesis will lead to a theory relating to techniques for planning and constructing a uniform current housing complex and for city planning that considers characteristics of the location as well as those relating to the types of residents and the surrounding environment.

References

[1] Hitomi Fujioka, Chiaki Tagami, Hironori Negoro, and Hirotomo Ohuchi “Research on spatial cognition in children with sketch map”, Proceedings of the 18th Annual Paper on Environmental Information Science, no. 18, pp. 7-12, Nov. 2004

[2] Satoshi Yamada and Hirotomo Ohuchi "Research on environmental awareness of residents living in the body housing a collection of skyscrapers",Proceedings of the planning system,AIJ, no.18 pp.1749-1757, May 2008

[3] Akira Ito, Setsuko Ouchi and Hirotomo Ohuchi “Study on spatial composition and structure of the city's image of children as an educational environment - about the relationship between real space images of children in Makuhari and Tsukishima -"Annual Meeting Proceedings, AIJ, pp. 171-172 , Sep 2010

[4] Setsuko Ohuchi and Hirotomo Ohuchi ”Study on urban space composition as an educationalenvironment and image structure of children-Relation between actual space and child image-" XXIII UIA World Congress Torino 2008

[5] Ikeda, Daisaku. Dialogue on Life, Tokyo, Ushio shuppansha, 1974.

[6] Makiguchi, Tsunesaburo The Geography of Human Life, Tokyo, Seikyo Bunko, 1971.

[7] Hirotomo Ohuchi, Satoshi Yamada and Setsuko Ohuchi “Study on Child Spatial Cognition Using Sketched Maps of Urban Housing Projects Centering on Educational Institutions” Journal of South China University of Technology 35, pp. 205-208, Oct. 2007

Automated Extraction and Geographical Structuring of Flickr Tags

Omair Z Chaudhry[1, William A Mackaness[2
[1)]Manchester Metropolitan University, Manchester UK
[2)]University of Edinburgh, Edinburgh, UK,
Emails: O.Chaudhry@mmu.ac.uk, william.mackaness@ed.ac.uk

Abstract— **The volume and potential value of user generated content (UGC) is ever growing. One such source is geotagged images on Flickr. Typically, images on Flickr are tagged with location and attribute information variously describing location, events or objects in the image. Though inconsistent and 'noisy', the terms can reflect concepts at a range of geographic scales. From a spatial data integration perspective, the information relating to 'place' is of primary interest and the challenge is in selecting the most appropriate tag(s) that best describe the geography of the image. This paper presents a methodology for searching among the 'tag noise' in order to identify the most appropriate tags across a range of scales, by varying the size of the sampling area within which Flickr imagery falls. This is applied in the context of urban environments. Empirical analysis was then used to assess the correctness of the chosen tags (whether the tag correctly described the geographic region in which the image was taken). Logistic regression was then used to build a model that could assign a probability or confidence value to each selected tag as being a appropriate geographic tag. The high correlation values achieved bodes well for automated environments - environments in which this methodology could be used to automatically select meaningful tags and hierarchically structure UGC in order that it can be semantically integrated with other data sources.**

Keywords-data mining; information retrieval; vernacular geography; granularity modelling.

I. INTRODUCTION

The geospatial web comprises multiply sourced data (both formal and unstructured). Formal geographies (provided by National Mapping Agencies (NMA) and Government Bodies) reflect an *administrative* view of geography. Whereas User Generated Content (UGC) reflects observations at different levels of detail, more qualitative in nature, and relating to ideas of 'place' (events and performance) rather than formal and systematic descriptions of space. These two types of data offer complementary and synergistic approaches to the mining and intuitive understanding of geographic information. The conflation of 'formal' and 'informal' (such as free access to NMA datasets via Open Street Map) reflects a blurring of this binary but data integration is far more than simple overlay. Much has been written on the need for semantic and ontological modeling in order to automatically conflate the qualitative and the quantitative [1, 2]. The difficulty being the vagueness omnipresent in the geospatial domain, the problematic notion of space and place, and the granularity inherent in the description of geographic concepts [3].

There is increasing interest in mining the 'geography' now stored on the web. Geography provides a context and an intuitive way of organizing digital information. The 'Geospatial web' reflects a capacity to search for documents based on references to the geographical (using geotags [4]), to model vernacular geographies [5-7] and to support web mapping technologies [8]. Research on the Geospatial web is fuelled by freely available user generated content (UGC) or Volunteered Geographic Information (VGI) [9]. Open Street Maps, Wikimapia, WikiLocation, Geonames are frequently cited examples of VGI, and in some contexts rival conventional ways of capturing geographic information [10]. But the very nature of UGC means that it is often inconsistent, incomplete and poorly structured [11]. Tags attached to images and videos on data sharing services such as Flickr, and YouTube may contain a number of references to places, objects and events but not in a form that can be readily understood except by people with some knowledge of the vocabulary used.

For example, for Fig 1, how might we extract the 'meaningful' information inherent in the images and tags, and how might we structure the geography implicit in this image in a way that facilitates its retrieval and use. Is the tag 'Paris' in Fig. 1 the name of a person, a world capital or a community in Ontario, Canada? This is example of non-geo/geo ambiguity similarly there can be geo/geo ambiguity for example 'London' in UK or 'London' in Canada. A number of techniques have been proposed to 1) automatically unambiguously extract place names, and 2) assign them spatial coordinates [11]. These two process are commonly referred to as geo-parsing and geo-coding respectively [12]. This paper describes a technique for automatically retrieving and visualizing 'meaningful' place names from a VGI dataset (specifically Flickr geo tagged images) at different spatial levels of detail.

Section 2 describes a methodology to mine information from a list of geotags and to sample data at different granularities in order to hierarchically structure the data. Section 3 uses data mining techniques for 'post selection' of the tags that seeks to filter out selected tags that are not geographical in nature. The section also presents visualization techniques used to convey the hierarchical structure and confidence in the selection of tags.

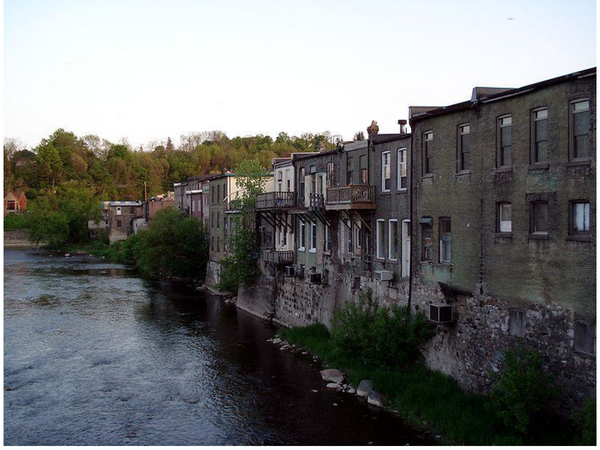


Figure 1. Geotags: Paris, Grand River, Nikon: Is Paris a person, a world capital, or a community in Ontario,Canada? Was Nikon the person who took the picture?

II. Flickr Images and their Tags

The aim is the extraction of meaningful place names and points of interest from among the tags that people associate with the Flickr images they take. Flickr is one of the biggest sources of images on the web. There are estimated to be 5.9 billion pictures available on Flickr. The Flickr website [13] suggests there are more than 153 million pictures that have been geo-tagged - pictures that have been assigned a geographic footprint (a latitude-longitude coordinate). Users are free to assign any number of tags to a picture. The tags can be city names, landscape descriptions, events, the camera used, dates, regions within a city, gardens, places and features of interest, indeed any adjective you care to imagine!

Various authors have presented techniques for extracting structured information from data [14-17]. An assumption common to these research efforts, is that the image tags variously connect the image with a particular geography of place and space (idea of ‘place semantics’ [14]). Among other things, the truth of such an assumption can be corroborated against the density of nearby images and diversity of image takers. The value in extracting such place semantics are well understood (e.g., improved search, intuitive (vernacular) descriptions of space, automated assignment of place semantics to untagged imagery). Most of the research has focused on extraction meaningful tags on the same level of spatial detail. Also such research uses manual comparison for testing accuracy of the approach. Here this research focuses on extraction of meaningful tags at *different* levels of detail and automatic assignment of confidence value (probability of correctness) by a model.

A. Accessing Flickr

Flickr provides a non-commercial API in order to access its dataset. The API provides a number of ways in which the Flickr images can be queried: by date, tags, geographic location, or groups for example. In addition there are a number of free Flickr API programming kits available. These kits are programming interfaces for different programming languages (notably C, Java, and Python [24]). These kits allow API queries to be embedded within user’s own code. For this research we used flickrj – a java Flickr API kit which was used to extract the Flickr dataset for the City of Edinburgh, Scotland.

In order to obtain all publicly geo-tagged images for the city of Edinburgh we overlaid a matrix of regular cells, each of 100m2 covering the whole city. A total of 134,986 images with its id, user tags, URL, user id, latitude, longitude values were thus obtained for the whole city of Edinburgh. There were a total of 20,400 100m2 cells covering the city of Edinburgh. Only 3,993 of those cells contained one or more Flickr image that were tagged (Fig. 2).

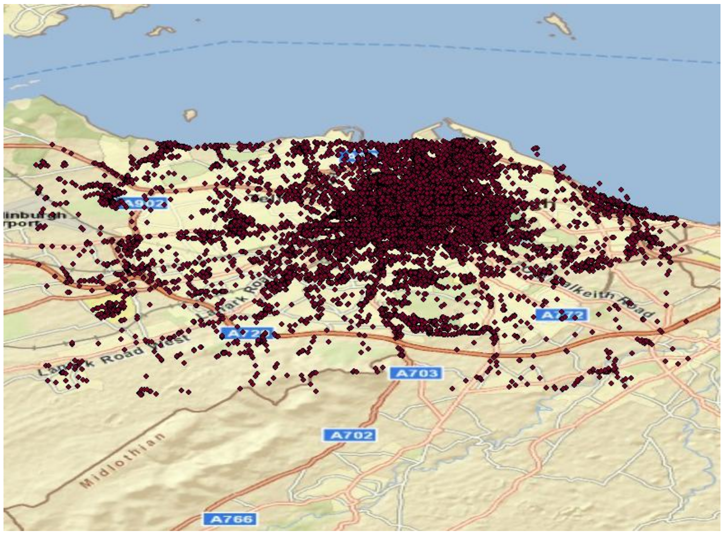


Figure 2. A map of Edinburgh showing the distribution pattern of Flickr images across the city.It shows the distribution is not homgenous across the whole city

B. Selection of ‘meaningful’ place names for each Cell

The next step was to assign to each cell, a place name or point of interest chosen from the associated image tags falling in that cell. For each cell we selected all the images and their respective tags. From these tags we then selected the most ‘appropriate’ tag for that cell. The simplest text analysis technique is to rank the tags according to frequency of occurrence and select the most frequently occurring. Unsurprisingly for many of the cells, the most frequently occurring tag was either ‘Edinburgh’ or ‘Scotland’ – not a tag that reflects the ‘local’ spatial granularity of a 100 m^2 cell! The second problem is one of ‘tag distortion’ arising from a single person, taking a relatively large number of images, and using the same tag to describe an event (rather than a place). For example one cell was named: ‘Elaine’s wedding’. This was because the tag was associated with 20 separate images spatially contained by that particular cell.

1) *Modeling the Local Context*

In order to incorporate the local context, we applied the TF-IDF algorithm (term frequency – inverse document frequency). TF-IDF is a well know information retrieval technique and is used to weight the importance of an individual term's contribution in relation to relevance in a search [11]. In our study for each tag contained by a cell we computed its term frequency (TF) by dividing the number of times the tag occurred within the cell by the total number of all tags within that cell. Inverse document frequency (IDF) was computed by taking the logarithm of the total number of cells that contain any tag (i.e. 3,993) divided by the total number of cells that contain that particular tag. TF-IDF is the product of TF and IDF. This product (TF-IDF) is the resultant weight assigned to the tag. The highest weighted tag is then selected for each cell. This approach ensures that those tags which are frequently inside one cell but occur rarely in other cells are given a high weight. So 'Princes Street' (a major high street in the city) will have a high weight, and 'Edinburgh' or 'Scotland' will have a low weight since they occur frequently within the cell as well as in the whole collection.

2) *Object View vs Subject View (User Frequency + TF-IDF)*

The TF-IDF approach identifies tags 'local' to a region, but it does not remedy the problem of 'tag distortion' (the example of 'Elaine's Wedding'). We can resolve (to large extent) this problem if we take an object's perspective of the tag, rather than a subject's perspective. We might realistically expect different people to use the same tag, thus corroborating the validity of that tag. In the example of 'Elaine's Wedding', it is very unlikely that other people would use this tag in the same cell. So by attaching importance to the number of different users who use a particular tag (the idea of collective intelligence), we might overcome the distorting effect of a single user attaching the same tag to multiple images falling in the same cell. So instead of using tag occurrences we use user frequencies associated with each tag in order to calculate TF-IDF weights. Using the user count reduces the TF for tags such as 'Elaine's wedding' and IDF ensures that tags with a high user count, such as 'Edinburgh' and 'Scotland', will have low IDF values. This results in low weights (TF-IDF) for both of these types of tags. All tags contained by a particular cell are then sorted in descending order and the tag with the highest weight is selected. There is a proviso: the tag is selected only if it has a 'user count' of at least two. The extra condition ensures that at least two distinct users have used the same tag. If this condition is not met then the next tag in the sorted list is checked and so on until both conditions are met. This process resulted in 3,951 cells being assigned a tag at the 100m2 spatial resolution for the city of Edinburgh.

C. *Hierarchical Structuring and Visualisation*

We can imagine people's understanding of the city to be hierarchical in nature (Fig 3), comprising high streets, shopping centers, and business districts, at one level, suburbs, districts, parks at another, all partonomically constituting the city [18].

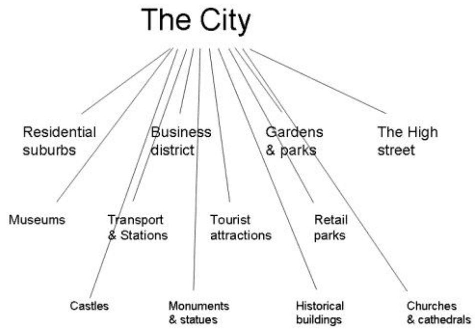


Figure 3. An example of conceptual view a city

Therefore, as a next step, we applied the same methodology but to increasingly larger cell sizes, covering the same region, in order to try and mirror a linked hierarchical structure. We applied the technique to grid cells with resolution: 500m2, 1000m2, 2000m2, 4000m2 covering the city of Edinburgh. Although the choice of scales (100m2 to 4000m2) is arbitrary but the presented approach for tag selection is applicable at any selected scale or shape and size of grid cell. Once the labels were selected for each cell at a specific level (100m2 to 4000m2) we aggregated adjacent cells if they had the same label. This created regions that shared the same tag. Fig 4 shows the result of applying this approach to the city centre of Edinburgh (The Castle and The Royal Mile). Fig. 4 also shows the selected tags as labels for each cell at different levels of detail. At each higher level the most dominant tag (highest TF-IDF weight) is selected as the label.

Upon inspection of the selected tags it was apparent that there was still 'noise' present among selected tags, most notably at the finest level of detail (100m2). Date tags are an example of such noise (for example '2007'). This happens because a tag such as '2007', will have less weight only if very few distinct users have used that tag or that it is common to a whole collection of cells. It is still possible that such tags have been used by a number of distinct users within the cell. Upon manual inspection of the selected tags at 100m^2 it was found that out of a total of 3,951 cells (100m^2) assigned a label, only 34% contained a meaningful place name or point of interest; the remainder was 'noise'. Similar manual inspections were carried out at all scales. As illustrated in Fig 5, the noise tags selected by this approach are highest at the most detailed level (100m^2). But, at lower levels of detail the TF-IDF weights of noise tags will be less as compared to non-noise tags because the spatial extent is larger and thus there are more images that have appropriate non-noise tags.

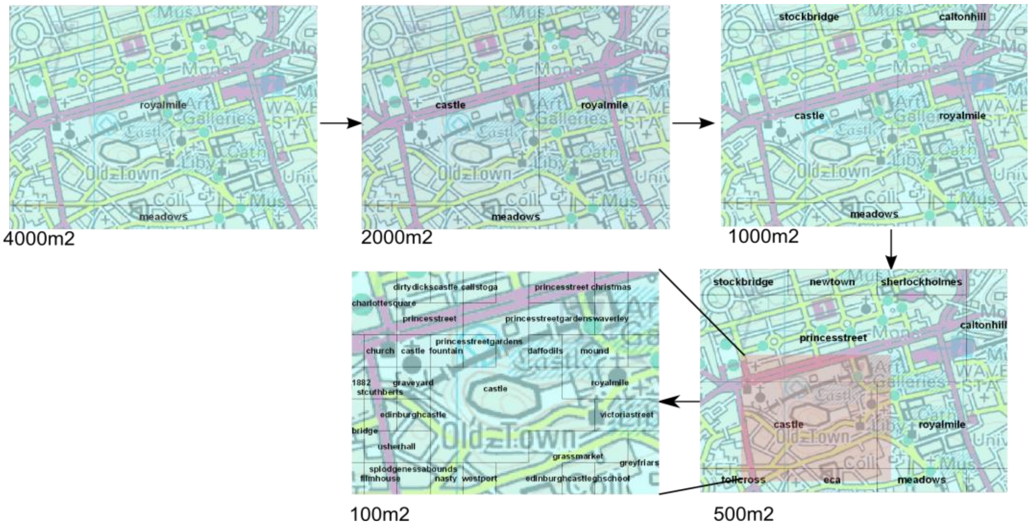


Figure 4. A conceptual view a city? Different tags at different levels of detail.

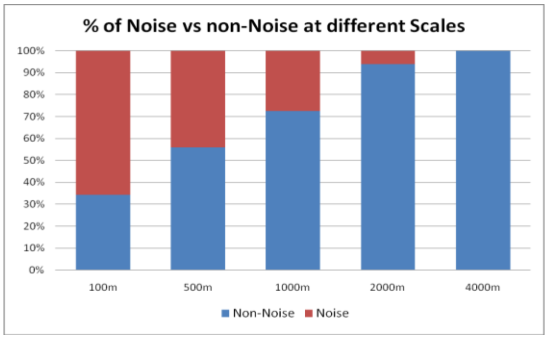


Figure 5. High amounts of noise at smaller cell size (Result of manual inspection)

III. Post Selection Refinement

In response to the problem of 'noise' (especially at the finest level of granularity), a data mining technique was explored in order to automatically filter out this noise. In the past, techniques such as 'stop words' and 'controlled vocabularies' have been proposed to remove such noise [19, 20]. In this research, we tested a data mining technique (logistic regression) in order to build a model using a number of (independent) variables computed for each selected tag from the source data (Flickr dataset) without utilizing any other external data. In essence the aim was to further refine the above approach such that a confidence value, representing the probability that it is not noise, can be attached to each selected tag. We used a manual classification to build and test the accuracy of the approach. We randomly selected 70 % of the manually classified cases at 100m2 to build the model. The remaining 30% of the manually classified data was used to assess the validity of the approach.

A. (Binary) Logistic Regression

Logistic regression is similar to multiple regression except that the dependent variable in the logistic regression is sampled as a binary variable i.e. noise (y=0) or non-noise (y=1). Logistic regression therefore models the probability of presence and absence for a given observed value among the predictor variables. The probability function can be written as: [21].

$$P(Y=1)=\frac{1}{1+e^{-(\alpha+\beta_1x_1+\beta_2x_2+....+\beta_nx_n)}} \qquad (1)$$

In Equation 1, y is the dependent variable, α is the intercept, β is the coefficient(s) of the independent variable(s) x. Equation 1 can be used to calculate the probability that the outcome (dependent variable) will be 1. In this research y is 1 if the selected tag is considered to be the correct label for a given cell, otherwise it is 0.

For each cell and its selected tag, we calculated a number of variables, (x_n). These included: the user frequency of a selected tag within the cell (x1); the user frequency of the selected tag in the whole collection (all cells) (x2); the selected tag frequency within the cell (x3); the selected tag frequency in the whole collection (x4); the total number of images contained by the cell (x5); the total user frequency for all the tags contained by the cell (x6); and the total raw frequency of all the tags contained by the cell (x7). Stepwise binary logistic regression was carried out in SPSS [22], randomly selecting 70% of the manually classified cases – the remaining 30% were used to the test the accuracy of the model.

Table I lists the selected variables (x1, x2 and x4) from the last stage of the stepwise logistic regression together with their coefficient values. Nagelkerke's R2 value for the model is 0.423. Table II lists the classification result after the final step of stepwise logistic regression. Table II shows the percentage of correctly identified cases from the 70% sample dataset is 81.1%, and the percentage of correct results for the remaining 30% of the sample dataset is 82.3%. The cut off value used in Table II to separate between cases classified as 0 or 1 is 0.5. This simply means that if the resultant probability for a tag is 0.51 it will belong to class 1 (non-noise) and if 0.49 it will belong to class 0 (noise).

TABLE I. SELECTED VARIABLES IN THE MODEL

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
x1	1.413	.090	246.095	1	.000	4.107
x2	.013	.002	46.500	1	.000	1.013
x4	-.001	.000	35.958	1	.000	.999
Intercept (α)	-2.738	.121	512.187	1	.000	.065

Once the model was built (Table I) for the most detailed scale (100m^2), we applied this model at all the remaining scales. The result from the model was evaluated against the manual inspection carried out in the previous section (Fig. 5) and is presented in Table III. As the scale reduces, the capacity of the model to correctly predict the result is significantly increased (Table III), especially for the true positive cases.

We linked the probability values calculated by the model to each tag and created an interactive tree view visualization in order to explore these hierarchal relationships in more detail. The tags are connected hierarchically via their spatial relationship – 'contained by' (Fig6). The number next to each tag name in Fig.6 show how confident the model is that the tag is not noise. The visualization is available as an applet at [23].

TABLE II. RESULT OF CLASSIFICATION FOR SELECTED AND UNSELECTED CASES AT 100M^2 WITH 0.5 CUTOFF VALUES

Observed		Predicted					
		Selected Cases			Unselected Cases		
		Ori_Class			Ori_Class		
		0	1	% Correct	0	1	% Correct
Ori_Clas	0	1697	112	93.8	751	37	95.3
s	1	410	549	57.2	172	223	56.5
Overall Percentage				81.1			82.3

TABLE III. EVALUATION OF LOGISTIC MODEL (TABLE I) AGAINST MAUL CLASSIFICATION AT LOWER LEVELS OF DETAIL (PROBABILITY CUT OFF VALUE IS 0.5 – THE SAME AS IN TABLE II)

	Class	0	1	% Correct
Scale: 500m2	0	220	29	88.35
	1	99	216	68.57
Scale: 1000m2	0	39	11	78.00
	1	21	111	84.09
Scale: 2000m2	0	1	2	33.33
	1	0	47	100.00
Scale: 4000m2	0	0	0	
	1	0	10	100.00

IV. CONCLUSION

The geo-tagged images generated by the public and freely accessible via a number of Web2.0 services such as Flickr offer great potential to understand people's perception of places and points of interest. A lot of research in Geospatial web has explored the use of flickr tags as a source for vernacular geography but there has been limited research in exploration of these images and tags at different levels of detail. This research has used data mining and text analysis techniques for selecting appropriate tags names as description of areas at different levels of detail. We have also presented a model used in post selection to calculate confidence (probability) values for each selected tag as a basis for assessing its likely correctness. The results were compared against manual inspection and it was observed that the range of the results correctly predicted by the model were from 80% at the most detailed level to 100% at the coarsest level. Future research will look into usage of clustering or road network partitions instead of arbitrary grid cells also threshold for selecting more than one tag for each region shall be addressed.

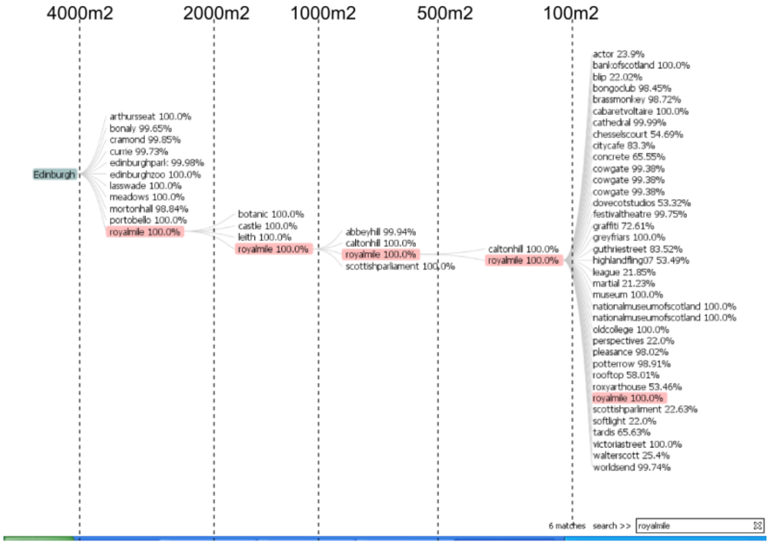


Figure 6. Tree view visualisation of selected tags and their confidence (probabilty) value as predicated by the model

ISBN: 978-1-61208-178-6

V. References

[1] S. Mustière and J. van Smaalen, "Database Requirements for Generalisation and Multiple Representations", in Generalisation of Geographic Information: Cartographic Modelling and Applications, W.A. Mackaness, A. Ruas, and L.T. Sarjakoski, Eds., Elsevier: Oxford, 2007. pp. 113-136.

[2] P. Agarwal, "Ontological considerations in GIScience". International Journal of Geographical Information Science, vol. 19, 2005, pp. 501-536.

[3] R.B. McMaster and K.S. Shea, "Generalization in Digital Cartography". Resource Publication in Geography: Washington D.C. 1992.

[4] R.S. Purves, et al., "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet". International Journal of Geographical Information Science, vol. 21, 2007, pp. 717-745.

[5] C.B. Jones, R.S. Purves, P.D. Clough, and H. Joho, "Modelling vague places with knowledge from the Web". International Journal of Geographical Information Science, vol.22, 2008, pp. 1045-1065.

[6] P. Lüscher and R. Weibel. "Semantics Matters: Cognitively Plausible Delineation of City Centres from Point of Interest Data". in Proc. 13th workshop of the ICA commission on Generalisation and Multiple Representation. 2010. Zurich, Switzerland.

[7] L. Hollenstein and R.S. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores". Journal of Spatial Information Science, vol.1, 2010, pp. 21-48.

[8] A. Scharl and K. Tochtermann, eds. "The geospatial web how geobrowsers, social software and the Web 2.0 are shaping the network society". Springer: London, 2007

[9] M.F. Goodchild, "Citizens as sensors: the world of volunteered geography". GeoJournal, vol. 69, 2007, pp. 211-221

[10] M. Zook, M. Graham, T. Shelton, and S. Gorman, "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake". World Medical & Health Policy, vol. 2, 2010, pp.7-33

[11] R.S. Purves, "Methods, Examples and Pitfalls in the Exploitation of the Geospatial Web", in The Handbook of Emergent Technologies in Social Research, S.N. Hesse-Biber, Ed., Oxford University Press: Oxford, 2011, pp. 592 -624.

[12] K.S. McCurley. "Geospatial Mapping and Navigation of the Web". in Proc 10th international conference on World Wide Web, Hong Kong: ACM, 2001

[13] Flickr. Available from: http://www.flickr.com/map/, 2011, Last accessed: 21 July 2011

[14] T. Rattenbury and M. Naaman, "Methods for extracting place semantics from Flickr tags". ACM Trans. Web, vol. 3, 2009, pp. 1-30.

[15] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs", in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM: Santa Barbara, 2006

[16] F. Girardin, F. Calabrese, F.D. Fiore, C. Ratti, and J. Blat, "Digital Footprinting: Uncovering Tourists with User-Generated Cotent". Pervasive Computing. vol. 7, 2008, pp. 36-43.

[17] S. Ahern, M. Naaman, R. Nair, and J. Yang. "World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections ". in Proc Seventh ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) ACM: New York, 2007

[18] O.Z. Chaudhry and W.A. Mackaness. "Utilising Partonomic Information in the Creation of Hierarchical Geographies". in Proc 10th ICA Workshop on Generalisation and Multiple Representation. 2007. Moscow, Russia.

[19] R. Pasley, P. Clough, R.S. Purves, and F.A. Twaroch, "Mapping geographic coverage of the web", in Proc. of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. 2008, ACM: Irvine, California.

[20] B. Croft, D. Metzler, and T. Strohman, "Search Engines: Information Retrieval in Practice", Addison-Wesley: Boston, 2009

[21] P.D. Allison, "Logistic Regression Using the SAS System: Theory and Application". Wiley Interscience:New York, 2001

[22] D.G. Kleinbaum and M. Klein, "Logistic Regression: A Self-Learning Text". 3rd ed., Springer: New York, 2010

[23] O.Z. Chaudhry. vailable from: http://www.omairchaudhry.net84.net/City_Viz/TreeView_Confidence.html. 2011 Last accessed: 20 Nov 2011

[24] Python http://www.flickr.com/services/api/ Last accessed: 20 Nov 2011

ISBN: 978-1-61208-178-6

From Point Clouds to 3D City Models: The Case Study of Villalba (Madrid)

Juan Mancera-Taboada, Pablo Rodriguez-Gonzalvez,
Diego Gonzalez-Aguilera*, Benjamín Arias-Perez
Land and Cartographic Engineering Department
University of Salamanca
Ávila, Spain
juaniyoperote@usal.es, pablorgsf@usal.es,
*daguilera@usal.es, benja@usal.es

David Hernandez-Lopez, Beatriz Felipe-Garcia
Regional Development Institute
University of Castilla-La Mancha
Albacete, Spain
david.hernandez@uclm.es, beatriz.felipe@uclm.es

***Abstract*—This article presents a practical study of the use of LIDAR (Light Detection and Ranging) data processing to transform point clouds into a three-dimensional city model compatible with CAD (Computer-Aided Design) graphic design systems. The article describes the methodology followed while concentrating on an increase in automation and reviewing the algorithms used. This case study demonstrates the importance of the LIDAR technology for 3D city modelling and notes several applications that may arise, especially in the context of urban and regional planning.**

Keywords—LIDAR; classification; reconstruction; 3D modelling; 3D city

I. INTRODUCTION

The airborne LIDAR system is a data capture method that can be used as an alternative or complement to photogrammetry. It also constitutes an effective tool for creating Digital Terrain Models (DTM) and Digital Surface Models (DSM) in urban areas. Although it is a technique that has yet to mature, LIDAR has advantages in certain situations when compared to aerial photogrammetry [1]. Example LIDAR applications include the following: urban areas, strip mines and landfills, snowy areas, dunes, marshes, wetlands, forests and areas with dense vegetation, waterways and water resources, the control and monitoring of coastal erosion and monitoring and managing natural disasters. Additionally, the LIDAR system is better suited to automating the detection of buildings than the current technique of extracting buildings from photogrammetry [2]. The demand for using LIDAR in these fields comes partially from the development of algorithms used to classify LIDAR point clouds. The scenarios that pose the greatest problems are complex urban landscapes, irregular building shapes (e.g., buildings with several floors, patios, stairs or squares) and discontinuities in the field (break-lines). In this sense, the proposed classification algorithms [3] can be sorted using one of the following methods:

A. According to the Data Structure

There are many algorithms that work with the raw point cloud [4-5], whereas other authors [6-9] resample the point cloud based on a mesh with the aim of classifying data in a more optimal and efficient way.

B. According to the Neighbourhood

Point to Point [10-11] is a method in which two points are compared to each other, and the discriminant function is based on the position of both points. If the obtained result is greater than a certain threshold, then one of the points is assumed to belong to a particular class. The greatest drawback to this method is that only one point is classified during each iteration.

Point to Points [4, 12-13] is a method in which the points neighbouring the point of interest are used to solve the discriminant function, and only one point is classified during each iteration.

Points to Points [5-7, 9] is a method in which several points are used to solve the discriminant function, and more than one point is sorted during each iteration.

C. According to the Initial Hypothesis

To use this method, neighbouring points must be adapted to a given parametric surface.

According to the hypothesis of the slope [10-11], where the slope or height difference is measured between two points. If the slope exceeds a certain threshold, then the highest point is sorted into a particular class. In [9], the initial hypothesis is a horizontal plane against which the differences in height are related. In [4-5, 7-9, 13], the discriminant function of the initial hypothesis is a parametric surface, which will act as a reference to establish the height differences among points.

D. According to the Calculation Method

One-step methods [10-11], in which the classification problem is solved linearly without requiring iteration, result in reductions in both computation time and dedicated memory.

Iterative methods [4-7, 9, 13], in which a nonlinear, and therefore iterative, approach is used, yield better results but consume more computational resources.

E. According to the Modification of the Point

Point Removal methods [4-5, 10-11, 13] remove points lying outside of the dominant function of the original cloud (irregular clouds).

Point Replacement methods [6-7, 9] replace the height of a point with a different height determined by interpolation. This type of approach is commonly used in uniform clouds.

The classification of LIDAR points in urban areas is conducted using automatic building extraction algorithms. Currently, the goal of total and automatic building classification has not been met, mainly due to the complexity of the urban scene [14]. Automatic building extraction can be conducted either using only LIDAR data or using LIDAR data supplemented with orthoimages. Studies have been conducted that combined both strategies, using the LIDAR data to classify buildings, and the images to identify and differentiate between vegetation areas and buildings [2, 15]. This strategy, however, has lacked horizontal accuracy in the detected buildings, especially along their edges [16]. As discussed in [17], it is difficult to extract a straight line accurately using only LIDAR data because the data resolution directly influences the quality of the geometric extraction [18]. To solve this problem, [19] used a methodology to extract buildings from LIDAR point clouds using the Hough´s Transformation [20]. Other authors [14, 21-24] have used integration techniques to incorporate both LIDAR data and images into the building extraction process. This strategy lends greater horizontal accuracy to the building detection. In particular, [21] applied a pixel-based classification strategy using a Normalised Digital Surface Model (NDSM) is used as an additional channel of aerial images, while [23] used aerial images and the point cloud to extract straight lines around the buildings by analysing the angle of the dominant line.

In this paper, a case study is presented that demonstrates the creation of a 3D city model from a LIDAR point cloud, with an emphasis on increased quality and automation. Following this brief introduction to LIDAR point classification algorithms and building extraction, the next section details the proposed LIDAR data processing methodology with special emphasis placed on the employed point classification algorithm and the building extraction. A practical case study focused on the town of Villalba (Madrid) is subsequently discussed, and finally, the most relevant conclusions and future perspectives are presented.

II. FROM POINT CLOUDS TO 3D CITY MODELS

To generate a DTM or DSM, extract buildings or model a 3D city, it is necessary to have a good classification of LIDAR data. This classification can take the form of either a simple discrimination between terrain points and non-terrain points or a more sophisticated discrimination of vegetation, buildings and invalid points. Points collected on flat surfaces are regularly distributed, and the differences in elevation between neighbouring points are smooth, linear and continuous. By contrast, the points obtained in rough terrain, wooded or urban areas, where there are several surface changes, have greater differences in elevation between neighbouring points and significant discontinuities in the data. Additionally, LIDAR data return a specific pattern depending on a surface's physical characteristics and materials. On metallic objects and glass, where the surfaces and reflections are small, the points usually appear in isolated groups. On roads and smooth surfaces, where the points are uniformly spaced and the differences in elevation between neighbouring points are small (less than 0.2 m), only a return echo is generated.

Depending on the type of roof, building surfaces are commonly smooth and regular, and they produce a single return (echo). Additionally, elevation differences between neighbouring point clouds will be considerable. In areas of vegetation, multiple echoes are commonly generated, and the resulting zone is characterised by an irregular distribution of points. There are also elevation differences between the clouds of neighbouring points. In water areas, the mirror-like surface behaviour reflects echoes away from the sensor, and there will be no collected information in these areas. Therefore, according to the number of returns (or echoes), an area that includes an echo can be classified as a ground surface, the roof of a building, or the top cover area of dense vegetation; similarly, if there are multiple echoes, the area can be classified as medium to high vegetation or a building edge. Finally, if there is no return, then the given area has a specular behaviour (e.g., water).

The following table (Table I) lists the possible LIDAR data point classifications.

TABLE I. DIFFERENT LIDAR DATA POINT CLASSIFICATIONS

Classes Of Points In Lidar Data			
Wrong information			
Low (Blunders)			
Ground			
Road			
Vegetation	High	Medium	Low
Building			

A. Point Cloud Classification

After considering all of the approaches mentioned in the introduction, it was decided that the methodology proposed in [12] be used for LIDAR data classification. This methodology is an automatic process requiring the user to only input critical parameters for data point classification. Additionally, this process is iterative and begins with an initial surface generated from a randomly chosen set of points. This set of points is triangulated and constituted as a reference surface (TIN - Triangulated Irregular Network). Subsequently, new points will be added only if they satisfy the established thresholds. The densification parameters of the TIN, the distance to the faces of the TIN and the vertices angles are derived from LIDAR data based on a simple statistical analysis using the minimum, median and maximum values of histograms.

Another important characteristic is the typology of the area; in forest areas, there are different characteristics and morphologies than in urban areas. In forest areas, variations in the terrain will be more continuous, whereas in urban areas, a flat surface with occasional discontinuities will be the norm. During each iteration, a point will be added to the TIN reference surface if it satisfies the distance threshold and angle criteria. These thresholds are adaptive and update after each iteration. This iterative process continues until there are no more points to add (Fig. 1 and Fig. 2).

The outline of the algorithm is as follows:

1) Establish the initial TIN parameters, distance thresholds (to the faces of the TIN) and angles (that are formed with the nodes).

2) Select points that define the initial surface from a random sample.

3) TIN iterative densification:

a) Calculate the parameters used for each iteration from the points added to the TIN.

b) Points are added to the TIN if they are within the threshold values set for distance and angle.

4) Repeat until all points have been classified as a terrain or object.

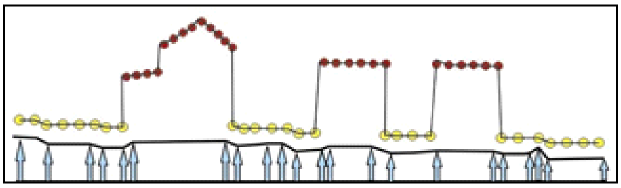


Figure 1. Example of building the adaptive TIN model of Axelsson

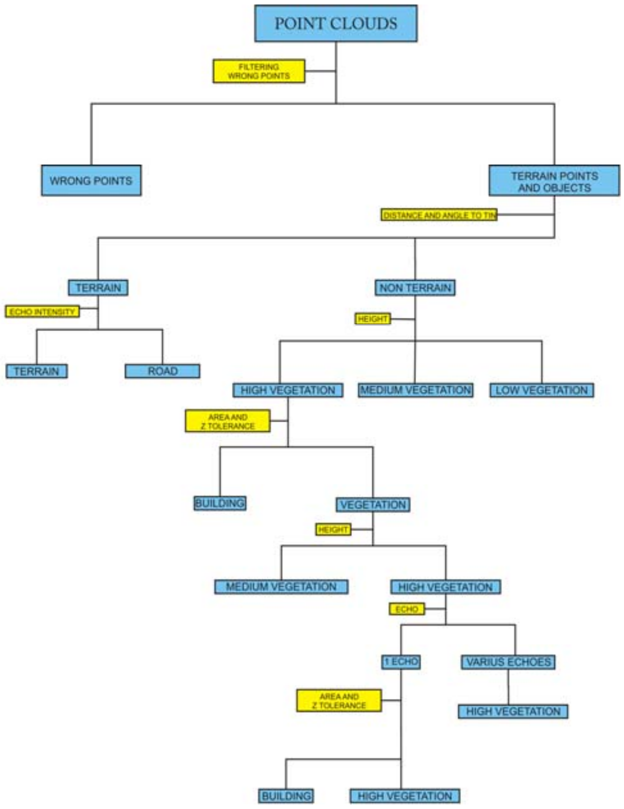


Figure 2. Workflow of the point cloud classification

B. *Automatic Building Reconstruction*

The automatic detection of buildings has advanced in recent years, but the available algorithms are still not fully autonomous because they always require human operator intervention. For this reason, the fully automatic acquisition of vector models for cities is still a challenge. The ultimate goal is to represent all of the entities in a city using a three-dimensional model while retaining the entities' physical characteristics of size and shape. Methods for extracting building models can be classified according to the input data: (i) reconstruction based on original data, without interpolation or generation of a regular grid [25]; (ii) reconstruction based on an interpolation of the original data [26]; (iii) reconstruction based on LIDAR data and image registration [21, 23]; or (iv) the direct extraction of parametric shapes such as planes, cylinders and spheres.

In this study, as mentioned in the introduction, the reconstruction was exclusively based on the use of the point cloud data without using any images as support for the automatic reconstruction [27]. This method was chosen due to a lack of suitable and geo-referenced images of the case study area that could support the automatic reconstruction process. For this reason, the available images were only used for radiometric mapping and manual edge correction.

There are many architectural objects that can be represented as flat shapes, cylinders and spheres, which allows the objects to be described using controllable parameters while also allowing them to be extracted using robust methods that detect groups in a parameter space. This study focuses on the extraction of planes, the most common elements in architectural construction. In the ideal case of a noise-free plane point cloud, all locally orthogonal surfaces should point in the same direction. Given a sufficiently reliable and discreet class, the plane extraction corresponding to the roofs of buildings was conducted using the Hough Transformation [20] extrapolated to a three-dimensional context and consistent in the parameterisation of a set of points defined initially in LIDAR space (*O'XYZ*) to a parameter space (*O'abd*).

Given a plane defined by the analytical equation $Ax + By + Cz + D = 0$, the equation can be transformed with reference to the z coordinate of the LIDAR point cloud as follows:

$$z = -\frac{A}{C}x - \frac{B}{C}y - \frac{D}{C} \qquad (1)$$

Therefore, a Z plane in the LIDAR space can be defined as a point *(a, b, d)* in the parameter space:

$$\mathrm{z} = \mathrm{ax} + \mathrm{by} + \mathrm{d}, \qquad (2)$$

with a =- A / C, b =- B / C, d =- D / C, and where xyz are the coordinates of a point belonging to the LIDAR space, ab are the coordinates of the point in the parameter space, and d is the distance from the point to the Z plane.

Although the classification process filters and optimises the building class, the number of points makes it impossible to undertake a raw parameterisation. Therefore, it was decided to use the robust estimator RANSAC (Random Sample Consensus) on a LIDAR point cloud in a three-dimensional space to find the best existing plane. For this purpose, RANSAC selects three random points from the building class and calculates the parameters of the plane that they constitute. It then detects all of the points of the LIDAR cloud belonging to the random plane using a certain threshold, usually the orthogonal distance to the plane. RANSAC repeats this process *N* times, each time comparing the resulting plane with the previously calculated one and keeping the plane that contains the most points. Ideally, the extracted plane will be obtained along with a list of possible out of plane points (outliers).

In particular, the algorithm needs the following three inputs:

1) The point cloud classified into the building class

2) A tolerance threshold according to *t,* which is the distance between the chosen plane and the rest of the points that takes into account the altimetric uncertainty associated with the LIDAR data.

3) The maximum number of probable points belonging to a single plane, which is deduced from the point density and the general characteristics of the object to be extracted.

Additionally, it is important to note that the number of RANSAC random combinations (*N*) can be considered an input parameter, or it can be calculated directly if a minimum probability of finding at least one set of observations is determined using the following equation:

$$P_{\min} = 1-(1-(1-\varepsilon)^{U})^{t_{\min}}, \quad (3)$$

where the minimum number of combinations is determined based on the number of unknowns, *U* (*U* = 3 in our case), and ε is the expected percentage of gross errors for a specified probability (in this case 95%). The number of random combinations can then be found as:

$$N(P_{\min},\varepsilon,U) = \frac{\ln(1-P_{\min})}{\ln(1-(1-\varepsilon)^{U})} \quad (4)$$

The following table (Table II) shows the necessary number of combinations to guarantee the correct solution under a certain probability and depending on the number of unknowns that we want to resolve [28].

TABLE II. NUMBER OF COMBINATIONS (N) REQUIRED IN RANSAC PROCESS FOR A GIVEN PROBABILITY (ε) AND A NUMBER OF UNKNOWNS (U). HIGHLIGHTING SHOWS THE NUMBER OF COMBINATIONS SELECTED IN OUR CASE.

N	ε =0.1	0.2	0.4	0.6	0.8	0.9
U = 1	2	2	4	6	14	29
2	2	3	7	18	74	299
3	3	5	13	46	373	2995
4	3	6	22	116	1871	29956
5	4	8	38	292	9361	299572
6	4	10	63	730	46807	2995731
7	5	13	106	1827	234041	29957322
8	6	17	177	4570	1170207	299573226

For the RANSAC algorithm to be successfully applied in a three-dimensional context (the LIDAR point cloud), the set of those points considered in the plane extraction will be excluded from the original point cloud after each iteration. This process is repeated until the number of points extracted or unmodelled falls below a certain threshold. In this way, every point belongs to only a single plane, and therefore, a point contributes to only the adjustment of the plane to which it belongs.

To obtain a plane extraction as accurately as possible, each RANSAC combination selected as a plane is iteratively adjusted through the use of a least squares calculation using all of the points belonging to the RANSAC selected combination. In this case, the least squares criterion is the orthogonal distance to the extracted plane.

Considering that most of the building roofs will not be constituted by a single plane, it is necessary to determine the intersection lines among planes and, thereby, the entire structure of the building eaves. For this purpose, the equation for the intersection between two planes is used. Given two planes, π_1:$A_1x+B_1y+C_1z=0$ and π_2:$A_2x+B_2y+C_2z=0$, their intersection may be defined by the line *l*. The direction vector of the line *l* is calculated using the cross product of the normal vectors of both planes:

$$\pi_1 x \pi_2=(A_1,B_1,C_1)x(A_2,B_2,C_2), \quad (5)$$

where (A_1, B_1, C_1) are the parameters of the normal vector to the plane π_1 and (A_2,B_2,C_2) are the parameters of the normal vector to the plane π_2.

As the axis of the line *l* is not uniquely defined by the vector obtained in (5), it is necessary to obtain a point p_0 = (x_1, y_1, z_1) that belongs to both planes and thus to the line *l*. This point is achieved by restricting the value of one of the coordinates (e.g., *z* = 0) and solving the resulting system with two equations and two unknowns.

Finally, the volume of the building is generated through an orthogonal projection over the DTM of the different edges of the extracted planes. Unfortunately, this automatic process yielded no definitive results and the model suffered from errors. Therefore, each building was checked, and multiple errors were removed using manual tools.

III. EXPERIMENTAL RESULTS: THE CASE STUDY OF VILLALBA

The case study was conducted on the town of Collado Villalba in the Province of Madrid (Spain). The centre of the town of Collado Villalba contains a consolidated area with a density of 2062.57 inhabitants/km^2. In this space, there are large areas with houses as well as areas with abundant vegetation. The project focused on an area of approximately 2 km^2 that contained gentle slopes, high vegetation and buildings of multiple dimensions and heights. This region yielded a cloud of approximately 4 million raw data points.

The LIDAR data used in this study were taken with the Leica ALS sensor 50_II (Table III). This sensor is an airborne laser scanner with a 95.8 kHz pulse rate (95800 pulses per second), an opening angle of 45° and a capture height of approximately 1000 m from the ground. The data are generated in the LAS1.0 free format, with an average density of 1.9 points per square meter, with an initial vertical accuracy of 12 cm. An average overlap between transversal scan passes of 32.33% was determined.

TABLE III. TECHNICAL CHARACTERISTICS OF THE DATA COLLECTED WITH THE 50_II SN48 ALS SENSOR.

Sensor	ALS 50_II SN48
Laser Pulse Rate Used	95800.00 Hz
Scan FOV (half angle)	22.50 (Deg.)
Point Density (average)	1.87 point/m^2
Estimated error Z	0.12 m
Overlap	32.33%
Terrain Elevation AMSL (minimum in survey area)	850 m
Terrain Elevation AMSL (maximum in survey area)	1050 m
Nominal Flying Height Above Minimum Terrain Elevation	1000 m
Nominal Flying Altitude	1850 m
Assumed GPS Error	0.05 m

The initial pre-processing tasks have been excluded from this article as they are outside of the article scope. The pre-processing comprised the identification of overlapped areas between runs, their alignment and the subsequent removal of redundant information for those points belonging to adjacent runs. Additionally, all of the points with a weak signal or systematic failures, points below the terrain class, or noise points such as birds and moving objects were deleted.

In this study, the data point classification was conducted using a bottom-up hierarchical strategy. This classification allowed an initial discretisation of the terrain points and non-terrain points and then focused on discerning non-terrain points into the following entities: vegetation (low, medium and high) and buildings. Specifically, in the case study, we begin with an initial terrain points model composed of the lowest elevation points. An initial TIN is generated based on Delaunay triangulation, and this TIN allows us to establish the reference surface. The triangles in this initial model are mostly below the true ground surface. The routine then begins to analyse the model iteratively from the bottom up, adding new points as it progresses.

Each added point makes the model of the soil surface closer to the true terrain. The parameters that during each iteration whether a point is added to the soil surface during each iteration are the angular parameter, consisting of the maximum angle between a candidate point and the nearest triangle vertices, and the distance parameter, which is the orthogonal distance between a candidate point and the closest triangular plane. These parameters are suitable for a terrain with smooth and continuous break lines. In the case study, these parameters were set as 6 degrees for the threshold angle and 1.4 m for the threshold distance. These parameters can be varied for adapting to other types of land if it is very abrupt. A total of 361,086 points were classified as terrain (Fig. 3), and the remaining non-terrain points were classified as either vegetation or building.

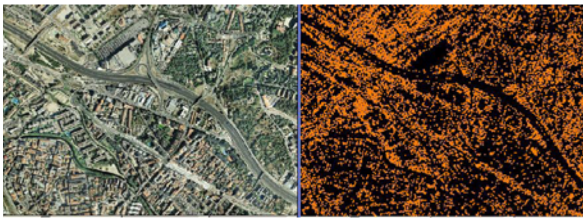


Figure 3. Results of the terrain points filtering process (right) and the validation with the orthophotograpy (left).

Considering that the points of the roads and highways have been classified as area-points due to the similarity between their characteristics and the ground-points, the two categories must be differentiated. This is accomplished using the response intensity, which depends on the material surface [4]. The reflectivity value of each point is represented as gray levels from 0 to 255, where 0 corresponds to no light incident on the sensor, and 255 is the maximum reflectivity. The average intensity of the reflected signal for a road corresponds to a gray level of 55.26, while the average for the ground is 148.13. In this case study, there was a double threshold of intensity values (40-100) that filtered out those points that belong to highways or roads. As a result, 185,430 points were classified as road (Fig. 4).

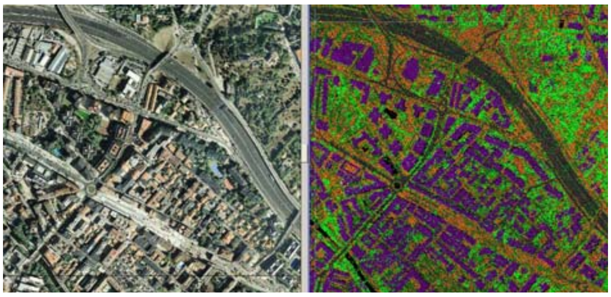


Figure 4. Classification of the road entity in black (right) and the validation with the orthophotography (left).

Subsequently, the non-terrain class was divided into the different types of vegetation entities. For that classification, a height threshold was used to classify a point into the following three possible types of vegetation: low vegetation (from 0.01 to 0.2 m), medium vegetation (from 0.2 to 3 m), and high vegetation (from 3 to 150 meters). In total, 50,543 points were classified as low vegetation, 331,291 points as medium vegetation, and 746,932 points as high vegetation (Fig. 5).

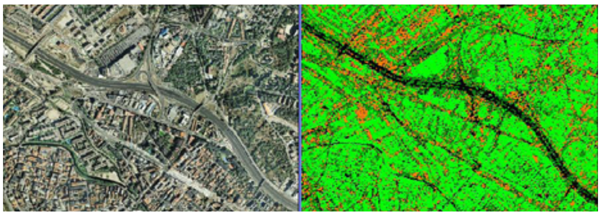


Figure 5. Classification of the vegetation entity represented in three shades of green (right) and validation with the orthophoto (left).

In regards to the building classification, it should be mentioned that the area of study is an urban area in which there are residential buildings between 3 and 5 plants with heights ranging from 8 to 24 meters. For the initial building classification, the following two thresholds were set: the minimum area of the building and the tolerance in height. For the initial classification process of the building class, an initial threshold was set at 40 m^2 for the building area and 0.65 m for the minimum height. Additionally, any points classified as a building had to only have a single return (echo). As a result of this process, 402,055 points were classified as belonging to the building class.

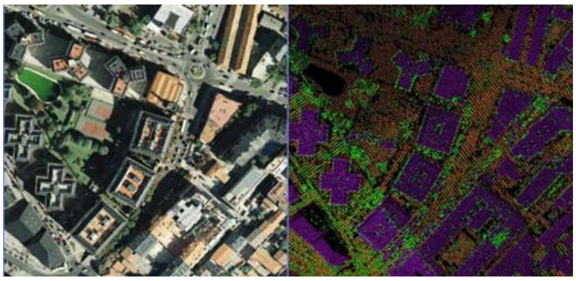


Figure 6. Classification of the buildings in purple (right) and validation with orthophotography (left).

As shown in Fig. 6, the established thresholds incorrectly classified the edges and the roofs of buildings as high vegetation. Therefore, manual intervention by the user was necessary to allow for a more accurate classification of the building class.

Once all of the data points had been classified, the next step was to apply the reconstruction algorithms to the buildings. This task, performed automatically, yielded incomplete results and numerous errors. It was necessary to manually supervise the process via photo-interpretation using the orthophotos as well as via a CAD cleansing task that fixed the closure and connection between the different planes of certain buildings. The following figures (Fig. 7-10) illustrate some of the common bug fixes that were required during the 3D building reconstruction. The data processing shown in this study case has been rather expensive in computational terms, i.e., data point classification needed 4 hours, reconstruction algorithms required 8 hours, and for CAD design 3 days were spent.

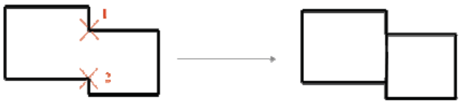


Figure 7. In this case, two buildings are defined instead of one.

.

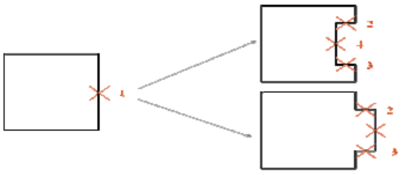


Figure 8. The correction of incoming or outgoing roofs

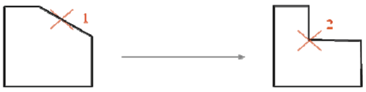


Figure 9. The correction of a building corner definition

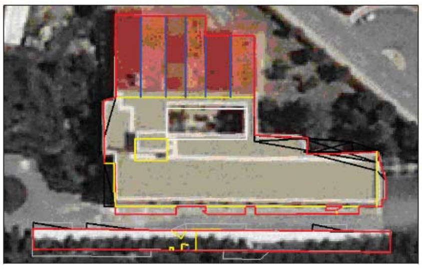


Figure 10. Common errors in the automatic vectorisation of the building edges and their corrections.

Fig. 11 presents the results of the building reconstruction after the photo interpretation and CAD debug phases.

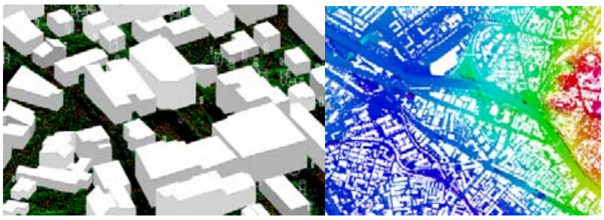


Figure 11. 3D CAD model of the buildings: detailed perspective view (left) and overview on the ground (right)

IV. Concluding Remarks and Future Perspectives

This study presented the generation of a 3D city model from LIDAR data taken over the urban centre of the city of Collado Villalba, Madrid (Spain). To obtain DTM and DSM of sufficient quality, it is crucial to utilise efficient algorithms that automatically classify the raw data points. Working with efficient algorithms is also critical for automatic building reconstruction and model triangulation. A classification methodology that operates according to the geometric features and proprieties of the data points has been presented alongside a robust strategy for extracting the roof planes of buildings. Although significant levels of automation have been attained, it is still necessary to manually correct certain errors relating to building vectorisation. The automatic vectorisation was ultimately successful and able to distinguish close buildings, although the time spent correcting model errors was very high and necessitates the further development of more efficient algorithms. Regardless of its shortcomings, a 3D city model was created that allows a multitude of applications in the field of urbanism and planning. In conclusion, the use of LIDAR data to create three-dimensional models of urban areas is valid provided that a sufficient point density (greater than 2 points/m^2) is utilised and the model is produced under the supervision of an operator capable of manually correcting any errors that arise.

In relation to the future perspectives, there is two relevant issues to be considered: (i) the quality of the building reconstruction process, which could be solved through the evaluation of different TIN interpolation methods, such as the inverse distance weighting (IDW), spline or Kriging algorithms; (ii) the accuracy assessment of final DEM and final building reconstruction by the use of Geomatics techniques (terrestrial laser scanner) as ground truth for evaluation purposes.

REFERENCES

[1] E. P. Baltsavias, "A comparison between photogrammetry and laser scanning," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 54, pp. 83-94, 1999.

[2] T. T. Vu, F. Yamazaki, and M. Matsuoka, "Multi-scale solution for building extraction from LiDAR and image data," International Journal of Applied Earth Observation and Geoinformation, vol. 11, pp. 281-289, 2009.

[3] G. Sithole and G. Vosselman, "Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 59, pp. 85-101, 2004.

[4] P. Axelsson, "Processing of laser scanner data--algorithms and applications," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 54, pp. 138-147, 1999.

[5] N. Pfeifer, P. Stadler, and C. Briese, "Derivation of digital terrain models in the SCOP++ environment," in OEEPE Workshop on Airborne Laserscanning and Interferometric SAR for Digital Elevation Models, Stockholm, Sweden, 2001.

[6] M. Brovelli, M. Cannata, and U. Longoni, "Managing and processing LIDAR data within GRASS," in Open source GIS - GRASS users conference, Trento, Italy, 2002.

[7] M. Elmqvist, "Ground estimation of lasar radar data using active shape models," in OEEPE Workshop on Airborne Laserscanning and Interferometric SAR for Digital Elevation Models, Stockholm, Sweden, 2001.

[8] M. Elmqvist, E. Jungert, F. Lantz, A. Persson, and U. Söderman, "Terrain modelling and analysis using laser scanner data," in ISPRS Workshop - Land Surface Mapping and Characterization using laser altimetry, Annapolis, Maryland, 2001, pp. 219-226.

[9] R. Wack and A. Wimmer, "Digital Terrain Models from Airborne Laserscanner Data-a Grid Based Approach," in Photogrammetric Computer Vision (PCV02), Graz, Austria, 2002, pp. 293-296.

[10] G. Sithole, "Filtering of laser altimetry data using a slope adaptative filter," in ISPRS Workshop - Land Surface Mapping and Characterization using laser altimetry, Annapolis, Maryland, 2001, pp. 203-210.

[11] M. Roggero, "Airborne laser scanning: clustering in raw data," in ISPRS Workshop - Land Surface Mapping and Characterization using laser altimetry, Annapolis, Maryland, 2001, pp. 227-232.

[12] P. Axelsson, "DEM generation from laser scanner data using adaptive TIN models," in XIXth ISPRS Congress, Amsterdam, The Netherlands, 2000, pp. 110-117.

[13] G. Sohn and I. Dowman, "Terrain surface reconstruction by the use of tetrahedron model with the MDL Criterion," in Photogrammetric Computer Vision (PCV02), Graz, Austria, 2002, pp. 336-344.

[14] D. H. Lee, K. M. Lee, and S. U. Lee, "Fusion of lidar and imagery for reliable building extraction," Photogrammetric Engineering and Remote Sensing, vol. 74, pp. 215-225, 2008.

[15] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Using the Dempster-Shafer method for the fusion of LIDAR data and multi-spectral images for building detection," Information Fusion, vol. 6, pp. 283-300, 2005.

[16] Y. Li and H. Wu, "Adaptive building edge detection by combining lidar data and aerial images," in XXIst ISPRS Congress, Beijing, China, 2008, pp. 197-202.

[17] L. Cheng, J. Gong, X. Chen, and P. Han, "Building boundary extraction from high resolution imagery and lidar data," in XXIst ISPRS Congress, Beijing, China, 2008, pp. 693–698.

[18] A. Sampath and J. Shan, "Building boundary tracing and regularization from airborne lidar point clouds," Photogrammetric Engineering & Remote Sensing vol. 73, pp. 805-812, 2007.

[19] G. Vosselman, "Building reconstruction using planar faces in very high density height data," in ISPRS Workshop - Automatic Objects from Digital Imagery, Munich, Germany, 1999, pp. 87-94.

[20] P. V. C. Hough, "Method and means for recognizing complex patterns," U.S. Patent 3.069.654, 1962.

[21] N. Haala and C. Brenner, "Extraction of buildings and trees in urban environments," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 54, pp. 130-137, 1999.

[22] L. C. Chen, T. A. Teo, Y. C. Shao, Y. C. Lai, and J. Y. Rau, "Fusion of LIDAR data and optical imagery for building modeling," in XXth ISPRS Congress, Istanbul, Turkey., 2004, pp. 732-737.

[23] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 62, pp. 43-63, 2007.

[24] N. Demir, D. Poli, and E. Baltsavias, "Extraction of buildings using images & LIDAR data and a combination of various methods," in Object Extraction for 3D City Models, Road Databases and Traffic Monitoring - Concepts, Algorithms and Evaluation (CMRT09), Paris, France, 2009, pp. 71-76.

[25] H.-G. Maas and G. Vosselman, "Two algorithms for extracting building models from raw laser altimetry data," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 54, pp. 153-163, 1999.

[26] F. Rottensteiner and C. Briese, "A New Method for Building Extraction in Urban Areas from High-Resolution LIDAR Data," in Photogrammetric Computer Vision (PCV02), Graz, Austria, 2002, pp. 295-301.

[27] G. Vosselman and S. Dijkman, "3D building Model Reconstruction from Point Clouds and Ground Plans," in ISPRS Workshop - Land Surface Mapping and Characterization using laser altimetry, Annapolis, Maryland, 2001, pp. 37-44.

[28] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. New York: Cambridge University Press 2003.

A Novel Health Monitoring System using Patient Trajectory Analysis: Challenges and Opportunities

Shayma Alkobaisi
Faculty of Information Technology
United Arab Emirates University
Al Ain, UAE
Email: shayma.alkobaisi@uaeu.ac.ae

Wan D. Bae
Mathematics, Statistics and Computer Science
University of Wisconsin-Stout
Menomonie, WI, USA
Email: baew@uwstout.edu

Sada Narayanappa
Global Navigation Services
Jeppesen, A Boeing Company
Denver, CO, USA
Email: sada.narayanappa@jeppesen.com

Cheng C. Liu
Engineering and Technology Department
University of Wisconsin-Stout
Menomonie, WI, USA
Email: liuc@uwstout.edu

***Abstract*—Continued advances and cost reduction in mobile devices such as smart phones made them widely used in our daily-life practices such as en route navigation and vehicle tracking. Health applications utilizing these battery-powered devices continue to grow, and so does the demand for effective modeling and analysis tools to support data collected by these devices. Health monitoring applications in particular became very popular these days. However, researchers must overcome many challenges, such as data acquisition, data scales and data uncertainty, in order to develop such applications. In this paper, we propose a novel health monitoring system that can interact with the patient and analyze the patient's moving trajectories combined with data of environmental conditions. We present a system architecture, and discuss ideas and challenges in developing the health monitoring system for asthma patients. This system can provide a better understanding of the effect of environmental factors on triggering health attacks and hence support individual-based health care.**

***Index Terms*—health monitoring, uncertain trajectories, environmental factors, asthma, road networks**

I. Introduction

Relations between negative health effects like asthma and lung cancer and elevated levels of the environmental factors, such as air pollution, tobacco smoke and humidity, have been detected in several large scale exposure studies [8]. Thus, public health care and service systems often require the ability to track, monitor, and analyze patients' trajectories and their relationships with several environmental factors in order to derive conclusions that will help in preventing and treating diseases. Health applications dealing with large volume of continuously moving data objects, such as humans and vehicles continue to grow. However, these applications present significant challenges in terms of data size, data scales, complex structures and relationships, uncertainty, and space and time constraints. Tracking moving objects has been a hot issue recently due to the large number of applications that depend heavily on it. However, individual monitoring of exposure to environmental conditions did not follow the same pace despite its great impact on public health; the general effect on earth has more been the concern. Limited research has been done on techniques for retrieving, storing and analyzing real-time data of patients along with the environmental conditions patients are exposed to.

The main objective of this research is to improve public health care through proposing ideas and directions to develop an effective and efficient real-time health monitoring system that can report potential health threats (e.g., asthma attacks) associated with environmental conditions, support individual's long-term health care management, reduce the cost, effort and time spent in traditional health visits to hospitals, and provide intelligent information that might be useful for improving public health care plans and strategies. Although we are targeting asthma in this paper as it is well known that asthma is highly affected by surrounding environmental conditions [2], [13], our proposed system can be used in improving the general well-being as well as targeting other diseases that are affected by the environment.

This paper focuses on the two main components in developing our proposed system that takes into account the correlation between the time and location of a patient and the level and time of exposure to negative environmental factors; these are patient trajectory tracking and environmental exposure measurement. The first component can be obtained by location tracking devices such as the GPS. The second component not only includes air pollution level but also other measurements, such as humidity and temperature levels, that can be normal to healthy people but not to asthma patients. Finally, we discuss challenges and opportunities to develop the proposed system and conclude the paper.

II. Motivation

A. Health and the Environment

According to the World Health Organization, asthma is now a serious public health problem with over 100 million sufferers worldwide. It continues to be one of the major causes of

ISBN: 978-1-61208-178-6

hospitalization of children in many countries. The number of reported adults and children diagnosed with Asthma in the U.S. in 2009 was 17.5 million and 7.1 million, respectively [17], [6]. Moreover, the number of visits (to physician offices, hospital outpatient and emergency departments) with asthma as primary diagnosis in 2007 was 17.0 million in the U.S. only [19]. In the same year, the number of discharges with asthma as first-listed diagnosis was 456,000 with an average length of stay being 3.4 days [14]. Although scientific advances have led to effective medical interventions to prevent morbidity of the disease, the burden in prevalence, mortality and health care use remains high.

Research has identified several factors associated with the development of asthma, such as exposure to traffic exhaust fumes, tobacco smoke, pesticides and changes in the weather, but none have proven to be the causative agent [2]. Rather the development is a combination of underlying susceptibility with environmental exposures [2], [13]. In addition, these environmental factors associated with asthma have been measured on a general basis, i.e., they are based on summarized data collected in large scale areas (e.g., city), and not based on individual exposure which would more accurately reflect the exposure to such factors at a specific location and time. Moreover, asthma triggers vary and can be very different from one patient to another. Thus, individual-based measurement of exposure is needed to be able to develop more accurate conclusions on causes of asthma attacks.

B. *Health Monitoring*

There exist some powerful health monitoring systems in this advanced era of information technology. These systems range from smart homes [9] that consist of several intelligent devices built in homes which are able to monitor and provide help to elderly and disabled people, to very tiny sensor chips that can be implanted in the body of patients to provide continuous monitoring of blood pressure, sugar level and other measurements [7], [21]. These types of health monitoring systems are very useful but limited to some conditions. For example, smart homes are useful for patients who spend most of their time at home (e.g., elderly). On the other hand, implanted sensor chips are able of continuously monitoring patients regardless of their locations. However, they are limited by high cost, patients' willingness for a device implant and patients' health condition. In addition, health monitoring may not work for some diseases using these sensors since the development of their attacks may not be measured by sensors. In the case of asthma, for example, it is not enough to depend on physical measurement of implanted sensors as many other factors play a role in triggering an attack. For example, it can be a result of environmental factors as stated earlier.

III. Asthma Monitoring System

This section proposes a health monitoring system that can track the trajectories of an asthma patient in a geographic region and various environmental factors associated with that region, and analyze these data along with the patient health level (peak flow level, for example). The system will be able to retrieve individual patient's location data and several environmental resource data through mobile phones or sensors. Integration of these spatio-temporal data can support the calculation of the patient's exposure to certain levels of the environmental conditions. Using statistical analysis and efficient data mining algorithms to retrieve intelligence information from relations between the patient's locations within time and various environmental conditions, the system can identify the potential asthma attack and provide useful information to the patient that may prevent asthma attacks. The main advantages of the proposed system can be summarized as follows: (1) Continuous monitoring and early attack detection, (2) Data analysis on individual-level to provide risk alerts, (3) Long term treatment based on spatio-temporal data analysis, and (4) Reducing time, effort and cost spent on emergency visits to hospitals and clinics.

A. *Architecture Overview*

The context and overall architecture of the proposed system are illustrated in Fig. 1. The system consists of data acquisition to acquire data and data integration from different sources to provide user and environment profiles. The internal reference data consist of maps, road networks, etc.

The system enables collecting data from users by various sources. For example, the user's mobile device connected to the GPS is a good source of user location and trajectory. In addition, the user's interaction with social networks or communication via mobile phone (phone number tied to business) etc., are possible data collection methods. The communication via mobile and sensor networks will provide other information of the user such as the current health condition of the user (e.g., users post that they are feeling difficulty in breathing or chest wheezing). These unstructured data can be used to predict the environmental conditions such as traffic, congestion, smoke level, etc. Environmental data, such as air pollution, tobacco smoke, temperature and humidity, can be obtained from standard sensors. Moreover, the user profile tied to user mobile devices provides intimate profile of the user.

Based on the input, area of interest, reference data trajectory and trend analysis, the system computes the interesting output based on the rules configured in the system. The system will provide the output to external systems such as health awareness broadcast systems, public safety systems, pollution control systems, etc. Several rules of filtering and inference provide specific configurations to change the system's behavior without having to change the entire system. Specifications on data acquisition and data analysis parts of the system will be discussed further in the opportunities section.

B. *Challenges*

A number of challenges exist in developing an online health analytical infrastructure that matches the proposed health monitoring system. The main challenge is the large scale of the proposed system; It requires major public participation/outreach efforts in order to (a) obtain active participation

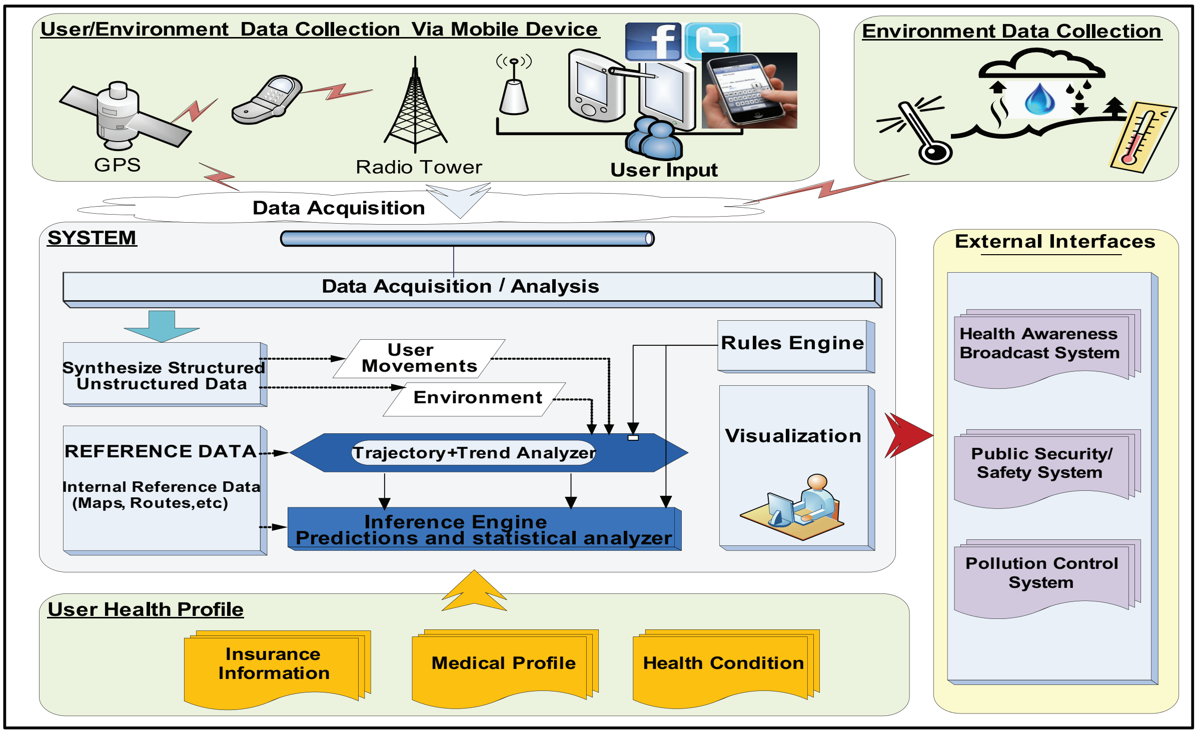

User/Environment Data Collection Via Mobile Device
GPS
Radio Tower
User Input
Environment Data Collection
Data Acquisition
SYSTEM
Data Acquisition / Analysis
Synthesize Structured Unstructured Data
User Movements
Environment
Rules Engine
Visualization
REFERENCE DATA
Internal Reference Data (Maps, Routes,etc)
Trajectory+Trend Analyzer
Inference Engine Predictions and statistical analyzer
External Interfaces
Health Awareness Broadcast System
Public Security/ Safety System
Pollution Control System
User Health Profile
Insurance Information
Medical Profile
Health Condition

Fig. 1. A health monitoring system using patient trajectory and environmental factors

of the users and (b) limit data distortions. The first can be addressed by providing a high quality human-machine interface to the system with user friendly options to interact with the system. In addition, patients need to be educated about the advantages of such a system and its potential role in saving their lives. The second challenge necessitates providing a lot of training to overcome the expected highly biased data collected. Methods of data training and testing exist and will be tested in the future phases of this project.

Many other technical challenges exist, we only list major ones here:

- Continuous tracking of data with limited resources (e.g., sensor cover radius, mobile device life time, etc.).
- Measuring environmental conditions with a fine scale.
- Accounting for uncertainties because of errors in device measurement and data sampling.
- Integrating different representations of spatio-temporal datasets (e.g., environmental data and trajectories).
- Capturing individual exposure to a certain environmental condition.
- Designing novel spatial join and search algorithms, and accessing methods to improve query performance.
- Developing visualization tools that can be used for data analysis.
- Developing data mining tools that are useful in making conclusions about integrated data
- Protecting and securing medical data of patients.

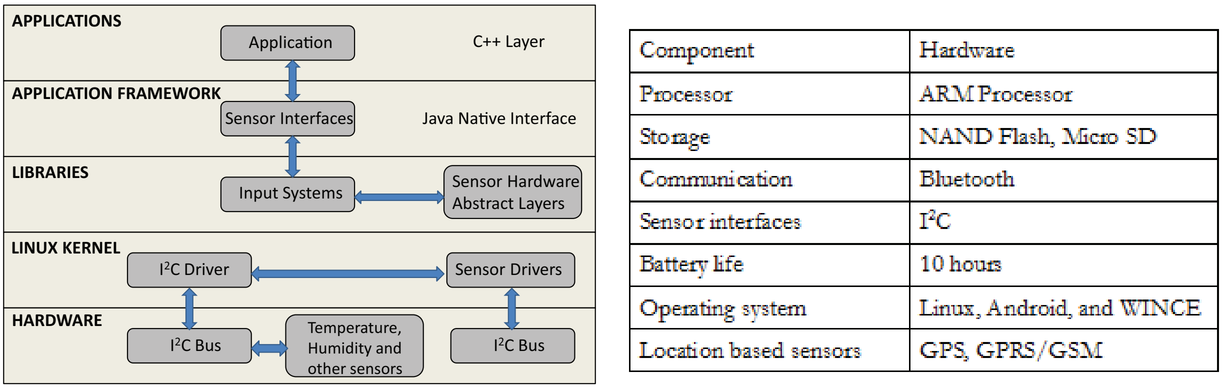
C. *Opportunities*

In this section, we discuss available opportunities as well as specifications about potential solutions to some of the raised challenges that are of high concern to us at this stage of our work. Two main data sources are required for our proposed system; measuring environmental factors using sensors and calculating the location/time of the patient using his/her trajectory data obtained by the GPS. The goal is to find some rules that will help in identifying meaningful relationship among these datasets that would be useful in measuring the effect of environmental conditions on the health of asthma patients. To achieve this goal, a geostatistical model of spatial autocorrelation need to be developed that would capture the correlation between a patient's trajectory dataset and the environmental factor datasets.

Improvement in technology has been greatly witnessed recently. For example, the development of sensors as well as GPS devices has more and more improved in terms of the miniaturization of the sensing/tracking devices and in terms of the battery life time. In addition, there is a great improvement in the wireless communication including Bluetooth, WiFi and 3G. Different wireless technologies address different system requirements such as coverage; the wireless infrastructure of the system should allow the use of diverse wireless and sensor devices in order to provide a complete support of the system's requirements.

Researchers have very recently proposed algorithms and models that allow the tracking of moving objects without consuming a lot of the tracking device's energy by adapting

ISBN: 978-1-61208-178-6



Component	Hardware
Processor	ARM Processor
Storage	NAND Flash, Micro SD
Communication	Bluetooth
Sensor interfaces	I²C
Battery life	10 hours
Operating system	Linux, Android, and WINCE
Location based sensors	GPS, GPRS/GSM

(a) Android sensor network

(b) Sensing devices

Fig. 2. Android sensor system architecture

to the movement and other parameters of the moving object [11]. The experimental results showed that their proposed system helped in acquiring trajectories of moving objects, yet consuming less battery energy of the tracking device.

D. SpatialGPSLogger: A Data Acquisition System

We now present our ongoing work towards the development of the presented system. We propose a location based data acquisition system called SpatialGPSLogger that uses the GPS on a smartphone device. Environmental data such as air pollution and tobacco smoke cannot be obtained by the smart phone device without proper wireless sensors. Most smart phone systems do not allow users to modify the internal structure and interface sensors. Therefore, a more flexible system than a smart phone is required to continue our study to collect environmental data. The embedded system and sensors as shown in Fig. 2 are used to implement our SpatialGPSLogger. The embedded system runs on an ARM processor and Android operating system and is capable of interfacing various types of sensors through its serial ports. We use this system to interface sensors to retrieve temperature, humidity, and biological signals of patients.

In our approach, SpatialGPSLogger selects a set of sensor data, which retrieves individual's location data through a mobile phone connection to GPS. We plan to port sensor drivers to the embedded system under the Android platform and create applications through the use of Linux kernel to retrieve temperature, humidity, orientation, accelerometer, light, magnetic field, proximity of patients, etc. The platform allows programmers to access raw data from the sensors through the Android sensor network to the application layer on the Android system as depicted in Fig. 2 (a). The application communicates with C++ layer through sensor Java native interfaces. The input system is a Linux framework for all sensors data that defines a standard set of events and it also interfaces to user space through Java native interface. I^2C is a two wire serial interface connection for sensor data to be accessible under Linux kernel. Fig. 2 (b) summarizes the sensing devices that are used to retrieve environmental data.

IV. Related Work

The progress in technology has inspired many to make use of it in monitoring environmental pollution. The authors in [18] used nanotechnology to develop an air pollution monitoring system. The system provides real-time interpolated maps of air quality using GIS which analyzes and displays data collected by "solid state gas sensors". As more signs show the negative effect of bad quality of the environment on health, more sources of assessing environment condition are developed such as "EnviroFlash" [3] which can notify patients about up to date air quality.

With the rapid development in sensor networks, location tracking devices and mobile networks, health care has made a great progress in utilizing these technologies in order to advance Telemedicine [16] applications as well as eHealth [10] services. Through these services, "virtual visits" of patients to doctors became possible. Some patients are now able to skip actual "face to face" visits with doctors for simple tasks like blood pressure and sugar level measurement assessment. In addition, commercial "emergency alert" systems made use of advances in technology to provide immediate help to patients with severe diseases such as heart problems and diabetes which might tackle patients as sudden attacks. An example is the "Invisible Bracelet" (iB) [1] which is supported by the American Ambulance Association that uses the mobile network technology to get immediate help.

Both environmental conditions and locations of moving objects are spatio-temporal data that are uncertain in nature. For example, the concentrate of a specific air pollutant vary from one region to another and from a given time to another and depends on the accuracy of sensors. Also, a GPS device may report a location of a moving object 2-5 meters away from the actual location. Authors in [12] discussed the spatio-temporal modelling of individual exposure to air pollution using data integration. They presented a case study example that uses a deterministic air quality model and a GPS to analyze collected data over an hour period.

Different technologies are available for located-based data acquisition but they device dependent. Still data collection relies on manual process and this hinders the analysis of individual trajectories and their relationships with environmental conditions. Hence an integration of different technologies is critical for an efficient and automated data acquisition of moving objects.

A patient's trajectory is represented as spatio-temporal data [20] and can be defined as space-time paths of a moving object constrained to road networks. Due to computation and database limitations and limited battery life of mobile devices, trajectories are modeled as discrete points that represent locations in time. This representation results in uncertainty in the location values recorded in the database of the moving object [4], [23]. Existing methods do not provide effective modeling for uncertainty of moving trajectories to support different types of queries to improve query processing.

An environmental condition (e.g., humidity-30%) is also a spatio-temporal measurement that can be modeled as a raster grid for discrete time stamps. Each cell in the grid represents the average value of the environmental factor at a given time for the area. This average value is associated with positional and temporal uncertainties because of the approximations and interpolations used in modeling [12]. These uncertainties in value and position can be characterized based on their probability distribution functions [15].

Data mining on moving objects is a process of extracting useful and interesting information from large sets of spatio-temporal datasets. Existing data mining techniques do not adequately fit the nature of moving objects that continuously change their properties in time and the complex relationships between them. Therefore flexible and scalable data mining tools are crucial in order to find accurate and sufficient information of moving objects' patterns, behaviors and trends. The integration (spatial joins) [5] between the environmental grid representation and the trajectory representation of a moving object will provide information about the exposure measurement of a patient to environmental factors. This linkage of data from different sources leads to ecological inference that has been classified as a special case of the change of support problem [22].

V. Conclusion

Health monitoring systems can be thought of as a natural extension to advances in health services and technology including mobile and sensor networks. Their main objective is to analyze real-time data collected from patients and other sources in order to provide individual-based care to patients. Such a system, if developed successfully, promises to reduce the cost, effort and time put in traditional health visits to hospitals.

In this paper, we proposed a health monitoring system that analyzes spatio-temporal data collected from patients. The system analyzes the effect of various environmental factors, such as humidity level, on the health of patients. The proposed health monitoring system gathers time/location data from patients through location aware devices such as GPS, and collects environmental data through sensors. Then, the system analyzes the data using a spatio-temporal integration model to derive conclusions that will help prevent or treat patients. Several intermediate tools are used such as data mining tools and visualization tools. Main challenges and opportunities in developing such a system were discussed.

References

[1] https://store.invisiblebracelet.org/home, Aug. 2011.

[2] http://www.cdc.gov/asthma, Aug. 2011.

[3] http://www.enviroflash.info, Aug. 2011.

[4] S. Alkobaisi, P. Vojtechovsk, W. Bae, S. Kim, and S. Leutenegger. The truncated tornado in tmbb: A spatiotemporal uncertainty model for moving objects. In *Proceedings of the 19th international conference on Database and Expert Systems Applications*, pages 33–40, 2008.

[5] W. D. Bae, P. Vojtechovsk, S. Alkobaisi, and S. L. nd S.H. Kim. An interactive framework for raster data spatial joins. In *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems*, pages 19–26, 2007.

[6] B. Bloom, R. Cohen, and G. Freeman. Summary health statistics for u.s. children: National health interview survey, 2009. *National Center for Health Statistics. Vital Health Stat*, 10(247), 2010.

[7] O. Boric-Lubeke and V. Lubecke. Wireless house calls: using communications technology for health care and monitoring. *Microwave Magazine, IEEE*, 3(3):43–48, 2002.

[8] B. Brunekreef and S. T. Holgate. Air pollution and health. *The Lancet*, 360:1233–1242, 2002.

[9] M. Chan, D. Estève, C. Escriba, and E. Campo. A review of smart homes–present state and future challenges. *Computer methods and programs in biomedicine*, 91:55–81, 2008.

[10] G. Eysenbach. What is e-health? *Journal of Medical Internet Research*, 3(2):e20, 2001.

[11] S. Fang and R. Zimmermann. EnAcq: Energy-efficient GPS Trajectory Data Acquisition Based on Improved Map Matching. In *ACM SIGSPATIAL GIS 2011*, pages 221–230, 2011.

[12] L. E. Gerharz and E. Pebesma. Accounting for uncertainties and change of support in spatio-temporal modelling of individual exposure to air pollution. In *geoENV 2010*, pages 13–15, 2010.

[13] M. I. Gilmour, M. S. Jaakkola, S. J. London, A. E. Nel, and C. A. Rogers. How exposure to environmental tobacco, smoke, outdoor air pollutants, and increased pollen burdens influences the incidence of asthma. *Environ Health Prospect*, 114(4):627–633, 2006.

[14] M. J. Hall, C. J. DeFrances, S. N. Williams, A. Golosinskiy, and A. Schwartzman. National hospital discharge survey: 2007 summary. *National Health Statistics Report*, (29), 2010.

[15] G. B. M. Heuvelink, J. D. Brown, and E. E. Loon. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, 21(5):497–513, 2007.

[16] M. M. Maheu and A. Allen. http://www.telehealth.net/glossary.html, Aug. 2011.

[17] J. Pleis, B. Ward, and J. Lucas. Summary health statistics for u.s. adults: National health interview survey, 2009. *National Center for Health Statistics. Vital Health Stat*, 10(249), 2010.

[18] O. Pummakarnchana, N. Tripathi, and J. Dutta. Air pollution monitoring and gis modeling: a new use of nanotechnology based solid state gas sensors. *Science and Technology of Advanced Materials*, 6:251–255, 2005.

[19] S. Schappert and E. Rechtsteiner. Embulatory medical care utilization estimates for 2007. *National Center for Health Statistics. Vital Health Stat*, 13(169), 2011.

[20] G. Trajcevski, O. Wolfson, F. Zhang, and S. Chamberlain. The geometry of uncertainty in moving object databases. In *proceedings of Int. Conference on Extending Database Technology*, pages 233–250, 2002.

[21] U. Varshney. Pervasive healthcare and wireless health monitoring. *Mobile Networks and Applications*, 12:113–127, 2007.

[22] L. J. Young and C. A. Gotway. Linking spatial data from different sources: the effects of change of support. *Stochastic Environmental Research and Risk Assessment*, 21(5):589–600, 2007.

[23] K. Zheng, G. Trajcevski, X. Zhou, and P. Scheuermann. Probabilistic range queries for uncertain trajectories on road networks. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 283–294, 2011.

Open source GIS Tools to Map Earthquake Damage Scenarios and to Support Emergency

Maurizio Pollino, Antonio Bruno Della Rocca,
Grazia Fattoruso, Luigi La Porta, Sergio Lo
Curzio, Agnese Arolchi
ENEA - National Agency for New Technologies,
Energy and Sustainable Economic Development
Casaccia Research Centre – Rome, Italy
Portici Research Centre - Portici (Naples), Italy
{maurizio.pollino, grazia.fattoruso, dellarocca, laporta,
sergio.locurzio, agnese.arolchi}@enea.it

Valentina James, Carmine Pascale
Consorzio T.R.E. - Tecnologie per il Recupero Edilizio
Naples, Italy
{valentina.james, carmine.pascale}@consorziotre.it

Abstract—**The latest improvements in geo-informatics offer new opportunities in a wide range of territorial and environmental applications. In this general framework, a relevant issue is represented by earthquake early warning and emergency management. In the recent years, the scientific community has recognized the added value of a geo-analytic approach in order to support complex decision making processes for critical situations, due to disastrous natural events like earthquakes. This paper describes the research activities concerning a GIS-based solution, which is aimed at the development of seismic Early Warning Systems (EWSs). In this context, an innovative open source GIS has been studied, implemented and integrated as component of the seismic EWS. Its architecture consists in: a geospatial database system; a local GIS application for analyzing and modelling the seismic event and its impacts and supporting post-event emergency management; a WEB-GIS module for sharing the geo-information among the public and private stakeholders and emergency managers involved in disaster impact assessment and response management.**

Keywords-GIS, Open Source, Spatial analysis, Early Warning Systems, Emergency Management

I. INTRODUCTION

Over the last 100 years, more than 1,100 disastrous earthquakes have occurred worldwide, causing more than 1,500,000 casualties: buildings collapsing is about 90% of direct deceases [1]. Those occurrences are clearly linked to world's population increase jointly with cities expansion. In particular, the urban growth process is often characterized by lack of planning and unsuitable land use: those factors contribute to dramatically amplify the damages due to seismic events. In this framework, Geographical Information (GI) technologies can play a fundamental role both in seismic risk assessment and in complex decision making in the course of critical situations [2], supporting natural disaster early warning and emergency management tasks. The need for related standard and effective spatial GUI (Graphical User Interface), geo-visual analytic tools, integrated geographic platforms (GIS), spatial data infrastructures has been outlined within several research works (see [3], [4], [5], [6], [7] and [8] among others).

As regards the early warning and emergency response issues related to seismic events, the recent advances in geo-informatics, in communication and sensor technologies have opened new opportunities. The up-to-date earthquake Early Warning Systems (EWSs) consist in seismic sensor networks connected to a central unit (operating centre, OC) by high-speed communication network. The kernel of the operating centre is a decision support system (DSS) that should enable the operators to make decisions and to disseminate EW. The generated alarm should be used to evacuate buildings, shut-down critical systems (e.g., nuclear and chemical reactors), put vulnerable machines and industrial robots into a safe position, stop high-speed trains, activate structural control systems and so on. Immediately following an earthquake, the operating centre should also support emergency response and rescue operations.

Until recently, the most common information available immediately following a significant earthquake is its magnitude and epicentre. However, the damage pattern is not a simple function of these two parameters alone, and more detailed information must be provided for properly ascertain the situation and adequately plan and coordinate emergency response. Just as an example, although an earthquake has one magnitude and one epicentre, it produces a range of ground shaking levels at sites throughout the region depending on distance from the earthquake, the rock and soil conditions at sites and variations in the propagation of seismic waves. Hence, GIS systems can support quick analysis of the situation immediately following an earthquake and facilitate critical decision making processes. Prototype systems, currently available in literature, have been on purpose developed ([6],

[7] and [9]) and are fundamentally based on commercial technologies [6] and [9].

In Paragraph II of this paper it is described the innovative free/open source GIS system [10] developed as integrated component of a seismic EWS. In particular, Par. II.B describes its architecture, consisting in a geospatial database system, a local GIS application for analysing and modelling the seismic event and its impacts, a WebGIS module for sharing the geo-information (listed in Par. II.C) and supporting post-event emergency management. Paragraph III is devoted to describe and discuss the results, in terms of expected seismic damage in structures and infrastructures, and the tools to support a rapid impact assessment and the disaster response. Finally, conclusions and future developments are reported in Par. IV.

II. Materials and Methods

A. Case of study and context

The research work here described is focused on the development of a methodology for a regional seismic risk analysis by using GIS technologies and methodologies. The study is part of a seismic hazard, vulnerability and risk analysis for the seismically active areas in the Campania Region (Southern Italy) (Fig. 1).

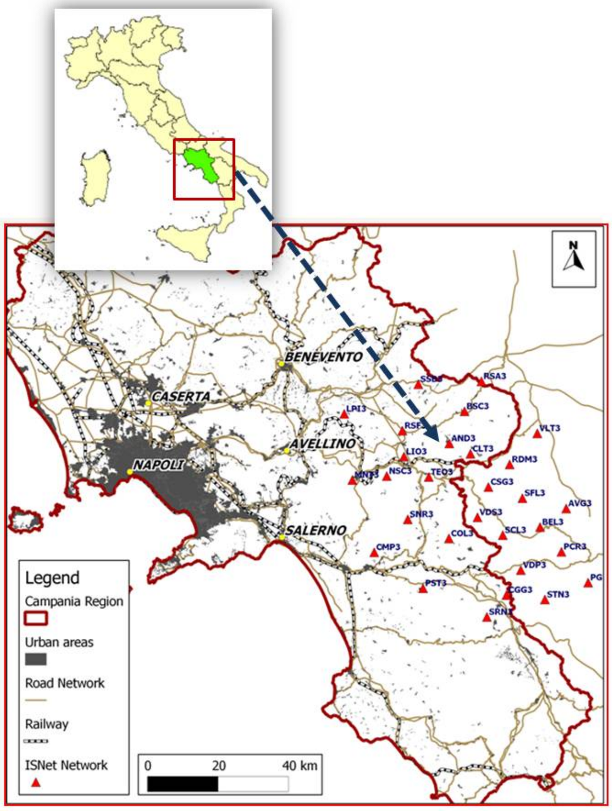


Figure 1. Geographic location of the study area

The main goal is to develop an integrated system for emergencies management in case of natural disasters, focusing on risk assessment and mitigation, early warning (EW) methodologies and post-event support activities. Further, the System is based on data coming from an existing seismic network located within the study area (ISNet, Irpinia Seismic Network [11]). The entire system improves a hybrid EWS based on a regional approach that assumes that a dense seismic network is deployed around the fault zone and a site specific approach, which aims to the protection of specific facilities.

The earthquake EWS is based on the different propagation speeds between seismic waves and signal transmission, since the alert is given only after the detection of phenomena indicating the generation of a possibly dangerous event and it has to reach the terminal before it starts damaging a given location this allows an alert time of that goes for seconds to tens of seconds.

The functions of a Seismic Alert Management System is related to two different phases of an event:

- Early Warning: 10-20 seconds after the main shock the system should predict the ground motion intensity, evaluate the epicentre and provide dissemination of information;

- Post event warning: 100-200 seconds after the main shock, the decision support system should address a preliminary scenario based on spatial interpolation of ground motion and then a detailed scenario, based on simulation of simplified source/propagation models.

The OC receives and elaborates information coming from monitoring systems (ISNet) and allows to activate a series of automatic security measures for sensible structures and infrastructures (e.g., high-speed railways, gas and electrical plants and installations, hospitals, strategic buildings, etc). The OC also coordinates the rescue operations in the immediate post-event phase. Moreover, the System has been designed not only to manage the emergency tasks, but also to provide a real-time monitoring of the vulnerability of structures and infrastructures within the area of interest.

The OC is supported by a GIS system that represents and performs the geographical information related to the event source (real-time and near real-time phases) and analyses in few minutes the expected damages on structures and buildings.

B. GIS System architecture

The geospatial analysis and visualization play a fundamental role in earthquake EW and post-event emergency management: to this purpose, the GIS has been integrated into the overall Project architecture as geographic interface of the OC. Consequently, basic information and thematic maps are stored and managed into a geospatial database purposely implemented, so that it is possible to display and query the data by means a map viewer. Fig. 2 shows the GIS logical architecture, developed by using free

ISBN: 978-1-61208-178-6

open source software (FOSS). It consists in the following modules (between brackets the FOSS used):

1. *Geodatabase Module* (PostgreSQL/PostGIS);
2. *GIS Module* (Quantum GIS);
3. *WEB-GIS Module* (MapServer).

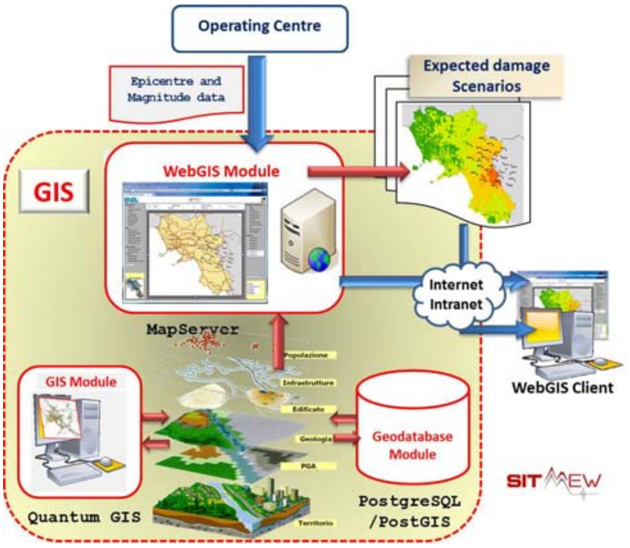


Figure 2. GIS architecture schema

Geodatabase Module has been designed to manage and integrate geospatial data provided as input to the system, including the alphanumeric data related to seismic events (e.g., magnitude and epicentre, recorded and processed by the OC) and specific geospatial data related to the area of interest (geology, vulnerability maps, urbanized areas, Census, etc.). The FOSS technologies chosen to implement this module was PostgreSQL/PostGIS (www.postgresql.org and http://postgis.refractions.net/).

The *GIS Module*, in direct connection with *Geodatabase Module*, is devoted to process geographical data and spatial information. By means of spatial analysis procedures and geo-processing operations, this module provides a complete and up-to-date description of the study area and, as final result, the maps of expected damage. After a comparative analysis between the main FOSS desktop GIS platforms available [10], the one chosen to implement the *GIS Module* was QuantumGIS (http://www.qgis.org/). The comparison was based on main functionalities, technology, geo-processing capabilities and interoperability with the other FOSS packages used for the modules *Geodatabase* and *WEB-GIS*.

Finally, the *WEB-GIS Module* was implemented by using the FOSS Mapserver (http://mapserver.org/): it allows the consultation of geo-spatial data stored in the system and support the management of activities during the immediate seismic post-event phase.

The main features of the GIS subsystem can be summarized in:

- Description and characterization of the study area;
- Production of thematic maps (e.g., expected damage scenarios) to support the management of near-EW and post-event phases;
- Consultation via intranet/internet to data and maps.

C. *Materials*

The *Geodatabase Module* has been purposely implemented to provide the spatial description of the study area of the Campania region (Fig. 1) and structured into different logical schemes (homogeneous for geographic data type). In detail, the following data (UTM-WGS84 reference system) have been used:

- Basic GIS Layers (Administrative boundaries, road network, railways, hydrograph, etc ...);
- Thematic Maps (hydrology, geomorphology, seismic classification, etc.);
- 1:25.000 Cartography;
- Census data;
- Digital Terrain Model (DTM, 20 m ground spacing);
- Geographic location and data of ISNet sensors;
- PGA (Peak Ground Acceleration) distribution maps;
- Data from parametric catalogue of damaging earthquakes in Italy (INGV, Italian National Institute of Geophysics and Volcanology).

Those layers and information have represented the basis of the spatial analysis carried out through the *GIS Module* and, along with the new maps produced, have been stored and managed into the *Geodatabase Module*.

D. *Spatial analysis*

As stated before, magnitude and epicentre are fundamental information available immediately after a significant earthquake. Considering other parameters such as rock and soil conditions, distance from the epicentre and variations in the propagation of seismic waves, it is possible to produce the ground shaking maps by means of spatial analysis and geo-processing tools. Then, such maps can be overlaid with inventories of buildings, critical facilities, transportation networks and vulnerable structures and provide a mean of prioritizing response.

The work here presented is based on several of these concepts in a simplified analysis over a fairly large region and exploits data from the parametric catalogue of damaging earthquakes in the Italian area, achieved by INGV [12]. The GIS system has been opportunely designed to process and achieve shake maps and multiple scenarios with different local magnitude (ML) for different epicentres. Considering the spatial data stored into the *Geodatabase* module and the geographic location of the ISNet sensors, the system has been structured to receive earthquake from the OC (epicentre and ML) in order to return PGA maps, vulnerability maps and expected damage scenarios. In this way, it is possible to have a preliminary assessment of the expected damages after a seismic event of given magnitude and epicentre. The model developed within the GIS systems takes into account the potential effects of the earthquake on manmade objects and population: for this reason, the vulnerability has been expressed in terms of macroseismic intensity I_{MCS}. In particular, PGA and I_{MCS} values have been correlated using the law proposed by Sabetta and Pugliese [13] to calculate the PGA distribution and to correlate the PGA to I_{MCS} [14]. In order to estimate the surface ground shaking in the region, the following attenuation relationship [13] has been used (1):

$$\log_{10}(Y) = a + bM + c\log_{10}(R^2 + h^2)^{1/2} + e_1S_1 + e_2S_s \pm \sigma \quad (1)$$

being Y the parameter to evaluate the PGA for this case study, M the magnitude (local), R the distance (from the epicentre) and σ the standard deviation of log Y. The parameters S_1 and S_2 refer to site classification and take the value of 1 for shallow and deep alluvium sites, and zero otherwise. The analysis don't take in account of site effects and the PGA has been calculated considering bed rock condition. To convert IMCS to PGA, the following equation (2) has been used:

$$\log PGA = 0.594 + 0{,}197 \cdot I_{MCS} \qquad (2)$$

III. RESULTS

A. *Vulnerabilty Index*

To evaluate the vulnerability is required a suitable inventory of the buildings in the region and well-defined relationship between earthquake motion (including local site effects) and both structural and non-structural damage. The estimation of buildings vulnerability is fundamental to provide a measure of their susceptibility to be damaged in consequence of specified seismic events. To obtain this information, as basic source has been used the inventory of the buildings extracted from ISTAT [15] Census data (2001). Those data, in table format, have been linked to the respective census section (in vector format) and processed using spatial analysis GIS functions. In this way, it has been possible to produce new GIS layers containing aggregated information about built-up density, structural typology (2 classes: Masonry or Reinforced Concrete), age of construction (7 classes), number of storeys (4 classes). Adapting the approach proposed by Giovinazzi and Lagomarsino [16] and using the above described data, thus, it has been calculated the vulnerability index I_v for each census section [17]. Firstly, the buildings have been basically distinct (Table I) in Masonry (M) or Reinforced Concrete (RC). Other information contained in the ISTAT data (number of floors and period of construction) have been instead used to correct the vulnerability index for each category and considered as behaviour modifiers (Table II).

TABLE I. VULNERABILITY INDICES FOR BUILDING TYPOLOGIES AND CONSTRUCTION AGE OVER THE STUDY AREA

	Construction age	I_V	
		Masonry	**RC**
1	Before 1919	50	-
2	1919 ÷ 1945	40	-
3	1946 ÷1961	30	20
4	1962 ÷1971	30	20
5	1972 ÷1981	20	20
6	1982 ÷1991	20	0
7	After 1991	20	0

TABLE II. VULNERABILITY INDEX MODIFIERS DEPENDING OF NUMBER OF STOREYS AND CONSTRUCTION AGE

Age **N. of storeys**	*<1919*	*1919-1945*	*1946-1961*	*1962-1971*	*1972-1981*	*1982-1991*	*>1991*
1	0	0	0	0	0	-6	-6
2	+5	+5	+5	+5	+5	0	0
3	+5	+5	+5	+5	+5	0	0
>4	+10	+10	+10	+10	+10	+6	+6

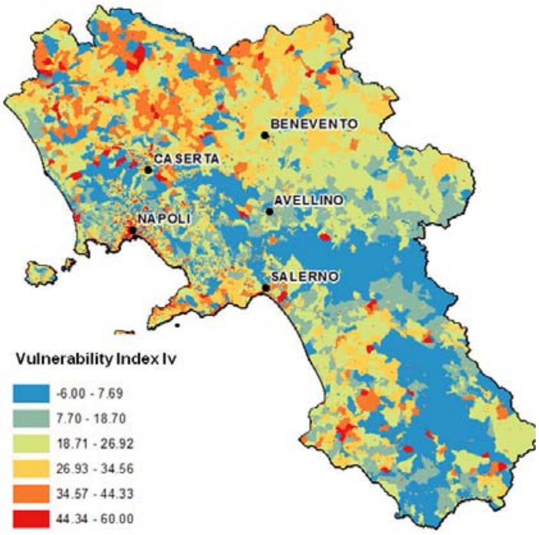


Figure 3. Map of the vulnerability index I_V: values for Census sections

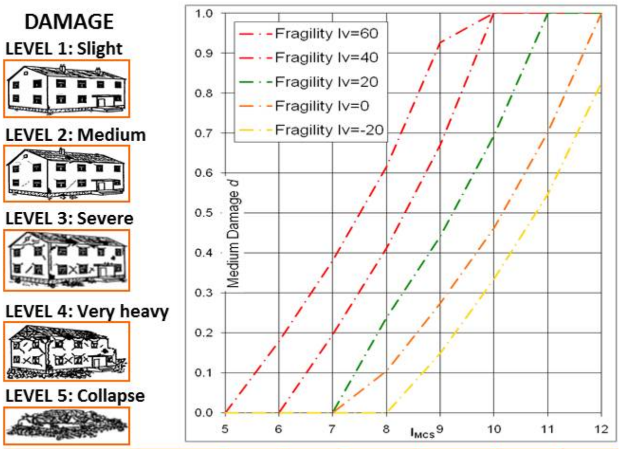
Those modifiers are used to increase or decrease in the I_v index, depending on the characteristics of the buildings within the area considered. Because each building has his intrinsic vulnerability and census sections may contain buildings with different values of the index, I_V (ranging from -6 to 60) has been calculated for each polygon as a weighted average of the values due to different building characteristics (Fig. 3).

B. *Maps of expected damage*

Despite the obvious approximations, the preliminary assessment of seismic vulnerability performed by the above described approach has the advantage of an extensive and prompt application, especially considering a large area like the Campania Region. After converting each map into an array of numeric values with square cell size 50 m spatial resolution, consistent with other data, it has been possible to process (by means spatial analysis modelling) the thematic maps representing PGA, I_{MCS} and I_V in overlay with the spatial representation of Census data above described. According to Giovinazzi and Lagomarsino [16], the damage d has been calculated (3) as:

$$d = 0.5 + 0.45\{\arctan[0.55(I_{MCS} - 10.2 + 0.05 \cdot I_V)]\} \quad [3]$$

The formula (3) expresses the relationships between I_{MCS} and damage *d*, according to the trend of fragility curves depicted in Fig. 4.



Levels of Damage	0	1	2	3	4	5
Factor of Damage (f_d)	0	0.01	0.1	0.35	0.75	1

Figure 4. Fragility curves and Vulneravility Index I_V relationships, in terms of mean damage

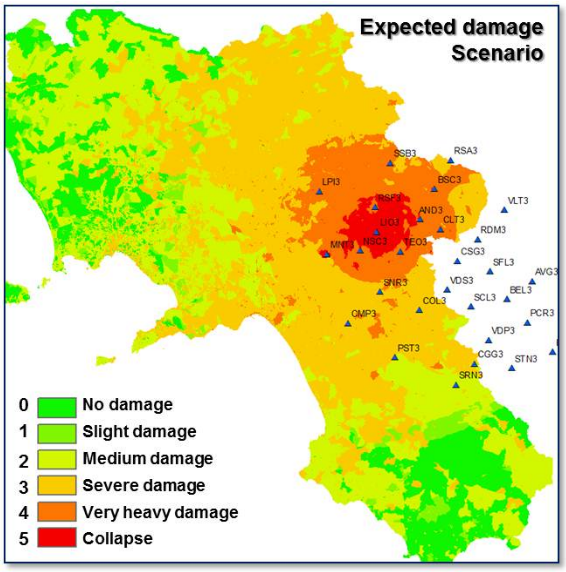


Figure 5. Scenario: example of map of expected damage (categorized according to different levels of damage)

Therefore, from a qualitative point of view, it is possible to establish a relation between I_{MCS} and *d* by differentiating the mean damage into 5 different levels (Fig. 4, left side). Then, the damage can be expressed by an a-dimensional parameter f_d (ranging between 0 and 1), in order to obtain a correspondence (Fig. 4, underside) between the levels of damage and the values of *d* calculated by means the formula (3). Using this approach, have been obtained the expected damage maps (Fig. 5) for each seismic event simulated: the variables are represented by epicentre coordinates and local magnitude ML.

C. *The WebGIS*

The primary goal of the *WEB-GIS Module* is to make geographic data and thematic maps available to specific end-users and, potentially, to the public. The application allows the end-user to view spatial data within a web browser, without a specific GIS Desktop software. This Module provides interactive query capabilities and integrates the GIS solutions with other technologies, according to server-side or client-side applications. Over 30 different WebGIS packages are available at present. Among these, the most popular and commercially successful are ESRI ArcIMS (www.esri.com/arcims), Intergraph GeoMedia WebMap (www.intergraph.com) and AutoDesk MapGuide (www.autodesk.com).

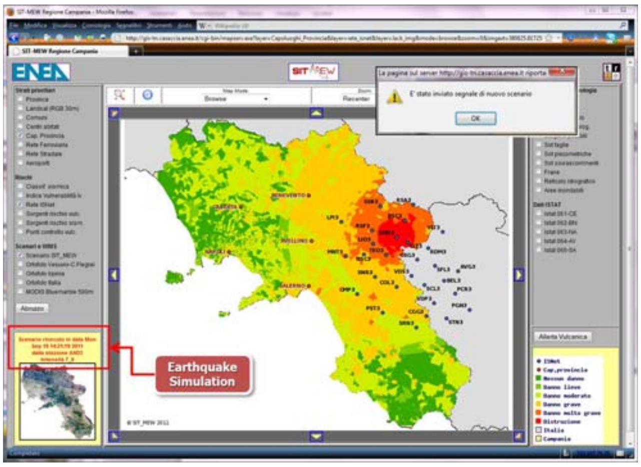

E' stato inviato segnale di nuovo scenario

Figure 6. WebGIS visualization of damage scenario (Earthquake simulation: epicentre in Andretta, ML 7.0)

UNM MapServer is a FOSS application developed by the University of Minnesota through a NASA sponsored project, that has been widely adopted. The package is a free alternative to other commercial applications and it is a good solution when highly customised applications are needed. MapServer is a Common Gateway Interface (CGI) programme that sits inactive on the web server. MapServer provides a scripting interface for the construction of web and stand-alone applications, adding WebGIS capability to popular scripting languages. The expected damage scenarios produced through the spatial analysis function above described are the main features of the *WEB-GIS Module* (Fig. 6). The Module allows to display these maps in overlay and in relationship to each other data stored into the *Geodabase Module*: basic geographic layers, geology, shaking maps, urban areas, Census data, PGA and I_V maps, Scenarios, etc.

IV. Conclusions and future development

A procedure for mapping and assessing buildings seismic vulnerability has been developed integrating spatial analysis and using geoprocessing tools [18]. The GIS approach, including the reclassification and overlay of each spatial data layer, has been applied to analyse the potential hazard that would result from a certain magnitude earthquake. In this way, it has been possible to define a conceptual model that, exploiting the GIS architecture here described, allows a quick management of damage scenarios. Maps of PGA, I_{MCS}, I_V and building characteristics reported in the Census data represent fundamental parameters to detect the areas (subdivided in parcels) that could probably face serious problems in consequence of total or partial collapses.

Further, the model described in this paper is fully functional and available to regional scale and the advantage of GIS methodologies is that the system is open and additional data can be integrated as soon as new information is available. In perspective, multi-source data and GIS integrated analysis can contribute to a better emergency planning, providing fundamental information for immediate response when future disasters will occur. A interactive DSS based on GIS approach could support the public government to address, in the near post-event phase, activities related to emergency management and damage evaluations for buildings and lifelines. Finally, the global architecture of the system will be enhanced also taking into account the implementation of a backup system, in order to manage and/or mitigate the effects potentially coming from a network failure (electricity, telecommunications, etc.).

Acknowledgment

The methodologies here described have been designed and developed within the context of “SIT_MEW” Project (*Sistema integrato di telecomunicazioni a larga banda per la gestione del territorio e delle emergenze in caso di calamità naturali comprensivo di Metodologie di Early Warning*), which is a R&D project co-funded by Italian Ministry of Education, Universities and Research (MIUR) in the framework of the “Grandi Progetti Strategici” (GPS) national program.

References

[1] Lantada, N., Pujades, L. G., and Barbat, A. H., “Vulnerability index and capacity spectrum based methods for urban seismic risk evaluation” Nat. Hazards, 51, 2009, pp. 501-524

[2] Cova, T. J., “GIS in emergency management”. In: Longley, P. A. et al. (eds). Geographical Information Systems, V 2: Management Issues and Applications, John Wiley & Sons Inc., New York, 1999, pp. 845-858

[3] Andrienko, N. and Andrienko, G., “Intelligent Visualization and Information Presenattion for Civil Crisis Management” Transaction in GIS, 11 (6):, 2007, pp. 889-909

[4] Charvat, K., Kubicek, P., and Talhofer, V., Konecny M., Jezek, J., “Spatial Data Infrastructure and geo-visualization”, In: Emergency management. Resilience of Cities to Terrorist and other Threats, Springer Science, 2008, pp. 443- 473

[5] ESRI White Paper: Geographic Information Systems Providing the Platform for Comprehensive Emergency Management, 2008

[6] FEMA 2008. HAZUS-MH Estimated Annualized Earthquake Losses for the United States (FEMA 366). Federal Emergency Management Agency, Washington, DC, April 200.

[7] Mueller, M., Dransch, D., and Wnuk, M., “Spatial GUIs for Complex Decision Support in Time-critical Situations” In: XXIII International Cartographic Conference, 4-10 August 2007, Moscow Russia, 2007

[8] Della Rocca, A.B, Fattoruso, G., Locurzio, S., Pasanisi, F., Pica, R., Peloso, A., Pollino, M., Tebano, C., Trocciola, A., De Chiara, D., and Tortora, G., “SISI Project: Developing GIS-Based Tools for Vulnerability Assessment” In: Sebillo, M., Vitiello, G., Schaferer. G. (eds) LNCD, Vol. 5188, Springer, Heidelberg, 2008, pp. 327-330, doi:10.1007/978-3-540-85891-1_37

[9] Theodoridis, Y., “Seismo-Surfer: A Prototype for Collecting, Querying, and Mining Seismic Data” In: Y. Manolopoulos et al. (eds.) PCI 2001. LNCS, vol. 2563,. Springer Berlin-Heidelberg, 2003, pp. 159--171

[10] Steiniger, S. and Bocher, E., “An overview on current free and open source desktop GIS developments” Int. J. Geogr. Inf. Sci., 23: 10, 2009, pp. 1345-1370

[11] Irpinia Sesimic Networ (ISNet). http://isnet.amracenter.com/ <retrieved: Nov., 2011>

[12] INGV, Italian National Institute of Geophysics and Volcanology. http://www.ingv.it/eng/ <retrieved: Nov., 2011>

[13] Sabetta F., and Pugliese A., “Estimation of response Spectra and Simulation of Non-Stationary Earthquake Ground Motion”. Bulletin of Seismology Society of America, Vol. 86 - No 2, April 1996, pp. 337-352

[14] Decanini, L., Gavarini, C., and Mollaioli, F., “Proposta di definizione delle relazioni tra intensità macrosismica e parametri del moto del suolo” In: Atti 7° Convegno Nazionale “L’ingegneria sismica in Italia”, Siena, vol. 1, 1995, pp. 63-72

[15] ISTAT, Italian National Institute of Statistics. http://www.istat.it/en/ <retrieved: Nov., 2011>

[16] Giovinazzi, S., and Lagomarsino, S., “Una metodologia per l’analisi di vulnerabilità sismica del costruito” In: Atti X Congresso nazionale “L’ingegneria sismica in Italia”, Potenza – Matera, 2001

[17] Borfecchia, F., De Cecco, L., Pollino, M., La Porta, L., Lugari, A., Martin, S., Ristoratore, E., and Pascale, C., “Active and passive remote sensing for supporting the evaluation of the urban seismic vulnerability” Italian Journal of Remote Sensing, 42(3), 2010, pp. 129-141, doi: 10.5721/ItJRS201042310

[18] Pollino, M., Fattoruso, G., Della Rocca, A. B., La Porta, L., Lo Curzio, S., Arolchi, A., James, V. and Pascale, C.: An Open Source GIS System for Earthquake Early Warning and Post-Event Emergency Management. In: B. Murgante, O. Gervasi, A. Iglesias, D. Taniar and B. O. Apduha (eds.) LNCS, vol. 6783, Springer Berlin-Heidelberg, 2011, pp. 376-391, doi: 10.1007/978-3-642-21887-3_30

[19] UNM MapServer. http://mapserver.org/ <retrieved: Nov., 2011>

Incorporating the Analyses of Urban Form into the Geocomputational Modelling

The Morphological Approach

Małgorzata Hanzl
Institute of Architecture and Town Planning
Technical University of Lodz
Lodz, Poland
e-mail: mhanzl@p.lodz.pl

***Abstract* - Geomatics allows for advanced modeling of phenomena in space and for the observation of process development in time. It enables comparative analyses of various aspects of urban-scape, including the social and human dimensions. At this time, the scope of urban data required by legal regulations is limited to basic issues, including land-use zoning, transportation and, cultural/natural values preservation. Successful sustainable urban environmental planning requires concentration on other aspects of city structure, with special emphasis on urban morphology analyses. Such an approach is more appropriate to the mixed used development and revitalisation processes, which take place in urbanised areas. The paper provides insights into the importance and challenges of using urban structure analysis from different perspectives: social, cultural, urban planning and design, and explains the main fields of contemporary urban morphological analyses, reviewing the geomatics use in this field.**

Keywords-urban planning; urban design; urban morphology analyses; geomatics; data modelling; INSPIRE

I. Introduction

The city - *"a whole"* in Plato's terms - should be considered holistically. The elements of the system are so complex that it is difficult to define them in an explicit way. Geomatics allows for comparative analyses of the distribution of different phenomena in space and for the observation of the dynamics of the processes of development. Contemporary urban planning practise and theory, presenting the holistic approach and underlining the importance of time [4], progressively develop the above qualities. The officially approved scope of urban analyses [1] restricts itself to the level of blocks and refers to the limited quantity of issues: mainly technical, including land-use zoning, transportation and cultural/natural values preservation. The objective to successfully plan the sustainable urban environment requires concentration on other aspects of city structure, with special emphasis on urban morphology analyses.

A. Urban morphology as a repository of socially defined space

When regarding the development of physical structures in relation to culture, the built form constitutes an important repository of cultural information, an artifact of cultures and societies, that created them in a given time [2], [11], [21], [26]. Hillier and Hanson [18] underline the relation of patterns of people movements and physical environment, introducing the concept of spatial logic of space. Thus, the analyses of existing and former urban structures provide an important tool for the creation of new structures, which not only follow the site genius loci and local tradition, but at the same time stay in compliance with the integral cultural patterns of social groups, e.g., the research by Gabaccia [15], cited after [26]. More contemporary research on the social production of space seeks to place the understanding of built form in the larger context of society's institutions and history [27], [26]. Proxemics relates to the human environment with the behavioral patterns proper for distinguished cultures. The above factors remain important when considering the constant displacement of people in an era of globalization, and the requirement to provide an environment, which suits their needs, while at the same time reducing the problems of social adaptation [16].

B. Complexity of urban design and planning

As Carmona et al. [6] suggest, urban design and planning should remain a holistic process including the approaches of different disciplines. The writing of Lefebvre [27] also points at the need for the unity of science in the description of urban systems. The need for the consilience of science is considered indispensable also by scientists of other disciplines, e.g., sociobiology [41].

The current domination of engineering, including individual transportation policies and reducing planning to the definition of land use zoning, leads to the ugliness of urban settlements. As Landry [24] and Florida [14] state, the factor of attractiveness in urban environment is particularly important in cities, which search the development of the creative industries. The development of theories referring to urban perception, starting with Lynch [28], [29] and Debord [9], follows on constantly (e.g., works of Amoroso [3], Kempf [23]). The use of neogeography and mash-ups allows also for the inclusion of lay experience of urban forms in the analyses of urban scapes (e.g., popular competitions for best stories about different cities). Looking at cities through the eyes of citizens plays an important role in the public participation and project implementation process. All of the above issues (attractiveness and perception) refer to the

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-178-6

urban forms. At the same time attractiveness and beauty remain issues of culture [13]. These complex issues are omitted in the common descriptions of urbanised space referring to a very limited quantity of factors. For some projects, especially realized in the mixed up city cores and the rehabilitation projects, a clear definition of planned land use contradicts the objectives of these projects. In these cases, the description of city structure assumes the form based approach, where the shape of volumes and their scale and relations in space prevail over the definition of functions.

II. Basic types of morphological analyses

Panerai et al. [34] distinguish three main groups of urban analyses: (1) the general ones which refer to the geographical level, including the analyses of the growth of cities, (2) the analyses of urban landscapes – the sequential ones, and (3) typological analyses, which refer to streets, parcels, and buildings.

Healey [17] points at the presence of the three main planning traditions, concentrated on: economy, formal issues, and political processes in planning. Studies of urban form are present in planning in different European countries although they belong to different disciplines, i.e. urban planning, urban design, architecture, and revitalisation projects. Mainstream urban planning at the European level covers the more general considerations, mainly of a geographical nature. The concentration on definition of land use and transportation systems, present in the official planning regulations, is also the consequence and reminiscence of the modernism era, when these two subjects constituted the main area of interests. These derive from modernist principles of separation of uses in the city, underlying the concept of the functional city, which constituted one of the most important rule of the Charte d'Athènes [7]. Planning theory and practice, which continue the approach rooted in the modernism tradition, find out its description in geocomputational models, which refer to a very limited number of issues, exposing the subjects, which are typically of a geographical nature: like land use, growth of cities, transportation, and usually omitting the analysis of urban form.

A. Analyses at the geographic level

The most important body of analyses, which finds its place in mainstream theory and in legal regulations at the European level [1], refers to the general description of the city structure, which is considered in geographical terms. The geographical descriptions usually concentrate on more general units (urban blocks, districts, etc.), although more detailed elements, like public spaces, streets, built up areas, are also present in some treatises, e.g., [25].

The analyses concern the level of metropolitan areas, cities and towns, or districts, the most detailed ones refer to urban blocks. The parcels are usually below levels of concern. Most of the analysis deals with urban functions, urban areas are classified according to their dominant usages and this continues since modernism [5], cited in [34], p.10. Within the city structure the distinguished parts are usually limited to the down-town and suburbs. Development studies, including growth simulations and tracing, constitute a considerable body of knowledge.

B. Analyses of the urban scapes

The analyses of the urban landscapes is the basic method of gathering data for urban projects in urban planning and urban design. Direct contact with the environment allows for observations and validation. The theoretical body for the studies is derived from Lynch's theory [29], in Polish architectural writings the theory of urban composition was developed by Wejchert [40], Szmidt [38] and Żurawski [44]. The theory was further developed in the Anglo-Saxon tradition, e.g., by Venturi et al. with regard to urban stripes [39]. The perception of cities in terms of sequence is not limited to the above authors. Sequential analyses were always present in architectural theory, their comeback started since Sitte [37]. In parallel, the continuation of the British Picturesque tradition was developed by Cullen [8]. Currently, concentration on the human perception of cityscape became quite a common approach, compare [23]. This group of analyses contains also psycho-geographical examinations of cityscape, e.g., [9], [33] or maps of East Paris [32].

C. Typological analyses

There are three basic elements of the urban landscape which are the subject of analyses: streets and public spaces, parcels, and buildings. These basic elements are interrelated, they create patterns, in which relations and hierarchies are also subject to analyses. Public spaces constitute the subject of planning since the very beginning. In the postmodern era, the analysis of public spaces is equalised with the examination of publicly accessible edifices and places (churches, commercial zones), as shown in the famous Map of Rome by Giambattista Nolli, 1748. Streets are examined as subject of sequential analysis, the example of which are transects drawings, compare, e.g., Project 360 degrees on Amsterdam city scapes, containing an assemblage of drawings and photo-collages of the physical materiality of the city, emphasizing signage and infrastructure [10], and as elements constituting patterns [31], [20].

One of the basic units of morphological analyses is parcel: the layout of parcels, their dimensions, shape, correlations. Patterns remain one of the main threads of urban morphology analyses. The requirement of description of the character of space needs the engagement of more detailed observations – concerning the elementary level, which refers to the issues typically, in the Anglo-Saxon praxis, connected with the discipline of urban design.

Lynch and Rodwin [30] distinguish the two general categories: flow systems and adapted space. They may be broken down into more categories, based on the following criteria: elements types, quantity, density, grain, focal organization, and generalized spatial distribution [30]. Element types as well as quantitative analyses are quite common concepts. A density may refer both to the physical structures (the intensity factor) and to the patterns of streets. Grain describes the extent to which the typical elements and densities are differentiated and separated in space. Focal organization refers to the key points, which may be the density peaks, concentrations of dominant building types, main open spaces or nodes of the circulation flows and their interrelations. The generalized spatial distribution is a kind of synthesis, which includes patterns of zones taken by different types of development and densities.

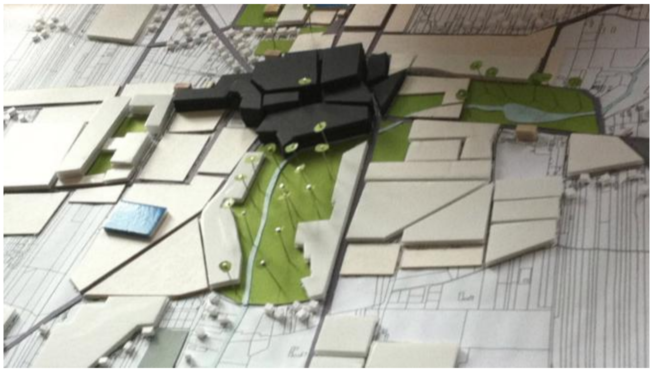


Figure 1. Central part of Brzeziny, close to Lodz – model by students of the second year of Architecture Engineering course, International Faculty of Engineering, Technical University of Lodz, for the Urban Design classes. Synthesis model presenting the generalised spatial distribution of densities and basic elements of town structure. Model by students: I. Sikirycka, D. Pogorzelska, M. Socha, A. Salamończyk, tutor: M. Hanzl

D. Analysis of morphological development

The comparison of the layout of the city structure in different periods allows for the tracing of the development. It allows also for the comparison of the intercultural differences.

This is particularly visible when analysing the cities - like Lodz, Poland, where the consecutive structures were the result of different urban theories and building cultures implementation. In case of Lodz ‘Old Town’ the former rural Polish town was at the end of XIX century converted, preserving part of its former fabric, into the heart of the Jewish district, constituting a part of a big industrial city. During the World War II, it was partly demolished by Nazis and the Jewish ghetto was created there. After the War the demolitions were continued and part of former structures were replaced with socio-realist and modernist structures, as shown in Figs. 2 to 5.

Figure 4 shows the perception analyses of the former Jewish district in Lodz, currently nonexistent, using the Lynch criteria [29]. The comparison to the current state proves the loss of street space definition in the eastern part of the area. The after war transformations utterly changed the former appearance of this part of the city. The land-use parameters remained the same, though the character of the district was altered.

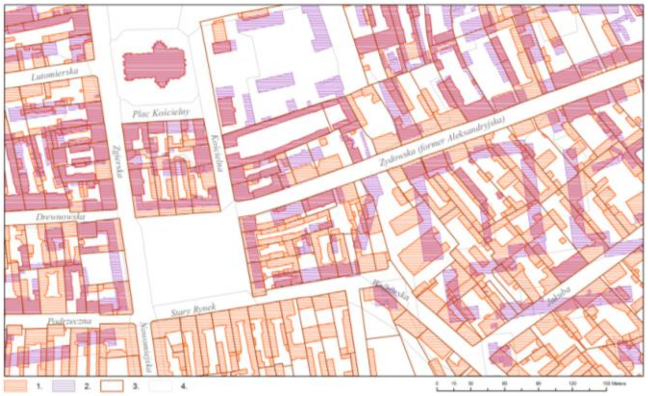


Figure 2. The comparison of built up area and parcellation in the central part of the Old Town of Lodz, Poland – the former Jewish District between the II World War and currently: 1. Buildings 1939, 2. Buildings 2010, 3. Parcellation 1939, 4. Parcellation 2010. Street names – 1939 (in the paranthesis – current names).

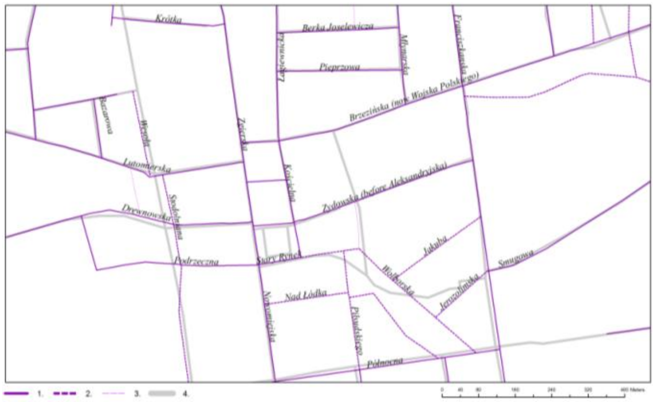


Figure 3. Street layout for the central part of the Old Town of Lodz, Poland – the former Jewish District between the II World War and currently: 1. Streets 1939, 2. Streets 1939, which do not exist anymore, 3. Passages, 4. Streets 2010. Street names – 1939.

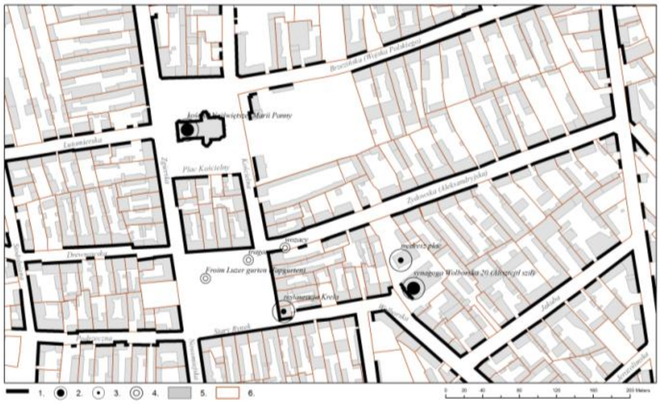


Figure 4. Nonexistent appearance of the central part of the old Jewish district: 1. frontages, 2. distant landmarks, 3. landmarks, 4. special places, 5. buildings in 1939, 6. parcellation in 1939

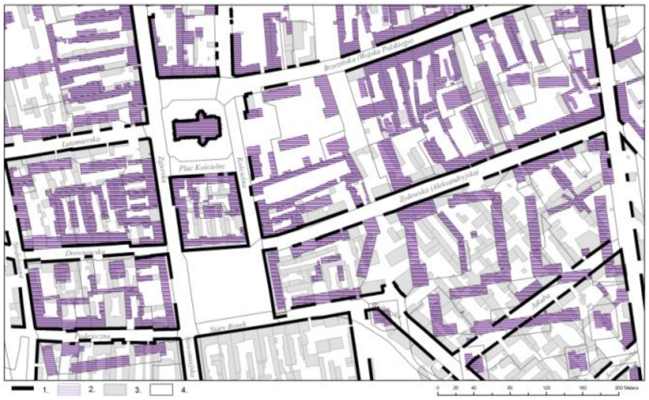


Figure 5. Not existing appearance of the central part of the old Jewish district contrasted with the contemporary figure-ground map: 1. lines of frontages – 1939, 2. buildings in 1939, 3. buildings in 2010, parcellation in 2010

III. Urban morphology and geocomputation

In the above context, traditional morphological analyses of urban structure, which remain as the basic tool in urban planning praxis, takes on importance and should be considered principal in geocomputational modelling, along with the description of land use definition, transportation policies, etc.

The basic unit of the morphological analyses is parcel: the layout of parcels, theirs dimensions, shape, correlations, patterns remain one of the main threads of urban morphology analyses. Descriptions usually concentrate on more general units: urban blocks, districts, etc., although more detailed elements had also been present, like public spaces, streets, built up areas [34]. As it was said, the requirement of description of the character of space needs the engagement of more detailed observations, involving the perception of urban form.

A. Geocomputation as a platform for urban analyses

The analyses of physical patterns of urban development should constitute the basis for other layers, describing the non-physical issues - as physical structure change progress is by definition slower than other processes.

The concept of sustainability and climate resilience assumes the conciseness of urban structures [22]. The need for extended application of geomatics in assessing the current structures seems obvious in this context.

Currently, the most popular theoretical approaches refer to figure background illustrations and land use data. The analyses of urban structure may assume the extended use of remote sensing. The commonly used photogrammetric refers to land use layer only, for example the commonly used thematic classifications, like the one by Anderson et al. at the U.S. Geological Survey [36]. The other most common example of remote sensing use refers to the analyses of growth processes of urban organisms [35]. The research of Yua et al. [43] provides an example of counting urban intensity, using LIDAR data. Another approach addresses the analyses of urban structure including the description of housing units [12].

B. Space Syntax and modelling of social behaviours

Space Syntax is a method of simulating human social behavior based on the analyses of spatial layout [18], [19]. The analyzed patterns include movements, vulnerability, and activity in buildings and urban settings. The simulations are based on a process, that informs human and social usage of an environment. A model of individual decision behavior, based on spatial affordances offered by the morphology of a local visual field, is consistent with the spatial configuration of movement patterns.

C. Development of morphological description of urban structure - suggestions

The commonly applied land-use analyses derived from the LIDAR data are useful in urban planning praxis, but not satisfactory. The production of planning documents by definition must refer to the property layer. The basic analysis of urban morphological structure used for the master plans preparation includes a description of much more features than land-use parameter only, this concerns, e.g., Polish urban planning practice. The list of the most basic descriptors is contained in the table below. The list may be extended to include other features or shortened in case, when some of them are not substantial. The list content differs in different country planning regulations.

TABLE I. Qualities of urban structure (examples)

Feature	**Description**
Land-use	Basic and complimentary uses
Parcellation	Minimum and maximum sizes of parcels and front widths
Urban parameters	Intensity of development, percentage of built up area
Location on plot	Buildings forming line of frontages, of different character: continuous or with breaks, Set back of building from the street
Character of development	Traditional or modernist, contemporary or of historical style, rich in/devoid of details, etc.
Form of development	
Building height	Number of stories, height [m]
Roof shapes	Slopes, kind of roof: flat or sloped, number of slopes, main ridge direction
Materials	Facades and roof materials
Fences	Materials, dimensions, shape
Details	Decoration

The above description refers to urban structures only. Although a similar list of features could be defined for streets and squares, their patterns, etc. [20], [30], [31] or for any other set of features referring to urban-landscapes of different cultures and periods. An example of an analysis of the city physical structure, following some of the principles listed above, is shown in the Figure 6. The analyses assumed elements of both the qualitative description, referring to the types of development, and the quantitative one – describing a level of completeness. GIS generalisation allowed for the semi-automatic production of consecutive maps showing different phenomena in the scale of the whole city.

ISBN: 978-1-61208-178-6

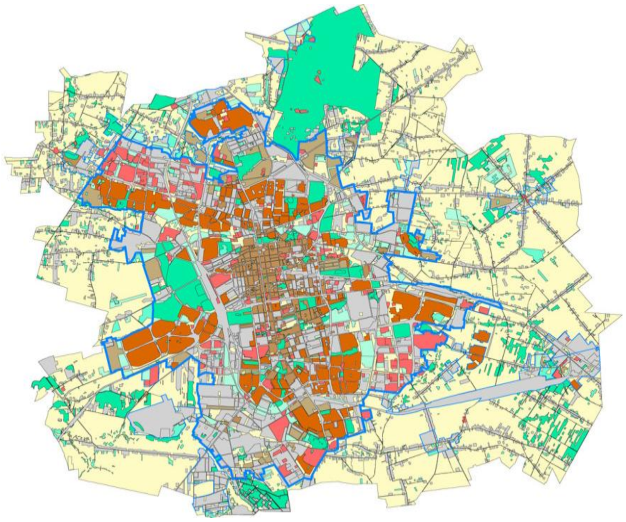


Figure 6. An example of the analyses following the principles mentioned in the paper - Study... Lodz [42]. Level of completion of urbanisation processes. The presentation assumed the generalisation of basic classes with the objective to get the map of most important structures types.

A GIS analysis may also show non-physical phenomena, which influences both the physical aspects and the perception of cityscape by inhabitants and visitors. The juxtaposition of different layers allows for the comparison of the physical appearance and, e.g., public transportation flows and nodes, like in Figure 7. Visual methods of analysis enhance the understanding of phenomena and helps in finding appropriate solutions.

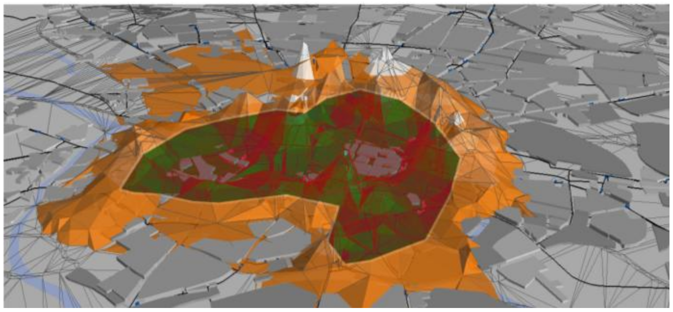


Figure 7. Wrocław center border drawn by capacity and location of tram transport nodes proceeded for 7-9 a.m. on workdays. Course: Architecture for Society of Knowledge, Warsaw University of Technology, seminarium GIS, student: Tomasz Kujawski, tutor: M. Hanzl. Further visualisations are available: http://system.asknow.eu/users/s_tomaszkujawski/ , retrieved: 09.2011

IV. Conclusions

Planning practice in the postmodern era requires references to urban form rather than to other aspects of urban development. The modernist definition of functions is not adequate and not sufficient, both in the process of analysis and decision making. The rehabilitation of brown-field development, as well as the down-town mixed uses areas, require more flexible treatment of land-use concept and concentration on urban form. The perception of urban landscapes, the concept of legibility and the semiotics of the environment requires conscious treatment of the formal issues in the processes of urban planning [28], [29], [39]. Similarly, concentration on urban form has become result of the desire to redevelop the cities as more concise, which is perceived as required from the point of view of ecology and sustainable development [22]. City morphological structure is also one of elements responsible for the possibility of social adaptation of members of different cultures [16].

The observed changes in urban planning theory and practice should influence alterations of the geomatics approach towards the environmental data, gathered for the purpose of urban planning, including the legal regulations and normative procedures concerning these issues. The current emphasis on land-use analyses should be replaced with form based approaches, which is enabled by the contemporary GIS and remote sensing technology. Currently, the tools encompassing 3D modeling are used mainly for presentation purposes. The requirements of progressing shift in urban planning towards more formal approaches should follow with their usage for analytical purposes. The basic element, and thus level of analysis required, should be the one of parcels. The technology allows for the integration of the two approaches – the main limitation seems to be lack of sufficient communication between urban planners and urban designers and geoinformatics professionals and researchers. The integration of approaches should result in increased easiness to fulfill the requirements of the analysis of urban environment and consequently urban policy preparation, which should also be reflected in the legal requirements.

References

[1] 2007/2/EC Directive of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), http://eur-lex.europa.eu, retrieved: 9.2011

[2] Ch. Alexander, S. Ishikawa, and M. Silverstei, Pattern Language, New York: Oxford University Press, 1977

[3] N. Amoroso, The Exposed City: Mapping the Urban Invisibles, New York: Routledge, 2010

[4] K. H. Bhabha, “Rem Koolhaas in conversation with Homi Bhabha”, keynote speaches at the Conference Ecological Urbanism: Alternative and Sustainable Cities of the Future, Harvard University, 2009, http://ecologicalurbanism.gsd.harvard.edu/conference.php, retrieved 12.2010

[5] J. Beaujeu-Garnier, Traité de géographie urbaine, Paris: Flammarion 1977

[6] M. Carmona, T. Heath, T. Oc, and S. Tiesdell, Public Places Urban Spaces The Dimensions of Urban Design, Oxford: Architectural Press, 2003

[7] Le Corbusier, La charte d'Athènes, Urbanisme, Une injonction à penser droit, Paris: Edition de Minuit, 1957, (selection of the 1943 edition)

[8] G. Cullen, The Concise Townscape, Oxford: Elsevier Architectural Press, 1971

[9] G. Debord, Psychogeographic guide of Paris, Bauhaus Imaginiste Ed., Dermark: Permild & Rosengreen, 1955(?), http://imaginarymuseum.org/LPG/Mapsitu1.htm, retrieved 09.2011

[10] F. Dresmé, Project 360°, student project realised in Hogeschool voor de Kunsten Utrecht, directed by 21bis 2006-2007, http://www.21bis.nl/project/26/3, retrieved: 9.2011

[11] R. Dubos, A God Within, New York US: Charles Scribner's Sons, 1972

[12] A. Duncan, A. Smith, and T. Crooks, From Buildings to Cities: Techniques for the Multi-Scale Analysis of Urban Form and Function, CASA Working Papers Series 155, 07.2010 http://www.casa.ucl.ac.uk/working_papers/paper155.pdf, retrieved: 10.2010

[13] U. Eco, Historia Piękna, Warsaw: Dom Wydawniczy Rebis Sp. z o.o. , 2006

[14] R. Florida, Who's your city?, New York: Basic Books, 2008

[15] D. R. Gabaccia, From Sicily to Elizabeth Street: Housing and Social Change Among Italian Immigrants, 1860-1930, New York: Albany State University, New York Press, 1984

[16] E. T. Hall, The Hidden Dimension, Garden City, New York: Doubleday, 1966

[17] P. Healey, Collaborative Planning: Shaping Places in Fragmented Societies, Vancouver BC: University of British Columbia, 1997

[18] B. Hillier and J. Hanson, The Social Logic of Space, Cambridge: Cambridge University Press, 1984

[19] B. Hillier, Space is the Machine, A configurational theory of architecture, Space Syntax 2007, http://www.scribd.com/full/17429763?access_key=key-17h1eg897r3ausi05ud3 , retrieved: 09.2011

[20] A. B. Jacobs, Great Streets, Boston: The MIT Press, 1995

[21] B. Jałowiecki and M. Szczepański, Miasto i przestrzeń w perspektywie socjologicznej, Warsaw: Wydawnictwo Naukowe Scholar, 2006

[22] M. Jenks and C. Jones, Dimensions of the Sustainable City (Future City), London: Springer, 2010

[23] P. Kempf, You Are the City: Observation, Organization and Transformation of Urban Settings, Kösel, Germany: Lars Müller Publishers, 2010

[24] Ch. Landry, The Creative City: A Toolkit for Urban Innovators, London, Sterling VA: Earthscan Publishing, 2000

[25] P. Lavedan, Géographie des villes, Paris: Gallimard, 1936

[26] D. L. Lawrence and S. M. Low, "The built environment and spatial form", Annual Review of Anthropology, Vol. 19, 1990, pp. 453-505

[27] H. Lefebvre, The Urban Revolution, Minneapolis, London: University of Minnesota Press, 2003

[28] K. Lynch, Good City Form, Cambridge and London: The MIT Press, 1994

[29] K. Lynch, Image of the City, Cambridge: The MIT Press, 1960

[30] K. Lynch and L. Rodwin, "A Theory of Urban Form", 1958 in City Sense and City Design, Writings and Projects of Kevin Lynch, T. Banerjee, M. Southworth, Eds. Cambridge, London: The MIT Press 1991, pp. 355-378

[31] S. Marshall, Streets and Patterns, New York: Spon Press, 2005

[32] Ch. Nold, "East Paris Emotion Map", 2008 http://paris.emotionmap.net/paris.pdf, retrieved: 09.2011

[33] Ch. Nold, Ed. Emotional Cartography, Technologies of the Self, 2009, http://emotionalcartography.net/, retrieved: 09.2011

[34] P. Panerai, J. Ch. Depaule, and M. Demorgon, Analyse urbaine, Marseille: Édition Parenthèses, 2009

[35] A. Schneider, M. A. Friedl, and D. Potere, "Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'", Remote Sensing of Environment 114, 2010, pp. 1733–1746

[36] N. M. Short, Sr., "Urban and land use applications: from Los Angeles to Beijing, Some Basic Principles and Examples" in Remote Sensing Tutorial, NASA, http://rst.gsfc.nasa.gov/Sect4/Sect4_1.html, retrieved: 09.2010

[37] C. Sitte: L'art de bâtir les villes, L'urbanisme selon ses fondement artistiques, D. Wieczorek, trans. Paris: Éditions du Seuil, 1996

[38] B. Szmidt: Ład Przestrzeni, Warsaw: Agencja Wydawnicza kanon, 1998

[39] R. Venturi, D. S. Brown, and S. Izenour, Learning from Las Vegas - Revised Edition: The Forgotten Symbolism of Architectural Form, Cambridge, London: The MIT Press, 2001

[40] K. Wejchert, Elementy Kompozycji Urbanistycznej, 2nd ed., Warsaw: Wydawnictwo Arkady, 1984

[41] E. O. Wilson, Consilience The Unity of Knowledge, New York: Random House Inc., 1998

[42] M. Wiśniewski et al., "Study on the preconditions and directions for the physical development of Lodz", 14 volumes, 1997-2002, urban planning document of Municipality of Lodz, Poland, unpublished

[43] B. Yua, H. Liub, J. Wua, Y. Hua, and L. Zhanga, "Automated derivation of urban building density information using airborne LiDAR data and object-based method", Landscape and Urban Planning, in press.

[44] J. Żurawski, O budowie formy architektonicznej, Warsaw: Wydawnictwo Arkady, 1962

Using GIS for Impact Analysis from Industries Installation

Camila Dasso Thomasi, Gerson Alberto Leiria Nunes, Márcio Medeiros Jugueiro, Diana Francisca Adamatti
Centro de Ciências Computacionais
Universidade Federal do Rio Grande
Rio Grande, Brasil
{camilathomasi, dfrabu, marcio.juguero, dianaada}@gmail.com

Abstract—**This paper presents a simulator that analyzes the impacts of the pollutants emission of a new industry insertion on a specific region (in our studies, at Rio Grande City – Brazil). This simulation calculates the pollutants concentration in the atmosphere, using CALPUFF non-stationary Gaussian model integrated to GIS. The goals are to assist people and responsible entities in the evaluation and air quality control.**

Keywords-social simulation; pollutants dispersion; air quality control; CALPUFF.

I. Introduction

The problems created by excessive emission of pollutants are not recent and they had significantly increased due the fast industrial growth. An alternative is to take some preventive measures by knowing the risk of the installation of a new industry on a region.

Using computational tools is an alternative to predict risks. by this way, these tools can create new possibilities for the control of air quality, because the simulations are able to predict the concentration of pollutants in the atmosphere.

In this context, the Geographic Information Systems (GIS) are emerging as robust tool, helping to organize and to have a better understanding about the results, because they have methods/objects to view, manipulate, synthesize and edit the georeferenced data [1].

In Thomasi et al. [2], GIS is used to evaluate the dispersion radius of total suspended particles and to analyze the potential risk of contamination of coastal ecosystems near to the industrial zone in Rio Grande city (Brazil), using the stationary model ISC [3]. However, a non-stationary model [4] provides a more realistic evaluation because the meteorological factors are under flotation most of the time.

The main goal of this study is to analyze the pollutants emission impacts of a new industry insertion in a specific region (in our studies, in Rio Grande City – Brazil). We are developing a simulation that calculates the pollutants concentration in the atmosphere, using the CALPUFF non-stationary Gaussian model [4] together to GIS [1]. It allows to view the damage size in areas near to the industry operations.

The paper is organized as follows: Section II presents the basic concepts of air pollution, the dispersion phenomenon and the requirements to air quality control. In Section III the tools used to implement the simulator are described. In Section IV is shown complete modeling of the simulator and the results. Finally, Section V presents the conclusions and future works.

II. Pollutants emission in the atmosphere

A. Atmospheric Pollution

Chemicals, even toxics, are not necessarily considered atmospheric pollutants, because to cause damage they have to reach a certain concentration. In this way, an atmospheric pollutant is any form of matter, that a given quantity, exceeds the limits (defined by a control agency), and it transforms the air improper [4] [5].

Therefore, the atmospheric pollution occurs when the air contaminants injure the well-being and health of people, and it cause harm to the environment [5].

B. The Dispersion Phenomenon

The dispersion mechanisms of pollutants in the atmosphere are governed by fluctuations in wind fields and turbulence [6].

The main meteorological factors that influence the atmospheric dispersion phenomenon are wind, temperature, high and low pressure and terrain. The meteorological factors in the region can contribute in a positive or negative way in the mixing of contaminants with clean air. These factors could cause a quick or slow dispersion pollutants.

According to Saraiva and Krusche [7], the main damaging weather conditions to the pollutants dispersion in our study region are: high pressure associated with light winds, low temperature and high humidity.

C. Air Quality

The air quality standards are a control strategy used to indicate the maximum concentration of pollutants that could be issued to preserve the health and well-being of the population, flora, fauna and the environment in general [5].

The pollutants group responsible for controlling air quality, basing to their frequency in the environment, is composed of: sulfur dioxide (SO_2), particulate matter (PM), carbon monoxide (CO), ozone (O_3) and nitrogen dioxide (NO_2). In this work only $SO_{2,}$ CO and NO_2 are considered to evaluate the emission of air quality..

Excessive exposure to pollutants causes damage. Therefore, the great importance of the emissions control is to preserve the life quality of people and also the environment preservation [4]. According to the resolution of CONAMA – The Brazilian National Environmental Council [5], there are

ISBN: 978-1-61208-178-6

specific levels of emissions that are allowed, ranging from the lowest level of "attention" to the higher level of "emergency", as shown in Table 1.

TABLE I. STANDARDS OF AIR POLLUTION BY BRAZILIAN NATIONAL ENVIRONMENTAL COUNCIL (CONAMA).

	LEVELS		
Parameters	**ATTENTION**	**ALERT**	**EMERGENCY**
Sulfer Dioxide - SO$_2$ (µg/m^3) - 24h	**≥ 800**	**≥ 1.600**	**≥ 2.100**
Carbon Monoxide - CO (µg/m^3) - 8h	**≥ 17.000**	**≥ 34.000**	**≥ 46.000**
Nitrogen Dioxide - NO$_2$ (µg/m^3) - 1h	**≥ 1.130**	**≥ 2.260**	**≥ 3.000**

III. USED TOOLS

This section describes the tools used to develop the simulator. These tools were chosen because of their availability, documentation, generation and manipulation of maps interatively.

A. MATLAB

The MATLAB [8] is a programming language with several libraries that allow us to perform a series of scientific calculations, statistics, solution of linear differential equations, engineering calculations, etc. Moreover, it has specific libraries to work with neural networks, filtering, bioinformatics, telecommunications, imaging, digital signal processing, automation, GIS and others.

In this work, to automate the processes of file management, data import, creation, manipulation and presentation of maps and their respective layers, we have chosen the GIS library (called "mapping toolbox"). This library was also important to perform post-processing data of the CALPUFF model and results analysis

B. Mapping Toolbox

The mapping toolbox [9] is a library of specific functions that allow us analyzing geographic data, create and manipulate maps. It imports both vector and raster type data. It also has support to the most common file formats, such as shapefile, GeoTIFF and DEM SDTS. In addition, it is possible to import data from WMS servers (Web Map Service). Thus, the developer can customize his application in subsection, side view, intersection and other methods.

The toolbox features allow us to develop customized solutions for various geographic problems. Some of these features allow that different layers data can be easily manipulated and presented in the same map.

Other features that also deserve attention are those that allow you to convert different types of coordinates, facilitating the use of data from different sources and to allow you to save all creation and manipulation of files that can later be analyzed by users of GIS software.

C. GIS - Geographic Information System

The GIS (Geographic Information Systems) [1] are softwares that manage, view and manipulate geographic data computationally. They allow that each set of data can be presented in different layers.

The layers are composed by a set of features of geographic objects. These objects have an infinite variety of shapes, but they can be represented basically by three different shapes: polygons, lines or points. Polygons represent things that have limits such as countries, states, cities or lakes. Lines represent narrow things such as streets, roads, rivers or railroads. Points are used to small things such as buildings, hotels, schools, fire hydrants or poles. The union of polygons, lines and points generates the vector data.

Therefore, GIS are used for better understanding the patterns, relationships and trends in the data. Many times, those information are not obvious to find in databases.

IV. THE PROPOSED SIMULATOR

This section presents a complete simulator modeling that uses a non-stationary model of pollution dispersion. The non-stationary models can be Eulerian, Lagrangian and Gaussian [10]. The Eulerian and Lagrangian models have a high computational cost due because they need to solve complex equations. It requires parallelization to obtain a satisfactory response time. On the other hand, the Gaussian models do some simplifications that it allows a faster response, making it attractive for real time applications.

This work uses the CALPUFF non-stationary Gaussian model that allows the prediction of risks caused by one or more industries in a specific period of time, varying the weather conditions in space and time [11].

Calpuff model is recommended by U.S. EPA (Environment Protection Agency) due to be working with a micro-region and it can handle complex three-dimensional wind fields. It is easy to setup and run for point sources with multiple parameters.

The Gaussian models do not have accuracy when we use a complex topography or emissions near the ground. In our experiments, these factors are not relevant because the simulated region has a flat terrain and their emission sources have a certain distance from the ground. The work aims to develop a tool to help the community and responsible entities in decision-making, visualization and evaluation of possible risks caused by the emission of pollutants from industries, without the concerned of accuracy.

The proposed simulator integrates the Gaussian CALPUFF model to MATLAB in order to automate their execution. The Matlab Mapping Toolbox saves and presents the output data in a vector layers and it can be manipulated in GIS softwares.

A. The Modelling of Proposed Simulator

In this stage of the project, the goal of the simulator is to predict what will be the risks in a specific region by the pollutants emission caused by the insertion of one or more industries. The result of the simulation is a layer with georeferenced points containing the concentration of certain pollutants. The simulator calculates the pollutants concentration resulting from one or more emissions.

The simulator is divided in two modules: CALPUFF model execution and maps' manipulation. According to the flow diagram, shown in Figure 1, the basic inputs to start the simulation are: the shape file containing the map of the interesting region and the output shape file names. The simulation returns the files with georeferenced points.

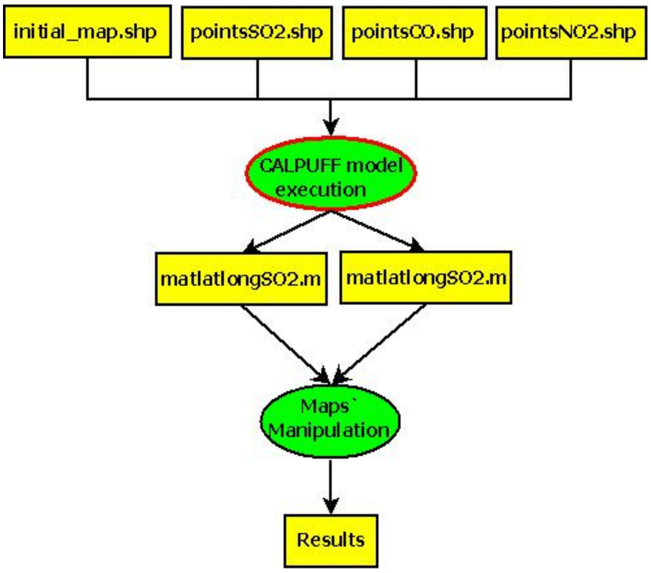


Figure 1. Data Flow Diagram to use the CALPUFF non-stationary Gaussian model.

In the CALPUFF model execution stage, all the settings and procedures for their execution are made and these steps are explained in greater details later. The output of this module has two georeferenced matrices containing latitude, longitude and pollutants concentration in the points. These matrices will be the inputs of the next step (maps manipulation) and they have the function of read, write, manipulate and present the vector files. Three structures are created, one for each pollutant to control the air quality: SO_2 (sulfur dioxide), CO (carbon monoxide) and NO_2 (nitrogen dioxide), containing the attributes shown in Table II. The elements of these structures are filled with the input values. After this procedure, three layers are created, one for each pollutant, which are the simulator output files.

TABLE II. STRUCTURE OF THE POINTS

Attribute	Type	Description
Geometry	*string*	Geometry of the point;
BoundingBox	[2x2 *double*]	Extreme of the point;
X	[1xN *double*]	X coordenates set in the point;
Y	[1xN *double*]	Y coordenates set in the point;
POLLUTION	*double*	Calculted pollution level by Gaussian model;
LOCAL_X	*double*	X coordinate of the location of the center point;
LOCAL_Y	*double*	Y coordinate of the location of the center point;

The last step of the simulation is to present the results, as shown in the Figure 2. The pollutants shape file is superimposed on the map of the interesting region, providing a better interface to analyze the results. In this map, the colored points refer to calculated concentration degree of the model. The Figure 3 is showing the value of the concentration pollution of each pollutant in µg/m^3.

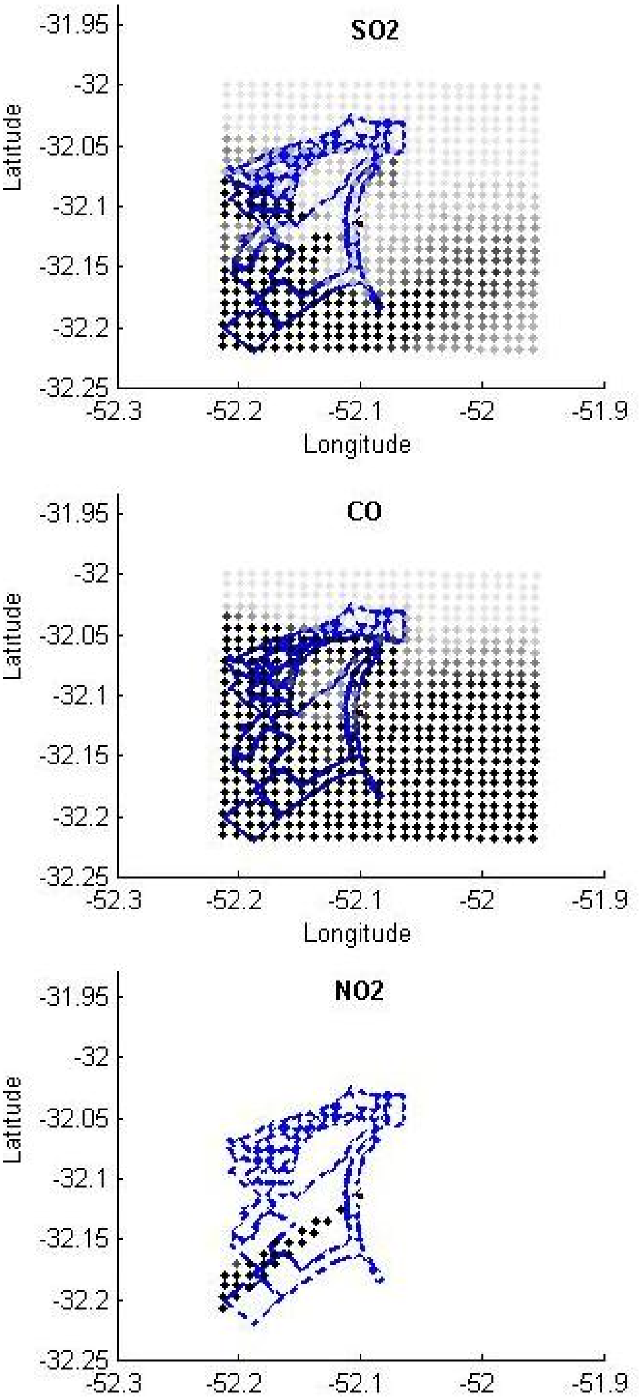


Figure 2. Results presentation.

B. Propagation Model of Pollution

The chosen model for the pollutants propagation in the atmosphere is the CALPUFF non-stationary Gaussian model. It is a puff Gaussian model used to simulate continuous puffs of pollutants emitted by a source to receivers according to the wind flow in the environment [12].

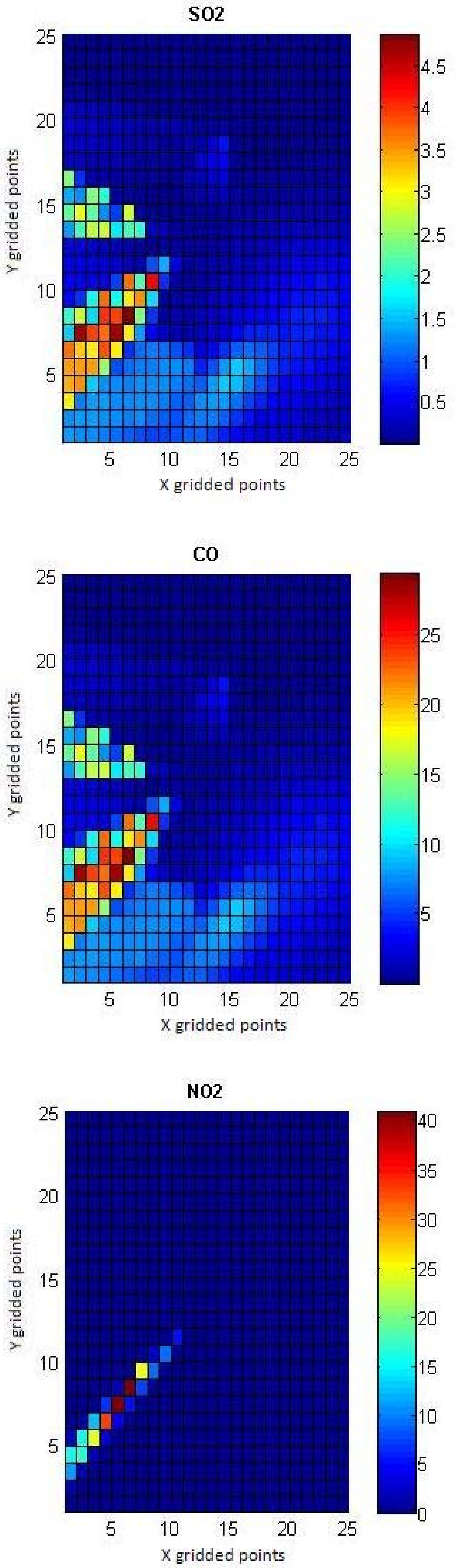


Figure 3. Georeferenced bidimensional arrays.

The plume, in these models, is represented by a set of discrete puffs of polluting material [6], as shown in Figure 4.

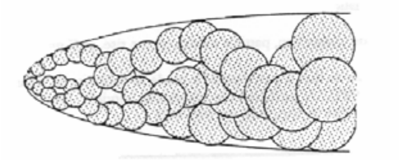


Figure 4. Representation of the plume in CALPUFF model

The puffs mass is dispersed according to a Gaussian distribution. The transport is made according to the trajectory of its mass center and the local wind vector speed [6].

The wind changes every hour, changing also the puffs path to the wind flow. While the puff is transported in the air, its pollutant concentration decreases. If the puff finds out other receptor, a new pollutant material is increased to it [12].

The Equation (1) is the basic equation to provide the contribution of a puff on a receiver [4]. To discover the total concentration of a receiver, it is necessary to add the concentration of neighborhood puffs of a particular receiver.

$$C = \frac{Q}{2\pi\sigma_x\sigma_y} g \exp\left[-d_a^2/\left(2\sigma_x^2\right)\right] \exp\left[-d_c^2/\left(2\sigma_y^2\right)\right] \quad (1)$$

where C is the ground-level concentration (g/m^3), Q is the pollutant mass (g) in the puff, σ_x is the standard deviation (m) of the Gaussian distribution in the along-wind direction, σ_y is the standard deviation (m) of the Gaussian distribution in the cross-wind direction, σ_z is the standard deviation (m) of the Gaussian distribution in the vertical direction, d_a is the distance (m) from the puff center to the receptor in the along-wind direction, d_c is the distance (m) from the puff center to the receptor in the cross-wind direction, g is the vertical term (m) of the Gaussian equation, [2].

C. *Related works*

In this subsection, we present some works that use CALPUFF in different domains. In Zhou et al. [13] is calculated the fraction of particulate matter emitted by power plants, using the CALPUFF model to estimate the risks for which the population is submitted to be exposed to these pollutants. MESOPUFF II method is used for simulation of chemical transformations. The dry deposition is modeled to gases and particles and the wet deposition for liquid precipitation. The internal plume increases are calculated as a function of some effects, as vertical wind shear and floating feather.

Levy et al. [14] make the assessment of emissions by power plants. It uses the NOAA'S RCU2 model to generate climatological data, spatial and temporal inputs to the CALMET grid. Simulation results compare the emission sources, in order to find the most polluted.

In Barsotti et al. [15], the CALMET is used to generate three-dimensional fields (wind and temperature) and two-dimensional (friction velocity, Obukhov length, atmospheric boundary layer height and so on). This work presents a CALPUFF based model called VOL-CALPUFF to simulate the launch, the transport and the volcanic ashes deposition.

D. CALPUFF Model Execution

The CALPUFF model execution in the simulator is responsible for setting up all files and parameters needed to run the model. With this analysis, is possible to observe the entire operation of the system.

Initially, the terrain configurations are made based in a file that has the *x* and *y* coordinates of all terrain and elevation. To obtain these information is necessary to run the TERREL.

TERREL is a preprocessor to create the terrain elevation data from multiple databases to a grid specified by the user [4]. This preprocessor has a control file where the scanned files are assigned. According to the flow diagram in Figure 5, the first thing is to obtain the terrain and topography grid, using files available in WebLakes [16]. The next step is to configure these files as input to the control file terrel.INP. Once that is done you can run the module *runTerrel*. This preprocessor provides as output a formatted file containing the coordinates in UTM X, Y and Z (elevation of the land).

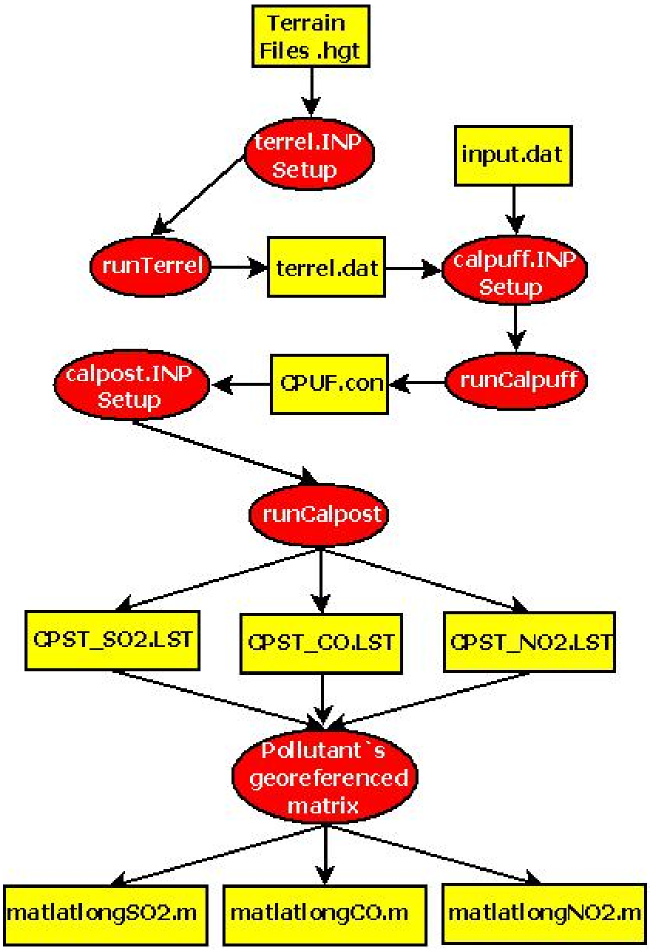


Figure 5. Steps of CALPUFF model execution

After obtaining the characteristics of the terrain, the next module is responsible for the configuration of the control file of the CALPUFF model. In this module, two files are needed as input: the terrain file (terrel.dat) and the region weather setting file (input.dat). The terrain file is provided from the preprocessor TERREL and the region weather file is a formatted file containing: day, month and year of data, simulation hours (1-24 hours), wind direction in degrees, air temperature (° C), Pasquill stability class [4] and the planetary boundary layer height (m). With these inputs and the configured control file calpuff.INP, it is possible to run the execution module of the CALPUFF (runCalpuff). This module generates its output as a binary file (*conc.dat*) that contains an average grid of pollutants concentrations, which were simulated in a period of time.

To calculate the concentrations, the conc.dat file must be executed in CALPOST postprocessor [4]. For that, firstly, the calpost.INP configuration module should be setted with the runCallpuff. After it, the runCalpost should be executed to the number of species issued. In this work, we have issued three species: SO_2, CO and NO_2. As output files are provided cpst_so2.LST, cpst_co.LST cpst_no2.LST and these files provide the concentration values in each grid point.

The last step is the creation of the georeferenced matrices that have the concentrations and converted UTM coordinates to latitude and longitude coordinates for each gridded point.

E. Simulation Scenario

We have chosen the industrial zone in Rio Grande city, in southern Brazil, to evaluate the model. The city has a large number of industries which emit a large quantity of pollutants in the atmosphere.

The simulator allows the user customizes the industry features to be inserted, according to their emission profile. The following features can be modified according to the desired simulation scenario: latitude and longitude, elevation, height and diameter of the stack, speed and temperature of the pollutant output and emission rates of pollutant source.

In our tests, the industry is located at latitude 52.1045S and longitude 32.1167 W. In Figure 6 are shown the neighborhood map of the Rio Grande city, highlighting the industrial zone. This region has the following characteristics: void elevation, because it is situated at sea level; 30 meters of stack height; 5 meters of stack diameter; output speed 5m/s; output temperature 195.3°C; and the following emission rates: 7.4 de SO_2, 44.39 de CO e 88.78 de NO_2, all in g/s.

The meteorological data in the simulator are: wind direction and speed, air temperature, and Pasquill stability class [4]. All these data were captured by the meteorological station at Universidade Federal do Rio Grande (FURG).

The used data correspond to a period of 24 hours in a day, with a propitious situation for pollutant dispersion, because these data present low pressure, light winds and medium temperatures.

F. Results Analysis

According to the scenarios described earlier, this section will analyze the consequences from emissions of the pollutants SO_2, NO_2 and CO in accordance with current legislation [4].

The chosen model (CALPUFF) provides the greatest values of pollutants concentration in a specific period. The highest values are compared with the data in Table I, that presents tolerable limits defined by law in Brazil, CONAMA [5].

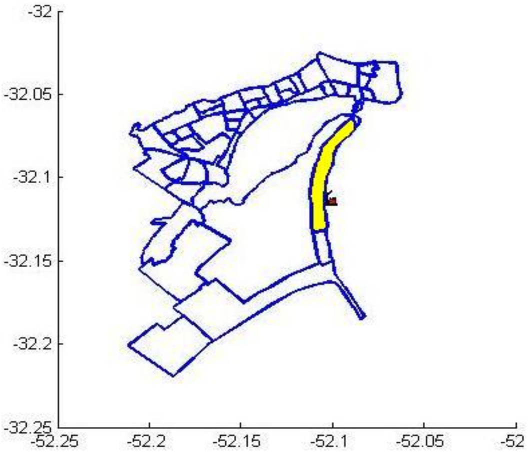


Figure 6. Industrial zone and industry location

The Figures 2 and 3 present the results of pollutants concentration to the SO_2 emissions (in 24 hours), the CO emissions (8 hours) and NO_2 emissions (1 hour).

According to the simulated results, the concentration of CO reached 28.019 mg/m³, the SO_2 reached 4.8623 mg/m³ and NO_2 reached 40.8490 mg/m³. Basing in Table I data, these values do not indicate a problem, because they do not even reach the lower limit (attention level). However, this is a hypothetical simulation in order to test the applicability of the proposed tool, because we had simulated just 24 hours and considered a single emission source. In real situations, the industrial interesting region has many industries with more than one source operating in continuous periods of time. Moreover, the effects of air pollution are cumulative. In this way, we believe that a simulation with a long period and many emission sources will create an attention level (or a worse situation).

Although the related work presented in subsection C helps in calculating the concentration of pollutants, both use the CALPUFF associated with complex weather forecasting models that have high computational cost.

Our results offer an estimate that is not so accurate due to the CALPUFF simplifications, but it allows application's real-time execution.

Furthermore, this paper presents graphic results that can be easily understood. The application allows saving the results generated in vectorial files. The application helps people and responsible entities in risk's evaluation using GIS software.

V. Conclusion and future works

This paper presented a simulator that use the CALPUFF non-stationary Gaussian model integrated with GIS. The idea is to analyze the air quality when new industries are inserted in a specific region.

As showed in section IV, the CALPUFF model is very complex to execute, but it presents complete data about pollutant dispersion.

Historically, the Rio Grande city (Brazil) has problems with air pollution and few works explore this scenario. The second reason to choose this region is because we live here and we believe that computational tools could help a better understanding about environmental problems and improve the people lives.

Our initial tests, using a hypothetical situation, just help to analyze how the CALPUFF model processes the data. As further work, we will execute the model for a long period (probably two years) and many emission sources (more than twenty). This new scenario will present a more realistic situation, and more comparisons will be done.

Acknowledgment

We thank FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul – Brasil) for funding research and development. Grant Number 10/0049-7.

This is a multidisciplinar project. Therefore, we will like thank you the Dr. Jonas Carvalho, to help in the understanding of CALPUFF model and the initial data of the meteorological station of Rio Grande.

References

[1] Câmara, G., Casanova, M.A., Hemerly, A., Medeiros, C.M.B. and Magalhães, G., “Anatomia de Sistemas de Informação Geográfica”, Curitiba, Editora SAGRES, p. 205, 1997.

[2] Thomasi, C. D., Nunes, G., Tolego, R., Jugueiro, M., Teixeira, P., Adamatti, D. F. and Tagliani, C., ”Um sistema para previsão de impactos gerados pela instalação de indústrias e sua influência sobre ecossistemas costeiros no extremo sul do Brasil”, In: III WCAMA 2011, pp. 1455-1464 Natal / RN, 2011.

[3] EPA - Environmental Protection Agency, ”User´s guide for the industrial source complex (ISC3), dispersion models volume II – Description of model algorithms”, p. 128, 1995.

[4] Scire, J.S., Strimaitis D.G. and Yamartino, R.J. “A user’s guide for the CALPUFF dispersion Model (Version 5)”, Earth Tech. Inc. 196 Baker Avenue, Concord MA 01742, p. 521, 2000.

[5] Conselho Nacional de Meio Ambiente (CONAMA), “Resolução N.º 003”, Brazil, 28/Jun/1990.

[6] Moraes, M. R., “Ferramenta para a Previsão de Vento e Dispersão de Poluentes na Micro-escala Atmosférica”, PhD. Thesis, Universidade Federal de Santa Catarina, UFSC, p. 166, Brasil, 2004.

[7] Saraiva, L. B. and Krusche, N., “Condições Atmosféricas Favoráveis à concentração de Poluentes em Rio Grande, RS”, resumos do II Congresso em física da camada limite planetária e modelagem de processos de dispersão. Santa Maria, RS. v.1. p.14. 2001.

[8] MathWorks, “MATLAB - Introduction and Key Features”, available: www.mathworks.com/products/matlab/description1.html accessed in: 28 Nov. 2011.

[9] MathWorks, “Mapping Toolbox Introduction and Key Features”, Available: www.mathworks.com/products/mapping/description1.html accessed in: 28 Nov. 2011.

[10] Vilhena M. T.,Carvalho J. C. and Moreira D. M., “Tópicos em turbulência e modelagem da dispersão de poluentes na camada limite planetária”, Porto Alegre, Editora da UFRGS, p. 260, 2005.

[11] Vallero D. A. “Fundamentals of air pollution”, Civil and Environmental Engineering Department Pratt School of Engineering Duke University Durham, North Carolina, p. 967, 2008.

[12] EPA – United States Environmental Protection Agency, “A Comparison of CALPUFF with ISC3”, Office of Air Quality Planning and Standards Research Triangle Park, NC 27711, EPA-454/R-98-020, p. 50, december 1998.

[13] Zhou Y., Levy J.I., Hammitt J.K. and Evans J.S., “Estimating population exposure to power plant emissions using CALPUFF: a case

study in Beijing, China", Atmospheric Environment, 37, pp. 815–826, 2003.

[14] Levy, J.I., Spengler, J.D., Hlinka, D., Sullivan, D. and Moon, D., "Using CALPUFF to evaluate the impacts of power plant emissions in Illinois: model sensitivity and implications", Atmospheric Environment 36, pp. 1063–1075, 2002.

[15] Barsotti, S., A. Neri, and J. S. Scire, "The VOL-CALPUFF Model for Atmospheric Ash 1 Dispersal: I. Approach and Physical Formulation", Istituto Nazionale di Geofisica e Vulcanologia, Journal of Geophysical Research, Pisa, Italy, p. 113, 2008.

[16] Lakes Environmental Software. Available: http://www.weblakes.com, acessed in: 28 Nov. 2011.

Distributed System Architectures, Standardization, and Web-Service Solutions in Precision Agriculture

Katja Polojärvi
School of Renewable Natural Resources
Oulu University of Applied Sciences
Oulu, Finland
e-mail: katja.polojarvi@oamk.fi

Mika Luimula, Pertti Verronen, Mika Pahkasalo
CENTRIA Research and Development
RFMedia Laboratory
Ylivieska, Finland
e-mail: {mika.luimula, pertti.verronen, mika.pahkasalo}@centria.fi

Markku Koistinen
Plant Production Research
MTT Agrifood Research Finland
Vihti, Finland
e-mail: markku.koistinen@mtt.fi

Jouni Tervonen
Oulu Southern Institute
RFMedia Laboratory
University of Oulu
Ylivieska, Finland
e-mail: jouni.tervonen@oulu.fi

***Abstract*—Effective precision agriculture requires the gathering and managing of geospatial data from several sources. This is important for the decision support system of automated farming. Distributed system architecture offers methods to combine these variable spatial data. We have studied architectural and protocol choices of distributed solutions and built up a demonstration implementation of the system. As one of the key factors of the system is interoperability, we applied standards for geospatial and agricultural data, a location-based service platform, and a technology of geosensor networks in the implemented system. We successfully demonstrated the interoperability and scalability requirements and functionality of the implemented system.**

Keywords- precision agriculture; spatial data; interoperability; distributed system architecture

I. Introduction

In automated farms, agricultural operations and processes are planned, executed, monitored, and assessed with the help of spatially referenced data. Spatial variability and changes in production conditions can be monitored by specific sensors, and thus can be taken into consideration when production inputs of precision agriculture (PA) are planned and executed. However, the realization of any expected higher productivity requires that decision-making and farming operations are based on spatial data of good quality and accuracy. Effective use of data also necessitates that data originating from internal (farm equipment and software) and external data sources (e.g., meteorological data) can be easily integrated and transferred between different hardware, software, and information systems [1][2]. Besides this, a refined and integrated analysis of the data acquired and the transformation of these data into information and knowledge useful for decision-making are also required [3].

The many existing data acquisition systems, documentation tasks, and precision farming applications result in a variety of data formats and interfaces, making data management complex in PA [1]. In addition to limitations in data exchange and communication between incompatible systems, lack of time, knowledge, or motivation of farmers to evaluate data from PA are among the key problems in data management, and result in great demand for automated and time-effective expert systems for the handling of precision farming data [3]. Because PA requires external services and cooperation with many partners, interaction and sharing of spatial data between a farmer and stakeholders must also be provided. Instead of data format conversions, coordinate transformations, and other forms of manual processing of heterogeneous data, effective information management is currently based on web-enabled spatial database systems, providing one of the most interoperable environments for distributed computing [4].

The purpose of the system design presented in this paper is to develop and implement the acquisition, storage, transfer, and management of spatial data produced in various operations of PA. In addition, one of the purposes is also assure the quality of spatial data by utilizing appropriate standards: ISO 11783 in agricultural vehicles and OGC’s (Open Geospatial Consortium) standards in wireless soil measurements. The development of the system is inspired by the architecture of a farm management information system (FMIS) designed in a Nordic project, ‘InfoXT’ [3]. We applied a location-based service platform and a technology of geosensor networks in the development of the system with the aim to demonstrate interoperable solutions for the FMIS. Implementation of communication and data interchange in the system is based on geospatial standards such as WFS

ISBN: 978-1-61208-178-6

(Web Feature Service), SensorML (Sensor Model Language), GML (Geographic Markup Language), and BXML (Binary Extensible Markup Language). Our previous experiments of BXML, SensorML, and ISO 11783-based data transfer are utilized in the development work.

In Section II, the main technologies and relevant standards related to distributed systems in agriculture will be discussed. Developing this kind of distributed system especially requires expertise related to location platforms and geosensor networks. Based on this introduction, the use of chosen standards will be further explained in this paper. Section III introduces the demonstration implementation of the system. The conclusion and future work are presented in Section IV.

II. RELATED WORK

In order to automate agricultural operations and processes we will need to build systems that are able to collect data from machines and from the field. Geosensor networks are among the new technologies that can be used to monitor agricultural phenomena. The technology is under intensive development and its application possibilities are gradually widening. Geosensor networks present a multidisciplinary research area, combining such fields as geoinformatics, computer sciences, telecommunications, and sensor technologies. Nittel et al. [5] have defined geosensor networks (GSN) as a sensor network that monitors phenomena in geographic space. This geographic space can range in scale from confined indoor conditions to highly complex outdoor environments.

Location platforms can be classified as device-centric, decentralized middleware-based, and centralized platforms [6]. Location-based service platforms can be considered as one form of distributed Geographical Information Systems (GIS). A distributed GIS can be defined as a network-centric GIS tool that uses a wired or a wireless network for providing access to distributed data, disseminating spatial information and conducting GIS analysis [7].

The main actors in this research field have been the Open Geospatial Consortium (OGC) research network and the National Aeronautics and Space Administration (NASA). According to OGC [8], their OpenGIS Interface Standard defines OpenLS Core Services, which form the Services Framework for the GeoMobility Server (GMS). These services cover Directory Service, Gateway Service, Location Utility Service, Presentation Service, and Route Service. GMS, as a location services platform, hosts not only these services but also Location Content Databases accessed through OGC Interfaces such as Web Map Server (WMS) and Web Feature Server (WFS). WMS produces maps of spatially referenced data dynamically from geographic information in a raster graphic format such as PNG (Portable Network Graphics), GIF (Graphic Interchange Format), or JPEG (Joint Photographic Experts Group), or in a vector graphic format, e.g., in SVG (Scalable Vector Graphics) format. Moreover, WMS generates a map when requested from spatial data stored in GML format, which is an XML (Extensible Markup Language) grammar for expressing geographical features [9].

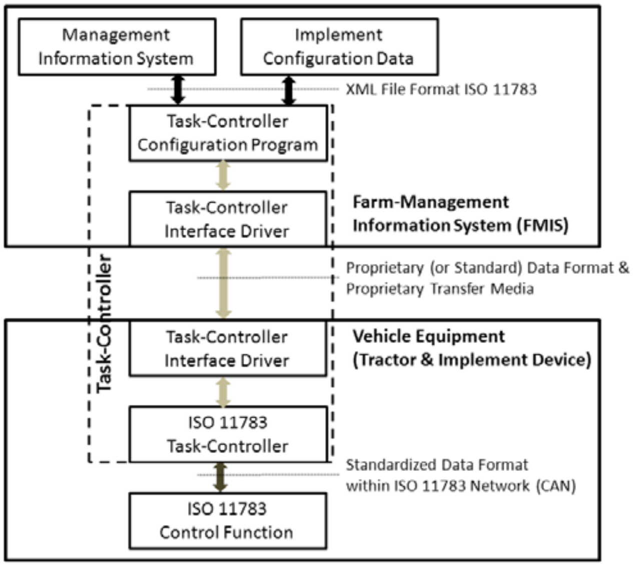

Management Information System
Implement Configuration Data
XML File Format ISO 11783
Task-Controller Configuration Program
Task-Controller Interface Driver
Farm-Management Information System (FMIS)
Task-Controller
Proprietary (or Standard) Data Format & Proprietary Transfer Media
Task-Controller Interface Driver
Vehicle Equipment (Tractor & Implement Device)
ISO 11783 Task-Controller
Standardized Data Format within ISO 11783 Network (CAN)
ISO 11783 Control Function

Figure 1. ISO 11783 [13] Task-Controller interfaces and transfer methods.

For geosensor networks, the main part of OGC's work is made under Sensor Web Enablement (SWE), which is a suite of specifications related to sensors, sensor data models, and sensor web services. The goal of SWE services is to enable sensors to be accessible and controllable via the Internet [10]. SWE consists of Observations and Measurements Schema (O&M), Sensor Model Language (SensorML), Transducer Markup Language (TML), Sensor Observations Service (SOS), Sensor Planning Service (SPS), Sensor AlertService (SAS), and Web Notification Services (WNS). XML is a key part of this infrastructure and all the services and content models are specified in XML schemas [11].

NASA has used OGC's SWE specifications in their geosensor networks and sensor web applications. Their EO-1 sensor web architecture is currently being updated to support an interface standard in development by the SWE group. This Earth Observing-1 Mission has been operating an autonomous, integrated sensor web linking dozens of sensor nodes in 24/7 operations since 2004. The use of sensor web technologies has been very successful with a number of cost-effective impacts on maintenance. These sensor web technologies have been used in the National Snow and Ice Data Center, and in the Cascade Volcano Observatory. In the first case, these technologies were applied to track ice formation and melting and use this information in EO-1 analysis. In the latter case, the goal was to acquire high-resolution infrared data of Mt. St. Helens [12].

An important standard for precision agricultural solutions is the so-called ISOBUS standard, i.e., the ISO 11783 standard family. The main focus of this standard is to define the bus and the control processes used in tractors and machinery for agriculture and forestry. It enables the hardware compatibility of tractors and different agricultural implement devices such as seeders, fertilizers, and harvesters. In addition, Part 10 [13] of the standard defines the data interchange between the farm management information system and the task controller (TC) process of the vehicle (tractor and implement device). The standard precisely defines the standardized data transfer and data transfer file in the XML format. According to the standard, the task controller is physically divided between the FMIS (farm management information system) and the vehicle. The ISO 11783 standard does not define the actual transfer media and the data format can be either a proprietary or a standardized data format. Figure 1 shows the proprietary version of data transfer where some parts of the Task-Controller are situated within the FMIS-subsystem instead of entirely at the vehicle equipment. The standardized option means that the ISO 11783 XML files are used in transferring between FMIS and vehicle equipment. In both options, the actual transfer media can be cable, transferable memory device, or some radio communication method.

The transfer of the actual XML files is not desirable when wireless radio communication methods are used since the larger data amount has the disadvantage of increased operational costs or a disturbingly long time spent on the data transfer. In our previous study [14], we have demonstrated the data transfer between FMIS and vehicle equipment using the radio protocol of IEEE 802.15.4 standard [15] and a compressed data format that is truly proprietary. The binary format XML compression methods based on the OGC's SensorML standard and BXML, the best practices specification [16][17] demonstrated in [18], are achievable for this purpose. Our suggestion is to use the compressed XML data format to achieve convenient wireless transfer media while simultaneously minimizing resource and cost requirements.

III. Implementation of the System

The system is implemented utilizing a location-aware system platform called Locawe that is developed at CENTRIA Research and Development, Ylivieska, Finland. The architecture, the use of geosensor network-related standards, several field experiments, and industrial pilots related to this platform have been introduced in [18][19]. The Locawe platform is a client-server solution for outdoor and indoor conditions. This platform consists of mobile units and servers for services like tracking and communication. It is possible to create user interfaces, which include video, location, and identification information on the map or on the floor plan [20]. In our previous experiments in agriculture reported in [14][18], the communication between GSN nodes and mobile devices was in a binary format based on OGC's SensorML standard and BXML best practices specification [16][17] (Fig. 2).

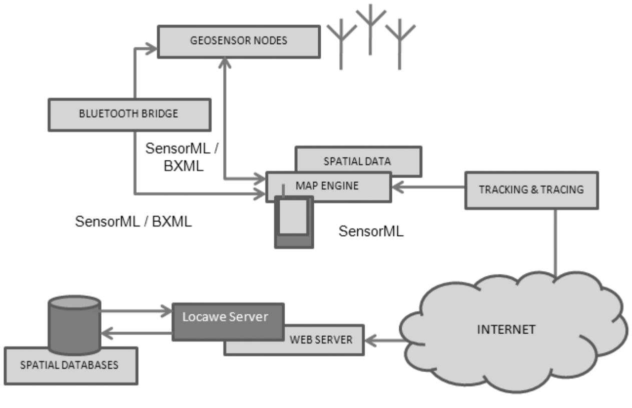


Figure 2. The use of GSN nodes with mobile devices in the Locawe Platform in previous studies.

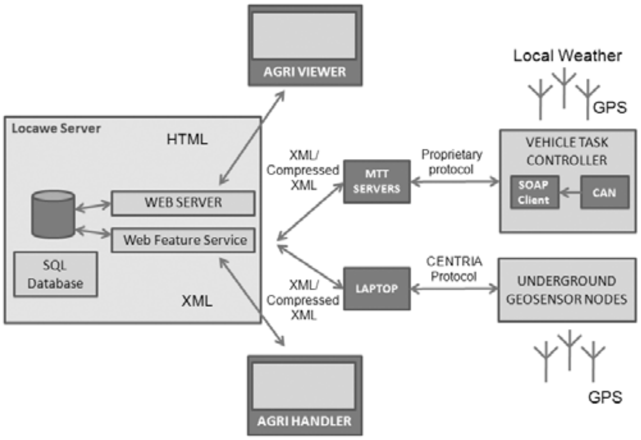


Figure 3. Agricultural sensor data delivery from various sources by utilizing OGC's WFS.

Figure 3 shows the implementation of the distributed system architecture. In the Locawe server, OGC's WFS standard acts as an interface between the SQL (Structured Query Language) database and applications. WFS messages contain basic XML definitions and action determinations. The data needed for executing queries are implemented in XML format, which can contain, e.g., GML type definitions. WFS requires login action to recognize user rights, which will determine what kind of queries the user can do. WFS manages the SQL database and correct data format. WFS messages can contain images in raw byte format or vector data in GML format.

We demonstrated the interoperability of the system by delivering spatial agricultural data to the server from various sources. The WFS interface manages data delivery from an underground geosensor network to a laptop as a gateway. It also handles data delivery to the Locawe server from the vehicle task controller via the servers of MTT Agrifood Research Finland. Spatial harvester data collected by MTT were also used as data in the demonstrations of AgriHandler and AgriViewer applications.

A. *Underground Geosensor Network*

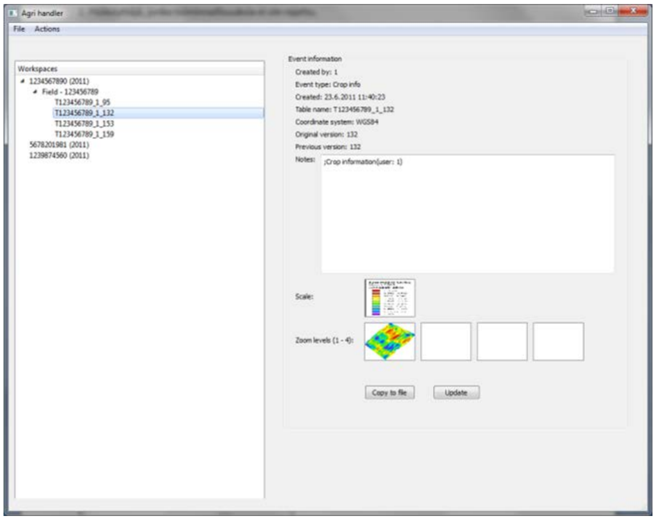
The main focus of the implementation of this study was the demonstration of architectural and protocol choices. The design of equipment of soil property measurements was beyond the scope of this paper. The proof of concept type and validation measurements were performed with wireless sensor nodes, which were neither optimized nor actually aimed for soil measurements. Our wireless sensor nodes used the 868 MHz frequency band version of the IEEE 802.15.4/ZigBee standard with the power level of 11 dBm. According to [21], this ISM (Industrial, Scientific, and

Medical) frequency band is the most suitable and practical compromise solution for wireless soil measurements. The wireless sensor nodes were underground at a depth of 30 cm. In our validation measurements, the range of 28.5 meters was achieved with favorable conditions, i.e., with relative low soil moisture. The long-term measurements of actual wireless underground soil scout prototypes are reported by [21][22]. These measurements showed that soil attenuation has a strong dependence on soil moisture.

Underground sensor nodes give measurement values to a main node device that is connected to a laptop. This PC and the main node communicate using a proprietary structure called the CENTRIA Protocol. It is designed for use in communications with low-rate devices. The used geosensors do not have any locating property, but the location of the sensors on the field can be set by the user with help of the developed PC software. For absolute location, the coordinates collected by a GPS (Global Positioning System) device on the field must be inserted in the PC where the application will use them automatically.

B. *User Interfaces*

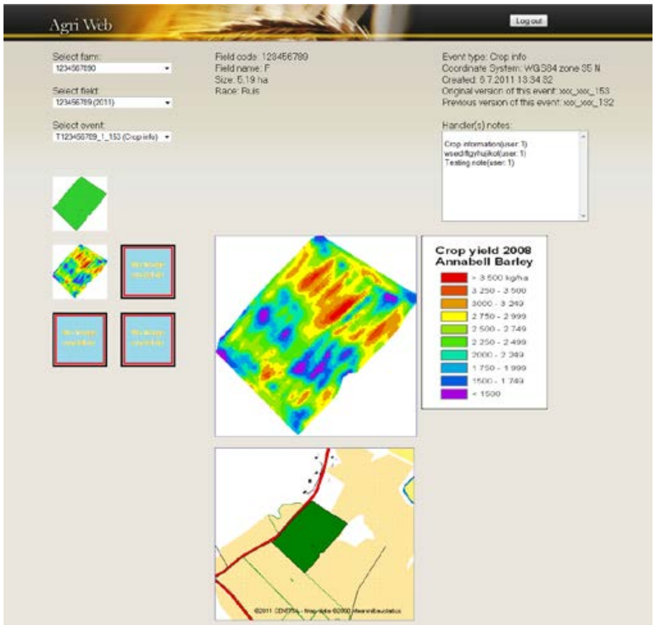
The demonstrated system has two applications as user interfaces for the database. AgriHandler application (Fig. 4) enables administration of farm and field data in the database through the WFS interface. Depending on user rights, the AgriHandler application can act as administrator tool or data handler tool with lower rights. An idea of the development of the application is to demonstrate how handled and analyzed precision farming data, e.g., interpolated nutrient balance maps, could be provided for a farmer (customer) as external information services. The quality and usability of handled and analyzed data is better compared to raw data. As a service provider, a data handler can get information to a local file and update the information in the database with the analyzed spatial data by messages over HTTP (Hypertext Transfer Protocol) POST to the WFS interface.



Workspaces
- 1234567890 (2011)
  - Field - 123456789
    - T123456789_1_95
    - T123456789_1_132
    - T123456789_1_153
    - T123456789_1_159
- 5678201981 (2011)
- 1239874560 (2011)

Event information
Created by: 1
Event type: Crop info
Created: 23.6.2011 11:40:23
Table name: T123456789_1_132
Coordinate system: WGS84
Original version: 132
Previous version: 132
Notes: ;Crop information(user: 1)
Scale:
Zoom levels (1 - 4):

Figure 4. AgriHandler application showing the event information.



Select farm: 1234567890
Select field: 123456789 (2011)
Select event: T123456789_1_153 (Crop info)

Field code: 123456789
Field name: F
Size: 5.19 ha
Race: Ruis

Event type: Crop info
Coordinate System: WGS84 zone 35 N
Created: 6.7.2011 13:34:32
Original version of this event: xxx_xxx_153
Previous version of this event: xxx_xxx_132

Handler(s) notes:
Crop information(user: 1)
wsedrftgyhujikol(user: 1)
Testing note(user: 1)

Figure 5. AgriViewer web application showing basic information and analyzed data related to the event.

These collected data of information (events) have own database tables and when the coordinate system is changed, a new table will be generated in the database. Inheritance of the events was taken into consideration in the database design. The application was demonstrated with crop yield data as vector points (MIF - MapInfo Interchange File) and images (PNG). The application also enables administration of geosensor network data. At the moment, AgriHandler supports WGS84 and ETRS-TM35FIN coordinate systems.

In the AgriViewer web browser (Fig. 5), end-users are able to view the same data explained above. The WFS interface is used only in authentication. Users have read-only rights in this web application.

C. *Vehicle Task Controller*

Dependable and efficient wireless communication may be challenging in mobile plant production in rural areas. However, this is an essential element in an approach where the assisting services supporting farming activities and knowledge management are based on a cloud computing paradigm.

In this demonstration (Fig. 6), the TC simulation was implemented giving high priority to the mobile connectivity challenges. Data packet size makes a difference in low mobile band width regions. The data collection itself is driven by GPS clock frequency, which is 5Hz when the signal strength is acceptable. Our approach to the dead network region dilemma is based on an idea to adjust the data transfer rate (from the TC to the cloud service) according to the mobile connection strength. In practice, this means that one collected data sample (that is, one data row) is sent to the service in every GPS clock cycle if the mobile connection strength is acceptable. In the case of mobile

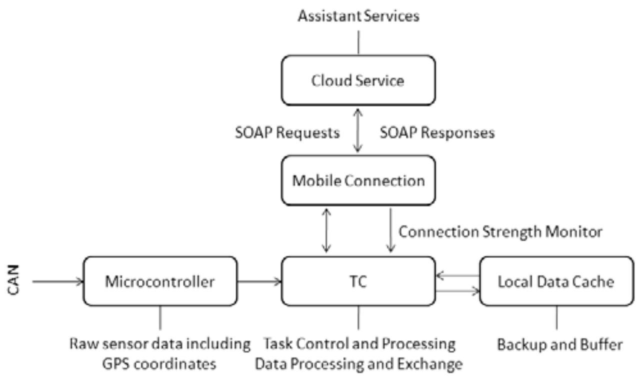


Figure 6. Main actors in real-time data collection of mobile farming.

connection strength dropping below the threshold level, the TC starts buffering data locally. The transactions continue when the connection is restored again. In this way, also near real-time data transfer is achieved, an essential factor in implementing, e.g., real-time advisory services.

In this demonstration, the mobile connection monitoring element was excluded since the data collection and sampling were simulated. The transferred data samples were based on actual harvest data. Adequate data sample transfer rates from cellular (3G) and Ethernet network were achieved for, e.g., remote monitoring and advisory service implements.

The sample TC implement is simple yet powerful enough to handle SOAP (Simple Object Access Protocol) communication and the extensive parsing of XML documents required for SOAP requests and responses. The client was implemented without using any of the specialized Java SOAP classes and it is based on an article by DuCharme [23].

Architectural components consist of the following:

- Main class for handling the data transfer task
- Class for handling the SOAP requests and responses
- Class for parsing XML documents required for SOAP requests and responses
- Class for providing appropriate log in and log out request XML documents
- Class for simulating the microcontroller that provides the data collection functionalities from the sensors

Total compiled size is 18.7 kb.

A data transfer sequence starts when the client requests a session ID by sending a login request to the service. If username is authorized, the session ID for further SOAP requests is returned. After that, the main class creates a data-fetching sample provider object and begins transactions using SOAP requests called "insert crop info". A data transfer sequence ends with a log out request.

IV. CONCLUSION AND FUTURE WORK

Basic functionality of the implemented distributed system architecture was demonstrated and tested. Interfaces were designed and implemented in such a way that requirements for interoperability and scalability were fulfilled. The use of chosen standards tackled both interoperability and quality requirements. Furthermore, in this paper the research focus was on proof of concept demonstration rather than actual performance evaluation. However, a full performance will be evaluated in a near-deployment phase.

In this demonstration, interoperability of the system was implemented with OGC's WFS interfaces, which managed delivery of spatial data as vector data and images. Besides the WFS interface, the implementation of the system could be developed further with OGC's Web Coverage Service (WCS) interface, which enables delivery of the spatial data in raster format.

Raw spatial data acquired in various operations of PA requires a lot of processing and analyzing before it can be utilized in decision-making. In the future, methods for automated handling and analyzing of spatial agricultural data should be developed to advance the information services provided for farmers. Implementation of the system architecture applying OGC's Web Processing Service (WPS) interface could provide functionalities for geospatial processing of data.

In the developed system, the usability of the graphical user interfaces was out of the focus. User acceptance and user experiences will be studied further in order to obtain applications that are fast and easy for farmers to utilize without special knowledge in the use of computers and databases.

ACKNOWLEDGMENT

The authors want to thank everybody who has contributed to discussions of the topic and made this study possible. This work was supported by several European Regional Development Fund Projects. The authors are grateful to all external financial supporters of these projects: Council of Oulu Region, City of Oulu, City of Ylivieska, Kerttu Saalasti Foundation, Ylivieska Sub-Region, Nivala-Haapajärvi Sub-Region, and Haapavesi-Siikalatva Sub-Region. The authors warmly thank the Croptech team of MTT Agrifood Research Finland, especially team leader L. Pesonen and researcher J. Kaivosoja.

REFERENCES

[1] G. Steinberger, M. Rothmund, and H. Auernhammer, "Mobile farm equipment as a data source in an agricultural service architecture," Computers and electronics in agriculture, vol. 65, Oct. 2008, pp. 238-246, doi:10.1016/j.compag.2008.10.005.

[2] E. Nash, P. Korduan, and R. Bill, "Applications of open geospatial web services in precision agriculture: a review," Precision Agric, vol. 10, Aug. 2009, pp. 546-560, doi:10.1007/s11119-009-9134-0.

[3] L. Pesonen, H. Koskinen, and A. Rydberg, InfoXT – User-centric mobile information management in automated plant production. Oslo, Norway: Nordic Innovation Centre, 2008.

[4] A. K. W. Yeung and G. Brent Hall, Spatial Database Systems. Dordrecht, The Netherlands: Springer, 2007.

[5] S. Nittel, A. Stefanidis, I. Cruz, M. Egenhofer, D. Goldin, A. Howard, A. Labrinidis, S. Madden, A. Voisard, and M. Worboys, "Report from the First Workshop on GeoSensor

Networks," SIGMOD Record, vol. 33(1), Mar. 2004, pp. 141-144, doi:10.1145/974121.974146.

[6] M. Zündt, A Distributed Community-Based Location Service Architecture. Doctoral Dissertation. Germany: Technical University of Munich, 2007.

[7] X. Chu and R. Buyya, "Service Oriented Sensor Web," in Sensor Network and Configuration: Fundamentals, Standards, Platforms, and Applications, N. P. Mahalik, Ed. Germany: Springer-Verlag, 2007, pp. 51-74.

[8] OGC – Open Geospatial Consortium, Inc., OpenGIS Geography Markup Language (GML) Endocing Standard, Version 3.2.1, Aug. 2007.

[9] J. Kangasharju and S. Tarkoma, "Benefits of alternate XML serialization formats in scientific computing," Proc. of the 2007 workshop on Service-oriented computing performance: aspects, issues, and approaches, 2007, pp. 23–30, doi:10.1145/1272457.1272461.

[10] A. Sheth, C. Henson, and S. S. Sahoo, "Semantic sensor web," IEEE Internet Computing, vol. 12(4), Jul. 2008, pp. 78–83, doi:10.1109/MIC.2008.87.

[11] M. E. Botts, G. Percivall, C. Reed, and J. Davidson, "OGC Sensor Web Enablement: Overview and high level architecture," Proc. of the 2nd International Conference on Geosensor Networks, 2006, pp. 175–190.

[12] S. Chien, D. Tran, A. Davies, M. Johnston, J. Doubleday, R. Castano, L. Scharenbroich, G. Rabideau, B. Cichy, S. Kedar, D. Mandl, S. Frye, W. Song, P. Kyle, R. LaHusen, and P. Cappaelare, "Lights Out Autonomous Operation of an Earth Observing Sensorweb," Proc. of the 7th International Symposium on Reducing the Cost of Spacecraft Ground Systems and Operations (RCSGSO 2007, AIAA), Jun. 2007.

[13] International Organization for Standardization, ISO 11783-10, Tractors and machinery for agriculture and forestry – Serial control and communications network - Part 10: Task controller and management information system data interchange, 2009.

[14] J. Tervonen, V. Sorvoja, M. Tikkakoski, I. Hakala, P. Weckström, M. Bialowas, H. Malinen, and M. Turpeenoja, "Control System Utilising Wireless Communication and GPS Position for a Direct Seeding Drill," Proc. of International Conference on Machine Automation (Smart Systems 2006 & ICMA 2006), Jun. 2006.

[15] The Institute of Electrical and Electronics Engineers, IEEE 802.15.4, Wireless Medium Access Control and Physical Layer Specifications for Low-Rate Wireless Personal Area Networks, Sep. 2006.

[16] OGC – Open Geospatial Consortium, Inc., OpenGIS Sensor Model Language (SensorML) Implementation Specification, Version 1.0.0, Jun. 2007.

[17] OGC – Open Geospatial Consortium, Inc., Binary Extensible Markup Language (BXML) Encoding Specification, Version 0.0.8, Jan. 2006.

[18] M. Luimula, Z. Shelby, J. Markkula, J. Tervonen, P. Weckström, and P. Verronen, "Developing Geosensor Network Support for Locawe Platform - Application of Standards in Low-Rate Communication Context," Proc. of the International Conference on Pervasive Services (ICPS 2009), ACM Press, Jul. 2009, pp. 73-82, doi:10.1145/1568199.1568211.

[19] M. Luimula, Development and Evaluation of the Location-aware Platform. Main Characteristics in Adaptable Location-aware Systems. Doctoral dissertation. Finland: Oulu University, Acta Universitatis Ouluensis, 2010.

[20] M. Luimula, K. Sääskilahti, T. Partala, S. Pieskä, and J. Alaspää, "Remote Navigation of a Mobile Robot in a RFID-augmented Environment," Personal and Ubiquitous Computing, vol. 14, 2010, pp. 125-136, doi:10.1007/s00779-009-0238-3.

[21] J. Tiusanen, "Wireless Soil Scout prototype radio signal reception compared to the attenuation model," Precision Agric, vol. 10, 2009, pp. 372-381, doi:10.1007/s11119-008-9096-7.

[22] J. Tiusanen, "Validation and results of the Soil Scout signal attenuation model," Biosystems Engineering, vol. 97, May 2007, pp. 11-17, doi:10.1016/j.biosystemseng.2007.02.005.

[23] B. DuCharme, A simple SOAP client - A general-purpose Java SOAP client, May 2001, <http://www.ibm.com/developerworks/xml/library/x-soapcl/index.html> 22.11.2011.

Footprint-Based 3D Generalization of Building Groups for Virtual City Visualization

Shuang He, Guillaume Moreau, Jean-Yves Martin
L'UNAM Université, Ecole Centrale Nantes, CERMA
Nantes, France
e-mail: {Shuang.He; Guillaume.Moreau; Jean-Yves. Martin}@ec-nantes.fr

Abstract - **This paper proposes a footprint-based generalization approach for 3D building groups in the context of city visualization. The goal is to reduce both geometric complexity and information density, meanwhile maintaining a rather recognizable shape. The emphasis is placed on converting 3D generalization tasks into 2D issues via buildings' footprints. In order to find suitable units for footprint projection and generalization, which should hold both semantic meaning and simple geometry, a meaningful partition is firstly introduced (from CityGML building models). For roof generalization, a new perspective is presented: to divide a building model into Top + Body, so that the Top part could be transplanted onto the extruded model by displacement. Two algorithms are developed for two types of building groups: one with a minor height difference and the other with a major height difference. For the former one, the outer units are detected and aggregated to represent the whole group. For the latter one, an iteration of aggregation is performed on subgroups. Each time the highest unit and its neighbors compose the subgroup. The algorithms are tested on two building groups and one part of a 3D city model.**

Keywords-3D generalization; building group; footprint; city visualization

I. INTRODUCTION

3D city visualization requires different representations of building models at different Levels of Detail (LoDs) to satisfy different scales and application needs. These LoDs should be generated automatically by specific generalization procedures. Generalization has a long history in cartography [1], with the goal of emphasizing the most important map elements while still representing the world in the most faithful and recognizable way. 3D building generalization in city visualization shares the same goal, but should consider both geographical and 3-dimensional information.

As discussed and listed in [2], unlike 2D maps that have standard official scale series, there are no generally agreed LoDs for 3D buildings. Including the four LoDs defined by CityGML (City Geography Markup Language) [3], the existing definitions of LoDs for 3D buildings only differentiate by 3D details. That is to say, they hardly respond to geographical generalization, like the generalization regarding a group of 3D building, where topological relations should also be considered. This seems to lead more attention to single building generalization.

A number of algorithms have been developed for 3D building generalization [4-13]. Most of those algorithms deal with single buildings [4-11]. Generalization of building groups is seldom addressed [12, 13]. In 3D city visualization, the goal of generalization is not only to simplify individual objects, but also to achieve better cognition by emphasizing important features. Thus, there rises a generalization need for building groups. Both 3-dimensional detail and geographical relations should be taken into account. More generalization operations like selection, aggregation, typification and their combinations are expected.

Footprint has been serving as the connection between 2D and 3D. Plenty of block models of buildings were extruded from cadastral maps using their footprints and heights. But more detailed models couldn't be acquired in this way. Therefore, a question rises here: how can we translate 3D building generalization issues into 2D scope for generalizing more detailed 3D building models?

This paper is organized as below: related work is first discussed in Section II. Section III introduces the idea of partitioning a 3D building model into suitable units for footprint-based generalization. Generalization algorithms for two types of building groups are presented in Section IV. Experimental results are given in Section V. Section VI concludes the paper.

II. RELATED WORK

Compared with map generalization techniques in 2D, generalization in 3D is still in its infancy [14]. Different from general 3D models, most 3D building models are already low-polygon objects, so generic geometrical simplification techniques from Computer Graphics seem to be of little use. Besides, parallel and orthogonal properties of buildings need to be respected during simplification. Therefore, algorithms for 3D building generalization need to be specifically designed [14]. Thiemann proposed to segment a building into basic 3D primitives [4], and to decompose the whole generalization process into segmentation, interpretation and generalization phases [5]. Mayer [6] and Forberg [7] developed scale-space techniques for simplifying buildings, partly based on the opening and closing morphological operators. Kada proposed to define parts of simplified buildings as intersections of half-planes [8] and to divide buildings into cells and to detect features by primitive instancing [9]. Without semantic information, these methods mainly detect building features based on pure geometry.

ISBN: 978-1-61208-178-6

By taking semantic information into account, Fan et al. [10] proposed a method for generalization of 3D buildings modeled by CityGML from LoD3 to lower LoDs. Their research showed that good visualization properties could be obtained by only using the exterior shell of the building model that drastically decreases the required number of polygons. Fan and Meng [11] extended their work to automatic derivation of LoDs for CityGML building models staring from LoD4. However, the above mentioned methods are all limited to generalization regarding single buildings.

Anders [12] proposed an approach for the aggregation of linearly arranged building groups. Their 2D silhouettes, which are the results of three orthogonal projections, are used to form the generalized 3D model. Guercke et al. [13] studied the aggregation of LoD1 building models in the form of Mixed Integer Programming (MIP) problems.

Techniques start emerging for generalizing 3D building groups in the context of city visualization. Glander and Döllner [14] proposed cell-base generalization by maintaining a hierarchy of landmarks. In each cell, only landmark buildings can be seen, the other buildings are replaced by a cell block. In the work of Mao et al. [15], buildings are divided into clusters by road network, and grouped with close neighbors in each cluster. However, only LoD1 buildings were handled.

III. Partition of Building Model

Our approach places the emphasis on translating 3D generalization into 2D scope. The strategy is to generate footprints of 3D buildings, perform 2D generalization on their footprints, and then extend the result to 3D. The main issue is how we extend the result to 3D without losing recognizable features like differentiated height and roof. Therefore, a meaningful partition is proposed at first, so that each footprint can carry feature information.

An implementation of partition is presented using building models encoded by CityGML [3], which supports coherent modeling of semantics and geometrical/topological properties. With semantically structured buildings models, generalization can be facilitated a lot. However, a building still can be structured in plenty of forms that make a uniform projection of footprint very difficult. Stricter rules are needed to form buildings into favorable partitions.

A. CityGML Building Model at LoD2

CityGML defines a standard for ontology of buildings at 4 different LoDs. At LoD1, 3D buildings are represented by block models with flat roofs. At LoD2, 3D buildings have differentiated height and roof structures. LoD3 models are detailed architectural models with openings like windows and doors. LoD4 completes a LoD3 model with interior structures.

This paper uses CityGML LoD2 building models for the partition and generalization. LoD1 block models are hardly recognizable; highly detailed models at LoD3 and LoD4 are too costly and normally only used for landmarks.

Before partitioning, we should be aware of the possible elements in such a model. So we draw a UML diagram of building model exclusively for LoD2 (Figure 1) according to the CityGML encoding standard [3]. The pivotal class is the abstract class *_AbstractBuilding*, which is specialized either to a *Building* or to a *BuildingPart*. Each can contain 3 types of properties: text attribute, pure geometry, and semantically structured geometry. The last one is the essential for our footprint-based generalization.

B. Partition Rules

The goal of partition is to get a well structured building in both semantic and geometric sense, so as to extract suitable unit for footprint projection and generalization. The unit should have meaningful geometry and be good for computation. The rules are introduced as below:

- If a *Building* is composed of unconnected segments, partition them into different *Buildings*.

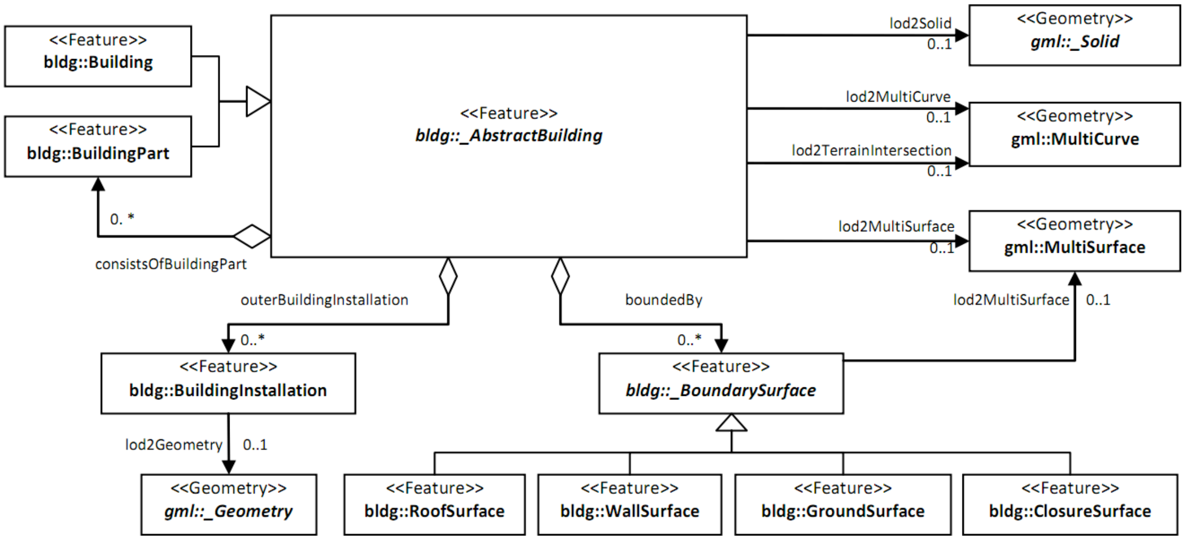


Figure 1. UML diagram of CityGML's building model at LoD2 (based on CityGML standard [3])

- If a *Building* is composed of structural segments differing in e.g. height or roof type, partition them into different *BuildingParts*.
- If a *Building/BuildingPart* has smaller components which are not significant as a *BuildingPart* (e.g. chimneys, dormers, and balconies), partition them into *BuildingInstallations*.
- If a *Building/BuildingPart* has geometries without semantic information, partition them into pure geometry.
- If a *Building/BuildingPart* has *_BoundarySurfaces* and includes *BuildingParts* at the same time, partition the *_BoundarySurfaces* into a new *BuildingPart*.
- If a *Building/BuildingPart* includes only one *BuildingPart*, aggregate the included *BuildingPart* into its parent *Building/BuildingPart*.
- If a *Building* has *_BoundarySurfaces*, there must be a *WallSurface* starting from and orthogonal to the ground plane; otherwise, partition this *Building* as a *BuildingInstallation* into another *Building*.
- If a *BuildingPart* has *_BoundarySurfaces*, there must be a *WallSurface* starting from and orthogonal to the ground plane; otherwise, partition this *BuildingPart* into a *BuildingInstallation*.
- If a *Building/BuildingPart* has unconnected or self-intersected *WallSurface*, partition it into more *BuildingParts*.

C. Beneficial Attributes

After employing the partition rules, beneficial attributes can be obtained:

- In a *Building* tree, all the leaf nodes must have *_BoundarySurfaces*; no branch nodes can then have *_BoundarySurfaces*.
- If there are *_BoundarySurfaces*, there must be a *WallSurface*; other types of surfaces are optional.
- A *WallSurface* must start from and orthogonal to the ground plane.
- The orthogonal projection of *WallSurfaces* of each leaf node form a simple polygon or polyline.
- Each leaf node only has one height.

A leaf node can contain text attributes, pure geometry, *BuildingInstallations* and *_BoundarySurfaces*, but only *_BoundarySurfaces* will be selected to form a basic unit of generalization. The term *unit* will be used in the following discussion, referring to a leaf node of a building tree only consisting of *_BoundarySurfaces*. Two examples are given in Figure 2.

IV. Generalization of Building Groups

In a building group, the adjacent buildings can be connected or disjoint. In this paper, we only deal with the ones with connected buildings.

There exists building groups with various features. They can hardly be generalized by a uniform method. Since height has significant influence on visual perception, we address two types of building groups in this paper: one with a minor difference in height (buildings that all look similar) and the other with a major difference in height (that include a significant building at city level).

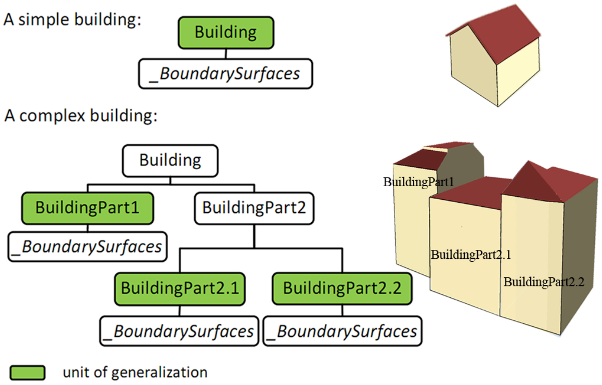


Figure 2. Two examples of building tree for generalization

A. Generating Footprints

Based on the partition discussed in the previous section, the first performed generalization operator is selection. For each *unit*, only *_BoundarySurfaces* are selected. Among *_BoundarySurfaces*, only *WallSurface* will be selected for generating footprint. However, the roof information will be lost during this projection of footprint. Another important issue is how we generalize roofs.

B. Handling Building Roofs

A common way of roof generalization is by primitive matching of different roof types. But type detection is a costly (most often manual) and uncertain process depending on the given types and lots of parameters. In CityGML building models, roof surfaces are separated from walls, but roof type is not always available in attributes. Even if given the roof type, the rebuilding of roof after extrusion would be another difficulty without knowing parameters.

Since an extruded model is usually a prism, if the original model could be divided into a top part and a prism body, the top part could be easily transplanted to the extruded model and could also be generalized with adjacent roofs. Therefore, we propose a way of dividing a building model into *Top* + *Body*. For a building, if all of its walls end at the top in the

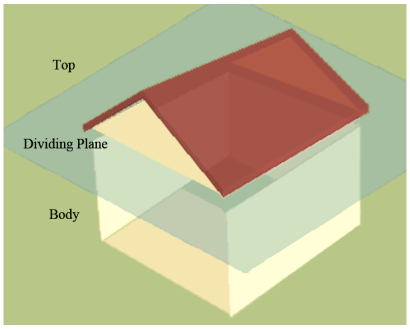


Figure 3. An example of dividing a building into *Top* + *Body*

ISBN: 978-1-61208-178-6

same horizontal plane, the *Top* only consists of its roof; otherwise, the *Top* consists of its roof and the end walls. An example is given in Figure 3.

C. Generalization of Building Groups with a Minor Difference in Height

For a building group with a minor difference in height, we believe its outer feature could represent the whole group to a certain extent, like in large scale visualization.

Therefore, inner *units* can be eliminated. Outer *units* can also be aggregated. If there are no inner *units*, aggregation can be directly performed on all *units*. If aggregated *units* have non-flat roofs, two levels of aggregation can be achieved. The original roof structures can be preserved based on the approach introduced in subsection A. They can be generalized to flat roofs as well.

Our generalization operations start from LoD2 but won't lead to LoD1 block models. Instead of using the term LoD, we use GeoLoD (Geographical Level of Detail) to denote the generalization results. Three GeoLoDs can be generated. At GeoLoD1, a building group is a prism model which conveys its outer shape. At GeoLoD2, a building group is a GeoLoD1 model added with differentiated roof. At GeoLoD3, a building group is represented by all its outer *units*. The main flow of the algorithm is depicted in Figure 4.

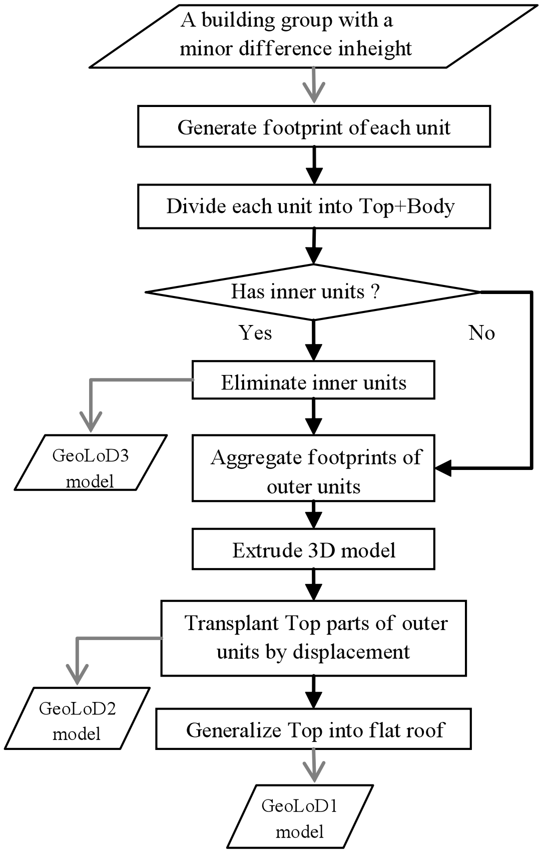

A building group with a minor difference inheight
Generate footprint of each unit
Divide each unit into Top+Body
Has inner units ?
Yes
No
Eliminate inner units
GeoLoD3 model
Aggregate footprints of outer units
Extrude 3D model
Transplant Top parts of outer units by displacement
GeoLoD2 model
Generalize Top into flat roof
GeoLoD1 model

Figure 4. The main flow of the generalization algorithm for a building group with a minor difference in height

D. Generalization of Building Groups with a Major Difference in Height

For a building group with a major difference in height, the generalization will be performed on its subgroups.

A subgroup is composed of a center *unit* and its adjacent neighbors. Each time the highest *unit* will be chosen from the unprocessed *units*. If this *unit* is much higher than its neighbor, they should not be aggregated. We use the term *coequal* in this paper to indicate that the height difference in two *units* can be ignored, that is to say, they can be aggregated. *Coequal units* are defined as below:

Given two units $\boldsymbol{U_1}$ *with height* $\boldsymbol{h_1}$ *and* $\boldsymbol{U_2}$ *with height* $\boldsymbol{h_2}$*, if they satisfy the constraint as in (1),* $\boldsymbol{U_1}$ *and* $\boldsymbol{U_2}$ *are coequal, where* $\boldsymbol{Th_1}$ *is a predefined variable as the threshold.*

$$1/\boldsymbol{Th_1} < \boldsymbol{h_1}/\boldsymbol{h_2} < \boldsymbol{Th_1}, \qquad (1)$$

When merging two adjacent and coequal *units*, either height of the original *units* can be assigned to the new *unit*. If the lower *unit* covers a rather large area, the new *unit* takes the lower height; otherwise, it takes the higher one. We propose a criterion as below:

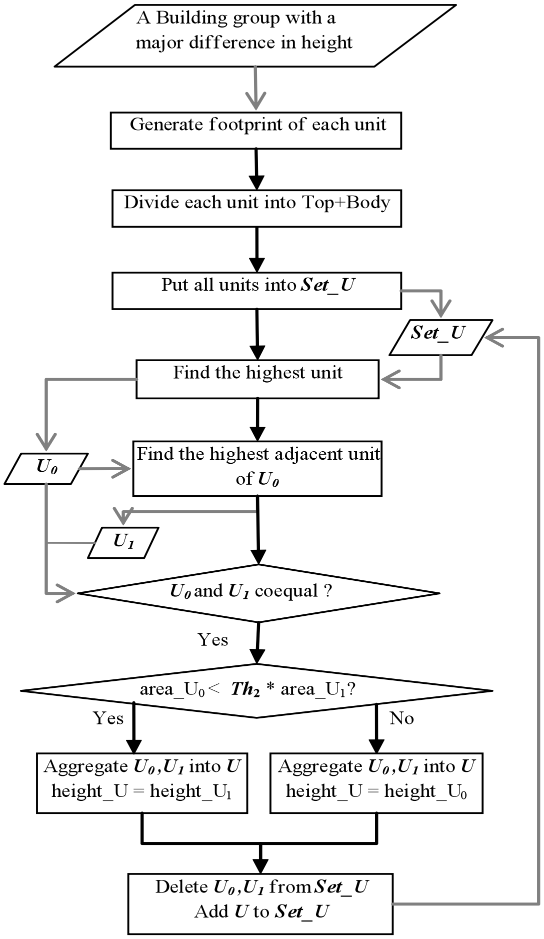

A Building group with a major difference in height
Generate footprint of each unit
Divide each unit into Top+Body
Put all units into Set_U
Set_U
Find the highest unit
U0
Find the highest adjacent unit of U0
U1
U0 and U1 coequal ?
Yes
area_U0 < Th2 * area_U1?
Yes
No
Aggregate U0,U1 into U height_U = height_U1
Aggregate U0,U1 into U height_U = height_U0
Delete U0,U1 from Set_U Add U to Set_U

Figure 5. The main flow of the generalization algorithm for a building group with a major difference in height

Given $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, *and* $\boldsymbol{h_1}$, $\boldsymbol{h_2}$ *(*$\boldsymbol{h_1} < \boldsymbol{h_2}$*) as the areas and heights of two coequal units, the height of new merged unit* $\boldsymbol{h_3}$ *is determined as in (2), where* $\boldsymbol{Th_2}$ *is a predefined variable as the threshold.*

$$h_3 = \begin{cases} h_1 ; a_1 \geq Th_2 \bullet a_2 \\ h_2 ; a_1 < Th_2 \bullet a_2 \end{cases}, \tag{2}$$

The main flow of the algorithm is depicted in Figure 5.

V. Experimentation and Results

The footprint-based generalization approach presented in Section IV will be tested on two sets of building groups.

A. Generalization of Building Groups with a Minor Difference in Height

The algorithm presented in Section IV-C is tested on a building consisting of 381 *units*. The results are shown in Figure 6, and the statistics are given in Table I.

TABLE I. STATISTICS1

Model	Footprint		3D model (percentage of the original)	
	Vertex	*Polygon*	*Vertex*	*Polygon*
Original	361	44	1565	381
GeoLoD3	186	18	843 (53.9%)	203 (53.3%)
GeoLoD2	121	3	679 (43.4%)	153 (40.2%)
GeoLoD1	121	3	590 (37.7%)	121 (31.8%)

B. Generalization of Building Groups with a Major Difference in Height

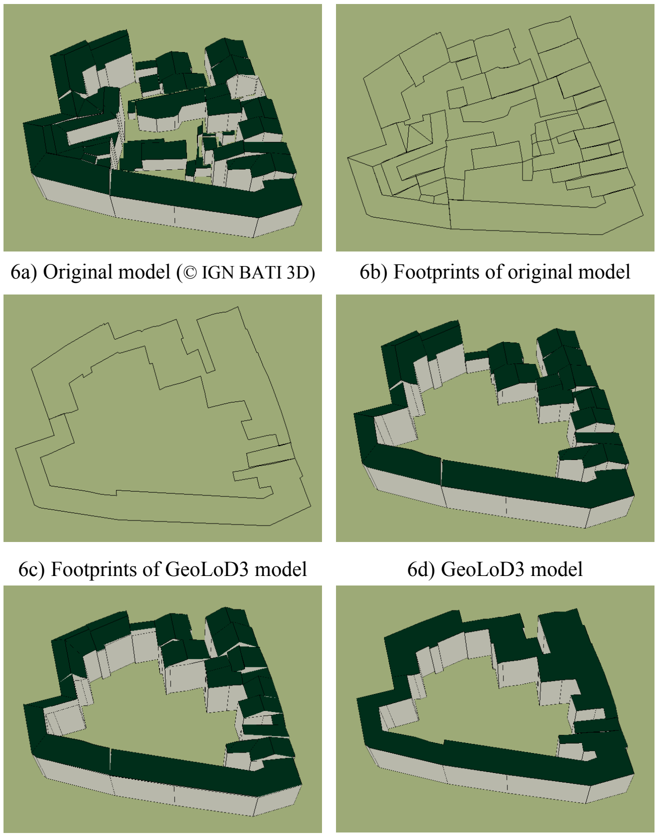
The algorithm presented in Section IV-D is tested on a building group consisting of 193 *units*. The results are shown in Figure 7, and the statistics are given in Table II. As for the threshold factor $\boldsymbol{Th_1}$ and $\boldsymbol{Th_2}$, we assign 2 to $\boldsymbol{Th_1}$, and 4 to $\boldsymbol{Th_2}$. Of course, other values could be assigned.

TABLE II. STATISTICS2

Model	Footprint		3D model (percentage of the original)	
	Vertex	*Polygon*	*Vertex*	*Polygon*
Original	193	19	879	211
Generalized	118	5	565 (64.3%)	116 (55%)

C. Generalization of Building Groups in 3D City Model

The approach is tested on a part of 3D city model of Nantes in France, which consists of 346 buildings (1536 *units*). The generalized result is shown in Figure 8. As we could see, both geometrical complexity and information density are reduced; meanwhile essential features are preserved and emphasized.

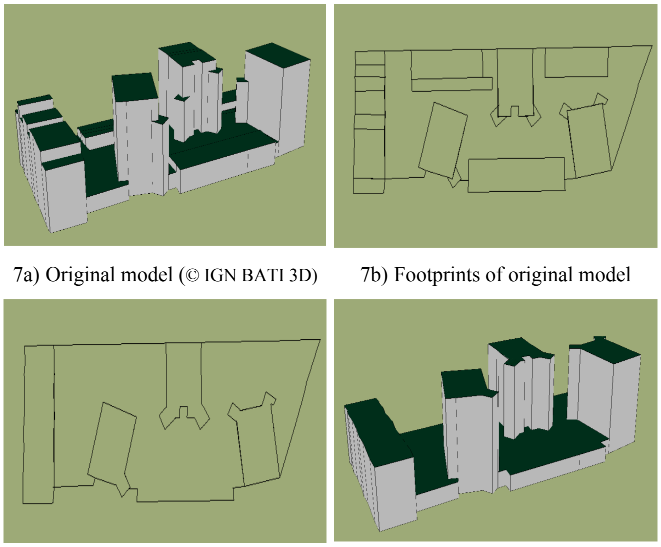


6a) Original model (© IGN BATI 3D) 6b) Footprints of original model

6c) Footprints of GeoLoD3 model 6d) GeoLoD3 model

6e) GeoLoD2 model 6f) GeoLoD1 model

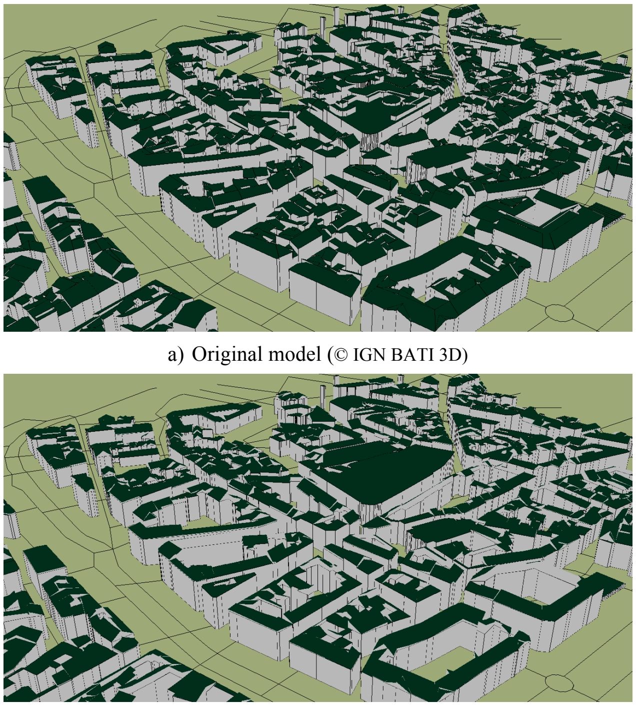
Figure 6. A generalization example of a building group with a minor difference in height



7a) Original model (© IGN BATI 3D) 7b) Footprints of original model

7c) Generalized footprints 7d) Generalized model

Figure 7. A generalization example of a building group with a major difference in height



a) Original model (© IGN BATI 3D)

b) Generalized model

Figure 8. A generalization example of 3D city model

VI. Conclusion

This paper presented a novel approach for generalizing 3D building groups. The goal is to reduce both geometric complexity and information density, meanwhile maintaining recognizability, which requires at least LoD2 models. The emphasis has been placed on translating 3D generalization issues into 2D scope via footprints. First of all, a meaningful partition was suggested so that each footprint can carry feature information. A set of partition rules was developed for partitioning the buildings modeled by CityGML at LoD2.

Footprint-based generalization is then confronted with the difficulty of roof generalization. Unlike the existing approaches such as primitive matching, we proposed to divide a building into *Top* + *Body*. Thus, *Top* part can be easily transplanted onto the extruded model by displacement.

Two types of building groups were addressed in this paper: one has major difference in height and the other has minor difference in height. For the former one, we believe its outer feature can represent the whole group to a certain extent. For the latter one, it should not be handled as a whole. An iterative aggregation process is performed by comparing the height and area of every two adjacent units starting from the highest one.

The approach was tested on two building groups and a part of 3D city model. Group generalization shows its advantage in reducing information density, e.g. by eliminating insignificant buildings. Different from the methods only handle LoD1 block models [13, 14], our approach can handle LoD2 models as well. Instead of aggregating detailed models directly into LoD1 blocks [15], our approach supports generalization of geographical LoDs, thereby achieving data reduction and maintaining recognizability at the same time.

However, only connected buildings have been handled and only two types of building groups have been addressed. More studies are needed for dealing with complex cases. Coarser levels (including addressing the issue whether 3D is still necessary) will also be studied.

References

[1] R. McMaster and K. Shea, Generalization in Digital Cartography. Washington, DC: Association of American Geographers, 1992.

[2] L. Meng and A. Forberg, “3D Building Generalisation”, in Generalisation of Geographic Information: Cartographic Modelling and Applications, chapter 11, W. Mackaness, A. Ruas, and T.Sarjakoski, Eds. Elsevier, 2007, pp. 211- 232.

[3] OpenGIS® CityGeography Markup Language (CityGML) Encoding Standard, Version: 1.0.0, OGC® Project Document, No. 08-007r1, 2008.

[4] F. Thiemann, “Generalization of 3D Building Data,” Proc. Geospatial Theory, Processing and Applications, IAPRS Vol. 34, Part B4, Ottawa, Canada, 2002, pp. 286-290.

[5] F. Thiemann and M. Sester, “Segmentation of buildings for 3D-generalisation”, Proc. Working Paper of the ICA Workshop on Generalisation and Multiple Representation, Leicester, UK, 2004.

[6] H. Mayer, “Scale-Spaces for Generalization of 3D Buildings”, International Journal of Geographical Information Science, Vol. 19, No. 8-9, Sept.-Oct. 2005, pp. 975-997.

[7] A. Forberg, “Generalization of 3D Building Data Based on a Scale-Space Approach”, ISPRS Journal of Photogrammetry and Remote Sensing 62, 2007, pp. 104-111.

[8] M. Kada, “3D Building Generalization Based on Half-Space Modeling,” Proc. ISPRS Workshop on Multiple Representation and Interoperability of Spatial Data, Hannover, 2006.

[9] M. Kada, “Scale-Dependent Simplification of 3D Building Models Based on Cell Decomposition and Primitive Instancing,” in Spatial Information Theory. Springer, 2007.

[10] H. Fan, L. Meng and M. Jahnke, “Generalization of 3D Buildings Modelled by CityGML,” in Advances in GIScience: Lecture Notes in Geoinformation and Cartography. Springer, 2009, pp. 387-405.

[11] H. Fan and L. Meng, “Automatic Derivation of Different Levels of Detail for 3D Buildings Modeled by CityGML,” Proc. 24th International Cartographic Conference, 2009.

[12] K.-H. Anders, “Level of Detail Generation of 3D Building Groups by Aggregation and Typification,” Proc. International Cartographic Conference, 2005.

[13] R. Guerche, T. Götzelmann, C. Brenner and M. Sester, “Aggregation of LoD 1 Building Models as an Optimization Problem,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 66(2), 2011, pp. 209-222.

[14] T. Glander and J. Döllner, “Abstract Representations for Interactive Visualization of Virtual 3D City Models,” Computers, Environment and Urban Systems, vol. 33, no. 5, 2009, pp. 375 - 387.

[15] B. Mao, Y. Ban and L. Harrie, “A Multiple Representation Data Structure for Dynamic Visualisation of Generalised 3D City Models,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 66(2), 2011, pp. 198-208.

A Web-based Environment for Analysis and Visualization of Spatio-temporal Data provided by OGC Services

Maxwell Guimarães de Oliveira, Cláudio de Souza Baptista and Ana Gabrielle Ramos Falcão
Laboratory of Information Systems - Computer Science Department
Federal University of Campina Grande (UFCG)
Campina Grande, Brazil
E-mails: maxwell@ufcg.edu.br, baptista@dsc.ufcg.edu.br, anagabrielle@copin.ufcg.edu.br

***Abstract*—The popularity of GPS devices has led to the quick increasing of spatial data volume in the web. Although there are several studies on spatial data sets, not many deal with the temporal variation that may exist on these sets. Most of the approaches that implement a visual analysis of spatiotemporal data still reveal limitations regarding its flexibility, usability and, mostly, generalization. In order to improve these limitations, we propose a new approach, a web-based spatiotemporal viewer and analyzer, which is domain-independent, deals with the temporal variation, and may be connected to any map server that implements the Web Map Service and Web Feature Service, specified by the Open Geospatial Consortium, and thus, it promotes interoperability. Furthermore, our approach includes data mining clustering algorithms, providing an intuitive visual analysis of the results. We performed a case study to validate the proposed solution and the improvements in spatiotemporal visual analysis. The results showed that the new proposed approach facilitates the analysis of spatiotemporal data by humans.**

Keywords—visualization; analysis; spatiotemporal; data mining; OGC services.

I. INTRODUCTION

The constant growth of the use of GPS-based (Global Positioning System) devices, such as smartphones, along with the ease of sharing the information from such devices in the internet have substantially increased the volume of spatiotemporal data in the web.

Such spatiotemporal data require visual and analytical tools in order to improve the decision-making process. These tools should be intuitive so that little time is spent obtaining relevant conclusions from the analysis process, such as information on predictions, recurrence patterns, and clustering.

Visualization techniques are well known for improving the decision support process [1], once they take advantage on the human skills for quickly perceiving and interpreting visual patterns [2][3]. However, it has been argued that the visualization resources provided by most of the existing geographic-based applications are not enough for decision support systems when used solely [4].

Furthermore, spatiotemporal data impose serious challenges for analytics. First of all, due to the geographical space complexity, that requires human involvement and his/hers sense of determination of space, place, and spatial relationships [5]. Secondly, due to the complexity of the temporal dimension. Time flows linearly, however, some events that occur over time may be periodically recurrent, with multiple cycles, forming hierarchical structures that overlap and interact with each other. Hence, temporal data analysis also requires human involvement [6].

It is necessary to perform analysis over data stored in heterogeneous and distributed data sources. In addition to the complex features of spatiotemporal data, the existence of heterogeneous sources results in an interoperability problem. Aiming at minimizing such problem, the Open Geospatial Consortium (OGC) [7] proposes standards for heterogeneous spatial databases connectivity services, such as the Web Map Service (WMS) and the Web Feature Service (WFS), broadly used by GIS (Geographic Information Systems) applications.

We propose in this paper a new approach, the GeoSTAT (Geographic Spatiotemporal Analysis Tool) system in order to address the problem of the lack of systems that may provide visualization and clustering techniques for large spatiotemporal datasets. GeoSTAT is a web-based environment that implements several spatiotemporal data visualization technique; interoperates with distributed data sources through OGC WMS and WFS services; and provides several data mining clustering algorithms proposed in the literature.

The main contribution of this work concerns the proposal of a visual approach for the analysis of spatiotemporal data that:

- Facilitates the exploration of the spatial and temporal dimensions;
- Provides interoperability and domain independence; and
- Integrates data mining algorithms into the visual analysis.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 focuses on the proposed environment. Section 4 addresses a case study to validate the proposed ideas. Finally, Section 5 concludes the paper and points out further work to be undertaken.

II. Related Work

There are many works based on spatial data visualization, but not many deal with the visualization of spatiotemporal data.

Reda et al. [8] present a tool that enables the visual exploration of the changes in dynamic social networks over time. It is an application focused on the domain of social networks visualization and it introduces a data structure based on a 3D cube for the spatiotemporal visualization.

Lu et al. [9] address a web-based system for the visualization of historical spatiotemporal data of the metropolitan area of Washington, D.C., in the United States. Apart from being an application based on a specific domain, their tool does not use georeferenced maps, dealing with the data visualization merely by several chart types.

He et al. [10] highlight another domain specific study. The authors developed a spatiotemporal data visualization system on the domain of oceans. It is a web-based system that uses 2D and 3D maps for the spatial visualization. The time visualization is static (has no animations), based on charts and triggered by user that chooses a region of interest in the map and a target time interval.

Chen et al. [11] implemented a tool that integrates several visualization techniques (GIS, self-organizing maps, hierarchical lists, periodical views, timeline views, etc.) for the criminal analysis domain. Although it is domain specific, this tool introduces interesting functionalities, like a time slider that allows time variations of the data over a georeferenced 2D map with basic interactivity tools such as pan and zoom. In spite of using several visualization techniques, the interface of the proposed application may get overloaded, making the user confused.

We have seen so far works that address spatiotemporal data visualization exploring several visualization techniques chosen according to specific domain. Next, we relate important works that propose several spatiotemporal visualization techniques.

Andrienko et al. [12] propose a framework based on the Self-Organizing Map (SOM) technique, combined with a number of interactive visualization techniques for the analysis of spatiotemporal data from two perspectives: spatial distributions that vary over time; and local time variation profiles distributed over time. This approach promises to be domain independent, gathering visualization techniques based on maps to enable the data analysis.

Compieta et al. [13] focus on issues related to the complexity of the manipulation, analysis and visualization of spatiotemporal data sets. The authors propose a spatiotemporal data mining system based on association rules and several techniques for the visualization and interpretation of georeferenced maps.

Andrienko et al. [14] argue that it is necessary to handle the time more effectively and list some characteristics that would be ideal for a good spatiotemporal data visual analysis system, such as treating and using both time and space characteristics, being visual, exploratory, scalable, collaborative, providing applicable methodologies for new and big data sets, and providing mechanisms for gathering evidences.

Considering the related works that involve spatiotemporal data visualization, most of them address specific domain solutions and do not present flexibility on obtaining the data, forcing the user to use solely the data source provided by the application.

It is necessary to conceive a spatiotemporal data visualization and analytical tool that is, mainly: flexible, to enable the user to manipulate data obtained from information sources that and to execute spatial or temporal queries over these data, according to the analysis criteria; practical, in the sense of providing an intuitive interface, with resources that assist the visualization and analysis of both spatial (map resources) and temporal features (charts, map animations); and generic, by providing all those resources for users interested on any spatiotemporal analysis domain.

It is then necessary to use visual analytical tools for spatiotemporal data, together with data mining algorithms that will enable the discovery of implicit knowledge. This is the aim of the GeoSTAT system.

III. The GeoSTAT System

This section introduces the GeoSTAT (Geographic Spatiotemporal Analysis Tool) system that implements the visualization and analysis of spatiotemporal data available on heterogeneous databases. We followed the guidance for good spatiotemporal visual analysis systems proposed by Andrienko et al. [14].

The GeoSTAT system is a web-based visualization tool, designed in three-tier architecture: Visualization, Control and Persistence. The first two tiers contain the core of our contributions. Figure 1 shows such architecture.

A. The Visualization Tier

The visualization tier is responsible basically for the user interface. The GeoSTAT viewer was implemented based on the Google Maps API [15] and provides, in addition to the basic map interactivity functionalities (drag, pan, zoom, information and scale), options for alternating between base map types (map, satellite or terrain), and adding map layers. The viewer enables the visualization of several map layers, whether they are spatial or spatiotemporal, simultaneously, regardless of the data source. For each map layer, it is possible to apply an opacity level (transparency) to allow a better visualization of the spatial information. It is also possible to execute spatial and non-spatial queries over the visible map layers. Figure 2 shows the GeoSTAT visualization screenshot with one spatiotemporal layer added to the map.

Figure 2 shows the presence of temporal controllers, in addition to the spatial data manipulation tools already presented. For the spatiotemporal map layers, the viewer offers some components such as the Temporal Slider (see component 1 in Figure 2) that allows (see component 1 in Figure 2) that allows data visualization according to their timestamps and over several possibilities of distinct tempo-

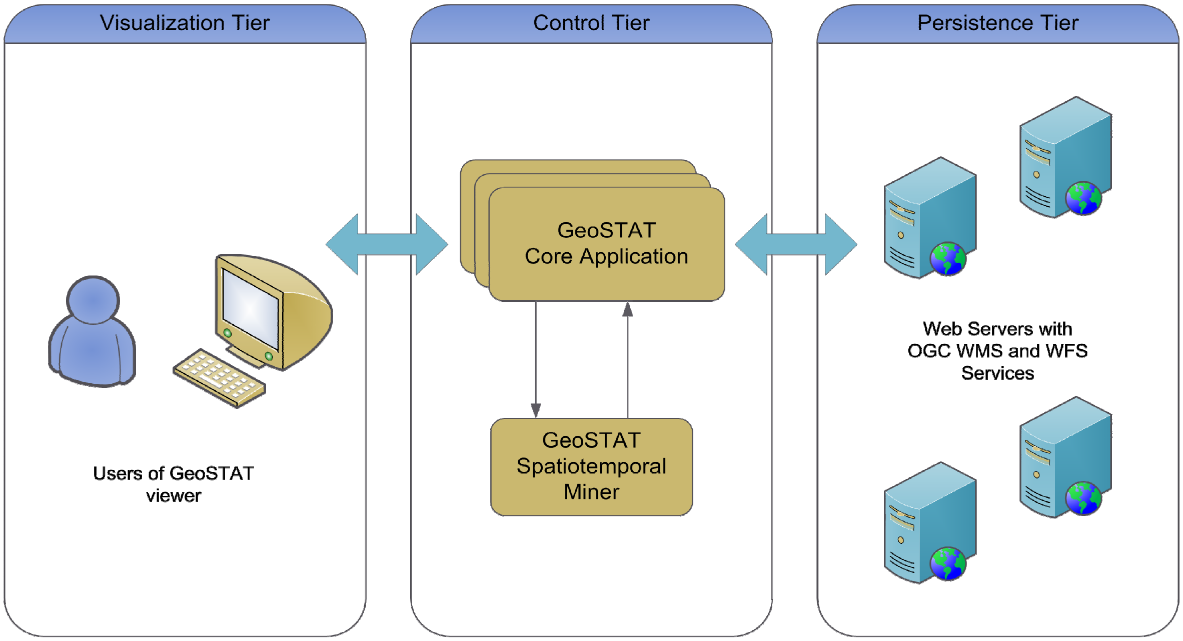


Figure 1. The GeoSTAT three-tier architecture.

-ral granularities, for example: day, month or year. It is also possible to apply temporal filters in order to reduce the number of data analyzed for relevant periods (moments or intervals), and to use interactive charts (see component 2 on Figure 2) to assist the analysis of the temporal distributions of the data.

A map layer is considered being spatiotemporal as from the moment of its inclusion in the application, the user indicates the temporal attribute of the layer.

Still at the visualization tier it is possible to execute and view the results of the clustering based spatiotemporal data mining on the spatiotemporal layers of the type POINT. Hence, the simultaneous visualization of a layer with real data and resulting data from the data mining processing may be achieved.

B. *The Control Tier*

The control tier is where the user requests from the visualization layer are treated and executed. This tier implements the application logic and direct communication with the data services that offer the spatial or spatiotemporal layers. It is responsible, mainly, for the spatial and temporal queries, resulting from the use of the Temporal Slider, and for the execution of the spatiotemporal data mining.

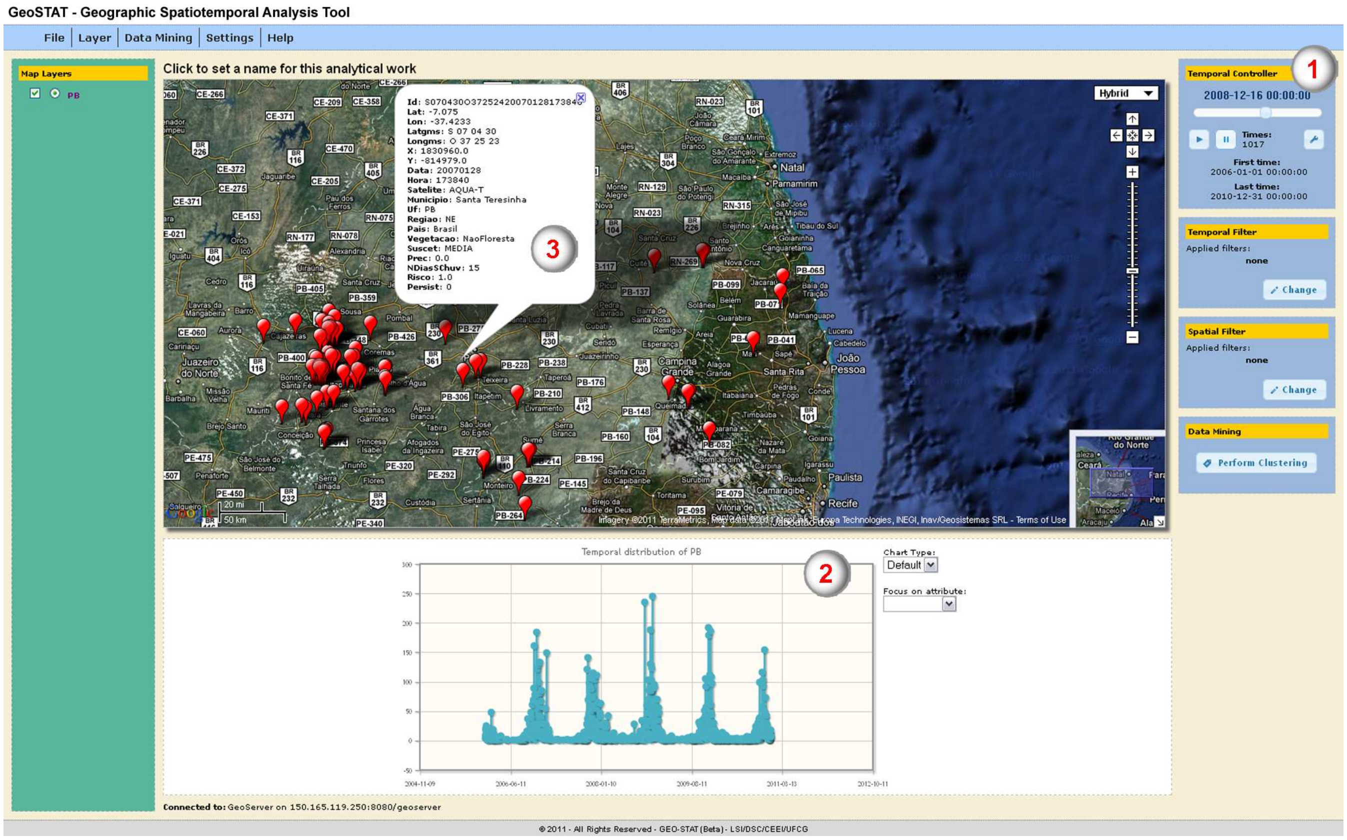
Inside the control tier there is a module responsible for the spatiotemporal data mining. This module integrates the clustering algorithms found in the Weka data mining library [16] and is extensible for the addition of other algorithms of the same nature. To provide such extensibility, we developed a communication interface that can be easily implemented for new algorithms and we established a data input format based on spatiotemporal points; and a data output format for the spatiotemporal clusters. The input format chosen was the Comma-Separated Values (shortly CSV), containing information on the spatiotemporal layer to be mined, such as the latitude, longitude and timestamp of the records. The output format chosen was XML (eXtensible Markup Language), containing relevant information over the generated clusters. An example of a generated cluster on the output format of the data mining module is presented in Code 1.

According to Code 1, each element of the type "cluster" is basically composed of an identifier, the number of instances (records) grouped (representing its density) and the cluster’s spatial and temporal elements. The spatial elements enable the visualization of the cluster on georeferenced circle format, with radius and center point clearly defined. The temporal elements specify the temporal granularity and the value.

Code 1. Snippet of the XML file that represents an example of a cluster generated by the data mining module.

```
<cluster>
    <id>1</id>
    <instances>144</instances>
    <spatial>
        <latCentroid>-8.253</latCentroid>
        <lonCentroid>-36.964</lonCentroid>
        <radius>0.14567</radius>
    </spatial>
    <temporal>
        <granularity>YEAR</granularity>
        <value>2010</value>
    </temporal>
</cluster>
```

ISBN: 978-1-61208-178-6


GeoSTAT - Geographic Spatiotemporal Analysis Tool
File
Layer
Data Mining
Settings
Help
Map Layers
PB
Click to set a name for this analytical work
Hybrid
3
1
2
Temporal Controller
Temporal Filter
Change
Spatial Filter
Change
Data Mining
Perform Clustering
Temporal distribution of PB
300
250
200
150
100
50
0
-50
2004-11-09
2006-06-11
2008-01-10
2009-08-11
2011-03-13
2012-10-11
Imagery ©2011 TerraMetrics, Map data ©2011 MapLink, Europa Technologies, INEGI, Inav/Geosistemas SRL - Terms of Use
20 mi
50 km

Id: S07043003725242007012817384
Lat: -7.075
Lon: -37.4233
Latgms: S 07 04 30
Longms: O 37 25 23
X: 1830960.0
Y: -814979.0
Data: 20070128
Hora: 173840
Satelite: AQUA-T
Municipio: Santa Teresinha
Uf: PB
Regiao: NE
Pais: Brasil
Vegetacao: NaoFloresta
Suscet: MEDIA
Prec: 0.0
NDiasSChuv: 15
Risco: 1.0
Persist: 0

2008-12-16 00:00:00
Times: 1017
First time: 2006-01-01 00:00:00
Last time: 2010-12-31 00:00:00

Applied filters: none

Applied filters: none

Chart Type: Default

Focus on attribute:

Connected to: GeoServer on 150.165.119.250:8080/geoserver

© 2011 - All Rights Reserved - GEO-STAT (Beta) - LSI/DSC/CEEI/UFCG

Figure 2. The GeoSTAT environment visualization screenshot with a spatiotemporal map layer.

C. *The Persistence Tier*

The persistence tier contains GeoServer map server that implements the OGC WMS and WFS services [17], ensuring the spatial DBMS (Database Management System) interoperability. The servers used by the users through GeoSTAT may be easily connected, through the information of the service URL and setting an alias to better identify it. With an established connection, the user has the option to perform map overlay of all the layers available from the OGC connected services.

IV. Case Study

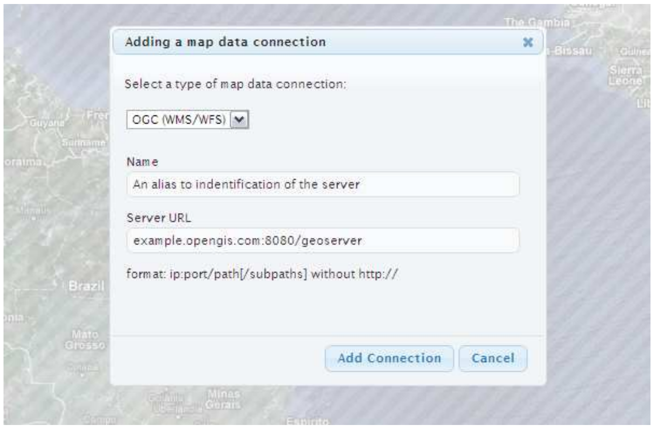
This section presents a case study in order to validate the solution proposed in this paper.

To explore the functionalities offered by GeoSTAT, we set up a web server with GeoServer version 2.0.3, that implements the WMS and WFS services. We also used the PostgreSQL DBMS version 9.0.4, with the PostGIS spatial extension version 1.5.

The spatiotemporal data set was obtained from the Brazilian National Institute for Space Research (INPE) [18]. resulting in 17,418 records.

The data set contains records of fire events detected by satellites in the state of Paraiba, located in the Northeast region of Brazil, during a period of five years (2006-2010), resulting in 17,418 records. We used spatial data obtained from the Brazilian Institute of Geography and Statistics (IBGE) to visualize the vector layers of the states of the Northeast of Brazil (9 polygons), and those from the cities of the Paraiba state (223 polygons),

Through GeoSTAT we may set up a data connection to the web server and obtain a list of available layers for inclusion, visualization and analysis. Figure 3 shows a screenshot for adding a new data connection. GeoSTAT may connect to any map servers that implement OGC WMS and WFS services. In the case study we used the GeoServer map server.



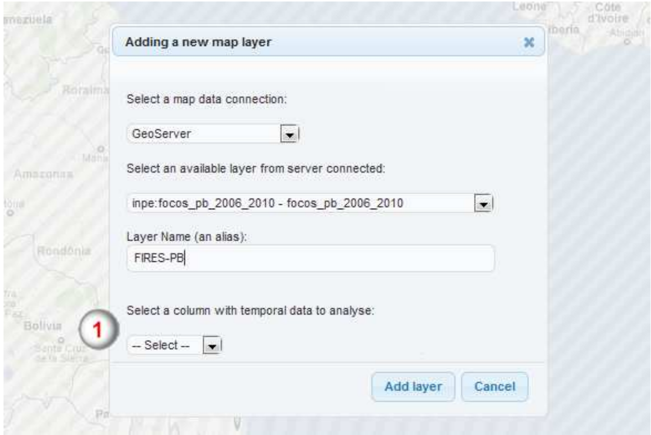
Select a type of map data connection: OGC (WMS/WFS)

Name: An alias to indentification of the server

Server URL: example.opengis.com:8080/geoserver

format: ip:port/path[/subpaths] without http://

Figure 3. Adding a spatial data connection.



Adding a new map layer ✖

Select a map data connection:

GeoServer ▾

Select an available layer from server connected:

inpe:focos_pb_2006_2010 - focos_pb_2006_2010 ▾

Layer Name (an alias):

FIRES-PB

Select a column with temporal data to analyse:

-- Select -- ▾

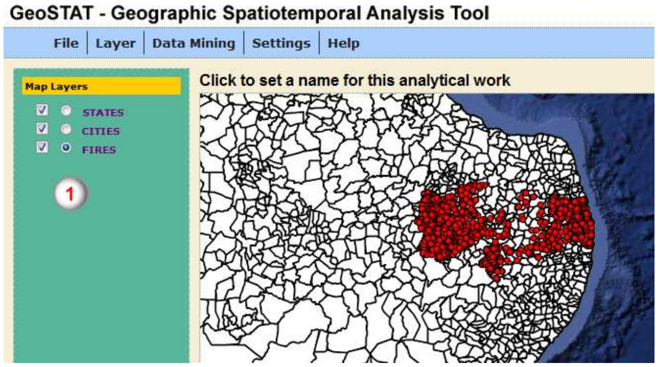
Figure 4. Adding a spatiotemporal map layer.

Figure 4 shows the screenshot of adding map layers. The layers available from the server are obtained by the *GetCapabilities* request, specified in the OGC WMS service. If a layer contains spatiotemporal data, the user has the option to select the attribute that contains the temporal dimension (see item 1 in Figure 4). The layer attributes are acquired using the *DescribeFeatureType* request, specified by the OGC WFS service.

The selection of the temporal attribute is an essential step to perform spatiotemporal analysis. By specifying this attribute, GeoSTAT provide temporal components to the user. Otherwise the system will treat the attribute as spatial and will make available only the spatial API.

In our experiment, we used the following layers: States (spatial), Cities (spatial) and Fires (spatiotemporal). Figure 5 shows the GeoSTAT viewer with these layers visible in the map. Map layers are obtained from the servers through the *GetMap* request, specified by the WMS service. We may control the visibility of layers (see item 1 in Figure 5). The "checkbox" corresponds to the control of the visibility of the layer, whereas the "radio button" to the active layer control (see item 3 in Figure 2).

Figure 6 shows the information component, which is enabled by clicking on the visible layer, e.g., the FIRES layer. This is done through the WMS *GetFeatureInfo* requisition. The layer opacity may be changed by clicking on the layer name.



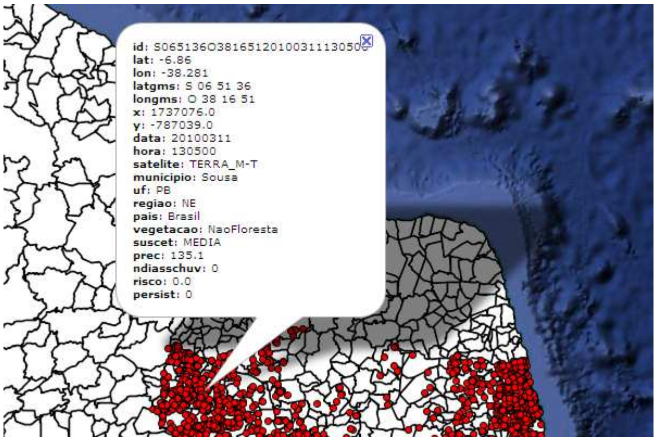
GeoSTAT - Geographic Spatiotemporal Analysis Tool

Map Layers

☑ ○ STATES
☑ ○ CITIES
☑ ● FIRES

Click to set a name for this analytical work

Figure 5. GeoSTAT viewer with three layers visible in the map.



id: S065136O3816512010031113050
lat: -6.86
lon: -38.281
latgms: S 06 51 36
longms: O 38 16 51
x: 1737076.0
y: -787039.0
data: 20100311
hora: 130500
satelite: TERRA_M-T
municipio: Sousa
uf: PB
regiao: NE
pais: Brasil
vegetacao: NaoFloresta
suscet: MEDIA
prec: 135.1
ndiasschuv: 0
risco: 0.0
persist: 0

Figure 6. An example of the information component.

After presenting the "Fires" spatiotemporal layer in the map, GeoSTAT displays a chart area with the distribution of such events over time (see item 2 in Figure 2). In the chart area, it is possible, for example, to classify the distribution chart using nonspatial and nontemporal attributes. In our experiment, we used the vegetation attribute to generate a new distribution chart showing types of vegetation over time (see Figure 7). Then the temporal slider is enabled so that the user may press the “play button” to start a map animation over time (see item 1 in Figure 2).

Figure 8 presents an example of the temporal animation component. There are interface components such as play and pause; and a temporal slider. The objects shown in the map are exhibited according to the specific time being played.

As the "States" and "Cities" layers include all of the Brazilian territory and we would like to focus on the region of fires analysis (Northeast region), we used the spatial filter to show on the map only the state of Paraiba and its 223 cities. The use of spatial filters is only possible due to the *GetFeature* request specified by the OGC WFS service, which enables to execute queries on the spatial layers. Then, the user may be interested in analyzing the detected fire events where the type of vegetation was "NoForest" and during the year of 2008. By analyzing the dynamic chart, we can see which type of vegetation registered the highest concentration of such events in that time period. To perform this operation, we first apply a spatial filter on the layer, in order to display in the map only the records associated with the vegetation type "NoForest", and then we

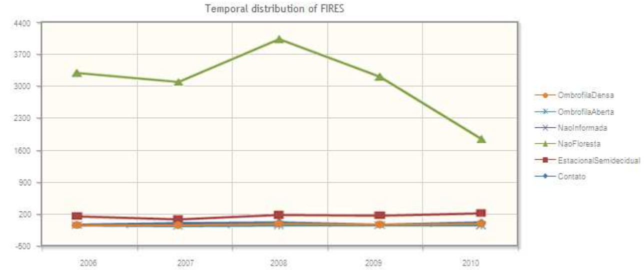
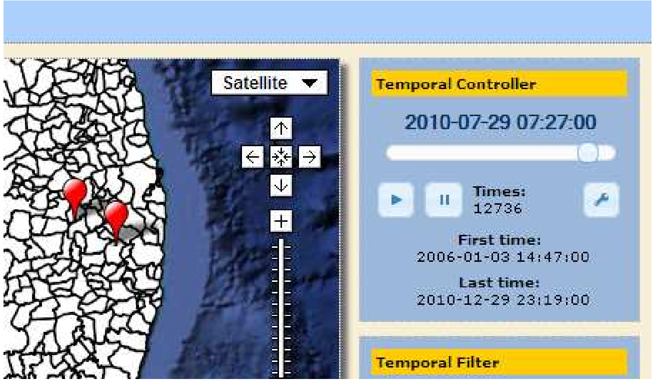


Figure 7. Fire distribution chart over time, classified by type of vegetation.

ISBN: 978-1-61208-178-6



2010-07-29 07:27:00

Times: 12736

First time: 2006-01-03 14:47:00

Last time: 2010-12-29 23:19:00

Figure 8. Using the temporal controller

apply a temporal filter for the time period between "2008-01-01" and "2008-12-31". That way, we reduced the result set to 4,010 records of fires to be analyzed. The results may be seen in the map shown in Figure 9a.

Finally, we execute the spatiotemporal data mining algorithms to the fire dataset aiming to find out relevant implicit patterns. For instance, we may detect that during a certain period of the year, a given type of vegetation is severally affected by fires in a specific geographic region (cluster). This knowledge may help decision makers to study fire defense and emergency services planning for future occurrences. We use the month granularity, through the density-based DBScan clustering algorithm [19], provided by GeoSTAT, with parameters epsilon = 0.1, minPoints = 2 and distance-type = ‘Euclidian distance’. The input values for the DBScan algorithm were established empirically. The resulting clusters can be seen in Figure 9b.

DBScan [19] is a density based cluster algorithm that groups fire spots using spatiotemporal neighborhood. Hence, the fires occurred in a given time interval and within a distance specified in the epsilon parameter will constitute a cluster. The "minPoint" parameter specifies how many neighbor points are necessary to obtain a cluster. Finally, the distance type parameter specifies the distance metric to be used; in our case we are using the Euclidian distance.

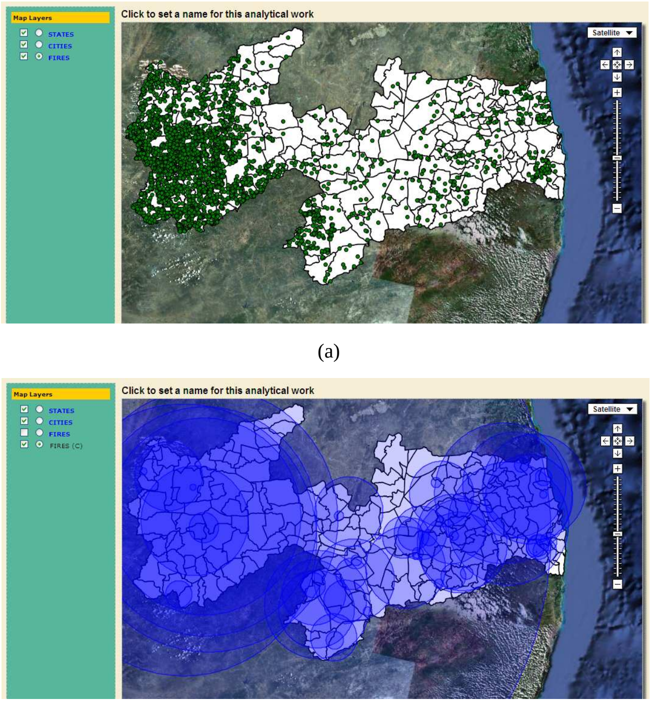
We validate our approach in a real scenario, and we could the usability and effectiveness of the proposed system.

V. Conclusion And Future Work

We addressed the question of spatiotemporal visualization and analysis, highlighting its utility in the presence of large data sets. We also pointed out the main limitations found in the related work.

We proposed a spatiotemporal visualization and analysis environment aiming to minimize the main limitations found, regarding the flexibility of the data sources, practicality of the analysis process and generalization of the domain. Our case study has shown that the proposed solution fulfills the requirements established, being valid not only in the domain adopted for the case study, but also on any other domains.

We built a DBMS independent environment, following the OGC guidelines specifying the WMS and WFS services that allow the interoperability between several data sources with the use of simultaneous map layers regardless of their origins, in a transparent way for the user.


Map Layers
STATES
CITIES
FIRES
Click to set a name for this analytical work
Satellite
Map Layers
STATES
CITIES
FIRES
FIRES (C)
Click to set a name for this analytical work
Satellite

(a)

(b)

Figure 9. Case Study results: (a) 4,010 fires where type of vegetation is equals to "NoForest" (transactional data) during the year of 2008. (b) density spatiotemporal clusters generated of these 4,010 fires (mined data).

As further work, we plan to improve GeoSTAT by incorporating a 3D viewer for spatiotemporal data. Also, the addition of the collaborative analysis concept, allowing several analysts to work in a shared and evolutionary manner is another interesting future work.

References

[1] W. Johnston, “Model visualization,” in Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002, pp. 223–227.

[2] I. Kopanakis and B. Theodoulidis, “Visual data mining modeling techniques for the visualization of mining outcomes,” Journal of Visual Languages and Computing, vol. 14, no. 6, 2003, pp. 543–589.

[3] N. Andrienko, G. Andrienko, and P. Gatalsky, “Exploratory spatiotemporal visualization: an analytical review,” Journal of Visual Languages and Computing, special issue on Visual Data Mining, vol. 14, no. 6, 2003, pp. 503–541.

[4] Y. Bédard, T. Merrett, and J. Han, “Fundaments of spatial data warehousing for geographic knowledge discovery,” in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 2001, pp. 53-73.

[5] G. Andrienko, N. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz, “Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research,” in Information Visualization, vol. 7, no. 3, 2008, pp. 173–180, doi:10.1057/ivs.2008.23.

[6] D. J. Peuquet, “Representations of space and time”, The Guilford Press, 2002.

[7] A. T. Kralidis, “Geospatial Open Source and Open Standards Convergences,” in G. B. Hall and M. G. Leahy (Eds.) Open Source Approaches in Spatial Data Handling, Berlin: Springer, 2008, pp. 1–20.

[8] K. Reda, C. Tantipathananandh, T. Berger-Wolf, J. Leigh, and A. Johnson, “SocioScape - a Tool for Interactive Exploration of Spatiotemporal Group Dynamics in Social Networks,” in Proceedings of the IEEE Information Visualization Conference (INFOVIS ’09), Atlantic City, New Jersey, 2009.

[9] C.-T. Lu, A. P. Boedihardjo, and J. Zheng, “Towards an Advanced Spatiotemporal Visualization System for the Metropolitan Washington D.C.” in 5th International Visualization in Transportation Symposium and Workshop, 2006, pages: 6.

[10] H. Yawen, S. Fenzhen, D. Yunyan and X. Rulin, "Web-based visualization of marine environment data," in Chinese Journal of Oceanology and Limnology, vol. 28, no. 5, Science Press, co-published with Springer-Verlag GmbH, 2010, pp. 1086–1094.

[11] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O’Toole, M. Chau, T. Cushna, D. Casey, and Z. Huang, “COPLINK: Visualization for Crime Analysis,” in Proceedings of The National Conference on Digital Government Research, 2003, pp. 1–6.

[12] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. V. Landesberger, P. Bak, and D. Keim, “Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns,” in Computer Graphics Forum, vol. 29, no. 3, 2010, pp. 913–922.

[13] P. Compieta, S. D. Martino, M. Bertolotto, F. Ferrucci, and T. Kechadi, “Exploratory spatiotemporal data mining and visualization,” in Journal of Visual Languages & Computing, vol. 18, no. 3, 2007, pp. 255–279.

[14] G. Andrienko, N. Andrienko, U. Demsarb, D. Dranschc, J. Dykesd, S. I. Fabrikant, M. Jernf, M.-H. Kraakg, H. Schumannh, and C. Tominskih, “Space, time and visual analytics,” in International Journal of Geographical Information Science, vol. 24, no. 10, 2010, pp. 1577–1600.

[15] Google Inc, “Google Maps Javascript API V2 implementation reference documentation,” 2010, available from: http://code.google.com/apis/maps/documentation/javascript/v2/reference.html 25.11.2011.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” in SIGKDD Explorations, vol. 11, no. 1, 2009, pp. 10–18.

[17] Open Geospatial Consortium, “GeoServer - a Java-based software server to view and edit geospatial data,” 2008, available from: http://geoserver.org/display/GEOS/What+is+Geoserver 25.11.2011.

[18] INPE, “Vegetation Fires - Fire Monitoring,” in Brazilian National Institute for Space Research, 2011, available from: http://sigma.cptec.inpe.br/queimadas/index_in.php 25.11.2011.

[19] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, pp. 226-231.

ISBN: 978-1-61208-178-6

A Heuristic-based Approach for Merging Layers Information in a GIS

Wilfried Segretier, Martine Collard, Enguerran Grandchamp
University of the French West Indies and Guiana
Laboratory of Mathématiques, Informatique et Applications (LAMIA)
Pointe-à-Pitre, Guadeloupe, France
{*wsegreti,mcollard,egrandch*}@*univ-ag.fr*

***Abstract*—Geographic Information Systems (GIS) help addressing geographical and environmental issues by providing information about a region or a city as a set of maps (layers), each one displaying information about a given theme like roads, vegetation, tourist spots or museums for instance. By combining different layers on a region, one can associate a given area to characteristics from their related themes. Indeed, the information from two or more layers might be merged and then transformed into a new layer as defined in map ”algebra”. When a theme vocabulary is organized as a taxonomy with concepts linked by *is-a* relationships, there are different ways to annotate an area with a concept depending on the level selected into the layer taxonomy. In this paper, we present an heuristic-based approach for an optimal merging of such layers in a GIS. Our goal is to generate new layers which sum up information from several themes in a most useful way. Two optimization criteria are considered, the average size of resulting areas and the average informative value of their resulting annotation. We demonstrate the validity of the proposed solution, firstly, on a formal example, and then, on a real world application.**

***Keywords*-Geographic Information Systems; Genetic algorithms; Geographic Knowledge Discovery.**

I. Introduction

Geographic Information Systems (GIS) [1] are providing powerful tools to capture, store, query, analyze and display geographically referenced data. They have proved to be particulary helpful in numerous domains thanks to their ability to handle and process multiple sources of information about geographic (or spatial) regions. They are increasingly used to support experts such as decision makers, geoscientists or environmental engineers for instance in their jobs.

In GIS, data are traditionally represented according one of the two standard systems, namely raster or vector systems and are stored as sets of maps (or layers) among which each one is dedicated to a theme. Thematics layers are also called projections since they project the real landscape according to a given theme such as streets, buidings, vegetation, precipitations or elevation. Layers can overlay one on the top of the others to form computer equivalents of physical maps. One research issue has been the problem of combining projections that do not line up. Tomlin [2] defined the Map Algebra, a vocabulary and conceptual framework for classifying ways to combine map data and produce new maps defined by raster data sets.

In this paper, we address the problem of combining layers too, but we investigate its semantic part related to themes. We consider a layer theme as a formal concept and we point out how hierarchies of concepts can be combined while combining layers. The *concept* paradigm has been commonly defined as a cognitive, abstract or symbolic representation of real objects or situations. Concepts may be built from or be part of others ones. They are often organized in a hierarchical structure that is the cornerstone of domain taxonomies and ontologies. For example, the concept of ”vegetation” can be extended by sub-concepts such as ”tropical rainforest” or ”boreal forest”. Concepts may annotate features, points, lines and areas on a map at different precision levels and according to their level in the hierarchy. For instance, lines can be annotated in the layer ”road” by different sub-concepts such as ”highway”, ”national highway” or ”trunk road” while surfaces can be annotated in the ”soil” layer by sub-concepts like ”rock”, ”grass” or ”sand”. As earliest works on ontology-based GIS we can cite the proposition of Fonseca and Egenhofer [3]. In this work, we focus on concepts annotating areas.

As for many other fields, the volume of data available in GIS has been growing significantly over the last decades. Simultaneously, techniques allowing to treat these data have been widely developped and improved. The Geographic Knowledge Discovery (GKD) domain [4] refers to the extension of Knowledge Discovery from Databases (KDD) where the data-objects are spatially referenced. It includes geographic data-mining, data selection, data preprocessing, data reduction, data enrichment and so on. Issues such as spatial planning, natural resources monitoring or risk prevention require to combine numerous thematic layers in order to produce useful information. GKD tools are thus fitted for such issues. When the number of related spatial areas and available layers is large, exhaustive approaches for the combination are not affordable due to their computational complexity. Heuristics such as stochastic methods provides alternatives to address this problem.

The purpose of this work is to present an heuristic-based approach for optimizing the merge of information sources related to different layers in a GIS. We propose to explore the possibilities offered by multiobjective genetic algorithms (GA) which are strong tools commonly used to address such complex problems.

In a first time, we give some definitions that formalize the context and the issue. They are inspired from Galton's work [5] on *aggregation* and *overlay* algebraic operations and are fitted to layers associated to hierarchies of concepts. These operations are defined to formalize the production of new layers. The *aggregation* allows to sum-up information in a layer while *overlay* permits to merge them. A solution of the GA is defined as a set of aggregate layers designed to be overlayed in order to evaluate the quality of the resulting layer according to two quality criteria.

The present work is a continuation of [6], which proposes to use satellite images (raster) as an information source in order to produce new information layers and then to combine them with other information layers.

This paper is organized in five sections. Section II presents formal definitions on space, layers and geographical operations that are used further on. Section III is devoted to our choices about the genetic algorithm involved. Section IV presents some of the experimental results obtained both on a synthetic dataset and on a realistic case dataset. Then, Section V gives the conclusions and perspectives of this study.

II. Formalization and definitions

In order to precisely define the context and the problem that we address, we propose in this section, a formal spatial framework.

A. Space

We consider the geographical space as an euclidean plane S. We use the term of *surface* to refer to any euclidean surface, i.e., two-dimensional topological manifold. Thus, a surface z is a set of points on the plane. We note $Z(S)$ the set of all surfaces in the space S (including the empty surface $\emptyset$). $Z(S) \subseteq \mathscr{P}(S)$, where $\mathscr{P}(S)$ is the power-set of S. Figure 1 figures an example with five surfaces $z1, z2, z3, z4, z5$.

B. Layers

In a similar way than Galton [5], we define a *layer* over a space S as an eventually partial function $f : S \rightarrow V$, where V is the *value set* of the layer. V may be ordered, unordered, finite, infinite, continuous, discrete, numeric, symbolic and so on. When it is unordered, discrete and symbolic, we call its elements *concepts*. In the following, we focus on such concept sets.

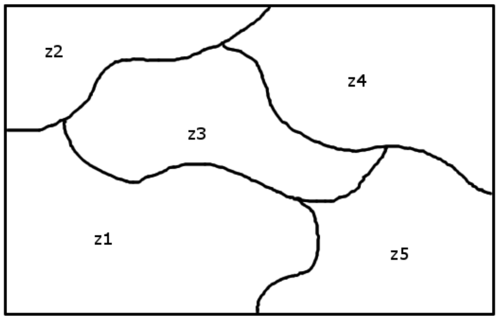


Figure 1. *Surfaces* on an euclidean plane

We say that a point $p \in S$ (resp a surface $z \in Z(S)$) is *annotated* by a concept $v \in V$ in the layer f if:
$f(p) = v$ (resp $\forall p \in z$, $f(p) = v$). By extension, if the surface z is annotated by the concept v, we write $f(z) = v$. We assume that it exists a unique tessellation of the space deductible from the function f of a layer that is $maximal$, i.e., such that there is no surface annotated by a concept c containing a given surface annotated by the same concept.

Given these definitions, it is possible to define some operations which can combine one or more layers in new ones.

The aggregation operation is defined by Galton [5] as follows. Given an equivalence relation E on V, the aggregation operation $f/_E$ is defined by the layer which annotates each point p by the equivalence class of $f(p)$. We have:
$f/_E : S \rightarrow V/_E$,
$p \longrightarrow [[f(p)]]_E$,
where $V/_E$ is the quotient set of V under E and $[[f(p)]]_E$ is the equivalence class of $f(p)$.
We extend this definition toward a hierarchical axis.

Hierarchical aggregation

We define the *hierarchical aggregation* by an aggregate layer as follows. Given a hierarchy $\mathscr{H} = \{H_1, H_2, ..., H_g\}$ on V, we consider subsets $V_{\mathscr{H}}$ of $\mathscr{H}$ such as:

- $\forall\, Hv_i, Hv_j \in V_{\mathscr{H}} \times V_{\mathscr{H}}$, $Hv_i \cap Hv_j = \varnothing$ and
- $\forall x \in V$, $\exists! i \in \{1, 2, ..., g\}$ so as $x \in Hv_i$

Thus, each $V_{\mathscr{H}}$ is a partition of V. The Figure 3 illustrates examples of such partitions where $\mathscr{H}$ is the hierarchy showed on Figure 2.

Let $\Re$ be the equivalence relation on V which quotient set is $V_{\mathscr{H}}$. We define an *aggregate layer* $f/_{\Re}$ *under the hierarchy* $\mathscr{H}$ as:

$$f/_{\Re} : S \rightarrow V_{\mathscr{H}}$$

$$p \longrightarrow [[f(p)]]_{\Re}$$

ISBN: 978-1-61208-178-6

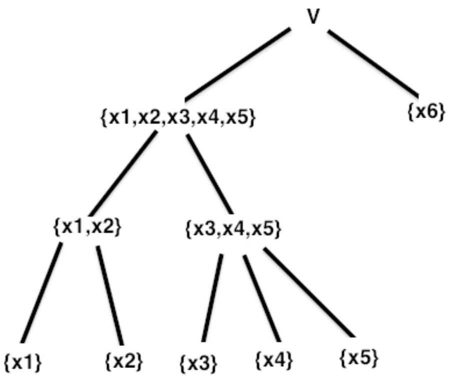


Figure 2. Example of hierarchy on a set

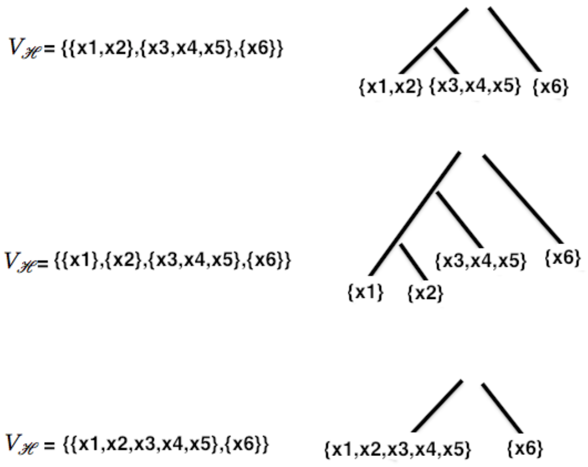

$V_{\mathscr{H}}$= {{x1,x2},{x3,x4,x5},{x6}}
{x1,x2} {x3,x4,x5} {x6}
$V_{\mathscr{H}}$= {{x1},{x2},{x3,x4,x5},{x6}}
{x1} {x2} {x3,x4,x5} {x6}
$V_{\mathscr{H}}$ = {{x1,x2,x3,x4,x5},{x6}}
{x1,x2,x3,x4,x5} {x6}

Figure 3. Examples of partitions of $\mathscr{H}$

Layers Overlay

Given n layers $f_1 : S \to V_1$, $f_2 : S \to V_2$, ..., $f_n : S \to V_n$ and $g : V_1 \times V_2 \times ... \times V_n \to V_f$, the *overlay* operation allows to define the new layer:

$$f_g : S \to V_f$$

$$p \longrightarrow g(f_1(p), f_2(p), ..., f_n(p))$$

If $x_1 \in V_1, x_2 \in V_2, ...x_n \in V_n$, we note $x_1 x_2 ... x_n$ the element of V_f associated with $x_1, x_2, ..., x_n$, i.e.,
if $f_1(p) = x_1$ and $f_2(p) = x_2$ and ... and $f_n(p) = x_n$ then $f_g(p) = x_1 x_2 ... x_n$.
If a function f_i defining a layer is partial on S, we note $\emptyset$ the image of the elements of S for which f_i is not defined. The symbol $\emptyset$ is not represented in the previous notation so if $f_2(p) = \emptyset$, we have $f(p) = g(f_1, f_2, f_3 ..., f_n) = x_1 x_3 ... x_n$.

Depending on the function g, the overlay type can be *union, intersection, symmetrical difference, identity,* etc. In the following we only consider the union overlay so that: $\forall X = (x_1, x_2, ..., x_n), g(X) = x_1 x_2 ... x_n$. Figure 4 shows an illustration of the union overlay.

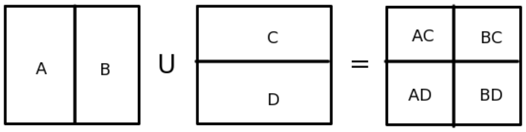


Figure 4. Union overlay on surfaces A,B,C,D of the same map

Union overlay on surfaces

It is assumed that the conjunction of surfaces is either a surface or the emptyset i.e.,
$\forall (s, s') \in (Z(S))^2$, $s \cap s' \in Z(S)$.

The union $\cup$ is defined as follows. Let us consider the layers: $f_1 : S \to V_1$ and $f_2 : S \to V_2$ and $s \in Z(S)$,

- if $\nexists s' \in Z(S)$ annotated by f_2 with $s \cap s' \neq \emptyset$, then $(f_1 \cup f_2)(s) = f_1(s)$,
- if $\exists (s'_1, ..., s'_p) \in Z(S)^p$ with each s'_i annotated by f_2 and $s'_i \cap s \neq \emptyset$, then $\forall i \in \{1, ..., p\}$ $(f_1 \cup f_2)(s'_i \cap s) = f_1(s'_i \cap s) f_2(s'_i \cap s)$.

C. *Optimization problem*

Let us take:

- S an euclidean (geographical) plane,
- n layers $f_1 : S \to V_1$, $f_2 : S \to V_2$, ..., $f_n : S \to V_n$,

The optimal union layer selection problem is a bi-objective combinatorial optimization problem which consists in applying an aggregation to m layers among $f_1, ..., f_n$ in order to obtain the union overlay layer $f : S \to V$ from these aggregate layers while trying to:

- maximize the average area of the resulting surfaces annotated by a concept in the union layer f,
- maximize the total number of concepts (i.e., considering all the input (aggregated) layers) that annotate these resulting surfaces in the overlay.

Obviously, these two objectives are antagonistic since the more concepts are numerous in layers, the less the resulting average area is large.

These two objectives can be treated using a scalar approach that requires to set parameters and return only one solution to the end-user. However, we chose to treat them separately in a multi-objective optimization approach in order to benefit from a diversified choice of solutions.

We have conducted experiments in an incremental way:

1) First, we have not considered any hierarchy on concepts. From a quantitative point of view, this approach can be very interesting as it allows to attain every

possible instance of aggregation in each layer, thus it results in optimal surface areas in the overlay layers. However, the results obtained may not always be useful for end users because some very distant concepts may be associated in a same class causing semantic inconsistencies. In the following we will refer to this method as the *free aggregation* method,

2) Secondly, we have followed a method consisting firstly in determining a hierarchy $\mathscr{H}_i$ under V and then achieving a **hierarchical aggregation**. Since the defined hierarchy is semantically sound, this method allows to eliminate the major drawback of the previous method, namely the lack of consistency of some solutions. In the following, we refer to this method as the *hierarchical aggregation* method,

Whatever the chosen method, the search for instances of aggregation for m among n layers can be summed up to a combinatorial search:

- In the first case, the total number of possibilities for a layer corresponds to the number of partitions of the concept set. It is given by the Bell number recursively defined as:
$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$$
with n the number of elements of the set. Then the total number of possibilities N_1 is given by the product of each layer number of possibilities.
- In the second case, the number of possibilities for a layer depends on the structure of the tree underlying the chosen hierarchy. It can be defined recursively for each node n as:
$$N_2(n) = \left(\prod_{s \in Sons(n)} N_2(s) \right) + 1$$
where $Sons(n)$ is the set of sons of n. Similarly, the total number of possibilities N_2 is given by the product of each layer number of possibilities. An important point to note is that the size of the search space is much smaller in this case than in the previous case, i.e., $N_1 >> N_2$.

As these numbers can increase rapidly with n_i, the number of concept of the i^{th} layer, stochastic methods are appropriate approaches to avoid exhaustive search which would often be impracticable. In the next section, we present the multi-objective genetic algorithm that we have implemented for this purpose.

III. Genetic Algorithm

As we have seen in the previous section, the optimal layer selection problem boils down to a multi-objective optimization problem. We decided to explore the solutions offered by Genetic Algorithms (GA) as they are simple, powerful and well used tools to solve combinatorial problems, particularly in the case of multi-objective issues [7]. However, more important than the choice of the heuristic is the definition of the problem as a combinatorial problem and the validation that such methods are useful to address it.

These algorithms are stochastic methods and use global search heuristics belonging to the family of evolutionary algorithms. They are inspired by evolutionary biology's main principles such as inheritance, mutation, selection and crossover. They allows to evolve a randomly chosen initial population until some defined criteria are reached (quality of solutions, number of generation, etc.).

In this section, we present the multiobjective genetic algorithm components that we implemented using the ParadisEO-MOEO framework [8]. We show our representation choices and genetic operators for both aforementionned methods, i.e., free and hierarchical aggregation.

All the problem-independent parts of the GA are based on the well known Non-Dominated Sorting GA-II (NSGA-II) [9] evolutionary multiobjective optimization method which is widely used for its low computational complexity and its ability to find good spreads of solution for a rather large range of problems. Table I gives an overview of its components.

Table I. NSGA II components overview

Components	NSGA II
Fitness assignment	Dominance-depth
Diversity assignment	Crowding distance
Selection	Binary tournament
Replacement	Elitist replacement
Archiving	none
Stopping Criteria	Max number of generations

A. Individual Encoding

For both methods, an individual can be encoded as the sequence of m layers that will overlay. The main difference between both representations lies on the way that instances of aggregation are represented for each layer.

1) Free aggregation: For the free aggregation method, a layer can be seen as an ordered sequence of n_i concepts, each one being associated with an equivalence class. Figure 5 figures the general structure of this representation.

2) Hierarchical aggregation: In this case, we represent each hierarchical aggregation for a given hierarchy by an integer value. The mapping between an integer and an instance of aggregation may be done in several ways, however, the important point to note is that it is done in a bijective manner so that each possible instance of hierachical aggregation

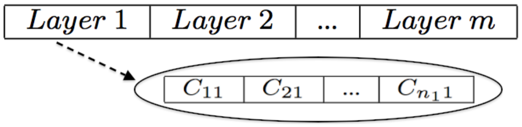


Figure 5. Individual encoding for free aggregation

corresponds to a unique integer value. Figure 6 illustrates the structure of this representation.

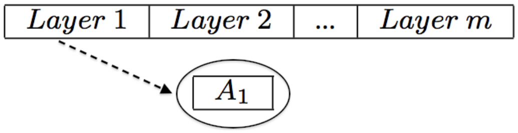


Figure 6. Individual encoding for hierarchical aggregation

B. Genetic Operators

1) Crossover: The crossover principle consists in mating chromosomes (individuals) -the parents- in order to obtain new ones -the offsprings- made with their genetic heritage. The main purpose of this operator is to diversify an existing population in order to improve it.

The main operator we used for both methods is a multi-point uniform [10] quad crossover operator which consists in choosing two parents and computing two offsprings with a given mixing ratio. Crossover points may be located either between any concepts of any layer for the free aggregation or simply between two layers for the second method. Figure 7 gives an illustration of this operator with one crossing point for the free aggregation.

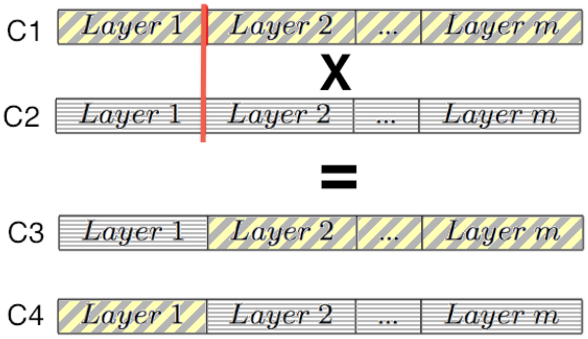
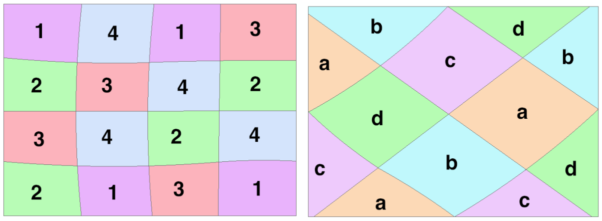


Figure 7. Crossover operator principle (one point)

2) Mutation: We used a uniform mutation operator which consists either in changing the equivalence class for a given concept (or a group of concepts which belongs to a same class) or in changing the hierarchical aggregation in a layer by modifying the integer which represents it.



(a) Layer 1

(b) Layer 2

Figure 8. Input layers

IV. Experimental results

In this section, we present and analyse some of the results that we obtained for both aggregation methods. First, a simple example is graphically showed for better understanding, then a more realistic case is presented. In the latter case, we focused on the hierarchical method and we checked various GA parameters (number of generations, operators probability, population size). The best results showed in the following were obtained with:

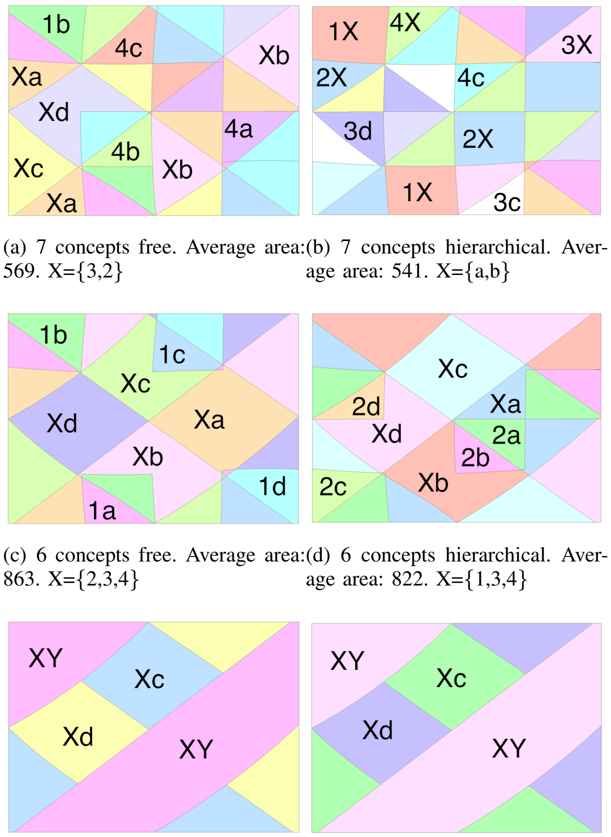
- 200 generations,
- 100 individuals population,
- a crossover probability of 0.7,
- a mutation rate of 0.001.

A. Simple case

We defined two simple layers (Layer 1 and Layer 2) as shown in Figures 8 a) and 8 b). Each one is composed of several surfaces annotated by four concepts. Surfaces of each layer have been designed to be slightly different from one to the other in order to introduce some local optima.

Figures 9 c), 9 e), 9 g) (left column of Figure 9) show some of the best solutions (with a number of concepts of 4,6,7) obtained with the free aggregation method while Figures 9 d), 9 f), 9 h) (right column of Figure 9) show the same results for the hierarchical aggregation method. On these figures, the concept that annotates a given surface is indicated. Figure 10 shows the Pareto sets obtained for both methods.

We can see that the best solutions founded by the GA differ from an aggregation method to the other for each number of concepts. The computation of the average area of each overlay layer shows that it is always greater or equal in the free aggregation case. As said in Section 2, this situation can easily be explained by the fact that the chosen hierarchy limits the number of reachable possibilities and thus reduces the possibility to find optimal solutions. However, as illustrated by Figures 9 e) and 9 f) which show the same solution for both methods, an optimal solution may be obtained by the hierarchic method (most certainly due to the reduced search space).



(a) 7 concepts free. Average area: 569. X={3,2}

(b) 7 concepts hierarchical. Average area: 541. X={a,b}

(c) 6 concepts free. Average area: 863. X={2,3,4}

(d) 6 concepts hierarchical. Average area: 822. X={1,3,4}

(e) 4 concepts free. Average area: 2775. X={1,2,3,4}, Y={a,b}

(f) 4 concepts hierarchical. Average area: 2775. X={1,2,3,4}. Y={a,b}

Figure 9. Different overlays of Layer 1 and Layer 2

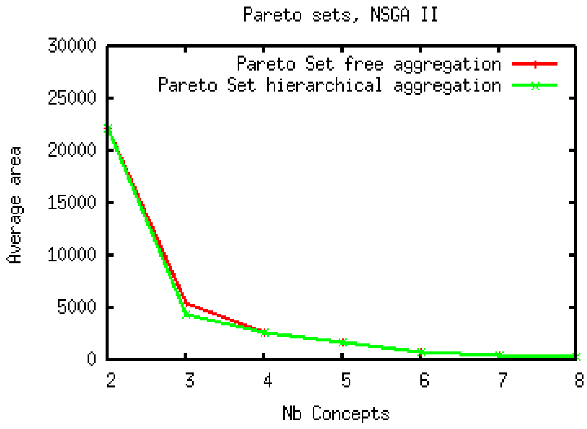


Figure 10. Pareto Sets

B. *Realistic case*

As stated in the introduction, this work is intended to tackle real world applications where the space and the numbers of concepts and layers are large enough to make exhaustive search impracticable. In the following, we show such a complex and realistic application which main characteristics are:

- high total space superficy (Guadeloupe F.W.I Island i.e., $1628, 43km^2$),
- 3 Layers,
- 10 concepts per layer,
- large tessellation of space per layer.

The size of the search space generated by this example is about 10^{15} in the free case ($B_{10} = 115975$) and may vary significatively from a set of hierarchies (a hierarchy for each layer) to another in the hierarchical case. In the following example the size of the space is about 10^6.

Table II presents the whole database containing the areal objects resulting from one pareto-optimal solution (with 13 concepts in the overlay layer) obtained by the hierarchical aggregation method. Each line is associated with a concept of the overlay layer. The first three columns show the original concepts values for each input layer of the related concept, whereas the last column gives the average area of the surfaces which it annotates. Lines are ranked in descending order with respect to the average area values. We can observe that the largest areas result from highly aggregated concepts (lines 1 to 9) while smaller aggregations may provide acceptable areas (lines 10 to 15) as well as very small areas (lines 21 to 25) which are obviously damaging for the maximization of the global average area value. More generally, we can see that the difference between the minimal and maximal number of initial concepts in the overlay concepts (respectively, line 11 → 12 concepts and line 2 → 21 concepts) can be relatively large whith the assumption that concepts of very different levels of abstraction may be included in a same solution. Since this situation could be problematic for some applications, it will be addressed in further works by introducing a distance parameter that may limit the difference of abstraction level.

Figure 11 shows a comparison between two typical Pareto sets obtained with both methods and the same GA parameters. We can see that the two curves are relatively close, however, the set obtained with the free aggregation is often better than the one obtained with the hierarchical aggregation, which is consistent with our first observations (Sections 2 and IV-A).

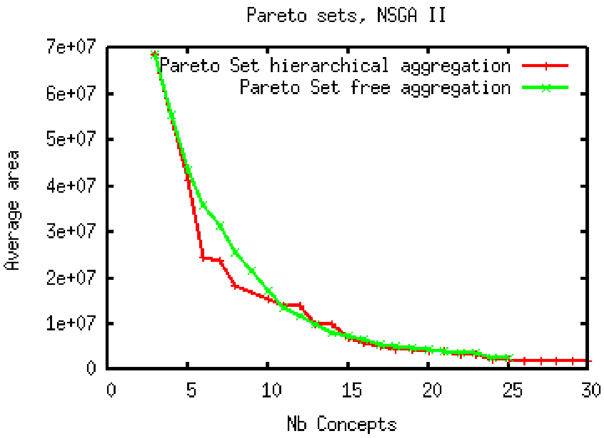


Figure 11. Realistic case Pareto Set example

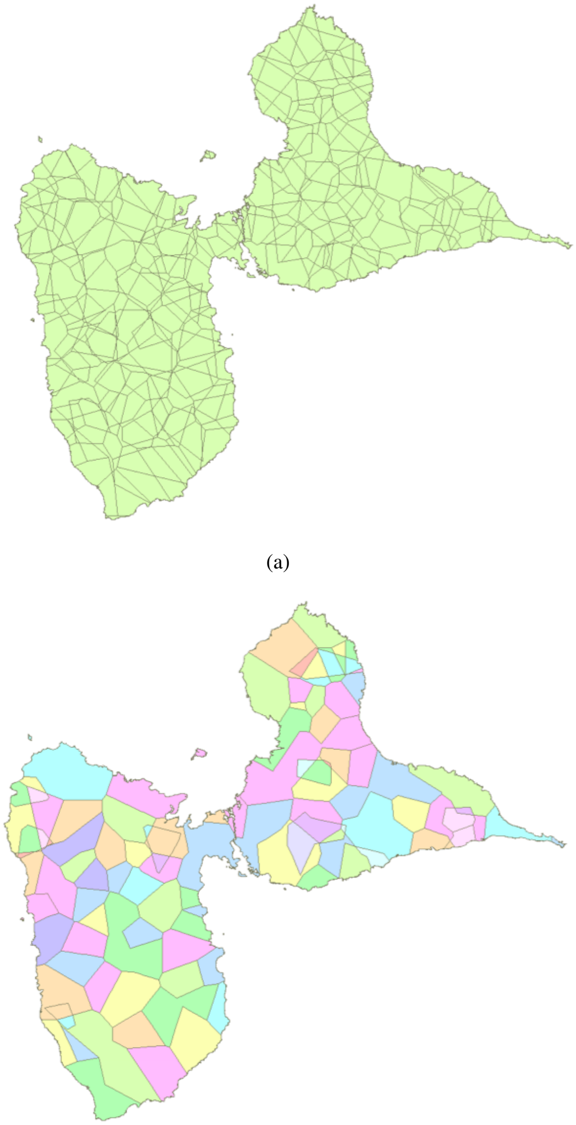
The free aggregation leads to better results concerning the two objectives, however it requires more iterations to converge and is not always consistent with semantic relations between concepts. Thus, the hierachical method turns to be a better compromise between complexity and efficiency as it allows to obtain good solutions - semantically acceptable - in a short time due to its reduced search space.

Focusing on the hierachical method, we can see that the average area starts with very high values but decreases very rapidly, between 3 and 6 concepts, while it is quite low and decreases more slowly since 15 to 30 concepts. Thus, the more interesting part of the curve from an end user point of view may be situated between 7 and 15 concepts where the compromise between the two objectives is not only in favor of one of them.

Figure 12 gives an illustration of a representative example. Figure 12 a) shows the tessellation of the space resulting from the raw overlay of all layers (i.e., whith no previous aggregation) while Figure 12 b) shows the tessellation of the space obtained with the optimal solution of Table II. It is easy to observe that the space is much less fragmented in the second case due to the aggregations made on each layer.

V. Conclusion and perspectives

In this paper, we have investigated the question of information selection for layers overlay in a GIS. We have presented a genetic algorithm-based approach allowing to efficiently find *overlay layers*. Candidate solutions have been previously aggregated before being overlayed then evaluated. We showed that the use of hierarchies for aggregations before overlaying layers represents a good tradeoff between complexity and efficiency when searching for solutions.



(a)

(b)

Figure 12. Graphical results on the realistic case

Our perspectives for further works include the use of other metaheuristics such as tabu search or simulated annealing, the definition of alternative aggregation methods which could be more efficient and the introduction of parameters such as a maximal distance between the number of initial concepts belonging to layers and thus to their combinations.

Acknowledgment

We thank the Claude-Emmanuel Blandin foundation and the Regional Council of La Guadeloupe for their financial support.

Table II. Areal objects for the pareto solution with 13 concepts

line	Layer 1	Layer 2	Layer 3	Area (Km^2)
1	[1-6],8,10	9	[1-10]	202.0
2	[1-6],8,10	1,2	[1-10]	189.8
3	[1-6],8,10	7	[1-10]	157.6
4	[1-6],8,10	4	[1-10]	155.5
5	[1-6],8,10	10	[1-10]	146.7
6	[1-6],8,10	8	[1-10]	146.6
7	[1-6],8,10	5	[1-10]	136.1
8	[1-6],8,10	3	[1-10]	127.9
9	[1-6],8,10	6	[1-10]	49.8
10	7	4	[1-10]	20.9
11	9	7	[1-10]	14.8
12	9	10	[1-10]	13.1
13	9	1,2	[1-10]	12.9
14	7	5	[1-10]	12.8
15	9	4	[1-10]	12.5
16	7	3	[1-10]	8.9
17	7	8	[1-10]	8.4
18	7	7	[1-10]	5.4
19	9	3	[1-10]	5.3
20	9	5	[1-10]	5.2
21	9	9	[1-10]	3.0
22	7	10	[1-10]	2.8
23	7	9	[1-10]	1.5
24	7	1,2	[1-10]	0.5
25	7	6	[1-10]	0.002

References

[1] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, *Geographic Information Systems and Science*. John Wiley & Sons, Jul. 2001, [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0471892750 <retrieved: November, 2011>.

[2] C. D. Tomlin, *Geographic Information Systems and Cartographic Modelling*. Prentice-Hall, 1990.

[3] F. T. Fonseca and M. J. Egenhofer, "Ontology-driven geographic information systems," in *Proceedings of the 7th ACM international symposium on Advances in geographic information systems*, ser. GIS '99. New York, NY, USA: ACM, 1999, pp. 14–19, [Online]. Available: http://doi.acm.org/10.1145/320134.320137 <retrieved: November, 2011>.

[4] H. J. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery, Second Edition*, 2nd ed. CRC Press, May 2009.

[5] A. Galton, "A formal theory of objects and fields," in *Spatial Information Theory*, ser. Lecture Notes in Computer Science, D. Montello, Ed. Springer Berlin / Heidelberg, 2001, vol. 2205, pp. 458–473, [Online]. Available: http://dx.doi.org/10.1007/3-540-45424-1_31 <retrieved: November, 2011>.

[6] E. Grandchamp, "Raster-vector cooperation algorithm for GIS," in *Proceedings of Geoprocessing*, Saint-Martin, 2010, p. 40062, [Online]. Available: http://hal.archives-ouvertes.fr/hal-00509489/en/ <retrieved: November, 2011>.

[7] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*, ser. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Chichester, 2001.

[8] A. Liefooghe, L. Jourdan, and E.-G. Talbi, "A Unified Model for Evolutionary Multiobjective Optimization and its Implementation in a General Purpose Software Framework: ParadisEO-MOEO," INRIA, Research Report RR-6906, 2009, [Online]. Available: http://hal.inria.fr/inria-00376770/PDF/RR-6906.pdf <retrieved: November, 2011>.

[9] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 182–197, 2002.

[10] D. H. Ackley, *A connectionist machine for genetic hillclimbing*. Norwell, MA, USA: Kluwer Academic Publishers, 1987.

Computationally Feasible Query Answering over Spatio-thematic Ontologies

Özgür Lütfü Özçep and Ralf Möller
Institute for Software Systems (STS)
Hamburg University of Technology
Hamburg, Germany
Email: {oezguer.oezcep,moeller}@tu-harburg.de

***Abstract*—Providing query answering facilities at the conceptual level of a geographic data model requires deduction, and deduction in geographical information systems (GIS) is a demanding task due to the size of the data that are stored in secondary memory. In particular, this is the case for deductive query answering w.r.t. spatio-thematic ontologies, which provide a logical conceptualization of an application domain involving geographic data. For specific logics (so-called lightweight description logics) and query languages (conjunctive queries) the query answering problem can be solved by compiling the ontology-based query into an SQL query that is posed to the database. Thus, ontology-based query answering becomes as feasible as standard database query answering. In the literature, this kind of query answering by compilation is formalized using the notion of first-order logic (FOL) rewritability. In this paper we show that lightweight description logics such as DL-Lite can be combined with spatial calculi such as the region connection calculus such that FOL rewritability is retained and the expressive power is sufficient for modeling important aspects of GIS data.**

***Keywords*-description logics; qualitative spatial reasoning; deductive query answering; FOL rewritability**

I. Introduction

In almost any area in which geographical information systems (GIS for short) are used, e.g., damage classification for flooding scenarios, development of eco systems in forestry, analysis of sociological and demoscopic aspects in urban areas, or semantic web applications over GIS data [1] there is a need to formalize relevant concepts and relations in a conceptual data model in order to answer queries. The preferred technical tool for providing a conceptualization is an ontology. Ontologies are represented in some logical language (e.g., a description logic) that has a formal semantics and allows for automated reasoning, in particular deductive query answering. The idea of exploiting a conceptualization over the domain is to provide a convenient query language for building applications, given some mappings of basic concepts and relations to expressive database queries are defined in terms of SQL. Mappings to SQL are required because the encodings of spatial information in databases might vary and might be rather cumbersome in practice. This holds in particular for GIS databases. As we will show in this paper, the conceptualization can be used to find GIS database queries (views) relevant for complete query answering.

For example, think of a planning problem for additional parks in New York, say. An engineering office responsible for this task uses geographical data, such as, e.g., the TIGER/Line® data—a well known free set of geographic data from the US Census Bureau. Among many others, the engineering consultants have declared concepts such as (i) a park that covers a lake (*Park+Lake* for short), and (ii) a park *Park4Playing* intended to denote all parks covering playing areas. Imagine the office has defined mappings from these concepts to SQL queries possibly involving GIS extensions in order to find instances of these concepts. Quite a large library of similar concepts and relations with mappings will emerge pretty soon. Moreover, assume that for quality assurance purposes the engineering office would like to formulate queries for identifying objects with certain design flaws, e.g., a park with a lake such that inside the lake there is a playing area. Intuitively, the query searches for parks except those for which there are objects on the righthand side of tuples in the relation *hasLake* and *hasPlayingArea* such that the locations of the objects are related by the proper containment relation. Note that *hasLake* and *hasPlayingArea* are relations declared in the conceptual domain model (see below for a more formal account).

We argue that despite the fact that it is not syntactically apparent, in order to find all parks that contain this type of design flaw (for being complete, that is) one also has to take into consideration the mappings that produce instances of *Park+Lake* and *Park4Playing*. The query rewriting process ensures that query answering w.r.t. the semantics of the ontology is implemented by answering a properly rewritten query using a standard SQL query (see [2] for details).

The goal of this paper is to demonstrate that the SQL queries relevant for complete query answering in our spatial domain can be automatically determined as well, and that qualitative spatial reasoning is required to achieve this. The only additional effort for finding relevant mappings is a very weak axiomatization for spatio-thematic concepts such as *Park+Lake* and *Park4Playing* in the form of necessary conditions that approximate the intended meanings. The axiomatization is part of the conceptual domain model.

For example, in case of *Park+Lake* we have a *Park* that *contains* a *Lake* such that the *Lake touches* the *Park* from within. The formalization of these notions is done with an ontology language, and the details of the formalization and its application to the sample scenario are given in this paper. The query is automatically compiled into a GIS database query, and the mappings are considered appropriately. Note that query compilation (or rewriting) is a method that cannot be reduced to simple macro expansion but requires reasoning over the axiomatization as demonstrated by the perfect rewriting algorithm of [2].

The contribution of our paper is the result that query rewriting can also be used in GIS scenarios, but rewriting has to be significantly enhanced because the logical language in which the ontology is represented has to be expressive enough in order to represent spatio-thematic concepts. We identify DL-Lite(RCC8) (Section IV) as an appropriate combined logic and show that it allows for the compilation of the ontology into the query, i.e., in technical terms, allows for FOL rewritability of query answering. The query language defined in Section IV and GeoSPARQL (see OGC draft specification, http://www.w3.org/2011/02/GeoSPARQL.pdf) have common features. But in contrast, DL-Lite(RCC8) allows for the controlled use of RCC8 relations in the ontology. Preparing the results of Section IV, Section II introduces the logical components of DL-Lite(RCC8) and Section III describes the obstacles in finding a combination.

II. Logical Preliminaries

The logic DL-Lite(RCC8) to be introduced in this paper is a combination of two logics described in the following subsections: the region connection calculus RCC8, which can model topological relations like that of containment; and a member of the family of lightweight description logics DL-Lite, which is well suited for reasoning over large databases and which can model many elements of the Unified Modeling Language (UML).

A. RCC8-calculus

The Region Connection Calculus (RCC) [3] is one of the most widely known qualitative spatial reasoning calculi that take regions and not points as the basic entities for representing spatial knowledge and reasoning about it. In the axiomatic representation of RCC [3], a primitive binary relation C is intended to model the connectedness relation between regions; C is therefore axiomatically restricted to be reflexive and symmetric. C is used to define different relations between regions that are termed *base relations*. One family of base relations, denoted $\mathcal{B}_{RCC8} = \{$dc (disconnected), ec (externally connected), eq (equal), po (partially overlapping), ntpp (non-tangential proper part), tpp (tangential proper part), ntppi (inverse of ntpp), tppi (inverse of tpp)$\}$ henceforth, is the building block of RCC8. Further calculi of RCC can be defined by considering other sets of base relations. The base relation dc is intended to model disconnectedness and is defined by $\mathsf{dc}(x, y)$ iff $\neg C(x, y)$. The other base relations are defined similarly [3]. The axioms imply that the eight base relations are jointly exhaustive and pairwise exclusive (JEPD property).

With the help of the base relations, real-world spatial configurations can be represented in the form of constraint networks, which can be efficiently processed by constraint satisfaction procedures. A network is defined by a set of formulas that have the form $r_1(a, b) \vee \cdots \vee r_k(a, b)$ where a, b are constants and $r_1, \ldots, r_k$ are base relations from $\mathcal{B}_{RCC8}$. These sentences are presented in the more succinct algebraic notation as $\{r_1, \ldots, r_k\}(a, b)$. The set of all possible disjunctions of base relations $Pot(\mathcal{B}_{RCC8})$ is denoted Rel_{RCC8}. With disjunctions of base relations, indefinite knowledge on spatial relations of regions can be expressed. The networks are labelled graphs derived from the formulas such that the vertices of the network are the constants used in the formulas, and edges (a, b) labelled $\{r_1, \ldots, r_k\}$ are derived iff $\{r_1, \ldots, r_k\}(a, b)$ is contained in the set of formulas.

A practically relevant question is whether a constraint network is satisfiable with respect to the RCC8 axioms. Testing satisfiability of networks can be carried out on the basis of path consistency algorithms [4]. These algorithms are based on composition tables. For every pair of base relations r_1, r_2 they contain an entry for the composition $r_1 \circ r_2$. In general, the composition $\circ$ of two relations r_1 and r_2 is defined as $r_1 \circ r_2 = \{(x, y) \mid \exists z.r_1(x, z) \wedge r_2(z, y)\}$.

The composition table for RCC8 [5, p. 45] is in fact a table of weak compositions. For two relations r_1, r_2 the weak composition $r_1; r_2$ is the minimal disjunction of base relations that cover their composition $r_1 \circ r_2$, i.e., $r_1 \circ r_2 \subseteq r_1; r_2$. For example the weak composition table entry for the pair $(\mathsf{tpp}, \mathsf{tppi})$ is $\mathsf{tpp}; \mathsf{tppi} = \{\mathsf{dc}, \mathsf{ec}, \mathsf{po}, \mathsf{tpp}, \mathsf{tppi}, \mathsf{eq}\}$. This composition table entry can be described by the following (implicitly universally quantified) FOL sentence:

$$\mathsf{tpp}(x, y) \wedge \mathsf{tppi}(y, z) \rightarrow \{\mathsf{dc}, \mathsf{ec}, \mathsf{po}, \mathsf{tpp}, \mathsf{tppi}, \mathsf{eq}\}(x, z)$$

The (weak) composition relation ; is defined for non-base relations $r_1 = \{r_1^1, \ldots r_1^k\}$ and $r_2 = \{r_2^1, \ldots, r_2^l\}$ in the usual way by pairwise composing the contained base relations: $r_1; r_2 = \bigcup_{1 \leq i \leq k; 1 \leq j \leq l} r_1^i; r_2^j$.

Testing the satisfiability of arbitrary RCC8 networks is NP-complete and thus computationally intensive [6] [7]. Rather than using the axioms of [3], which are based on the relation C, we use axioms that directly state that the eight base relations $\mathcal{B}_{RCC8}$ have the JEPD property, together with axioms corresponding to the composition table and the axiom $\forall x.\mathsf{eq}(x, x)$. This theory is named Ax_{RCC8} and is shown in Figure 1. Adapting the term of an ω-admissible domain [8], we call Ax_{RCC8} an *ω-admissible theory*.

B. DL-Lite + UNA

DL-Lite denotes a family of lightweight description logics that are tailored towards reasoning over ontologies with large sets of data descriptions. We will focus on the member of

- $\{\bigvee_{r \in \mathcal{B}_{RCC8}} r(x,y)\} \cup$ (joint exhaustivity)
- $\{\bigwedge_{r_1, r_2 \in \mathcal{B}_{RCC8}, r_1 \neq r_2} r_1(x,y) \rightarrow \neg r_2(x,y)\} \cup$ (pairwise disjointness)
- $r_1(x,y) \wedge r_2(y,z) \rightarrow r_3^1(x,z) \vee \cdots \vee r_r^k(x,z) \mid r; s = \{r_3^1, \ldots, r_3^k\}\} \cup$ (weak composition axioms)
- $\{\mathsf{eq}(x,x)\}$ (reflexivity of eq)

Figure 1. Ax_{RCC8}. Formulas are implicitly universally quantified

$$\begin{array}{lll} R & \longrightarrow & P \mid P^- \\ B & \longrightarrow & A \mid \exists R \\ C & \longrightarrow & B \mid \neg B \\ \textit{TBox:} & & B \sqsubseteq C, (\text{funct } R), R_1 \sqsubseteq R_2 \\ \textit{ABox:} & & A(a), R(a,b) \end{array}$$

Figure 2. DL-Lite

the DL-Lite family allowing functional roles, role hierarchies and role inverses. The syntax of concept descriptions, axioms (a set of axioms is called TBox), and assertions for describing data (a set of assertions is called ABox) is given in Figure 2. Here P is a role symbol, A a concept symbol and a, b are constants. Moreover, in order to keep query answering complexity low, the interplay of functionality and inclusion axioms is restricted in the following way: If R occurs in a functionality assertion, then R and its inverse do not occur on the right-hand side of a role inclusion axiom. The semantics of the logic is defined in the usual first-order logic style in terms of relational structures, or interpretations $\mathcal{I}$, that satisfy axioms and assertions, with the additional constraint of the unique name assumption (UNA): Different constants are mapped to different elements in the domain of the interpretations. The UNA is needed for FOL rewritability [2, Theorem 6.6].

An ontology $\mathcal{O}$ is a tuple $(Sig, \mathcal{T}, \mathcal{A})$, with a signature Sig (i.e., set of concept symbols, role symbols and constants), with a TBox $\mathcal{T}$, and with an ABox $\mathcal{A}$. An ontology is satisfiable iff there exists an interpretation satisfying $\mathcal{T}$ and $\mathcal{A}$. Given an interpretation $\mathcal{I}$, checking whether $\mathcal{I}$ satisfies $\mathcal{T}$ and $\mathcal{A}$ is called model checking (and $\mathcal{I}$ is called a model if satisfiability is given).

Given an ontology, query answering is a decision problem directly relevant for practical applications. An *FOL query* $Q = \psi(\vec{x})$ is a first-order logic formula $\psi(\vec{x})$ whose free variables are the ones in the n-ary vector of variables $\vec{x}$; the variables in $\vec{x}$ are called *distinguished variables*. If $\vec{x}$ is empty, the query is called boolean.

Logics of the DL-Lite family have the remarkable property that checking the satisfiability of ontologies as well as answering queries w.r.t. ontologies can be reduced to model checking. Since in the logical perspective a relational database is nothing else than an interpretation (or a finite part of the canonical model, aka Herbrand model, to be more precise), DL-Lite thus offers the possibility to keep data descriptions as a virtual ABox in a relational database and reduce consistency checks and query answering to SQL queries (first-order logic formulas) w.r.t. the database. These properties of DL-Lite are formally described by the term *first-order logic rewritability* or *FOL rewritability* for short.

Some definitions are required to explain this in detail. Let $\vec{a}$ be a vector of constants from the signature of the ontology. The semantics of n-ary FOL queries with respect to an interpretation $\mathcal{I}$ is given by the set $Q^{\mathcal{I}}$ of n-ary tuples $\vec{d}$ over the domain $\Delta^{\mathcal{I}}$ such that $\mathcal{I}_{[\vec{x} \mapsto \vec{d}]} \models \psi(\vec{x})$. The semantics of FOL queries w.r.t. an ontology $\mathcal{T} \cup \mathcal{A}$ is given by the set of certain answers $cert(Q, \mathcal{T} \cup \mathcal{A})$. This set consists of n-ary tuples of constants $\vec{a}$ from Sig such that $\psi[\vec{x}/\vec{a}]$ (i.e. the formula resulting from $\psi(\vec{x})$ by applying the substitution $[\vec{x}/\vec{a}]$) follows from the ontology.

$$cert(\psi(\vec{x}), \mathcal{T} \cup \mathcal{A}) \quad = \quad \{\vec{a} \mid \mathcal{T} \cup \mathcal{A} \models \psi[\vec{x}/\vec{a}]\}$$

Two well investigated subclasses of FOL queries are *conjunctive queries (CQ)* and *unions of conjunctive queries (UCQ)*. A CQ is a FOL query in which $\psi(\vec{x})$ is an existentially quantified conjunction of atomic formulas $at(\cdot)$, $\psi(\vec{x}) = \exists \vec{y} \bigwedge_i at_i(\vec{x}, \vec{y})$. The UCQs allow disjunctions of CQs, i.e., $\psi(\vec{x})$ can have the form $\exists \vec{y_1} \bigwedge_{i_1} at_{i_1}(\vec{x}, \vec{y_1}) \vee \cdots \vee \exists \vec{y_n} \bigwedge_{i_n} at_{i_n}(\vec{x}, \vec{y_n})$. We conceive a UCQ as a set of CQs. The existential quantifiers in UCQs are interpreted in the same way as for FOL formulas (natural domain semantics) and not with respect to a given set of constants mentioned in the signature (active domain semantics).

With the technical notions introduced so far we are in a position to give the definition for FOL rewritability. In the following, let the canonical model of an ABox $\mathcal{A}$, denoted $\mathit{DB}(\mathcal{A})$, be the minimal Herbrand model of $\mathcal{A}$. *Checking the satisfiability of ontologies is FOL rewritable* iff for all TBoxes $\mathcal{T}$ there is a boolean FOL query $Q_{\mathcal{T}}$ such that for all ABoxes $\mathcal{A}$ it is the case that the ontology $\mathcal{T} \cup \mathcal{A}$ is satisfiable just in case the query $Q_{\mathcal{T}}$ evaluates to false in the model $\mathit{DB}(\mathcal{A})$. *Answering queries from a subclass $\mathcal{C}$ of FOL queries w.r.t. to ontologies is FOL rewritable* iff for all TBoxes $\mathcal{T}$ and queries $Q = \psi(\vec{x})$ in $\mathcal{C}$ there is a FOL query $Q_{\mathcal{T}}$ such that for all ABoxes $\mathcal{A}$ it is the case that $cert(Q, \mathcal{T} \cup \mathcal{A}) = Q_{\mathcal{T}}^{\mathit{DB}(\mathcal{A})}$.

For DL-Lite it can be shown [2] that the satisfiability check is FOL rewritable. Let $\mathcal{T} = \{A \sqsubseteq \neg B\}$ and $\mathcal{A} = \{A(a), B(a)\}$, then the satisfiability test is carried out by answering the query $Q_{\mathcal{T}} = \exists x. A(x) \wedge B(x)$ w.r.t. $\mathit{DB}(\mathcal{A})$, resulting in the answer yes and indicating that $\mathcal{T} \cup \mathcal{A}$ is unsatisfiable. Moreover, answering UCQs in DL-Lite can be shown to be FOL rewritable [2]. FOL rewritability of satisfiability is a prerequisite for answering queries because in case the ontology is not satisfiable the set of certain answers is identical to all tuples of constants in the signature.

The main technical tool for proving the rewritability results is the chase construction known from database theory. The idea is to "repair" the ABox with respect to the constraints formulated by the positive inclusion axioms $\mathcal{T}_p$. The essential property of the canonical model $can(\mathcal{O})$ resulting from the chasing process is that it is a universal model of $\mathcal{T}_p \cup \mathcal{A}$ with respect to homomorphisms, i.e., $can(\mathcal{O}) \models \mathcal{T}_p \cup \mathcal{A}$ and $can(\mathcal{O})$ can be mapped homomorphically to all models of $\mathcal{T}_p \cup \mathcal{A}$. As existentially quantified positive sentences are invariant under homomorphisms, this property has the consequence that every UCQ Q posed to $\mathcal{T}_p \cup \mathcal{A}$ can be answered by computing $Q^{can(\mathcal{O})}$.

The idea of introducing the concept of FOL rewritability is motivated by the demand to enable computationally feasible reasoning services over large ABoxes. Because the size of the TBox (and the queries) is small with respect to the size of the ABoxes, computational feasibility is measured with respect to the size of the ABox alone, thereby fixing all other parameters (TBox, query respectively). The resulting type of complexity is called *data complexity*. Aiming at FOL rewritability is indeed a successful venture with respect to computational feasibility. This is due to the fact that the data complexity of answering FOL queries w.r.t. DL-Lite ontologies is in the low boolean circuits complexity class AC^0, which, roughly, is the class of problems that can be decided instantly (in constant time) with the help of polynomially many processors.

III. Obstacles for combining DL-Lite and RCC8

The NP-completeness of satisfiability tests for RCC8-constraint networks poses a severe problem when trying to define tractable or—even stronger—FOL rewritable spatio-thematic description logics that use the RCC8-calculus as the spatial domain. The main challenge in constructing a computationally tractable logic is to restrict the way the spatial domain can be accessed from within the logic; one has to control the "flow of information" from the spatial domain to the thematic domain of the underlying lightweight logic. For example, reducing the thematical component of the logic $\mathcal{ALC}(RCC8)$ of [8] to DL-Lite is not enough to define a combined logic that allows for FOL rewritability.

As testing the satisfiability of arbitrary RCC8 constraint networks is not FOL rewritable, the envisioned combination of some lightweight DL with the RCC8 domain cannot be expected to be FOL rewritable in the standard sense of FOL rewritability as recapitulated in Sect. II-B. Consider, e.g., the simple boolean query $Q = \mathsf{ntpp}(a^*, b^*)$, which asks whether regions a^*, b^* in the database are related such that a^* is a non-tangential proper part of b^*. The composition axiom for the pair $(\mathsf{ntpp}, \mathsf{ntpp})$ states that ntpp is a transitive relation; but the transitiveness condition can not be compiled into a finite FOL query. Intuitively, at least one would have to take into account all ntpp-paths from a^* to b^*, i.e., one would have to query the database for all $n \in \mathbb{N}$ with queries Q_n of the form $Q_n = \exists x_1^* \dots \exists x_n^*.\mathsf{ntpp}(a^*, x_1^*) \wedge \dots \wedge \mathsf{ntpp}(x_n^*, b^*)$, because the database may be of the form $\{\mathsf{ntpp}(a^*, c_1^*), \mathsf{ntpp}(c_1^*, b^*)\}$ or of the form $\{\mathsf{ntpp}(a^*, c_1^*), \mathsf{ntpp}(c_1^*, c_2^*), \mathsf{ntpp}(c_2^*, b^*)\}$ etc. Therefore, we define the following completeness and consistency condition for ABoxes and weaken the notion of FOL rewritability of satisfiability to FOL rewritability of satisfiability with respect to these ABoxes. An ABox $\mathcal{A}$ is called *spatially complete* iff the constraint network contained in $\mathcal{A}$ is a complete and satisfiable constraint network. A special case is a network in which there are no disjunctions but only base relations used for labeling edges. In practice, these networks can be computed from (consistent) quantitative geometric data.

Another obstacle for FOL rewritability with respect to query answering is the expressiveness of the query language. Though conjunctive queries are weaker than FOL queries, they allow for querying unnamed objects and building joins that are not treelike. We will therefore consider a weaker query language (GCQ^+ queries below) that is similar to the language of grounded conjunctive queries.

IV. Combinations of lightweight DLs with RCC8 allowing for FOL rewritability

We consider the following extension of DL-Lite, denoted DL-Lite(RCC8), in which concepts of the form $\exists U_1, U_2.r$ may appear on the right-hand side of TBox axioms and in which only the attribute loc is allowed to be functional. The semantics $U^{\mathcal{I}}$ of role chains $U = R \circ loc$ with respect to an interpretation $\mathcal{I}$ is given by role composition of $R^{\mathcal{I}}$ and $loc^{\mathcal{I}}$. The interpretation $C^{\mathcal{I}}$ of concepts of the form $C = \exists U_1, U_2.\{r_1, \dots r_k\}$ for $r_i \in \mathcal{B}_{RCC8}$ $(1 \leq i \leq k)$ is given as follows:

$$\begin{array}{rcl} C^{\mathcal{I}} & = & \{d \in \Delta^{\mathcal{I}} \mid \text{There are } e_1, e_2 \text{ with} \\ & & \quad (d, e_1) \in U_1^{\mathcal{I}} \text{ and } (d, e_2) \in U_2^{\mathcal{I}} \\ & & \quad \text{such that } (e_1, e_2) \in r_1^{\mathcal{I}} \text{ or} \dots \text{ or } (e_1, e_2) \in r_k^{\mathcal{I}}\} \end{array}$$

The restriction for concepts of the form $\exists U_1, U_2.r$ in Figure

$$\begin{array}{rcl} R & \longrightarrow & P \mid P^- \\ U & \longrightarrow & loc \mid R \circ loc \\ B & \longrightarrow & A \mid \exists R \mid \exists loc \\ C & \longrightarrow & B \mid \neg B \mid \exists U_1, U_2.r \text{ for } r \in Rel_{RCC8} \\ & & \text{and not } (U_1 = U_2 = loc \text{ and } \mathsf{eq} \notin r) \\ \textit{TBox:} & & B \sqsubseteq C, (\text{funct } loc), R_1 \sqsubseteq R_2 \\ T_\omega & = & Ax_{RCC8} \end{array}$$

Figure 3. The combined logic DL-Lite(RCC8)

3 assures that we do not get empty concepts from the beginning (without any interesting deduction); clearly, $\exists loc, loc.r$ denotes an empty concept with respect to Ax_{RCC8} if r does not contain the relation eq. We could also handle empty

concepts in the rewriting algorithms, but deciding to exclude empty concepts facilitates the rewriting process.

Excluding the special case that $U_1 = U_2 = loc$, one can see that concepts of the form $\exists U_1, U_2.r$ on the right side of TBoxes are not relevant for satisfiability checks; the reason is that at least one of U_1 or U_2 will contain a role symbol that leads to totally new regions, which cannot be identified by regions already taken into consideration. In short, DL-Lite(RCC8) does not essentially generate new potential inconsistencies with ABoxes in comparison with the potential inconsistencies of the pure DL-Lite part because DL-Lite(RCC8) offers only a weak means for restricting the models of the ABox. Therefore it is possible to use the satisfiability check of pure DL-Lite ontologies. The resulting proposition, which states that checking the satisfiability of DL-Lite(RCC8)-ontologies with spatially complete ABoxes is FOL rewritable is a corollary of [2, Theorem 4.14].

Proposition 1: Checking the satisfiability of DL-Lite(RCC8)-ontologies whose ABox is spatially complete is FOL rewritable. (Proofs of the propositions can be found in the accompanying technical report [9].)

Proposition 1 provides a prerequisite for rewriting queries with respect to ontologies in DL-Lite(RCC8). The query language for which the rewriting is going to be implemented is derived from grounded conjunctive queries and will be denoted by GCQ^+. This query language is explicitly constructed for use with DL-Lite(RCC8) and so provides only means for qualitative spatial queries. But it could be extended to allow also for quantitative spatial queries.

Definition 1: A *GCQ^+ atom w.r.t. DL-Lite(RCC8)* is a formula of one of the following forms:

- $C(x)$, where C is a DL-Lite(RCC8) concept without the negation symbol and x is a variable or a constant.
- $(\exists R_1 \dots R_n.C)(x)$ for role symbols or inverses of role symbols R_i, a DL-Lite(RCC8) concept C without the negation symbol, and a variable or a constant x
- $R(x, y)$ for a role symbol R or an inverse thereof
- $loc(x, y^*)$, where x is a variable or constant and y^* is a variable or constant intended to denote elements of the ω-admissible theory Ax_{RCC8}
- $r(x^*, y^*)$, where $r \in \mathit{Rel}_{RCC8}$ and x^*, y^* are variables or constants intended to denote elements of Ax_{RCC8}

A *GCQ^+ query w.r.t. DL-Lite(RCC8)* is a query of the form $\tilde{\exists}\vec{y}\vec{z^*} \bigwedge C_i(\vec{x}, \vec{w}^*, \vec{y}, \vec{z}^*)$ where all $C_i(\vec{x}, \vec{w}^*, \vec{y}, \vec{z}^*)$ are GCQ^+ atoms and $\tilde{\exists}\vec{y}\vec{z^*} = \tilde{\exists}y_1 \dots \tilde{\exists}y_n \tilde{\exists}z_1^* \dots \tilde{\exists}z_m^*$ is a sequence of existential quantifiers that have to be interpreted w.r.t. the active domain semantics.

With respect to this query language it is possible to show that a TBox in the combined logic DL-Lite(RCC8) can indeed be compiled into a UCQ and thus into an SQL query—if one presumes that the ABox is spatially complete.

Proposition 2: Answering GCQ^+ queries with respect to DL-Lite(RCC8)-ontologies whose ABox is spatially complete is FOL rewritable.

This proposition can be proved by extending the proof of Theorem 5.15 in [2]. The main component of our proof is a reformulation algorithm that is an adaption of the algorithm PerfectRef [2, Fig. 13] for reformulating UCQs w.r.t. DL-Lite ontologies to our setting in which GCQ^+ queries are issued to DL-Lite(RCC8) ontologies.

The original algorithm PerfectRef operates on the positive inclusion axioms of a DL-Lite ontology by using them as rewriting aids for the atomic formulas in the UCQ. For example, if the TBox contains the positive inclusion axiom $A_1 \sqsubseteq A_2$ (A_1 is a subconcept of A_2), and the UCQ contains the atom $A_2(x)$ in a CQ, then, among the CQ with $A_2(x)$, the rewritten UCQ query contains a CQ in which $A_2(x)$ is substituted by $A_1(x)$. In our adaption of PerfectRef, we integrate GCQ^+ atoms of the form $\exists U_1, U_2.r(x)$ into the overall reformulation process. The relevant implications of GCQ^+ atoms of the form $\exists U_1, U_2.r(x)$ that we have to account for are the following:

- The conjunction of concept $\exists R_1 \circ loc, loc.r_1$ and $\exists loc, R_2 \circ loc.r_2$ is a subconcept of $\exists R_1 \circ loc, R_2 \circ loc.r_3$ where $r_3 \in \mathit{Rel}_{RCC8}$ is a superset of composition table entries $r_1^i; r_2^j$ for $r_1^i \in r_1$ as left and $r_2^j \in r_2$ as right argument. I.e., if the formula $\exists R_1 \circ loc, R_2 \circ loc.r_3(x)$ occurs as a conjunct during the rewriting of a CQ, then it can be replaced by a conjunct of $\exists R_1 \circ loc, loc.r_1(x)$ and $\exists loc, R_2 \circ loc.r_2(x)$ in a new CQ for all $r_1, r_2 \in \mathit{Rel}_{RCC8}$ such that $r_1; r_2 \subseteq r_3$.
- If $\exists U_1, U_2.r_1(x)$ occurs as conjunct in the query and $B \sqsubseteq \exists U_1, U_2.r_2(x)$ with $r_2 \subseteq r_1$ is in the TBox, then create a new CQ in which $\exists U_1, U_2.r_1(x)$ is substituted by $B(x)$.
- If $\exists U_1, U_2.r_1(x)$ occurs as conjunct in the query and $B \sqsubseteq \exists U_2, U_1.r_2(x)$ with $r_2^{-1} \subseteq r_1$ is in the TBox, then create a new CQ in which $\exists U_1, U_2.r_1(x)$ is substituted by $B(x)$.
- If $\exists R_1 \circ loc, U_1.r(x)$ occurs as a conjunct in the query and $R_2 \sqsubseteq R_1$ is in the TBox, then create a new CQ by substituting $\exists R_1 \circ loc, U_1.r(x)$ with $\exists R_2 \circ loc, U_1.r(x)$.

As query answering in DL-Lite(RCC8) is FOL rewritable, queries like those from the scenario of the engineering office can be answered in a complete way by transforming them into SQL queries and getting the answers from the underlying database. The TBox of the engineering office may contain the following axioms, which formalize the necessary conditions for parks with lakes and playing areas, respectively, within DL-Lite(RCC8).

$$\begin{array}{rcl}
\mathit{Park{+}Lake} & \sqsubseteq & \mathit{Park} \\
\mathit{Park4Playing} & \sqsubseteq & \mathit{Park} \\
\mathit{Park{+}Lake} & \sqsubseteq & \exists \mathit{hasLake} \circ loc, loc.\mathtt{tpp} \\
\mathit{Park4Playing} & \sqsubseteq & \exists \mathit{hasPlAr} \circ loc, loc.\mathtt{tpp}
\end{array}$$

The ABox $\mathcal{A}$ is derived virtually by mappings from GIS data in a database; think of mappings for *Park+Lake* and

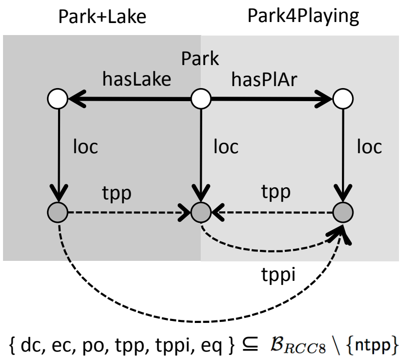

Park+Lake
Park4Playing
Park
hasLake
hasPlAr
loc
loc
loc
tpp
tpp
tppi
{ dc, ec, po, tpp, tppi, eq } ⊆ $\mathcal{B}_{RCC8} \setminus$ {ntpp}

Figure 4. Interpretation satisfying the queries Q, Q', and Q''.

Park4Playing that produce an object a as instance of *Park+Lake*, *Park4Playing*, respectively. That is, assume the following: $\{\mathit{Park{+}Lake}(a), \mathit{Park4Playing}(a)\} \subseteq \mathcal{A}$.

The query asking for all parks with lakes and playing areas such that the playing area is not contained as island in the lake can be expressed as the following GCQ^+ (see Figure 4):

$$\begin{aligned} Q \quad = \quad & \mathit{Park}(x) \wedge \\ & \exists \mathit{hasLake} \circ \mathit{loc}, \mathit{hasPlAr} \circ \mathit{loc}.(\mathcal{B}_{RCC8} \setminus \{\mathsf{ntpp}\})(x) \end{aligned}$$

The reformulation algorithm introduced above produces a UCQ that contains, among Q and others, the following CQ according to the first rewriting rule in the extended reformulation algorithm presented above

$$\begin{aligned} Q' \quad = \quad & (\exists \mathit{hasLake} \circ \mathit{loc}, \mathit{loc}.\mathsf{tpp})(x) \wedge \\ & (\exists \mathit{loc}, \mathit{hasPlAr} \circ \mathit{loc}.\mathsf{tppi})(x) \end{aligned}$$

Due to the RCC composition table the relation between the location of the lake and the location of the playing area referred to in Q' is $\{\mathsf{dc}, \mathsf{ec}, \mathsf{po}, \mathsf{tpp}, \mathsf{tppi}, \mathsf{eq}\}$ (see Section II-A), which implies a relation $\mathcal{B}_{RCC8} \setminus \{\mathsf{ntpp}\}$ between the lake and the playing area as stated in Q. Thus, using the fact that $\exists \mathit{loc}, \mathit{hasPlAr} \circ \mathit{loc}.\mathsf{tppi}$ can be rewritten to $\exists \mathit{hasPlAr} \circ \mathit{loc}, \mathit{loc}.\mathsf{tpp}$ in combination with the subconcept rewriting rule for $A_1 \sqsubseteq A_2$ (see above) we get another CQ

$$Q'' = \mathit{Park{+}Lake}(x) \wedge \mathit{Park4Playing}(x)$$

Using the mappings of *Park+Lake* and *Park4Playing* to SQL, the final query to be posed to the database is obtained. This query captures the object a mentioned above such that query answering is complete and all objects with design flaws are found by taking the complement of Park w.r.t. the result set of Q.

V. Conclusion

The query language GCQ^+ allows for the SQL compilation of queries w.r.t. a DL-Lite(RCC8) conceptualization (TBox) for a geographic application domain (Propositions 1 and 2). In order to find all relevant mappings to SQL, the TBox is used to provide an axiomatization of the concepts used in the domain. In order to provide for complete query answering this formalization needs only to be quite weak as shown in the example. DL-Lite(RCC8) is not expressive enough to define sufficient conditions for concepts like that of a park containing a lake in terms of quantitative data. However, as we have argued, given the mappings of concepts (relations) to SQL, only necessary conditions on spatio-thematic concepts need to be formulated in order to automatically construct SQL queries that provide for complete query answering. The process of query rewriting requires reasoning w.r.t. the TBox and the axioms Ax_{RCC8}, and reasoning algorithms are indeed combinatorial w.r.t. query and TBox size (but we have small TBoxes and queries). However, given a compiled query, query answering is tractable in data complexity, and hence feasibility of ontology-based query answering in GIS applications is achieved.

References

[1] R. Grütter, I. Helming, S. Speich, and A. Bernstein, "Rewriting queries for web searches that use local expressions," in *Proceedings of the 5th International Symposium on Rule-Based Reasoning, Programming, and Applications (RuleML-2011 – Europe)*, ser. LNCS, N. Bassiliades, G. Governatori, and A. Paschke, Eds., vol. 6826, 2011, pp. 345–359.

[2] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodríguez-Muro, and R. Rosati, "Ontologies and databases: The DL-Lite approach," in *Semantic Technologies for Informations Systems – 5th Int. Reasoning Web Summer School (RW-2009)*, ser. LNCS, S. Tessaris and E. Franconi, Eds. Springer, 2009, vol. 5689, pp. 255–356.

[3] D. A. Randell, Z. Cui, and A. G. Cohn, "A spatial logic based on regions and connection," in *Proceedings of the 3rd International Conferecence on Knowledge Representation and Reasoning*, 1992, pp. 165–176.

[4] A. K. Mackworth, "Consistency in networks of relations," *Artif. Intell.*, pp. 99–118, 1977.

[5] J. Renz, *Qualitative Spatial Reasoning with Topological Information*, ser. Lecture Notes in Computer Science. Springer, 2002, vol. 2293.

[6] B. Bennett, "Modal logics for qualitative spatial reasoning," *Logic Journal of the IGPL*, vol. 4, no. 1, pp. 23–45, 1996.

[7] J. Renz and B. Nebel, "On the complexity of qualitative spatial reasoning: a maximal tractable fragment of the region connection calculus," *Artif. Intell.*, vol. 108, no. 1-2, pp. 69–123, 1999.

[8] C. Lutz and M. Miličić, "A tableau algorithm for description logics with concrete domains and general TBoxes," *J. Autom. Reasoning*, vol. 38, no. 1-3, pp. 227–259, 2007.

[9] Ö. L. Özçep and R. Möller, "Combining lightweight description logics with the region connection calculus," Institute for Softwaresystems (STS), Hamburg University of Technology, Tech. Rep., 2011, available online at http://www.sts.tu-harburg.de/tech-reports/papers.html.

O*-W: An Efficient O* algorithm to Process Optimal Sequence Traversal Query in a Complete Directed Graph

Qifeng Lu
MacroSys LLC.
Arlington, United States
qilu1@vt.edu

Kathleen Hancock
Center for Geospatial Information Technology, Virginia Polytechnic Institute and State University
Arlington, United States
hancockk@vt.edu

Abstract— **Optimal Sequence Traversal Query (OSTQ) is a graph based query that retrieves a minimum cost path that starts from a predefined vertex, traverses a set of given vertices, and ends at a predefined vertex. One of its applications is trip planning in the GeoProcessing domain in transportation. With sound theoretical support for optimally efficient, optimal, and greedy solutions, best first search in the artificial intelligence domain is effective to process such trip planning queries. O* is an existing bivariate best first search framework and its special case O*-SCDMST was proposed recently to retrieve optimal solutions for such a query. In this paper, we propose O*-W, a novel O* algorithm that uses a globally admissible heuristic to obtain optimal solutions for such a query. The performance of O*-W and O*-SCDMST in a complete directed graph whose vertices only contain the origin, the destination, and the vertices of interest, and is studied through a set of experiments, and the result demonstrates that when the number of vertices of interest is up to 15, on average, O*-W reduces computation time by more than one order of magnitude when compared to O*-SCDMST. More importantly, on average and in worst cases, O*-W increasingly outperforms O*-SCDMST when the number of points of interest increases.**

Keywords- Bivariate Best First Search, Heuristic, OSTQ, O*, O*-SCDMST, O*-W

I. Introduction and Background

Optimal Sequence Traversal Query (OSTQ) is a graph based query that retrieves a minimum cost path that starts from a predefined vertex, traverses a set of given vertices, and ends at a predefined vertex [1]. The graph, g-G, may include normal vertices that are neither vertices of interest, nor the given origin or destination. In the GeoProcessing domain in transportation, an OSTQ corresponds to a trip planning query that asks for a shortest or quickest path that traverses a set of locations with the given origin and destination. For example, a user may start from his/her office, go to a supermarket, a book store, a restaurant, a movie theater, and end at home.

Similar to a travelling salesman problem [2], OSTQ is NP hard [3][4]. The naïve method that retrieves the optimal solution would compute all possible order combinations of the given points of interest. The time complexity is O(n!), where n is the number of locations of interest.

O*, a bivariate best first algorithm that uses problem domain knowledge to guide the search for the optimal solution to an OSTQ in a g-G graph, was recently proposed [1]. As exact solutions, two special cases of O*, O*-SCDMST [1] and O*-Dijkstra [1], were proposed to process OSTQ in a fully connected directed graph whose vertices only contain the origin, the destination, and the vertices of interest. Such a graph is defined as *g-G*. The O*-SCDMST is proved to be optimal in a directed graph that obeys the triangle inequality and more efficient than O*-Dijkstra.

In a g-G graph, based on given rules, O* incrementally searches paths most likely to lead towards the goal state until it finds a path of minimum cost having traversed vertices of interest to the goal. O* uses a vertex's identification plus its VisitList, a sorted list that consists of a sorted sequence containing traversed vertices of interest along the path, to describe a state in the search.

A vertex in O* may have multiple states, and each state represents a different set of traversed vertices of interest. Different from states in a single-variate best first search such as A*, these states may not be compared to each other and removed accordingly unless some special state pruning rules unique to bivariate best search are followed [1].

At each step, O* uses a distance-plus-cost heuristic function, defined as *f(s)* for a state *s*, to determine the order in which the search visits states in the graph [1]:

$$f(s)=g(s)+h(s) \qquad (1)$$

where

g(s) is the cost function of the path from the initial state to the current state, and

h(s) is the global heuristic that estimates the distance from the current state to the goal state, traversing the remaining vertices of interest.

O* first takes the paths most likely to lead towards the goal state, which means the lower the *f(s)*, the higher is the priority for a state to be expanded.

Whenever an equal *f(s)* occurs, the state with a larger VisitList will be the next to expand. Otherwise, one is randomly selected. State A's VisitList, vl_A, is larger than State B's VisitList, vl_B, i.e., $vl_A>vl_B$, if the length of vl_A is longer. In other words, the path from the start state to A traverses more vertices of interest than that to B.

For an *N*-point traversal problem, O* first generates the source state that contains the given vertex and an empty VisitList. For all the states in the open list, the algorithm expands the state with the lowest f(s) value, and its children states are generated. A child state always inherits the VisitList of its parent whenever the child is not a vertex of interest; otherwise, the child's VisitList will be incremented by adding the vertex to it. The process continues until a goal state whose vertex is the final goal and VisitList contains all the vertices of interest or no solution is found. Once a goal state is reached, the algorithm will retrieve the obtained path using a data structure called backpointer, the combination of vertex identification and VisitList, to recursively obtain the parent until the origin state is reached.

Different heuristics in O* will result in different O* algorithms. The recently proposed O*-SCDMST is a special case of O* in that it uses a Semi-Connected Directed Minimum Spanning Tree (SCDMST) [1] to compute h(s) in a directed graph [1] and retrieves optimal paths.

In bivariate best first search, the global admissibility of a heuristic guarantees the solution be optimal. Global admissibility means that the global heuristic h(s) is admissible, i.e., its value is always smaller than or equal to the actual cost of the minimum-cost path from the current state to the goal state.

In this paper, O*-W, a novel O* algorithm that uses a globally admissible heuristic *H-W*, is proposed to retrieve optimal solutions for OSTQ query in a complete directed graph. Its performance is studied against O*-SCDMST, and the result shows that when the number of vertices of interest is up to 15, on average, O*-W reduces computation time by more than one order of magnitude when compared to O*-SCDMST. In addition, O* increasingly outperform O*-SCDMST on average and in worst cases.

The remaining of the paper is organized as follows. First, related work is introduced. Then, the O*-W algorithm is presented, followed by the algorithm to retrieve the heuristic for O*-W. Next, experiments and results are discussed. At last, the conclusion is provided.

II. Related work

This section provides a review of the state-of-the-art research on Traveling Salesman's Problem (TSP) and OSTQ. TSP is similar to OSTQ and the only difference is that TSP's origin and destination are the same while OSTQ's are not necessarily the same.

A. *Travelling Salesman Problem (TSP)*

The earliest TSP research is in Euclidean space that searches for a shortest round-trip route to traverse each city exactly once with all cities directly connected to each other. Solutions using dynamic programming [5], nearest neighbor [6], iterative algorithms such as 2-OPT, 3-OPT, and n-OPT [7], colony simulation [8], simulated annealing [9], Branch and Bound approach [10], were proposed to solve the problem, either exactly or approximately, and the result is a Hamiltonian cycle that visits each vertex exactly once and returns to the starting vertex. These algorithms may be adjusted to process OSTQ. However, none of these algorithms is within the best first search domain, which is the major interest of this paper.

B. *Bivariate Best First Search*

The concept of bivariate best first search was first proposed in [11] to address the deficiency of a single-variate best first search to process multiple categories of interest. It uses multiple variables to specify a state to be evaluated and expanded. L#, a generalized best first search that evaluates the promise upon a state in a similar form as A*, was proposed, together with a set of novel concepts in best first search, including local heuristic, global heuristic, local admissibility, and global admissibility [11]. As an instance of L#, the bivariate best-first-search C* was provided to processes Category Sequence Traversal Query (CSTQ) in a graph, which asks for a minimum cost route that starts from a given origin, traverses a set of ordered categories of interest that includes multiple objects in each category, with one selection from each category, and ends at a given destination [11].

As another instance of L#, O* was proposed to process OSTQ in a graph [1]. It is different from C* because their state definitions are different, and adopted data structures are different as well. O*-SCDMST and O*-Dijkstra are two special cases of O* to retrieve optimal solutions for fully connected directed graphs.

III. O*-W: An Efficient O* Algorithm to Process OSTQ in a Complete Directed Graph

In this section, a global heuristic, *H-W*, as the estimate from the current state to the goal state is proposed. The O* algorithm uses *H-W* as the heuristic to retrieve optimal solutions for OSTQ is named *O*-W*.

Consider a complete directed graph, *G(V,E)*, where *V* and *E* are the set of vertices and edges, respectively, and a starting vertex *s* and an ending vertex *e* in *V*. *V* only contains *s*, *e* and the vertices of interest. Associated with each edge *(i,j)* from vertex V_i to vertex V_j in *V* is a cost *c(i,j)*. Let $|V|=n$ and $|E|=m$, *n* is larger than 2, and all the edge costs in *G(V,E)* obey the triangle inequality, which means for any triangle composed by three vertices in *V*, the sum of the costs of any two sides must be greater than or equal to the cost of the remaining side. *W*, an edge set, is defined to contain the following edges: 1) for any vertex *v* except *s* and *e* in *V*, *W* contains its incoming edge from a vertex v_x other than *e* and its outgoing edge to a vertex v_y other than *s* where v_x is not the same as v_y and the sum of the costs of these two edges are the least of all its incoming edge and outgoing edge combinations*; 2) for s, W* contains *s*'s least-cost outgoing edge, le_s, whose end vertex is not *e*; and 3) for *e*, *W* contains *e*'s least cost incoming edge, le_e, whose end vertex is not *s*. Accordingly, the total number of such edges in *W* is *2(n-1)*. Assume the total cost of these edges in *W* is *S*. Now *H-W* is defined as the following in such an environment:

For an OSTQ, assume all vertices except *s* and *e* in *V* from *G* are remaining vertices of interest, *e* is the goal state, and *s* is the current state, then the global heuristic *H-W* is *S/2*. When no vertices other than *s* and *e* exist in V, then *H-W* is *c(s,e),* the cost of the edge from *s* to *e*.

Theorem 1: *H-W* is globally admissible.

Proof:

1) When no vertices other than *s* and *e* exist in *V*, *H-W=c(s,e)*. Since *c(s,e)* is the same as the actual cost, then *H-W* is globally admissible.

2) The following is to prove that *H-W* is globally admissible when more than two vertices exist in *V*.

Assume an optimal path, *p*, is defined as ($v_0, v_1, \ldots, v_{n-1}$) with the sequence of visited vertices of interest, its cost is *c(p)*, and $v_0=s$ and $v_{n-1}=e$. First, it is clear that 1) any vertex *v* other than *s* and *e* along *p* will have both an incoming edge and an outgoing edge, 2) the other end vertex of its incoming edge, v_a, is not *e*, 3) the other end vertex of its outgoing edge, v_b, is not *s*, and 4) v_a and v_b are not the same either because the graph is complete and obeys the triangle inequality. It is also true that the number of edges along such an optimal path is *n-1*. For any two consecutive edges *(j-1,j)* and *(j,j+1)* from vertex v_{j-1} to vertex v_j and then to v_{j+1} along *p*, based on the definition of the edges in *W*, the sum of the costs of both edges are not smaller than the sum of the cost of the incoming edge to *j* and the cost of the outgoing edge from *j* in *W*. Therefore, for any two consecutive edges starting from a vertex v_k on the partial path ($v_0, v_1, \ldots, v_{n-3}$) of *p*, the sum of the two edges' costs is never smaller than the sum of the costs of two edges of the vertex v_{k+1} from *W* (one to v_{k+1} and the other from v_{k+1}).

Now assume all vertices on the partial path ($v_0, v_1, \ldots, v_{n-3}$) of *p* are taken into account, then any edge (*j-1,j*) (*j>1* and *j<n-1*) along *p* is repeated once while each of edge *(0,1)*, i.e., $s \to v_1$, and edge *(n-2,n-1)*, i.e., $v_{n-2} \to e$, is counted only once. Correspondingly, all edges except edge le_s and le_e in *W* are taken into account and each is counted once. In addition, based on the definition of le_s and le_e, we know *c(0,1)*, the cost of edge *(0,1)*, is never smaller than $c(le_s)$, the cost of le_s; and *c(n-2,n-1)*, the cost of edge *(n-2,n-1)*, is never smaller than $c(le_e)$, the cost of le_e. Accordingly, assume the total weighted cost of the *n-1* edges along *p* is *c(ce)* (the cost of an edge that repeats once is doubled when to calculate the total cost),i.e., $c(ce)=c(0,1)+2c(1,2)+\ldots+2c(n-3,n-2)+c(n-2,n-1)$, then the following inequations exist:

$c(0,1)>= c(le_s)$ (2)

$c(n-2,n-1)>=c(le_e)$ (3)

$c(ce)>=S- c(le_s)- c(le_e)$ (4)

since $c(0,1)+c(n-2,n-1)+c(ce)=2c(p)$ (5), based on inequations (2), (3), and (4) and equation (5),

$2c(p)>=S$, and thus

$c(p)>=S/2$.

Based on case 1) and case 2), the proof is complete.

Since *H-W* is globally admissible, based on Lemma 2 in [1], O*-W that uses globally admissible heuristics retrieves optimal solutions.

It is clear that *O*-W* is a special case of O*. Accordingly, it follows the same process as O* as discussed in Section 1. O*-W is proved globally admissible in a complete directed graph while O*-SCDMST in a fully-connected graph. Both Graphs obey the triangle inequality, and the first is a special case of the second. The other difference between O*-SCDMST and O*-W is that *O*-W* uses *W-H* as the global heuristic while O*-SCDMST uses the cost of the SCDMST tree.

IV. THE YUMEI ALGORITHM TO RETRIEVE *H-W*

The Yumei algorithm in Figure 1 is proposed to retrieve the edge set W for *H-W*. After W is obtained, the half of the total cost of its edges is *H-W*.

The time complexity of Yumei is $O(|V|^2)$ and spatial complexity is $O(|V|^2)$.

```
Input: A complete directed weighted graph with vertices V and edges E (V only contains the starting vertex s, and the ending vertex e, and vertices of interest).
Initialize: $V_{new}$ = {s},  W = {}
Calculate $le_s$ and $le_e$, $V_{new}$ = {s,e}, W = { $le_s$, $le_e$ }
Repeat until $V_{new}$ = V:
  Calculate edge (u, v) from E whose value is the minimum among all edges to v and u is not e.
  Calculate edge (v, w) from E whose value is the minimum among all edges from v and w is not s.
  If u=w:
        Calculate the edge (u',v) from E whose cost is the second least among all edges to v and u' is not e.
        Calculate the edge (v,w') from E whose cost is the second least among all edges from v and w' is not s.
        If cost(u',v)+cost(v,w)>cost(u,v)+cost(v,w'):
          Add v to $V_{new}$, add (u, v) and (v,w') to W
        Else:
          Add v to $V_{new}$, add (u', v) and (v,w) to W
  Else:
        Add v to $V_{new}$, add (u, v) and (v,w) to W
Output: $V_{new}$ and W
```

Figure 1: The pseudo code for Yumei algorithm

V. EXPERIMENT AND RESULT ANALYSIS

The purpose of the experiment is to test the performance of O*-W to retrieve optimal paths in complete directed graphs. O*-SCDMST is used as the baseline.

A. Data Set

An asymmetric TSP problem (Fischetti) with 34 points of interest [12], corresponding to vertices of interest in O*, is used as the data set for this experiment. The problem is a special case of Vehicle Routing Problem, and thus an asymmetric TSP [12]. The data set contains the edge costs between any two points in the generated complete directed graph that only contains the starting vertex, the ending vertex, and the vertices of interest. In this experiment, a set of sample OSTQ problems is generated from this data set. First, the number of points of interest consecutively changes from 2 to 15. Second, for each number of points of interest, 30 problem samples are randomly generated, i.e., the origin, the destination, and the points of interest in each problem sample are randomly selected from the 34 points. Consequently, a set of 420 problems is generated.

B. Performance Measures

To study the performances of the two algorithms, the following performance measures are identified.

Minimum Process Time (MinPT): the minimum time required to obtain a solution for each number of points of interest (seconds);

Maximum Process Time (MaxPT): the maximum time required to obtain a solution for each number of points of interest (seconds);

Average Process Time (APT): the average time required to process a query over all runs (seconds).

C. Results and Discussion

The results are presented in Table 1. *O*-S* represents O*-SCDMST, and *NPoI* represents the number of points of interest, i.e., the number of cities to traverse.

Figure 2 through Figure 4 are provided to visualize the performance measures provided in Table 1. Notice that the Y-Axis in Figure 3 and Figure 4 uses logarithmic scale.

In addition, it is observed that the cost of each optimal path retrieved by O*-W is the same as O*-SCDMST, which demonstrates that O*-W also retrieves optimal solutions for OSTQ processing.

Based on MinPT, O*-W can retrieve the optimal path within 2 seconds for an OSTQ of 15 NPoI. However, based on MaxPT, it may still require 871.88 seconds for another query with the same number of points of interest. This implies that O*-W's performance depends on how closely the selected H-W heuristic approaches to the actual cost.

Based on MinPT shown in Figure 2, O*-W outperforms O*-SCDMST over all runs.

Based on MaxPT shown in Figure 3, O*-W outperforms O*-SCDMST when NPoI is larger than 4. This is due to the fact that O*-W requires additional time to compute the H-W heuristic, and when NPoI becomes larger, this additional time is no longer a dominant factor, instead, the obtained heuristic expedites the overall search process. In addition, for these worst cases, O*-W increasingly outperforms O*-SCDMST when the number of points of interest increases.

Based on APT shown in Figure 4, O*-W increasingly outperforms O*-SCDMST when the number of points of interest increases. On average, O*-W can process OSTQ of up to 14 NPoI within 1 minute.

In Figure 3 and Figure 4, it is noticeable that O*-W is sub-exponential in time complexity. It is hard to decide from Figure 2 whose Y-Axis does not use logarithmic scale.

Table 1: Performance results for O*-W and O*-SCDMST (in seconds)

	MinPT		*MaxPT*		*APT*	
NPoI	**O*-S**	**O*-W**	**O*-S**	**O*-W**	**O*-S**	**O*-W**
2	0.00	0.00	0.04	0.02	0.00	0.00
3	0.00	0.00	0.00	0.02	0.00	0.00
4	0.00	0.00	0.01	0.02	0.00	0.00
5	0.00	0.00	0.03	0.02	0.01	0.01
6	0.00	0.00	0.15	0.09	0.04	0.02
7	0.00	0.00	0.77	0.39	0.17	0.07
8	0.05	0.00	2.21	0.36	0.53	0.16
9	0.05	0.02	7.62	1.98	1.78	0.47
10	0.04	0.03	51.96	6.68	5.63	1.15
11	0.18	0.16	314.78	52.60	29.03	6.07
12	0.39	0.14	234.94	40.94	66.82	10.78
13	3.54	0.52	1109.16	199.05	245.07	30.78
14	2.06	0.87	3900.67	537.26	548.08	58.34
15	3.37	1.39	20857.75	871.88	2204.14	144.32

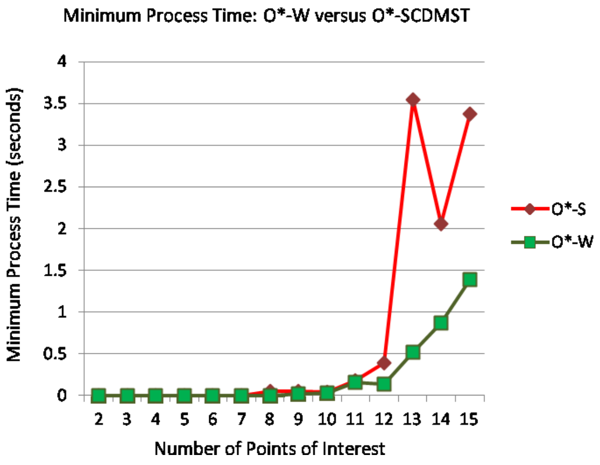


Figure 2: Minimum process time over different number of traversed points of interest: O*-W versus O*-SCDMST

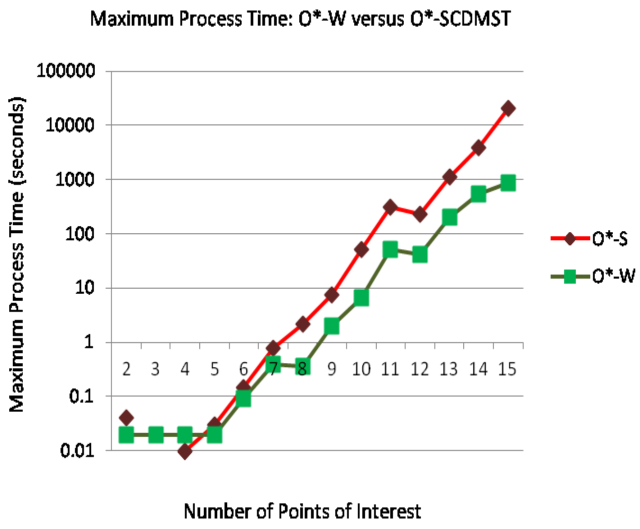


Figure 3: Maximum process time over different number of traversed points of interest: O*-W versus O*-SCDMST

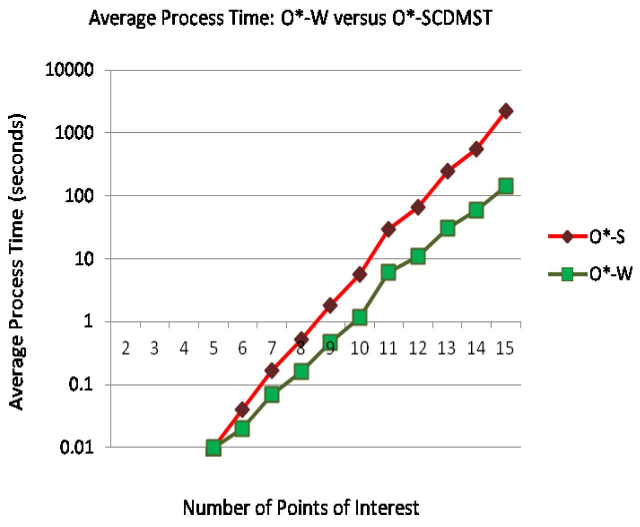


Figure 4: Average process time over different number of traversed points of interest: O*-W versus O*-SCDMST

VI. CONCLUSION

This paper proposed a novel O* algorithm, O*-W, to retrieve optimal solutions for OSTQ processing in a complete directed graph. It uses a globally admissible heuristic, H-W, to effectively and efficiently prune states. According to the experiment result, on average and in worst cases, O*-W increasingly outperforms O*-SCDMST. Even when the number of points of interest is not larger than 15, O*-W can still outperform O*-SCDMST by more than one order of magnitude, measured in either MaxPT or APT, which indicates O*-W works significantly better than O*-SCDMST.

The data size used in the experiments is rather small. To address OSTQ processing of a large data set, since it is known that the complexity to compute the pair-wise distances between all pairs of via-vertices is polynomial and much lower than the complexity of the corresponding OSTQ problem, in existing geoprocessing practices, typically these distances are pre-computed and stored in a database to expedite the process for querying any pair-wise distances.

REFERENCES

[1] Q. Lu and K. Hancock. O*: A Bivariate Best First Search Algorithm to Process Optimal Sequence Traversal Query in a Graph. geoprocessing, pp.53-61, 2011 Third International Conference on Advanced Geographic Information Systems, Applications, and Services, 2011.

[2] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 1997

[3] Michael R. Garey, Donald S. Johnson, and Larry Stockmeyer. *Some simplified NP-complete problems*. Proceedings of the sixth annual ACM symposium on Theory of computing, p.47-63. 1974.

[4] Wikipedia. *Travelling Salesman Problem*. http://en.wikipedia.org/wiki/Traveling_salesman_problem, last retrieved on September 20, 2011.

[5] M. Held and R. M. Karp. A Dynamic Programming Approach to Sequencing Problems, Journal of the Society for Industrial and Applied Mathematics 10(1) (1962): pp. 196–210.

[6] D. J. Rosenkrantz, R. E. Stearns, and P. M. Lewis II. An Analysis of Several Heuristics for the Traveling Salesman Problem. SIAM Journal on Computing. 6 (1977): pp. 563–581.

[7] J. L. Bentley. Fast Algorithms for Geometric Traveling Salesman Problems. ORSA Journal on Computing 4, (1992), pp. 387-411.

[8] Ma. Dorigo. Ant Colonies for the Traveling Salesman Problem. Université Libre de Bruxelles. IEEE Transactions on Evolutionary Computation, 1(1) (1997):pp. 53–66.

[9] E.H.L. Aarts and J. Korst. Simulated Annealing and Boltzmann Machines: A stochastic Approach to Combinatorial Optimization and Neural Computing. John Wiley & Sons, Chichester, 1989.

[10] J. Clausen and M. Perregaard, On the Best Search Strategy in Parallel Branch-and-Bound - Best-First-Search vs. Lazy Depth-First-Search, Proceedings of the Parallel Optimization Colloquium, (1996).

[11] Q. Lu and K. Hancock. C*: A Bivariate Best First Search to Process Category Sequence Traversal Queries in a Transportation Network. geoprocessing, pp.127-136, 2010 Second International Conference on Advanced Geographic Information Systems, Applications, and Services, 2010.

[12] Robert C. Prim: Shortest connection networks and some generalizations. In: Bell System Technical Journal, 36, pp. 1389–1401,1957

An Architecture for Geographic Information Systems on the Web - webGIS

Mariano Pascaul, Eluzai Alves, Tati de Almeida,
George Sand de França, Henrique Roig
Geosciences Institute
University of Brasilia, UnB
Brasilia, Brazil
mariano.pascual@gmail.com

Maristela Holanda
Department of Computer Science
University of Brasilia, UnB
Brasilia, Brazil
mholanda@cic.unb.br

*Abstract***— Geographic Information Systems for the web (webGIS) are being implemented for different purposes. In this context, one of the greatest challenges is to integrate different sources of geographic data, as well as the visualization of this information using maps in an interactive environment. This paper presents a proposal for architecture for the webGIS with interoperability between different sources of heterogeneous data, as well as the visualization of maps in different formats with components implemented with Web 2.0 technology. The architecture was validated through a case study that implemented a webGIS to academic research at the Geosciences Institute of the University of Brasilia.**

Keywords-Geographical Information System; webGIS; Geographical database; Map Visualization.

I. INTRODUCTION

Geographic data, having been collected, is now available in a wide variety of formats. Geographic data is available in files, databases or Geographic Information Systems (GIS) [1]. A GIS is frequently defined as the combination of a database management system, a set of operations for exploring data, and a graphic display system that is used for geospatial analysis. These GIS analyses have the main purpose of supporting decision making and modeling some of the possible consequences of those decisions [2][8][16][20]. GIS environments are also cartographic tools that facilitate building maps and examining the impacts of changes to the maps interactively [1][3][5][9][11][12].

Currently, GIS on the web (webGIS) is being developed, and one challenge in that environment is interoperability among heterogeneous databases. For interoperability of the data, the web services technology is being used [15]. The standard set by the OGC (OpenGIS Consrtium) proposes the open service architecture of web GIS to support data-interoperability. And, it suggests the GML (Geographic Markup Language) based on XML to exchange the data between the web client and the web GIS [6][7]. REST [22] technology is also used to support interoperability with geographical databases.

For the visualization of maps in an interactive way, Web 2.0 technology is being applied through different components of RIA (Rich Internet Application). As is observed in [10], this technology is being applied for the development of web-GIS. Web mapping applications such as Google Maps, Google Earth, Microsoft Bing Maps and Yahoo Maps are usually considered good examples of Web 2.0 [15].

This paper presents an architecture for webGIS based on Web 2.0 and interoperability among different geographical data. The architecture is based on web services and can be used with open or owner map and database servers.

The content of this paper is divided into the following sections: 2 – basic concepts about webGIS are presented; 3 – the proposed architecture is defined; 4 – related works, which are analyses; 5 – Case Study, where the architecture was used to develop the GIS for the Geosciences Institute of the University of Brasilia; and finally, 6 – the conclusions.

.

II. WEBGIS

Web GIS is any GIS that uses Web technologies. The simplest form of webGIS should have at least a server and a client, where the server is a Web application server, and the client is a Web browser, a desktop application, or a mobile application [15][4].

With regard to the architecture of a webGIS, the architecture based on three layers is most commonly used: User Interface Layer, Application Server Layer and Database Layer [17][19][21]. Some authors considered four layers, where the integration layer is added on the architecture webGIS, which is based on web services [14].

The User Interface layer serves as a graphic user interface (GUI) to present the result of spatial data, allowing the end users to interact with the backend services

The Application Server layer communicates with multiple data sources via the data integration layer, and interacts with end users to analyze and manipulate data coming from data provider services

The Database layer of data provider services, is a set of remote data provider services for data sharing. Each data provider service offers a set of interfaces through which client applications can pull remote data in and manipulate the data.

III. PROPOSED ARCHITECTURE

The model proposed (Figure 1) presents an abstract architecture for a webGIS. In this model we observe a set of classes developed that integrate with one another and servers of interoperability, and web services in the treatment and insertion of information, as well as in the availability of data for the final user.

ISBN: 978-1-61208-178-6

The use of web services and the development of classes that read and organize the overlay of data provided with servers of owner interactive map or of open software, provide a hybrid tool that can use WMS, WCS or REST for the presentation of layers on the web in a single RIA application utilizing web services.

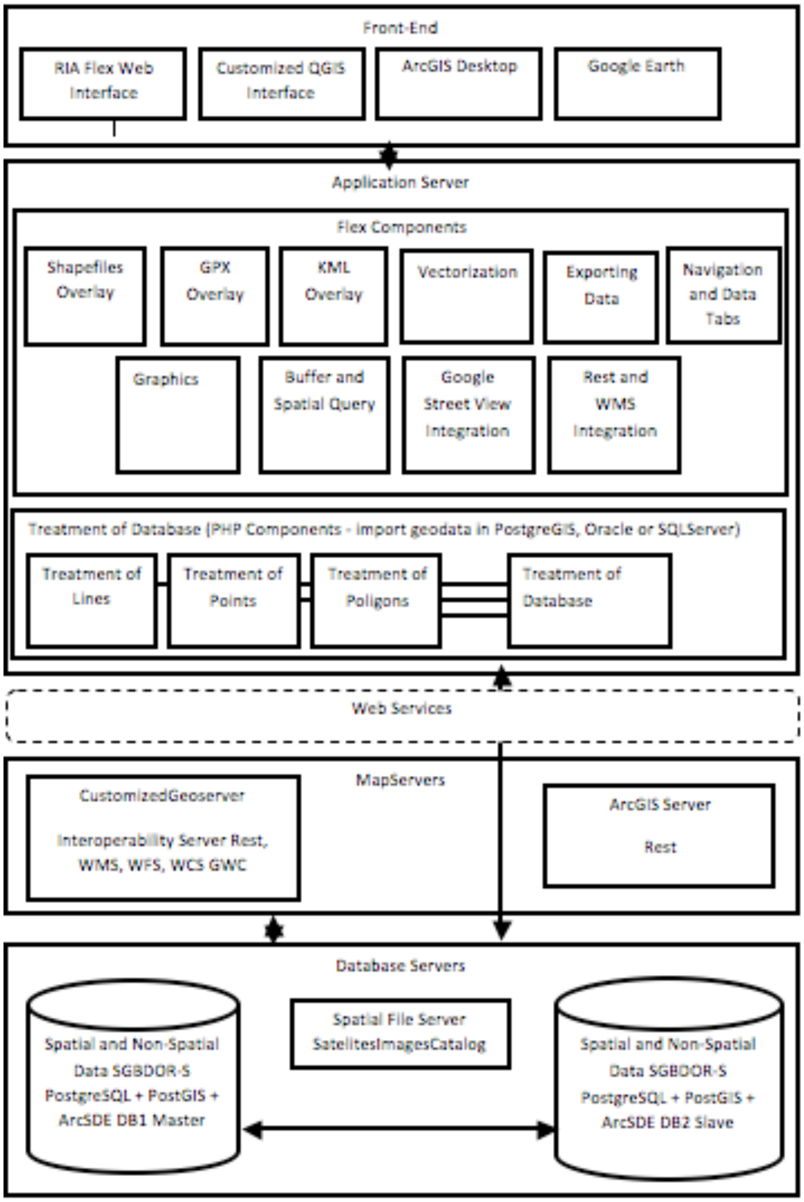


Figure 1. Abstract Model of the proposed architecture

Each component of the abstract architecture is described as follow.

- Shapefile Overlay

Component that enables the opening of shapefile files and their attributes, providing the overlay of geometries available in .shp files in the web Flex environment, the constant attributes in .dbf files are presented in a tabular format and related to the correspondent geometry. The features that can be viewed are: line, point, or polygon. In order to do this simply click on the tool of the folder with the zipped file containing minimally formatted .shp, .shx and dbf. Files. In case the set of files is projected in a system distinct to the map base, the inclusion of the compacted file .prj is necessary for the system to complete re-projection. In the construction of this component were used methods for the reading of attributes of the DBF written in ActionScript and also for reading the files in .shp format about the same language.

As it is necessary to open the files in a zipped format a unzipping package was used on files also written in ActionScript.

- GPX Overlay

The great quantity of Garmim brand GPS field equipment users was an enormous motivator in the development of this component, which makes the overlay of files downloaded directly from equipment of this type possible. All of the waypoints available to track overlays of a line where the initial point is presented in green, the path in yellow and the end in red are presented. The attributes given to the paths are shown in a floating table. The standard open GPX format was considered in the construction of the component.

- KML Overlay

This class provides the opening and comparison of the routes generated in Google Map platforms, and also in simple files in KML format, the available attributes in the referred to file are also presented in splints through floating panels.

- Vectorization

The class that serves vectorization and release in distinct database formats such as lines, points and polygons. Algorithms for the calculation of the area and perimeter are also in this class, if sending to the database is not necessary, the user can opt to save the designed features on a local disc. This component acts directly linked to the components in PHP for recording in database, the Flex interface, responsible for the instruments of vectorization accessing through web-services, routines that insert data in a determined data base already established. After the insertion one can, on moving or updating the map, visualize the geometry already inserted. This is possible because the feature is available as a service on the map base, facilitating the availability through ArcGIS Server as well as the Geoserver Java. For the case of use presented here the Post GIS type data was used directly.

- Graphics

Tools for which geographic database tables, or simply web services about them, can be chosen by way of selection boxes, from which one can select numerical attributes and vectorize an area on the screen resulting in pizza or bar form graphics. These are interesting analytical tools for statistical census data, however, with geographic presentation and selection.

- Exporting Data

Totally integrated components or Geoserver Java interactive server maps, can be made through this class, with the selection and later exportation and download of layers or a part of them in diverse formats such as shapefile, csv, pdf, XLS, KML, KMZ, JPG, PNG and others. Through this component, whatever is selected by the user on the screen, with the exception of the shapefile and Google formats, will be presented for download with a quadrant. Thinking a step ahead, as of yet developed, will be the component that can make the clip using the Geoserver server.

- Buffer with Spatial Query

Tools of bufferization through which complete features or a part of their registers can be used to create buffers. After the conclusion of this step the user of the class can once again select features, however, now to execute special queries about the buffer presented on screen. The result presented can be saved so that it is not necessary to complete the whole process again. Methods for selection of registers and services for generating the buffer were used, soon after the execution of the consultation and selection of attributes of geographic features about which one need obtain information, a spatial consultation is done, which outlines the occurrences of information inside the areas of buffers intersected.

- Google Street View Integration

A developed component that integrates with the Google Street View Platform, so that the information can be visualized in two dimensions, and through which one can navigate on maps, in addition to visualizing, in a part of the frame, the entire Google Street platform base, which always synchronizes the observation points with the street navigation.

- Integrated Overlay REST and WMS

This is the most important component developed, since it makes possible the integration of different formats of interactive maps hybridizing the framework, through which one can make use of standard REST web services available through the software ArcGIS Server as well as the Geoserver and makes the overlay transparent to the user of standard WMS and WFS web services. Actually tasks such as this are already available in API 2.x of the ESRI.

- Navigation and Data Tabs

A component of presentation and formatting grids was built aimed at improving the visual aspect of register lists extracted from the database. There is a great interaction between the navigation and consultation of attributes related to the geometry visualized.

- Treatment of Geographic Database

This component aims at adopting the tool of a set of classes capable of treating incoming information from the interface that will be sent to the database, and through this one can select which type of database will be used and the classes will be the interactions necessary for a correct treatment of distinct types between the manufacturers of SGBDs. Treatment of Geographic Features: set with the minimum rules necessary to avoid classical errors at the moment of vectorization, such as the creation of polygonal ties. Through this, perimeter and area are also calculated and different symbologies can be attributed to the design features, such as, completing the recording in databases or in the text format to be saved in a locale and overlay. These classes were developed using PHP language in standard MVC and object orientated, and web-services were made available, which it was necessary to send the textual and geographic information, vectorized on screen by way of the Flex interface, and the information of authentication aiming at increasing the degree of security of the tool since, these components interact directly with the database chosen.

A. *Development of the Architrcture*

For the development of the SIG, various program languages were chosen, including: Flex, Action Script, PHP , Java Script and Ajax. Java programming language was used in the customizations done for the Geoserver software, and was the same language in which the software was written.

In the development of the SIG web interface, the standard RIA, the Flex and Action Script was used. The API for Flex of ESRI was used as the core of the application. Beginning with the basic navigator, diverse components were used in the solution. Among them components capable of doing the overlay of layers in WMA and WCS format provided by software such as Geoserver, turning the navigator maps in SIG hybrid, capable of consuming data originating in the ArcGIS Server and/or the Geoserver. Another point to be considered is the framework developed in PHP language, oriented to objects and in standard MVC—Model View Controller. This server not only does dynamic construction of the electronic forms, but also treats data stemming from vectorization and, afterward, stores them in special extensions available in databases such as PostGIS of the PosgreSQL and Spatial of the Oracle.

Other components, such as the integration of the Google platform, with overlay for shapefile, KML and GPX bufferizaiton, spatial and graphic queries were developed and incorporated in the application.

One of the advantages presented by the set of components developed is not only the creation of the components that approximate the web platform of a GIS client/server environment, but the concept of classes makes the use of technologies with platforms based on free software viable, such as, Geoserver or others that generate web services in standard OGC, including WMS and WCS through the overlay of layers. Other advantages visualized were the speed with which the applications were created based on the set of codes developed.

The customization completed in the software Geoserver, together with the system customization done on the Geoserver software, aims at accelerating the process of making web services for maps available so that they can be incorporated quicker in the context of the application, although it is not necessary to have such a module to make use of the Geoserver software together with the system.

IV. Related Work

The proposal presented in this paper is the architecture of webGIS, which has the characteristic of components based on web 2.0 for visualization of spatial data. Our proposal has a layer of interoperability with free and own mapping server and different Geo-DBMS.

Shunfu Hu [19] presented an architecture for development web-based GIS applications. The webGIS was based on Microsoft Visual Basic. Microsoft Internet Information Server (MIIS) was employed as web Server and ESRI MapObjects Internet map Server as map Server. Unlike our proposal, the architecture present in [19] is not interoperable.

Boucelma et al. presented [1] a WFS-based mediation system for GIS interoperability. The functional architecture of the geo- graphic mediation system is mainly composed of three layers: a GIS mediator, Web Feature Servers (WFS) and data sources. In [1], the integration of query capabilities available at the sources and a geographical query language to access and manipulate integrated data is addressed. Differently, our architecture integrates the data source and the components Treatment of Database and Buffer and spatial query.

Majchrzak and More, in [10], cover how Web 2.0 technologies can be used to develop GIS through interaction with users. In [10], the aide volunteers in disaster situations is presented, using Google technologies. Our proposal is an abstract with different front-end, map server and Geo-DBMS, which can be used with Google technologies or others.

Zongyao and Yichun proposed, in [14], a service-oriented architecture for spatial data integration (SOA-SDI) of a large number of available spatial data sources that are physically sitting at different places, of which the development web-based GIS systems were based on SOA-SDI, allowing client applications to pull in, analyze and present spatial data from those available spatial data sources. Lu, in [18], defined a GIS platform architecture as a multi-layer architecture that integrated the web service, Servlet/JSP functions and GIS APIs based on the framework of J2EE infrastructure. The GIS system can be accessed by many different computers in networks with different kinds of operating systems. It is a distributed, platform independent system architecture. The data are stored and managed with EJB. Frehner and Brandli [20] presented the Virtual Database that consists of a framework for web-based retrieval, analysis, and visualization of spatially related environmental data based on the integration of distributed data. This architecture is based on web services. The proposal presented in [14][18][20] has the interoperability properties, however, our proposal supports more geographic data formats.

V. Case Study

The development of a public consulting system of academic research built on a GIS (Geographic Information Systems) paradigm, composed of an interface of interactive maps available on the Internet, using geographic databases, servers of interactive maps that generate services on the web and the use of languages and technologies of the latest generation for the layer of presentation and interaction with users. In Figure 2, we present the initial vision of the system.

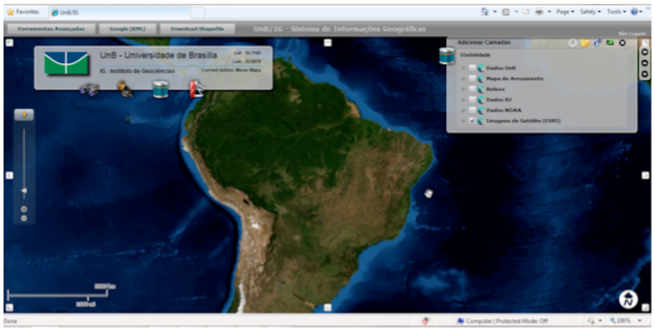

Favorites
UnB/IG
Page
Safety
Tools
Ferramentas Avançadas
Google (KML)
Download Shapefile
UnB/IG - Sistema de Informações Geográficas
UnB - Universidade de Brasília
IG - Instituto de Geociências
Current Action: Mover Mapa
Adicionar Camadas
Visibilidade
Dados UnB
Mapa de Arruamento
Relevo
Dados RJ
Dados NOAA
Imagens de Satélite (ESRI)
N
Done
Computer | Protected Mode: Off
100%

Figure 2. Initial View of the System

Various geographic web services are integrated, such as wind velocity and direction services of the American NOOA, with the rest platform of the ESRI as a mosaic of world images and services of maps from diverse Brazilian institutions. Thus, many services were in standard WMS and WCS, while services in standard REST were integrated guaranteeing then, a good degree of interoperability and sharing.

In Figure 3, we see the geographic features presented with the richest detail when using zoom tools on the map; at this level of zoom, one can see rural properties in salmon tones, as orange highlights the human settlements, and airstrips, train tracks and the main rivers can also be seen at this level. An altimeter base completes the background. The features not used as web services were stored in SGBD PostgreSQL with PostGIS.

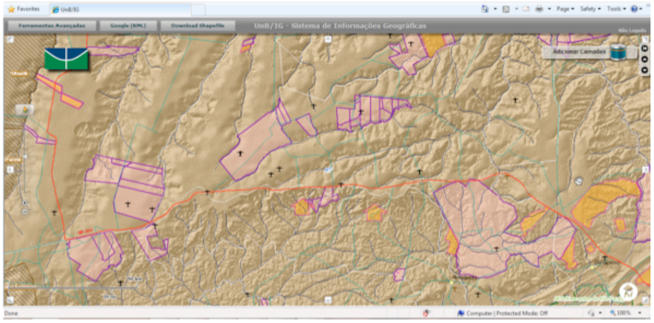


Figure 3. Zoom of detailed information of the geography mounted

The transparency of the maps can be altered and diverse layers of information of the same geographic area can be presented; this is a resource that Flex technology provides, which is useful and has great visual impact. In Figure 4, on selecting the area of the city of Rio de Janeiro we have the system fusing local data with data from web services of other institutions, such as: the Pereira Passos Urban Institute of Rio de Janeiro and the Brazilian Institute of the Environment and Renewable Natural Resources—IBAMA.

In Figure 5, we see the integration with the Google Street View platform. A point of observation was located in front of the Metropolitan Cathedral of the city of Rio de Janeiro; one can see in the upper part of the screen a higher view of the area and in the lower part a 3-D view obtained by the Street View platform of the same region and with the same direction of the arrow pointed.

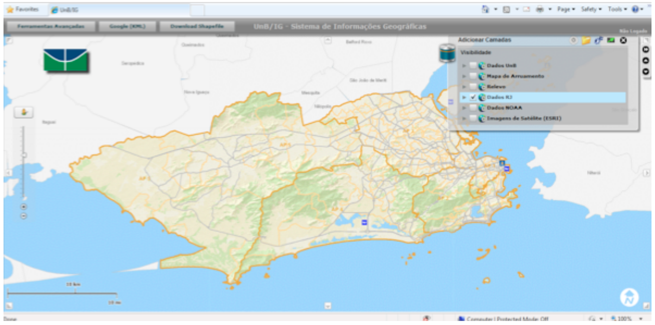


Figure 4. Zoom of the city of Rio de Janeiro

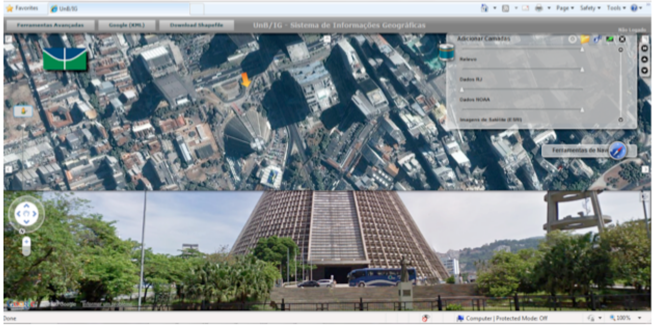


Figure 5. Integration Google Street View

Research tools of academic works were developed, and through these one can find scientific articles or studies completed in the research area.

Tools such as buffer and spatial queries were implemented and integrated; below shows the two operating together, according to data presented in Figure 6.

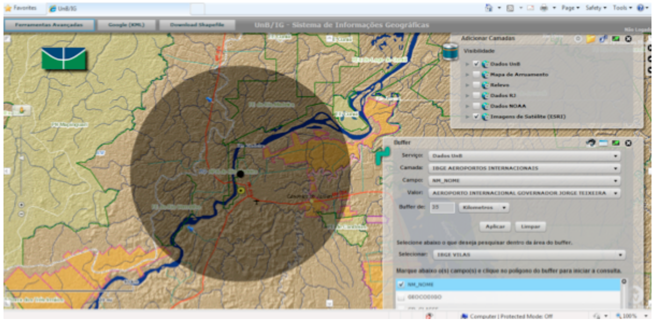


Figure 6. Buffer Tool and Spacial Query

Spatial graphic generating tools permit an evaluation of areas excessively inventoried for a determined resource or the identification of areas in need of particular studies, or identification of the needy areas of the given study, as presented in Figure 7.

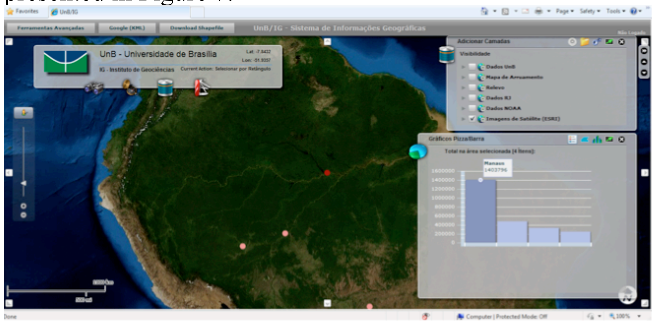

UnB/IG - Sistema de Informações Geográficas
UnB - Universidade de Brasília
IG - Instituto de Geociências
Adicionar Camadas
Visibilidade
Dados UnB
Mapa de Arruamento
Relevo
Dados RJ
Dados NOAA
Imagens de Satélite (ESRI)
Gráficos Pizza/Barra
Manaus

Figure 7: Spatial consultancy tools for generating graphics

VI. CONCLUSION

The architecture presented in this paper proposes a tool of rapid implantation and availability of geographic data through the web, with a set of services made available, which can easily integrate data of structured systems such as, consumption of information originating from diverse data sources.

The architecture proposed responded well to the system of application, much as the same being used for the development of other webGIS. Recently IBAMA adopted the architecture proposed in this work in the implementation of webGIS for the monitoring of its field supervision operations; through the set of components available, the information of operations and navigation of airships and roadways, always in a geo-referenced way are presented. Similarly, the effect of agents or the quantity of apprehension and automations are informed through structured systems and presented in a spatial way on the platform.

The next step is to implement the proposal architecture independent of API ESRI.

REFERENCES

[1] O. Boucelma and M. Essid. "A WFS-Based Mediation System for GIS Interoperability". Proc. ACM GIS'02, 2002, pp. 23-28, doi:10.1145/585147.585153.

[2] J. Arenas and H. Zambrano. "Web-based GIS Applications for Government". Proc. 3rd ICEGOV International Conference on Theory and Practice of Electronic Governance, Nov. 2009, pp. 383-384. doi:10.1145/1693042.1693125.

[3] M. Ostergren, J. Hemsley, M. Belarde-Lewis and S. Walker. "A Vision for Information Visualization in Information Science". Proc. iConference, Fev. 2011, pp. 531-537, doi: 10.1145/1940761.1940834.

[4] T. Bressan. Desenvolvimento e integração de um ambiente SIGWEB com ferramentas de software livre. Master Thesis. Federal University of Santa Maria Brazil, 2010.

[5] G. Câmara and G. Queiroz. Geographic Database Book. INPE Press. 2010.

[6] D. Kim and M. Kim. "Web GIS Service Component Based On Open Environment". Proc. IGARSS Geoscience and Remote Sensing Symposium. IEEE Press, Jun. 2002 pp. 3346-3348, doi:10.1109/IGARSS.2002.1027178.

[7] OGC. OpenGIS® Geography Markup Language (GML) Encoding Standard 2007.

[8] S. Dragiéevié and S. Balram. "A Web GIS collaborative framework to structure and manage distributed planning processes". Journal of Geographical Systems, Spring-Verlag, vol. 6, 2004, pp. 133-153, doi: 10.1007/s10109-004-0130-7.

[9] M. Rhyne. "Visualizing Geospatial Data". Proc. ACM SIGGRAPH 2004, doi:10.1145/1103900.1103931.

[10] A. Majchrzak and P. More. "Emergency! Web. 2.0 the Rescue!". Communications of the ACM, vol. 54, n. 4, April 2011, pp. 125-132, doi:10.1145/1924421.1924449.

[11] A. Longley, F. Goodchild, J. Maguire and J. Rhind. Geographical Information Systems and Science. 2nd Edition. John Wiley & Sons, 2001.

[12] P. Rigaux, M. Scholl and A. Voisard. Spatial Databases with Application to GIS. Elsevier Science, 2002.

[13] A. Lbath and F. Pinet. "The Development and Customization of Gis-Based Applications and web-based GIS Applications with the CASE Tool Aigle". Proc. 8th ACM international symposium on Advances in geographic information systems, 2000, pp. 194-196, doi: 10.1145/355274.355307.

[14] S. Zongyao and X. Yichun. "Design of Service-Oriented Architecture for Spatial Data Integration and Its Application in Building Web-based GIS Systems". Geo-spatial Information Science. vol. 3, n. 1, 2010, pp. 8-15, doi:10.1007/s11806-010-0163-7.

[15] P. Fu and J. Sun. Web GIS: Principles and Applications. ESRI Press 2010.

[16] R. Wolfgang. "Principles and Application of Geographic Information Systems and Internet/Intranet Technology". Proc. New Information Processing Techniques for Military Systemns, pp. 1-10, 2000.

[17] J. Baumann. "Future of Web GIS: An Interview with Pinde Fu". GeoConnection International Magazine, April 2011.

[18] X. Lu. "An Investigation on Service-Oriented Architecture for Constructing Distributed Web GIS Application". Proc. IEEE International Conference on Services Computing, pp. 191-197, 2005, doi:10.1109/SCC.2005.27.

[19] H. Shunfu. "Web-Based Multimedia GIS for the analysis and visualization of spatial environmental database". Proc. Symposium on Geospatial Theory, Processing and Applications, 2002.

[20] P. Tigaux, M. Scholl and A. Voisard. Spatial databses: with application to GIS. Editora Morgan Kauffman. 2002.

[21] M. Frehner and M. Brandli. "Virtual database: Spatial analysis in a Web-based data management system for distributed ecological data". Environmental Modelling & Software, vol. 21, 2006, pp. 1544-1554, doi:10.1016/j.envsoft.2006.05.012.

[22] K. Page, D. Roure and K. Martinez. "REST and Linked Data: a match made for domain driven development?". Proc. I Second International Workshop on RESTful, 2011, doi:10.1145/1967428.1967435.

D-WCPS: A Framework for Service Based Distributed Processing of Coverages

Michael Owonibi, Peter Baumann
Center for Advanced Systems Engineering (CASE)
Jacobs University
Bremen, Germany
{m.owonibi,p.baumann}@jacobs-university.de

***Abstract*— Distributed, service-oriented systems is often used today for geospatial data access and processing. However, it is difficult to find methods for easy, flexible, and automatic composition and orchestration of workflow of geo-services. More promising is the Open Geospatial Consortium (OGC) Web Coverage Processing Service (WCPS), which offers a multidimensional raster processing query language with formal semantics; we believe that this language contains sufficient information for an automatic orchestration. Based on this, we present the D-WCPS (Distributed WCPS) – a framework in which coverage processing workflow can be dynamically distributed among several WCPS servers. Every server can schedule and execute a query using information from its local WCPS registry. Each local registry, in turn, is mirrored across all servers. Some other contributions of this paper include query optimization algorithms, tuple-based parallelism, registry synchronization techniques. Several servers can, therefore, efficiently share data and computation with respect to dynamic, resource-aware coverages processing.**

Keywords - geoprocessing, query processing, distributed query processing, service registry, query optimization

I. INTRODUCTION

Current trend specifies the use of distributed, service-oriented systems for geospatial data access and processing. Some of the motivations for this include

- Availability high-speed networks.
- Computation intensiveness of either a single or workflow of geoprocessing tasks such that distributed processing pays off.
- Server limitations in terms of processing capability and applications installed.
- Distribution of dataset across several data centers.
- Increase in the geo-application requirements, which vary from simple 2-D and 3-D map visualization and download, to complex computation such as statistical analysis, data mining, image classification and ocean, atmosphere, and climate modeling.

Geo-services are usually, standardized by the Open Geospatial Consortium (OGC) and some of the standardized services include the OGC Web Coverage Service (WCS) for coverage data access [5]**;** the OGC Web Coverage Service-Transactional (WCS-T) used for adding, updating and deleting coverages on a server [4]**;** the OGC Web Processing Service (WPS) which defines a generic service that offers any sort of geoprocessing functionality over a network[25]**.** It turns out very difficult at least to find methods for a flexible, automatic orchestration of geo-services. Hence, geo-service orchestrations are typically, performed manually, such as in the case of cascading OGC WPS and Web Map Service (WMS) requests, and process-based compositions e.g., BPEL which hardwires the processing configuration. We claim that this is due to the lack of an explicit, machine-understandable semantics of these services. The Web Coverage Processing Service (WCPS) [23], however, accomplishes interoperability by defining a language for server-side processing of multi-dimensional spatial-temporal data. This language has a formal semantics definition, hence is machine readable and semantic web ready. To this end, this paper addresses the efficient answering of queries on large, complex spatial-temporal data sets distributed across a number of heterogeneous computing nodes. The aim is that incoming query requests, expressed in WCPS, are automatically split into sub-requests which then are processed by suitable nodes in the network to collectively produce the final result for the client. Task distribution can be based, among others, on the individual node capabilities and availability, and load situation, network capabilities, and source data location. Sample use cases for distributed processing include site suitability studies or statistical analysis using data available from different servers, and climate reconstruction using climate modelling algorithms.

The rest of the paper is structured as follows: Section II introduces the WCPS; related work is presented in Section III; we describe the distributed WCPS framework in Section IV, implementation and performance evaluation is presented in Section V and we conclude the paper in Section VI.

II. WEB COVERAGE PROCESSING SERVICE

WCPS specifies the syntax and semantics of a query language (service request) which allows for server-side retrieval and processing of multi-dimensional geospatial coverages representing sensor, image, or statistics data [23]**.** The term “coverage”, in the general definition of OGC [20] and ISO [21], encompasses any spatial-temporally extended phenomenon. As currently overarching query languages in this generality are not sufficiently understood, WCPS focuses on raster data. The raster data is a gridded multi-dimensional array of some dimensionality, and some extent (spatial-temporal domain) where each grid cell value represent information. Sample raster data include 1-D sensor time series; 2-D satellite imagery; 3-D x/y/t image time series, and 4-D atmospheric model data. WCPS queries are given as expressions composed from coverage-centric primitives. The grouping of these primitives is shown below:

- Geometric operations extract some subset of cells which together again form a coverage. Trimming retrieves a sub-coverage whose dimensionality remains unchanged. Slicing delivers a cut-out with reduced dimensionality.
- Induced operations apply cell type operations to all cells in coverage simultaneously. This includes arithmetic, trigonometric and logical operations, etc
- Coverage summarization includes aggregation operations like count, average, min, max etc.
- All of the above functions actually represent convenience operators which can be reduced to a coverage constructor, an aggregator, or a combination thereof.
- Scaling and reprojection constitute non-atomic function.
- Data format encoding specifies how results are to be shipped back to the client. The list of such encodings includes formats like TIFF, NetCDF, or SEG-Y.

Shaped in the style of XQuery and SQL, the WCPS defines a declarative, set-oriented query language for a coverage processing workflow. The overall structure of WCPS is as follows:

```
for       $c_1$ in ($C_{1,1}$, $C_{1,2}$,…, $C_{1,m}$),
          $c_n$ in ($C_{2,1}$, $C_{2,2}$,…, $C_{2,m}$),
          …,
          $c_n$ in ($C_{n,1}$, $C_{n,2}$,…, $C_{n,m}$)
where     booleanExpr($c_1$, …, $c_n$)
return    processingExpr($c_1$, …, $c_n$)
```

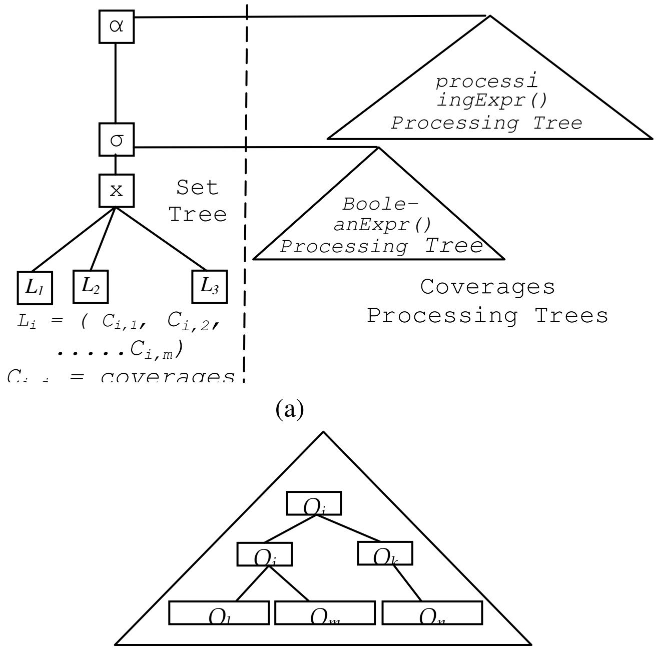
This can be seen as a nested loop where each c_i is bound to the $C_{i,j}$ coverages in turn. For each variable binding, the “where” predicate `booleanExpr()` is evaluated first. Only if the boolean expression evaluates to true will the `processingExpr()` will be evaluated on the current variable assignment and its result element will be added to the result list. We introduce WCPS by way of example.

Assume a WCPS server offers 3-D satellite image time series stacks, S1 and S2, plus a 2-D bitmask M with the same spatial extent as the time series cubes. Then, the following query returns those cubes where, in time slice T, threshold value V is exceeded in the red band somewhere within the mask area:

```
for       s in ( S1, S2),
          m in ( M )
where     some(s.red[t(T)]> m and m>0)
return    encode( s/max(x), "netcdf" )
```

The subsetting operation in square brackets specifies a cut along axis `t`. The aggregator expression `some()` conflates this to a single Boolean value. Those result cubes which pass this filter are shipped to client in NetCDF format.

The WCPS query processing model is based on adapted rasdaman query processing model [27]. It consists of a set processing tree, and coverages processing trees as sub-trees as shown in Figure 1(a). The set processing trees specifies the assignment of coverages to coverage iterator. The leafs of the set tree are the coverage lists. The set tree contain three relational operations – the cross (cartesian) product (*x*) of the different coverage lists, the selection (σ) of tuples from the resulting cross product table based on the predicate defined in the *“where”* clause, and the application (*α*) where the coverage processing expression is evaluated on the current iterators' coverages assignment. On the other hand, the coverage processing expressions, i.e., *booleanExpr()* and *processingExpr()*, are trees of coverage processing operators O_x as shown in the sample query tree in the Figure 1(b)


α
σ
x
Set Tree
processi ingExpr() Processing Tree
Boole-anExpr() Processing Tree
L_1
L_2
L_3
L_i = ($C_{i,1}$, $C_{i,2}$, ...$C_{i,m}$)
$C_{i,j}$ = coverages
Coverages Processing Trees
O_i
O_j
O_k
O_l
O_m
O_n

(a)

(b)

Figure 1: WCPS Query Tree

Due to the fact that the semantics of WCPS service request is known both to the client and servers, [22] opined that automatic service dispatching, chaining and optimizing is possible, as a WCPS server is able to automatically extract portions that are best resolved locally, distribute other parts to other suitable servers, re-collect results, package them, and return them without any human interference.

In this paper we described our distributed WCPS (D-WCPS) system wherein several servers can collaboratively and dynamically execute geo-processing tasks specified as WCPS query transparently. Servers can therefore share data, load and applications. We also describe the means of orchestrating the composed service efficiently in a fault-tolerant manner. Depending on the objective of the servers, different scheduling algorithms can be used for decomposing a query in such infrastructure. Hence, we do not specify the details of any particular scheduling algorithm in this paper.

III. RELATED WORKS

In classical distributed database systems (DBMS), grid computing systems, and Service Oriented Architecture

(SOA), a mediator-based method typically used for distributed query processing (DQP). In this approach [26], a mediator's registry stores and integrates the data sources schema, statistics and properties. It also contains the server properties of all the data servers**.** Global queries (queries which address more than one server) are directed to the mediator which parses and translates them into a query tree. Using information from the registry, the query tree is optimized logically and physically using both single and multi-node algorithms. The query tree operators are, later on, scheduled, and execution code is generated and run for the scheduled tree. Usually, the orchestration and integration of partial results from other servers are done in the mediator.

Some of the DBMS-based middleware using this approach include DISCO[3], Garlic [15], Hermes [28], TSIMMIS [12], Pegasus [18]. Also, several levels of support for a mediator-based data integration is provided by major database vendors such as IBM (using DB2 Propagator), Sybase (using Replication Server), Microsoft (SQL Server), and Oracle (using Data Access Gateways) [1][8]. Classical systems like R*[10] integrates data from several databases, while SDD-1[24], present mechanisms for distributed join processing algorithms on a homogenous set of servers.

Similarly, grid-based DQP typically consists of several mediators using one registry. Some of the generic grid-enabled query processors include Polar [13], OGSA-DQP[19], SkyQuery[30], CoDIMS-G [31]; and GridDB-Lite [29]. The main task of GridDB-Lite is the selection and transfer of distributed scientific data from storage cluster to compute clusters. Polar* is a distributed object oriented query processor which accesses remote data using remote operation calls. OGSA-DQP and CoDIMS-G are based on service oriented grid. OGSA-DQP extends the concepts in POLAR by automatically composing a static, per-client DQP service instance from a set of manually selected resources. CoDIMS-G, on the other hand, profiles services in order to select the resources to use, and adaptively reschedules query operators based on runtime conditions. SkyQuery provides an implementation of a mediator-based DQP for distributed astronomic data in a SOA. Its mediator dynamically tests for performance of the servers before scheduling its query.

However, with respect to distributed coverage processing, we note the following

- In DBMS, grid and SOA-based DQP, many of the scheduling algorithms, and parallelization, optimization, and execution model deals with relational and xml databases as opposed to coverages. For instance, a typical DBMS-based DQP will focus on scheduling relational join operators on servers based on different join algorithms. However, in D-WCPS, scheduling the costly coverage processing operators takes precedence over comparatively less expensive join operators.
- The centralized mediator used for the registry management and query execution constitutes a performance bottleneck and single point of failure.
- Concepts and framework of D-WCPS is based on SOA while DBMS-based DQP is typically not.
- Because the grid deals with heavy weight, long running scientific applications, the costs of the grid-based DQP scheduling and execution overhead is relatively small compared to the query execution cost. However, WCPS is typically, a medium weight application whose running time varies between milliseconds to minutes. Hence, cost of overhead of grid based methods in D-WCPS is significantly large and inefficient.

A component common to all distributed systems is the registry. These can go by different names such as Universal Description Discovery and Integration in SOA, OGC Catalog Service for the Web in OGC web services, federated database catalog in distributed databases, Monitoring & Discovery System in Globus Grid etc. To ease management overhead, registries are usually centrally located. However, besides constituting a performance bottleneck and single point of failure, this architecture is not usually efficient. Another proposal uses decentralized Peer to Peer registry based on distributed hash table [6]. Although this system scales up and is resilient to failure, they response times for a query is usually large. Also, its registry's management is complex, and it does not efficiently support range queries. Similarly, some other registries are based on meta-directory architecture whereby a node stores meta-information about the distributed registries [2]. This has the advantage of scaling, however, performance is still an issue and it is prone to single point of failure. Hence, we proposed the use of mirrored registry for D-WCPS. This has the advantage of being highly efficient in its query processing because it is available locally on every server. However, synchronization of all the servers for transaction based queries (updates, insert, delete) can be costly. As it is not expected that the rate of transaction in D-WCPS is going to be high, this cost will not be significant. Besides, since the database is mirrored on all servers, the registry is more resilient to failure, and the central server's performance bottleneck is removed.

Overall, our work focus on processing coverages and more emphasis is laid on optimizing, scheduling and executing coverage processing operators rather than relational operators. We also present an orchestration model which is based on recursive nesting of queries. Furthermore, we specify the use of inter-tuple parallelism in query execution. Lastly, we use an architecture where every participating server can serve as a mediator.

IV. D-WCPS Framework

On a high level, we present D-WCPS - a framework where several WCPS servers can share the computation of a service request. Every server in this framework has a local copy of the global WCPS registry of data and services. Information from this registry is used to decompose a global query request into a distributed query request. As shown in Figure 2, every server runs WCPS and registry services, and these in turn, are made up of several components. The procedure of composition and execution of distributed WCPS is adapted from distributed DBMS-query processing. After the global query is received by any of the servers:

- The parser transforms the query to a query tree.

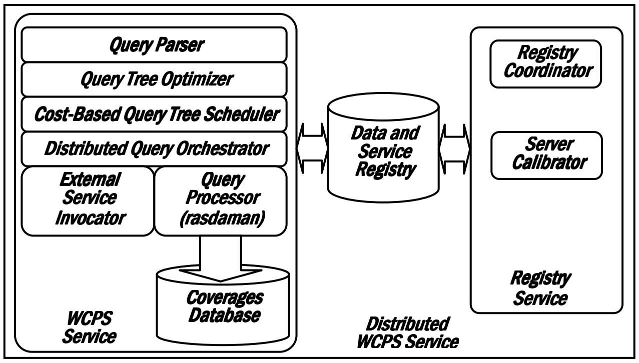


Figure 2: Components of a D-WCPS Server.

- The coverages metadata and location, and the servers to use for the distributed processing together with their capacities (e.g., CPU speed) and restrictions (e.g., maximum memory available) are then determined by querying the local registry.
- Based on the query structure, data sizes and location, the coverage processing query tree is optimized for distributed execution by rearranging its operators using different sets of equivalence rules.
- Using the server capabilities and initial data locations, the query is decomposed to a distributed query such that fulfils an objective function. Several objective functions exists and these include minimizing execution time, total data transferred, and total time spent on all servers; maximizing throughput; and load balancing on different servers. Decomposition of the query tree involves the scheduling of the query tree operators on different servers. This is an NP-complete problem, hence, the use of heuristic-based algorithms [26].
- After the scheduling, the global query is re-written into a distributed query (query with scheduling information) which is then executed.

A. *Inter-Tuple Parallelism in D-WCPS*

From Section II, the WCPS query tree consists of a set tree and coverage processing trees. The set tree creates a table of cartesian product of the coverage lists bounded to each coverage iterator in the query. Then, the predicate expression and/or processing expression are executed for each tuple in the table. Compared to the coverage processing operation, the cartesian product operation is cheap, hence, it is done on the server which receives the query. However, the execution of the coverage processing tree is distributed. Moreover, since each tuple in the cross product table can be processed independently of others, we parallelise the processing of each tuple in the table i.e., the optimization, scheduling and execution of a query are parallelized on a tuple-by-tuple basis. So, given the sample query below

```
For   a in (X, Y),
      b in ( Z )
where max(a) < avg(b)
return encode(cos(a)+min(b),"hdf")
```

Table 1: Cartesian Product Table of WCPS Query.

Tuples	Iterator a	Iterator b	Query Materialization
1	X	Z	Predicate : max(X) < avg (Z)
			Processing Expression: Encode(cos(X) + min(Z), “hdf”)
2	Y		Predicate : max(Y) < avg(Z)
			Processing Expression: Encode (cos(Y) + min(Z), “hdf”)

A cartesian product in Table 1 is created whose processing is then parallelized tuple-wise.

B. *Optimization*

In order to increase the efficiency of the execution of the D-WCPS, the global query is optimized by re-writing of the query based on equivalence rules. By query re-writing, we mean the re-arrangement of the ordering of operators of a query tree. This is done using a two-staged approach – applying single-node optimization before a multi-node optimization. We apply the optimizations on the coverage processing tree, hence, any reference to operator from henceforth implies coverage processing operator. The single node optimization assumes the query is executed on a server and is based on rasdaman query re-writing rules. The rationale and proof of the optimizations can be obtained from [7]. Overall, the idea is to minimize the size of the data processed by an operator. This is because, smaller input data for an operator implies

- Less processing work would be done by the operator
- Reduced data transfer time between an operator and its operand.

Some of the single node optimization includes pushing down of domain reducing, and aggregation expression down the query tree; using the associative and distributive properties of expression to re-write queries such that data transferred is minimized. For instance, assuming O_G represents coverage subsetting operation and O_C represents other coverage processing operations. By pushing down of the geometric operation as shown in Figure 3(a), the cost of executing O_C, and transferring data between O_C and O_G will be smaller. Similarly, in Fig 3(b), if operators *a, and b* are associative, their operands can be re-arranged such that data processed at and transferred from operator *b* is minimized.

The aim of multi-node optimization is to prepare a query

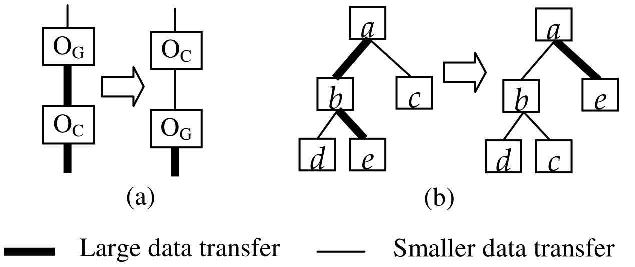
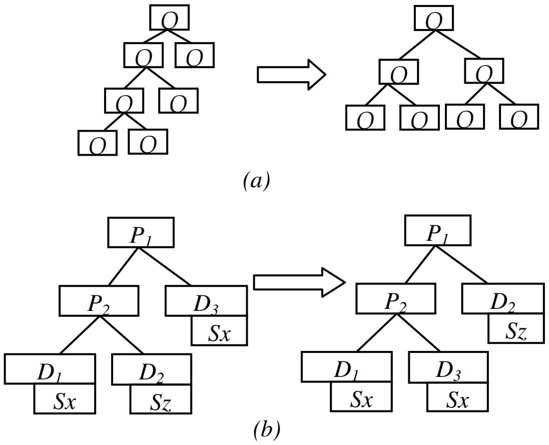


Figure 3: Single Node Query Optimization.



(a)

(b)

Figure 4: Multi-node Optimization.

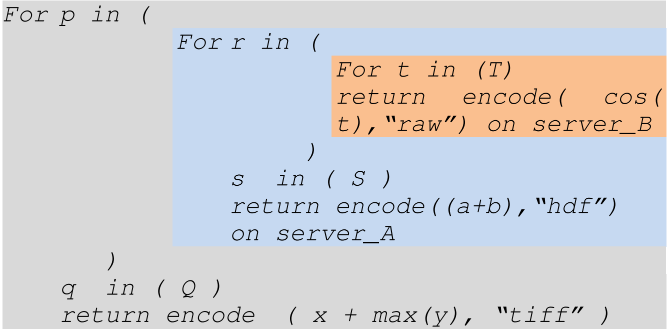
tree for distributed execution. And one of the optimizations we introduce here aim at transforming a left deep tree to a bushy tree (Figure 4(a)) using the associative and distributed properties of the operators in the tree. This is because left deep tree would have a high through time; irrespective of scheduling algorithm used. Besides, more operators can be executed in parallel in bushy tree. Another optimization we used here is to use the bring operators which have data on the same host as close together as they can possibly get on the tree. This ensures that data will not be transferred just for integration purposes when there is no performance gain. As an example, consider the tree in Figure 4(b), where D_1, and D_3 are located on server *Sx* and D_2 on server *Sz*. If operators P_1 and P_2 are associative, operand D_2 will be interchanged with D_3 in order to ensure D_1 and D_2 are close to each other.

C. Scheduling

After the optimization of the query tree, the operators of the coverage processing trees are scheduled. Several algorithms exist for scheduling DAGS [32][9]. The choice of algorithm by a server is specified by used by the objective the server wants to fulfill. Different scheduling algorithms can therefore be used in D-WCPS. However, it has been shown that Heterogeneous Earliest Finish Time (HEFT) algorithm [11] is one of the best algorithms [9] in systems consisting of several heterogeneous servers.

D. Distributed Query Modeling and Its Orchestration

A P2P orchestration model, whereby WCPS servers recursively invoke other servers with a distributed query request, is used for executing D-WCPS. After a server receives a distributed query request, it invokes other servers with partial query requests as specified in the query, executes its local query requests and integrates the results. The integration may involve writing a temporary copy of the data due to the fact that partial query results can be larger than the main memory. The distributed query request is composed by the server which receives the initial global query. Therefore, other servers used for executing the distributed query need not run the scheduler again except if there is a change in the conditions in the network, and the query needs to be adapted.



```
For p in (
        For r in (
                For t in (T)
                return  encode(  cos(
                t),"raw") on server_B
              )
          s  in ( S )
          return encode((a+b),"hdf")
          on server_A
     )
  q  in ( Q )
  return encode  ( x + max(y), "tiff" )
```

Figure 5 : Nested Distributed WCPS Query.

The WCPS query syntax is modified to support such distributed execution [17]. In the introduced modifications, a coverage iterator will not only bind to coverages, but can as well bind to partial queries with a specified execution host. For example, Figure 5 shows a distributed WCPS query with different levels subquery nesting. The server that receives the query invokes *server_A* with its inner subquery, and *server_A* in turn invokes *server_B* with its inner subquery.

The quality of D-WCPS schedule generated initially can deteriorate if conditions in the network change during the execution its execution. Two major runtime disruption addressed in D-WCPS are server overload, and server unavailability. In the case that an overloaded server receives a partial query, the overloaded server reschedules the query for other servers. Similarly, if a server that is critical (it has some data or operations which is not available on other servers) to the query is unavailable, the query execution is terminated, otherwise, the partial query is rescheduled by the server which invokes the request on the unavailable server.

E. The Registry

The architecture of the registry adopted for a framework depends on the trade-offs between requirements such as efficiency, scalability, simplicity, availability, fault-tolerance, ease of management, flexibility, allowance for redundancy, rate of update, support for range or singleton query, ability to easily classify the registry entries. In D-WCPS, emphasis is placed on efficiency, fault-tolerance, and easy configuration. In addition, we envisage a system with a slow rate of update and which can scales to thousands in terms of services registered, and each server, in turn, can hold thousands of coverages. In this respect, we propose mirrored registry architecture. Because each server has a local copy of the registry, querying it for query decomposition information is very efficient compared to if the registry were to be external service. Furthermore, each local registry in the network is kept in sync with the others,

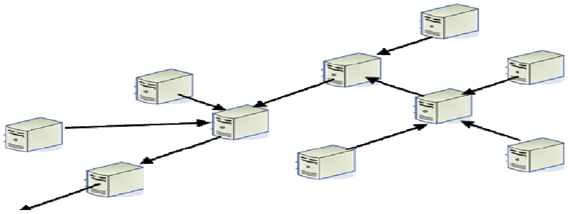


Figure 6: Mirrored Registry Synchronization.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-178-6

and updates are only made when a new server joins the network, a new data is added to /deleted from a server, a server's properties change, or when the properties of some data saved is changed. Therefore, the challenge in the networks is keeping servers in sync with minimum effort.

The arrangement of the servers in the mirrored registry network is based a hierarchical topology for. Each server can only join the network from one server i.e., a server has only one gateway but it can have several backup gateways. And several servers can join the network through a server. In this paper, all the servers that share a gateway are child servers of their gateway server. A server can only receive a message from either its gateway or any of its child servers. When a server receives or generates a message, it propagates it to all the servers connected to it (child servers and gateway) except the source of the message. In this way, messages get sent to all the servers without looping as a broadcast will do. Furthermore, every server that intends to join the network can register with any server in the network. In order not to overload a server, a server has the maximum number of child servers it can have, and if any intending server wants to register with a server and its maximum number of child servers has been reached, it forwards the registration information to its child servers in a round-robin fashion. In case the gateway of a server becomes unavailable, the server can re-register with any other gateway server.

Three interfaces are exposed by the registry service namely register, update, and info interface. The registration interface is used for registering and de-registering servers. The update is used to insert, delete or update information about servers or data, and the info interface is used for management e.g., receiving and sending keep-alive messages, informing other peer servers that their gateway server is dead.

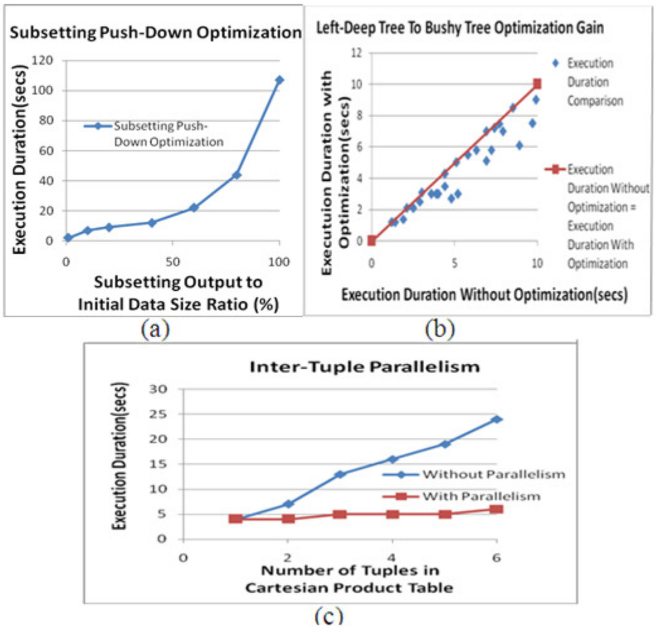
F. Calibrating and Profiling

Every server has a calibration engine which is used to measure its capacities for coverages processing. Information gathered by the engine are published into the service registry, and these include the read, write, and coverage processing speed and overhead; available memory; number of simultaneous processes that can run without degrading the servers performance; set of preferred servers (the set of servers a server will prefer to use in for distributed processing), and the network speed to these servers; and lastly, the average network speed to all servers. We obtain the read and write speed and overhead by writing some test data to and reading some test data from the database. The I/O speed and overhead depend on type of disk systems used (e.g., whether it is RAID system, virtual disk system etc), their speed and configurations, file/database system used and their configurations, etc. Similarly, [14][7] opines that the quoted speed of systems in terms of MIPS, FLOPS, QUIPS, clock rate cannot be used to determine the performance of a system due to factors such as inter-processor communication, I/O performance, cache coherence and speed, memory hierarchy etc. Therefore, we define the speed of processing of a system with regards to coverage processing as the speed it takes a system to do a copy of a char-typed pixel from one memory location to another. The calibration engine obtains this value by running a set of standardized query against the database. Furthermore, due to fact that systems nowadays have several physical and/or virtual processors, systems can run several queries simultaneously, without any significant degradation of its performance. Hence, we determine the maximum number of queries a system can run without performance loss.

V. IMPLEMENTATION AND EVALUATION

An experimental D-WCPS infrastructure was set up consisting of 30 WCPS servers which are heterogeneous with respect to processing, read, write and network speed. A server was chosen as the initial D-WCPS server with which other servers have to register before they can start registering other servers. For this setup, we choose minimization of execution duration as our objective. Therefore, a modified form of HEFT algorithm [11] based on a coverage processing cost model [16] was used to schedule the operators of any query received by any of the servers. Using query trees with different types of structural properties, operators, and initial data distribution, we evaluate the performance of our framework with respect to some of the query processing and optimization techniques.

In Figure 7(a), we highlight the gains of pushing down subsetting operators in a distributed processing system, given the percentage of the total initial data retrieved by the subsetting operator. Due to the large processing and inter-server data transfer costs, the smaller the percentage gets the smaller the execution duration. Similarly, Figure 7(b) compares the query execution times when a left deep tree is transformed to bushy tree and when it is not. In some of the cases, the execution duration does not change, but in many others, it is reduced. The performance gain of inter-tuple parallelism is also shown in Figure 7(c).



(a) (b)

(c)

Figure 7: D-WCPS Performance Evaluation Graphs

VI. CONCLUSION

The OGC WCPS offers a multidimensional raster processing query language with a formal semantics which contain sufficient information for automatic orchestration. Based on this, we present an infrastructure for distributed execution of WCPS query by dynamically composing services from a query request. Every server in the network has a local copy of the WCPS registry, and servers synchronize the updates to the registry with each other. Using the information from the registry, and tuple-wise based parallelization, servers can optimize, decompose and execute a received query. Finally, we present the model for a decentralized orchestration of the distributed WCPS query. Several servers can, therefore, efficiently collaborate in sharing data, computation, loads and other tasks with respect to dynamic, resource-aware coverages processing.

ACKNOWLEDGMENT

Many thanks to the Earth System Science Research School (http://earth-system-science.org/), and the European Union funded “EarthServer: Ad-hoc Analytics on Big Data” project (http://earthserver.eu/) for supporting this research.

REFERENCES

[1] A. Gounaris(2005). Resource aware query processing on the grid. Thesis report, University of Manchester, Faculty of Engineering and Physical Sciences.

[2] A. Kassim, B. Esfandiari, S. Majumdar, and L. Serghi, A flexible hybrid architecture for management of distributed web service registries, in Communication Networks and Services Research (CNSR), vol. 5, 2007

[3] A. Tomasic, L. Rashid, and P. Valduriez, Scaling heterogeneous database and design of DISCO, in Proceedings of the 16th International Conference on Distributing Computing Systems (ICDCS), Hong Kong, May 1996.

[4] A. Whiteside (ed), Web Coverage Service (WCS) —Transaction operation extension OGC 07-068r4 Version: 1.1.4 , 2009.

[5] A. Whiteside, and Evans J.D. 2008. Web Coverage Service (WCS) Implementation Standard, 07-067r5.

[6] C. Schmidt, and M. Parashar, “A Peer-to-Peer Approach to Web Service Discovery,” World Wide Web Journal, Vol. 7, No. 2, June 2004, pp. 211-229.

[7] D. J.Lilja. 2000. Measuring computer performance: a practitioner's guide, Cambridge University Press, New York.

[8] D. Kossmann. 2000. The State of the Art in Distributed Query Processing. ACM Comput. Surv., 32(4):422–469, December 2000.

[9] F. Dong, and S.K. Akl, Scheduling algorithms for grid computing: State of the art and open problems, Technical Report No. 2006-504, School of Computing, Queen’s University, Kingston, Ontario, Canada, January 2006.

[10] F. Mackert, and M. Lohman, R* Optimizer Validation and Performance Evaluation for Distributed Queries, Proceedings of the 12th International Conference on Very Large Data Bases, p.149-159, August 25-28, 1986.

[11] H Topcuoglu, S Hariri and M.-Y Wu, Task scheduling algorithms for heterogeneous processors, 8th IEEE Heterogeneous Computing Workshop (HCW '99) (1999), p. 3–14.

[12] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J.D. Ullman, V. Vassalos, and J. Widom, The TSIMMIS approach to mediation: Data models and languages, Journal of Intelligent Information Systems, vol. 8, no. 2, 1997.

[13] J. Smith, A. Gounaris, P. Watson, N. Paton, A Fernandes, and R. Sakellariou. Distributed query processing on the grid. International Journal of High Performance Computing Applications, 17(4), 2003.

[14] J.L. Gustafson, and Q. O. Snell, HINT: A New Way To Measure Computer Performance, Proceedings of the Twenty- Eighth Hawaii International Conference on System Sciences HICSS-95.

[15] M. J. Carey , L. M. Haas , P. M. Schwarz , M. Arya , W. F. Cody , R. Fagin , M. Flickner , A. W. Luniewski , W. Niblack , D. Petkovic , J. Thomas , J. H. Williams , and E. L. Wimmers, Towards heterogeneous multimedia information systems: the Garlic approach, Proceedings of the 5th International Workshop on Research Issues in Data Engineering-Distributed Object Management (RIDE-DOM'95), p.124, March 06-07, 1995.

[16] M. Owonibi and P. Baumann, A cost model for distributed coverage processing services. In Proceedings of the ACM SIGSPATIAL International Workshop on High Performance and Distributed Geographic Information Systems 2010. ACM, New York, NY, USA.

[17] M. Owonibi, and P. Baumann, 2010: Heuristic Geo Query Decomposition and Orchestration in a SOA. InProceedings of the 12th International Conference on Information Integration and Web-based Applications & Services (iiWAS '10).

[18] M.C. Shan, Pegasus architecture and design principles, in Proceedings of the ACMSIGMOD International Conference on Management of Data, Washington, DC, USA, June 1993.

[19] M.N. Alpdemir, A. Mukherjee, A. Gounaris, N.W. Paton, P. Watson, and A.A.A Fernandes, OGSA-DQP: A Grid Service for Distributed Querying on the Grid. EDBT 2004. LNCS, vol. 2992, pp. 858–861. Springer, Heidelberg (2004).

[20] N.n. 2007 Abstract Specification Topic 6: Schema for coverage geometry and functions. OGC 07-011.

[21] N.n., 2008 Geographic Information - Coverage Geometry and Functions. ISO 19123:2005.

[22] P. Baumann and S. Keens, (2007). OWS-4 Workflow IPR Workflow descriptions and lessons learned OGC 06-187r1 version0.0.9. OGC Discussion Paper, OGC.

[23] P. Baumann, (ed.),2008. Web Coverage Processing Service (WCPS) Language Interface Standard. OGC 08-068r2.

[24] P. Bernstein , N. Goodman , E. Wong , C Reeve , and J Rothnie, Jr., Query processing in a system for distributed databases (SDD-1), ACM Transactions on Database Systems (TODS), v.6 n.4, p.602-625, Dec. 1981 [doi>10.1145/319628.319650].

[25] P. Schut (ed.), 2007. OpenGIS Web Processing Service. OGC 05-007r7 version 1.0.0.

[26] R. Ramakrishnan, and J. Gehrke. 2007. Database Management Systems (Third Edition), McGraw Hill, 2007.

[27] R. Ritsch, "Optimization and Evaluation of Array Queries in Database Management Systems". PhD Thesis, TU Muenchen, 1999.

[28] S. Adali, K.S. Candan, Y. Papakonstantinou, and V.S. Subrahmanian, “Query caching and optimization in distributed mediator systems,” in Proceeedings of ACM SIGMOD International Conference on Management of Data, Montreal, Canada, June 1996.

[29] S. Narayanan, T. Kurc, U. Catalyurek, and J. Saltz, Database support for data-driven scientific applications in the grid. Parallel Processing Letters. v13 i2. 245-271.

[30] T. Malik, A Szalay, T. Budavari, and A. Thakar. SkyQuery: A web service approach to federate databases. In CIDR,2003.

[31] V. Fontes , B. Schulze , M. Dutra , F. Porto , and A. Barbosa, CoDIMS-G: a data and program integration service for the grid, Proceedings of the 2nd workshop on Middleware for grid computing, p.29-34, October 18-22, 2004, Toronto, Ontario, Canada [doi>10.1145/1028493.1028498].

[32] Y. Kwok and I. Ahmad, Static scheduling algorithms for allocating directed task graphs to multiprocessors, ACM Computing Surveys (CSUR), v.31 n.4, p.406-471, Dec. 1999 [doi>10.1145/344588.344618]

Modeling and Querying Mobile Location Sensor Data

Iulian Sandu Popa
PRISM Laboratory, University of Versailles
45, avenue des Etats-Unis
78035 Versailles, France
Iulian.Sandu-Popa@prism.uvsq.fr

Karine Zeitouni
PRISM Laboratory, University of Versailles
45, avenue des Etats-Unis
78035 Versailles, France
Karine.Zeitouni@prism.uvsq.fr

***Abstract*—Moving objects databases are an important research topic in recent years. A lot of work dealt with modeling, querying and indexing objects that move freely or in networks. However, a moving object – such as a vehicle - could report some measures related to its state or to its environment, which are sensed throughout his movement. Managing such data is of major interest for some applications such as analyzing driving behavior or reconstructing the circum-stances of an accident in road safety, or identifying, by means of a vibration sensor, the defects along a railway in maintenance. However, this management is not covered by the existing approaches. In this paper, we propose a new data model and a language to handle mobile location sensor data. To this end, we introduce the concept of spatial profile of a measure to capture the measure variability in space, along with specific operations that permit to analyze the data. We also describe their implementation using object-relational paradigm.**

Keywords-spatiotemporal databases; modeling; moving objects; query language; sensor data flows

I. Introduction

Integrating mobile technology and positioning devices has led to producing large amounts of moving object data every day. A wide range of applications like traffic management, location-based services (LBS), relies on these data. Besides, a moving object (MO) can easily be equipped with sensor devices that report on its state or on its environment. The generated mobile sensor data are meaningful for many applications such as reconstructing the circumstances of an accident in road safety, identifying defects from vibration sensors along a railway in maintenance, or analyzing the exposure to hazard (e.g., pollutant) along a trip. As an example, in the field of road safety, the observation of natural driving behavior (on normal route for usual journeys) tends to replace the tests on simulators or those limited to dedicated circuits. Known as "naturalistic driving", these studies are based on data collected on a large scale and over a significant period of time [12].

However, studies reported in the literature have limited the volume of data and the possibilities of their exploitation. As emphasized in a report [12] on a naturalistic driving campaign by the administration of U.S. Highway Safety, a large-scale database would be very useful to researchers and engineers to study the driving behavior and contribute to improving the vehicle equipment and road planning. The challenge of a large-scale study is the management of a large mass of spatio-temporal data. A database system that supports this type of data and efficient querying is needed. We aim to study and develop such a database management system. This subject is closely related to the field of moving objects databases (MOD). However, the moving object, e.g., a vehicle, is associated with additional measures (speed, acceleration, steering wheel angle, etc.) recorded throughout his trip. These measures are variable in space and in time.

For the type of applications we address, the measures are more important than the mere spatio-temporal location. However, most work on mobile object databases consider only the location of the moving object and cannot be generalized to measures ranging along a spatio-temporal trajectory. Moreover, although these values initially correspond to a temporal data stream, their variation is more dependent on their location in the network than time. For example, the variation of speed is usually constrained by the geometry of the road and the speed limit. Also, the temporal analysis of different trajectory data is irrelevant because on one hand they are asynchronous and on the other hand, this comparison makes sense only if these paths overlap in space. Therefore, we must capture the spatial variability of these measures and allow its manipulation through the data model and the query language.

To our knowledge, there is no such proposal in the related work. Nevertheless, among the works on MOD, the one proposed in [7] provides a solid basis for modeling and querying MOs. The idea of representing the temporal variation of the location or scalar values in a continuous way permits a good abstraction of moving objects. We extend this approach to capture the continuous spatial variation of scalars. The extension of the existing algebra consists in a new set of types and several classes of operations. The types capture the variation in space of any measure, which includes mobile sensor data as a particular case. New operations are needed to operate on the measures.

This paper provides the following contributions. First, we present a new concept of "space variant measure” and show its usefulness in the context of a naturalistic driving study. Second, we create a data model as an extension of the model proposed by Güting et al. [6]. Finally, we extend the existing query language with new classes of operations that are necessary in this novel application context. Besides the aforementioned application, the proposed model is interesting for other applications that generate and/or operate geo-localized data streams. This is the case of rail, air or sea transportation. The measures can be observed or calculated and can be related to a trajectory of an object or a location. Thus, the proposed model permits to reason about the speed of a MO, on the legal speed or inclination of a road or on the adherence at each location of the road depending on the weather. Also, it allows to model the fine data on the mobility (where, when and at what speed) of vehicles, freight, or persons and it meets the needs of management applications for fleets or road traffic, logistics and design of mobile networks.

The rest of this paper is organized as follows: we summarize the related work in Section II. Section III describes the proposed model. It presents the new types and the corresponding operations, and demonstrates its usefulness by expressing some query examples. Section IV discusses several aspects of the implementation. Finally, Section V concludes and offers directions for future work.

II. Related Work

The management of MOD has received particular attention in the recent years due to the advances and the omnipresence of mobile and geo-location technologies, such as cellular phones or GPS. Many works focus on modeling and language. We mention the work undertaken in the project Chorochronos [7][11] and the approach of the Wolfson's team [21]. Güting's book is a summary of progresses in this area [8]. Pelekis et al. summarize the data models for MOs in [17].

The target applications impact the model and the language in these proposals. We distinguish two types of applications. LBS applications rely on continuous or predictive queries, which are evaluated based on the current positions of MOs. The pioneers are [21] whose model MOST (Moving Objects Spatio-Temporal) describes databases with dynamic attributes that vary continuously over time. They also propose the so-called Future Temporal Logic to formulate predictive queries.

The second type of applications concerns the analysis of complete spatio-temporal trajectories, using queries combining temporal and spatial intervals. The work of Güting [7] is an important reference point today. Various implementations exist, as in SECONDO [6] and STAU [14], then in HERMES-MDC [15]. STAU is the first implementation to be based on object-relational database extensibility by providing a spatio-temporal data cartridge for Oracle [13].

However, these studies do not take into account the specific behavior of MOs, such as vehicles moving on a road network or trains on a rail network. This aspect is essential for many applications, including those considered here. Indeed, given a network, a constrained trajectory can be represented by the relative positions on the network edges (i.e., the road segments). Once more, the most comprehensive proposal is the one in [6]. Although the non-constrained (two-dimensional) model can be applied to the constrained trajectories, this is unwise for several reasons. The first is that the 2D model does not capture the relationship between the trajectory and the network space, while this information is essential for analysis. The second is that it limits the representation of the trajectory, estimated by linear interpolation between the reported positions, while the MO follows in fact the geometry of the network. In addition, the constrained model allows for dimensionality reduction by transforming the network in a 1D space by juxtaposing of all line segments [18]. This leads to better storage and query performance than with the free trajectory model. Finally, in the constrained model the trajectories can be easily described with a symbolic model as a sequence of traversed lines and time intervals, which is less detailed but more intelligible and more compact.

In this paper, we focus on managing historical data of objects moving in networks. The most comprehensive proposal to model the historical MO is, in our view, the framework of Güting [6]. Indeed, this proposal covers the abstract modeling, language and implementation. Moreover, it explicitly models the constrained MOs and the relative position on the network. As discussed below, our proposal is based on this model and extends it with specific data for mobile sensors. Therefore, we summarize this model and list the used notations in the rest of this section.

Güting et al. propose an algebra defined by a set of specific types (see Table 1) and a collection of operations on these types [6][7]. The types are: scalar types (*BASE*), 2D space types (*SPATIAL*), network space related types (*GRAPH*), scalar or spatial types varying in time (*TEMPORAL*). Examples of types are: *real*, *point* (2D position), *gpoint* (position on the network), *gline* (line on the network), *moving(point)* (2D position varying in time) and *moving(gpoint)* (network position varying in time). All the base types have the usual interpretation. For example, if we note with A_{α} the carrier set (definition domain) for the type α, then for the *real* type the carrier set is: $A_{real} = R \cup \{\perp\}$, where $\{\perp\}$ is null (or undefined). The time is isomorphic to the real numbers. The *range* data types are disjoint intervals and are used to make projections or selections on the *moving* types. Spatial types describe entities in the Euclidean space, while for the *GRAPH* types the space is represented by a network space. 2D types mainly correspond to standard definitions [9].

TABLE I. The types defined in [6][7]

Set of types	Type constructor
→ *BASE*	*int*, *real*, *string*, *bool*
→ *SPATIAL*	*point*, *points*, *line*, *region*
→ *GRAPH*	*gpoint*, *gline*
→ *TIME*	*instant*
BASE ∪ *SPATIAL* ∪ *GRAPH* → *TEMPORAL*	*moving*, *intime*
BASE ∪ *TIME* → *RANGE*	*range*

GRAPH types depend on the underlying network. Basically, the proposed model defines a network as a set of routes and junctions between these routes. A location in the network is a relative position on a route. It is described by the identifier of the route, a real number giving the relative position and the side of the road. This is directly related to the concept of linear referencing widely used in transportation applications and implemented in systems such as Oracle [13]. The types *gpoint* and *gline* are represented in this manner. Finally, from the *BASE*, *SPATIAL* and *GRAPH* types, they derive the corresponding temporal types, using the type constructor *moving*. The temporal types are functions or infinite sets of pairs (instant, value). Such an infinite representation conceivable in the abstract model cannot be implemented directly. In [6], a discrete representation is proposed for these types. We will discuss this aspect in more detail in Section IV of the paper.

A collection of operations is defined on the above data types. To avoid the proliferation of operations, one operator applies to several types. A set of non-temporal operations is first defined. Then, a process called *lifting* allows generating the corresponding temporal operations. Thus, all operations on non-temporal types are extended to the temporal types. Finally, specific operations are added to manage the temporal types. In the context of constrained network trajectories [6], some new operations, such as **distance**, have been adapted for *gpoint* and *gline* types (e.g., distance by route). New classes of operations are also added to analyze the interaction between the network and the 2D space, as well as specific operations such as computing shortest paths in the network. One can refer to [6][7] for more detail.

Besides, sensor data modeling was also considered from the angle of exchange formats [1]. This concerned static sensors. Recently, a draft has been initiated to exchange Moving Object Snapshots including velocity and acceleration parameters [16]. But, unlike SOS, it does not cover other measures. MauveDB [2] proposes model-based views in opposition to using raw data, in the context of environmental sensors. None of the previous work does capture the continuous variability in time and space of the moving sensor measures.

III. The Proposed Model

In this section, we present first a real application that has motivated our work (Section III-A). Then we introduce the new data types (Section III-B) and a collection of operations (Section III-C). A query scenario is used as an example throughout this paper (Section III-A and III-D).

A. *Motivation and Examples*

As indicated in the introduction, naturalistic driving studies have become popular in the last years. These studies are based essentially on data gathered in normal (natural) driving conditions. Such studies became economically possible thanks to the existing equipment in modern vehicles. Indeed, the large number of in-vehicle sensors is accessible via an interface (CAN bus) to which it is possible to connect an in-vehicle data logger. The CAN bus provides access to several measures including speed, acceleration, steering wheel angle, the action on the breaking or gas pedals, etc. The recording device can also receive data streams from other sources, such as a GPS sensor or radar (giving the distance to adjacent vehicles). This provides a comprehensive data source on natural driving on the road. The in-vehicle recorded data can provide valuable information on the use and utility of the driver assistance systems (ABS, ESP, etc.) and can highlight near-accident (near-crash) situations. Moreover, according to the principle of black boxes on airplanes, it will provide information prior to an accident.

INRETS (French acronym for “National Institute for Research on Transport and Safety”) has developed a data logger (DIRCO) for naturalistic driving campaigns [3]. This is an on-board recording device connected to the vehicle’s CAN bus. It records measures such as: vehicle speed, speed of each wheel, longitudinal acceleration, odometer, steering wheel angle, brake pedal (0/1), ABS (0/1), etc. DIRCO offers the possibility of connecting other data sources as well, e.g., a GPS, an inertial station measuring the 3D acceleration and angle of the vehicle. DIRCO acquires each data stream as a time sequence. The data from a source are stored in a specific file and each record is a tuple: $(t_i,\alpha_i^1,\alpha_i^2,\ldots,\alpha_i^n)$ where t_i is the i^{th} time instant and, α_i^k is the i^{th} value provided by the k^{th} sensor. As a detail, DIRCO allows sampling rates at very high frequencies of up to 10 ms cycles. The data flows from different sources are asynchronous.

While it may function as a black box for vehicles in order to reconstruct the circumstances of an accident, DIRCO is primarily a research tool that can help analyzing the driving behavior, the vehicle safety and diagnose problems related to road infrastructure. Its 16GB of flash memory allows data acquisition, camera off, for several months. A simple scenario is to equip several vehicles such as buses or cars with DIRCO, retrieve and centralize these data and then analyze it in order to identify behavioral patterns of driving.

This type of approach is also appropriate to the evaluation of recently emerged ADAS (Advanced Driver Assistance Systems). Whether the system is already well known as a GPS or speed control device, or it is an experimental system such as obstacle detection, all require an accurate and extensive assessment of their impact on driving. The European Commission is funding since 2008 large-scale projects to evaluate mature technologies in the category of "intelligent transportation" systems. A particular aspect of these projects is the recourse to systematic collection of driving data with devices similar to DIRCO e.g., the project euroFOT [22]. Given this kind of application, one can easily understand the importance of developing a database adapted to the characteristics of these data, such as the data volume or the geo-localized and temporal data features. The different types of studied systems induce a large variability in the methods of analysis and often involve a high level of required detail (e.g., situations of near-accident). Some indicators can be calculated by using common database management systems, but sometimes at the cost of heavy programming and prohibitive computational time. In addition, no system seems at present able to manage speed profiles (or any other information) measured at different times and positions but on the same road. However, a large number of queries in this context need this kind of approach. The concept of (spatial) profile is introduced in Section III-B

To illustrate the contribution of our model in this context, we refer throughout the paper to the following typical queries:

Q1. What is the acceleration profile along a given route segment for a given trip?

Q2. What is the difference between the vehicle's speed profile and the speed limit along a road segment?

Q3. How many times was the ABS enabled for a given trip?

Q4. What are the trips where the practiced speed exceeds a specified speed profile (e.g., the speed limit) by a certain value and what is the difference?

Q5. What is the ratio between speed and engine RPM for a given trip?

Q6. What is the average profile of acceleration for all vehicles passing through a certain road section (e.g., curve)?

Q7. Calculate the maximal speed profile of all vehicles passing through the indicated road section.

Q8. Find the practiced speed profile (85th percentile of the passing vehicles) on a road before and after the installation of a speed camera.

Q9. What is the average profile of the fuel consumption on a road before and after the installation of a traffic calming device (e.g., a speed cushion)?

Q10. What is the minimum and maximum profile of fuel consumption on a road, and what is its difference with the profile of the studied driver?

Modeling temporal sequences is feasible by using functions over time [7], but it is not useful for the above type of analysis. Indeed, the measures from the trips are collected at different times and comparing these profiles makes sense only if they were measured in the same place. What matters is not the time at which the measure was recorded, but rather where it took place on the road. The concept of spatial profile of a measure (e.g., speed, acceleration) reflects the relationship between the measure and the space. However, this notion of profile is not defined and cannot be derived in the model of Güting or any other model. It is therefore necessary to extend the existing model with new data types. Moreover, the above queries demand specific operations on the measure profiles. These operations, which were not necessary in the context of analyzing only the MO trajectories, are of major importance in this context.

B. *Introduction of New Data Types*

Like the algebraic model in [6] described above, our model includes a spatio-temporal type to model the trajectory of the MO, and temporal types to model the data generated by sensors. A temporal type is a function of time to base types (e.g., real, int). It expresses the variability of sensor measures from the temporal point of view.

However, the temporal view is not sufficient to model the *data from mobile sensors*, since the measures are often closely related to space. For completeness, the model should describe beside the evolution over time, the spatial evolution of the measures. To this end, we extend the model of [6]. We introduce a new concept describing the spatial profile of measures. The idea is to have a set of data types that allow modeling the evolution of a measure in space. This concept is divided into two categories: *SVARIANT* to describe the profile in a two-dimensional space, and *GVARIANT* for the profile along the network. *SVARIANT* (i.e., spatial variant) and *GVARIANT* (i.e., graph variant) represent two classes of data types.

We associate to these two classes of data types two new type constructors called smoving and gmoving (see Table 2). The type constructor smoving stands for *spatial moving* and allows modeling the evolution of a measure in the 2D space, whereas gmoving describes the evolution of a measure in a network (graph) space. The type constructors smoving and gmoving apply to *BASE* data types, i.e., int, real, string, bool. Hence, *SVARIANT* contains data types such as smoving(int), smoving(real), and, similarly, the class *GVARIANT* regroups data types such as gmoving(real), gmoving(bool), etc.

TABLE II. NEW DATA TYPES

Set of types	Type constructor
BASE → *SVARIANT*	smoving, inpoint
BASE → *GVARIANT*	gmoving, ingpoint

The definitions of these type constructors are given below using the notation of [7]:

Definition 1: Given α a *BASE* type having the carrier set A_α, then the domain of definition for *smoving*(α) is defined as follows: $A_{smoving(\alpha)} = \{ f \mid f : \overline{A}_{point} \to \overline{A}_\alpha$ is a partial function and $\Gamma(f)$ is finite$\}$, where $\overline{A}_\beta = A_\beta \setminus \{\perp\}$ and $\Gamma(f)$ denotes the set of maximal continuous components of the function f.

Definition 2: Given α a *BASE* type having the carrier set A_α, then the domain of definition for *gmoving*(α) is defined as follows: $A_{gmoving(\alpha)} = \{ f \mid f : \overline{A}_{gpoint} \to \overline{A}_\alpha$ is a partial function and $\Gamma(f)$ is finite$\}$, where $\overline{A}_\beta = A_\beta \setminus \{\perp\}$ and $\Gamma(f)$ denotes the set of maximal continuous components of the function f.

Since this paper focuses on constrained movement, we only detail the second category of types in the sequel. These definitions state that a spatial profile of a measure is a partial function. Each value f in the domain of $gmoving(\alpha)$ is a function describing the evolution in the network (graph) space of a *BASE* value. The gmoving type constructor describes an infinite set of pairs (*position, value*), where the position is a gpoint. The inpoint and ingpoint type constructors represent a single pair (*position, value*). Figure 1 presents a spatial profile of a real measure on a given road. The x-axis represents the relative position on the road that can vary between 0 and 1.

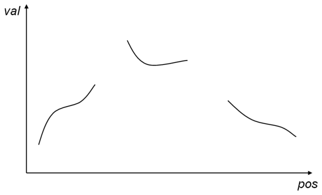


Figure 1. Example of spatial profile of a real value.

The condition "$\Gamma(f)$ is finite" means that f consists of only a finite number of continuous components. For example, the profile in Figure 1 has 3 continuous components. This condition is needed as a precondition to make the design implementable. It also ensures that projections of *gmoving* objects (e.g., on the spatial axis) have only a finite number of components.

The spatial and the temporal profile of a measure represent two complementary views of a measure varying in space and in time. The temporal profile is useful to compare data from different sensors (on the same vehicle) at the same time, e.g., Q5 in the query scenario in Section III-A. The spatial profile is useful to compare data from the same sensors on different vehicles at the same locations, e.g., Q6-Q10 in the query scenario in Section III-A.

Note that it is not practical to model the sensed values as a function on both time and space, since these two dimensions are not independent. Indeed, space is a function of time, which is captured in the spatio-temporal trajectory of the MO holding the sensors. At the same time, the spatial profile of a measure is necessary as motivated in Section III-A and by the query scenario, yet there are no data types in the existing data models [6][7] for such profiles.

Note also that the definition of spatial profiles imposes that for a given MO trajectory there is no overlapping between trajectory portions. This constraint is expected to hold in most cases. However, the self-overlapping trajectories have to be split into non-overlapping parts so that the associated sensor values fit the proposed model.

The presented model is an *abstract model*, which means that in general the domains or carrier sets of its data types are infinite sets. To be able to implement an abstract model, one must provide a corresponding *discrete model*, i.e., define finite representation for all the data types of the abstract model. This is done by the *sliced representation* introduced in [6]. Thus, a time dependent or spatial dependent value is represented as a sequence of slices (see Figure 4) such that within each slice the evolution of the value can be described by some "simple" function (e.g.,

TABLE III. EXAMPLES OF OPERATIONS FOR THE NEW DATA TYPES

Class	Operation	Signature
Projection to Domain/Range	**trajectory**	$gmoving\,(\alpha) \to gline$
	rangevalues	$gmoving\,(\alpha) \to range(\alpha)$
	pos	$ingpoint \to gpoint$
	val	$ingpoint \to \alpha$
Interaction with Domain/Range	**atpos**	$gmoving\,(\alpha) \times gpoint \to ingpoint$
	atgline	$gmoving\,(\alpha) \times gline \to gmoving\,(\alpha)$
	present	$gmoving\,(\alpha) \times gpoint \to bool$ $gmoving\,(\alpha) \times gline \to bool$
	at	$gmoving\,(\alpha) \times \alpha \to gmoving\,(\alpha)$ $gmoving\,(\alpha) \times range(\alpha) \to gmoving\,(\alpha)$
	atmin	$gmoving\,(\alpha) \to gmoving\,(\alpha)$
	atmax	$gmoving\,(\alpha) \to gmoving\,(\alpha)$
	passes	$gmoving\,(\alpha) \times \beta \to bool$
Basic Algebraic Operations	**sum, sub, mul, div**	$moving\,(\alpha) \times moving\,(\alpha) \to moving\,(\alpha)$ $gmoving\,(\alpha) \times gmoving\,(\alpha) \to gmoving\,(\alpha)$
Calculations	**mean[avg], min, max**	$moving\,(\alpha) \to real$ $gmoving\,(\alpha) \to real$
	no_transitions	$moving\,(int) \to int$ $gmoving\,(int) \to int$
Aggregates	**min_agg, max_agg, sum_agg, avg_agg**	$\{moving\,(\alpha)\} \to moving\,(\alpha)$ $\{gmoving\,(\alpha)\} \to gmoving\,(\alpha)$
	percentile	$\{moving\,(\alpha)\} \times real \to moving\,(\alpha)$ $\{gmoving\,(\alpha)\} \times real \to gmoving\,(\alpha)$
	count_agg	$\{moving\,(\alpha)\} \to moving\,(int)$ $\{gmoving\,(\alpha)\} \to gmoving\,(int)$

linear functions or quadratic polynomials). More details on the sliced representation are given in Section IV of the paper.

C. Introduction of New Operations

As for the type system definition, we use the operations in the algebra of Güting et al. [6] as a starting point. By introducing new types, we have to (i) extend the existing operations and (ii) add new specific operations for the target application type.

In order to extend the existing operations to the new types, we use a similar process with the *temporal lifting*, described in Section II. The temporal lifting permits generating from a non-temporal operation with the signature $\alpha_1 \times \alpha_2 \times ... \times \alpha_n \rightarrow \beta$, the temporal equivalent operation having the signature $\alpha_1^{'} \times \alpha_2^{'} \times ... \times \alpha_n^{'} \rightarrow moving(\beta)$ where $\alpha_i^{'} \in \{\alpha_i, moving(\alpha_i)\}$. Each of the arguments can become temporal, which makes the result temporal as well. We adopt this principle to generate the equivalent space variant operations. We propose a *spatial lifting* for the non-gvariant non-temporal operations. The operation induced by the spatial lifting is available for a signature $\alpha_1^{'} \times \alpha_2^{'} \times ... \times \alpha_n^{'} \rightarrow gmoving(\beta)$, where $\alpha_i^{'} \in \{\alpha_i, gmoving\ (\alpha_i)\}$.

We have also defined new operations that apply to *GVARIANT* and *TEMPORAL* set of types, which are necessary in this context. Table 3 presents a non-exhaustive list of the new operations, i.e., extended from the existing ones or newly introduced. We describe in this section some operations. Other operations are explained with the example queries in the next section. There are five classes of operations. The first two classes correspond to the extension of existing operations (i.e., spatial lifting), while the last three classes are new types of operations. Moreover, the first four groups represent conventional operations, i.e., those who take as input one or more objects (values) in accordance with their signature and return an object (a value). The last class includes aggregate operations, i.e., that return a single result based on a group of objects (similar to aggregates in the relational model).

The first class of operations comprises the projection in the network or value (range) domains. Thus, **trajectory** returns the network path of a trip. The operation **rangevalues** performs the projection in the range and returns one or several intervals of base values. Operations **val** and **pos** return respectively the value or the network position for an *ingpoint* type, which is defined as a pair (*gpoint*, *value*). The second class of operations concerns the interaction with the domain (network space) and range (values). They make selections or clippings according to criteria on one of the axes of variation (network space or values). Thus, **present** is a predicate that checks if the input object is defined at a given position in the network. Finally, the predicate **passes** allows one to check whether the moving value ever assumed one of the values given as a second argument.

The third class of operations considers the basic algebraic operations ('+', '-', '.' and '/'), which we include in non-gvariant non-temporal collection of operations. Therefore, they become subject to temporal and spatial lifting. We use named functions, i.e., **sum**, **sub**, **mul** and **div**, as for all defined operations. These operations are useful for the analysis of sensor measures. For example, they can calculate the difference between the speed profiles of two MOs on the common part of their trajectories, or return the difference between the practiced speed and the speed limit on a route. These operations take as input two functions of the same type (*GVARIANT* or *TEMPORAL*) and calculate a result function of which the definition domain is the intersection of the input objects' domains. For the division, the parts where the operation is not defined, are also eliminated from the domain of the result function.

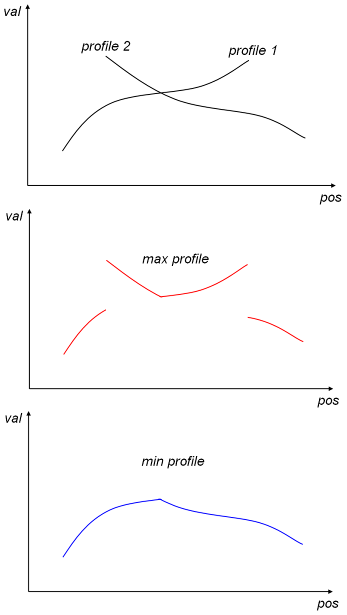


Figure 2. Example of using max_agg (second graph) and min_agg (third graph) on two profiles (first graph).

The fourth class of operation addresses the same categories of types, i.e., *GVARIANT* or *TEMPORAL*. The specified functions are: **mean**, **min**, **max** and **no_transitions**. Each of these operations takes as input a function of time or space and returns a value representing the aggregate of the input function. Their utility is to calculate an average or an extreme value for any measure, given a temporal or spatial interval.

The last class of operations concerns the aggregates. Aggregate operations return a single object result given a set of objects of the same type (see Figure 2). Unlike the previous class, these operations define aggregations of a group of objects. Some of these aggregates return an object of the same type as the input type, e.g., the average (**avg_agg**), the minimum (**min_agg**) and maximum profile (**max_agg**). The aggregate **count_agg** returns the number of profiles in the definition domain in the form of a *moving(int)* or *gmoving(int)* object. Finally, the function **percentile** computes the profile below which is found a certain percentage of profiles in the input set. The definition domain of the result function for an aggregate operation is the union of the domains of the aggregated functions. The usefulness of aggregate operations is shown in the queries Q6 to Q10.

The proposed collection of operations is only a basis, however rich, of functionality. Other operations may be added to meet specific needs of some applications. Thanks to the advances in the extension capabilities of the existing DBMS, these types and operations can be easily integrated into the DBMS. Then, it becomes possible to use them through the standard SQL language. Besides, the problem of query optimization must be addressed. This is exactly the plan we followed to implement our data server for MOs with sensors.

D. Query Examples

The great interest of using the extension capabilities of a DBMS is to easily integrate new types and operations in the SQL standard interface. The query examples in this section are based on a relational schema with one table that contains information on vehicle trips as follows:

vehicle_trip(mo_id:*int*, trip:*moving(gpoint)*,
g_speed:*gmoving(real)*, t_speed:*moving(real)*,
g_acceleration:*gmoving(real)*, t_acceleration:*moving(real)*,
g_RPM: *gmoving(real)*, t_RPM: *moving(real)*,
g_odometer:*gmoving(real)*, t_odometer:*moving(real)*,
g_ABS:*gmoving(bool)*, t_ABS:*moving(bool)*,
g_breakSwitch:*gmoving(real)*, t_breakSwitch:
moving(real))

In addition to the spatio-temporal trajectory, i.e., the "trip", the table contains sensor data reporting the speed, acceleration, RPM, odometer, ABS and brake pedal state. These data are modeled by functions of space (prefixed with g_) and of time (prefixed with t_). The parameters are prefixed with the symbol "&" and could be either given by the user at runtime, or existing from previous calculations.

Q1. What is the acceleration profile along a given route segment for a given trip?

SELECT **atgline**(g_acceleration, &aGline)
FROM vehicle_trip
WHERE mo_id = &anID

The operation **atgline** returns the acceleration profile restricted to the sub-space specified by the geometry aGline given as parameter.

Q2. What is the difference between the vehicle's speed profile and the speed limit along a road segment?

SELECT **sub**(g_speed, &legalSpeed)
FROM vehicle_trip
WHERE **inside**(**trajectory**(&legalSpeed),
trajectory(g_speed))=1

The difference between two functions describing measure profiles is calculated using the operation **sub**. An indexed predicate as **inside** could accelerate the query response time. This operation has two *gline* parameters and checks if the first is included in the second. To obtain the projection in space of a measure profile, we use the operation **trajectory**.

Q3. How many times was the ABS enabled for a given trip?

SELECT **no_transitions**(g_ABS)/2
FROM vehicle_trip
WHERE mo_id = &anID

This query simply illustrates the use of **no_transitions,** which is applicable to discrete *BASE* types (e.g., *bool*, *int*) and returns the number of transitions for a given discrete function.

Q4. What are the trips where the practiced speed exceeds a specified speed profile (e.g., the speed limit) by a certain value and what is the difference?

SELECT mo_id, **sub**(g_speed,&legalSpeed)
FROM vehicle_trip
WHERE **intersects**(**trajectory**(&legalSpeed),
trajectory(g_speed)) = 1 AND
max(**sub**(g_speed,&legalSpeed)) > &threshold

There are two new operations in this query. First, the predicate **intersects** is similar to **inside**, the only difference being that it only searches for an intersection between the two *gline* parameters and not for inclusion. Second, the operation **max** is an aggregate of a function. We use it to verify if the maximal value of the function given as parameter is above a certain threshold value. As in the previous query, the parameter for **max** is represented by the difference between the practiced and legal speed profiles.

Q5. What is the ratio between speed and engine RPM for a given trip?

SELECT **div**(t_speed, t_RPM)
FROM vehicle_trip
WHERE mo_id = &anID

This query shows the usefulness of basic algebraic operations for comparing temporal profiles of the same MO. The profile obtained by dividing the vehicle speed to the engine RPM can be used to detect the behavior regarding the gear shifting of a driver.

Q6. What is the average profile of acceleration for all vehicles passing through a certain road section (e.g., curve)?

SELECT **avg_agg**(g_acceleration)
FROM vehicle_trip
WHERE **inside**(**trajectory**(&aCurve), **trajectory**(trip))=1

We determine with this query the average acceleration profile of all vehicles passing through the indicated route section. The function **avg_agg** generates a new *gmoving(real)* object from the set of objects of the same type, passed as a parameter, i.e., all tuples of the table that match the predicate in the WHERE clause.

Q7. Calculate the maximal speed profile of all vehicles passing through the indicated road section.

SELECT **max_agg**(**atgline**(g_speed, &aRoad))
FROM vehicle_trip
WHERE **intersects**(**trajectory**(g_speed), &aRoad) = 1

First we find all the trips that intersect the given road. For these trips, we select by the function **atgline** the speed profile that corresponds to the road. Then we aggregate the resulted profiles in order to obtain the maximal profile, by using the aggregate **max_agg**.

Q8. Find the speed profile actually practiced (85th percentile of the passing vehicles) on a road before and after the installation of a speed camera.

SELECT **percentile**(**atgline**(g_speed,&aRoad),85)
FROM vehicle_trip
WHERE **intersects**(**trajectory**(g_speed),&aRoad) = 1
AND **inst**(**initial**(trip)) < &instalationDate

The query finds the speed profile (85th percentile) before installing a speed camera. A similar query should be posed to find the same profile after the installation of the camera. As for the query Q7, we filter the trips by retaining only those that intersect the given road, and that begin before the installation date of the camera. To do this we use the combination of functions **inst** and **initial** that return the start date of a trip. Finally, we apply the **percentile** function on all selected profiles. The second parameter of this function is the n^{th} percentile.

As we can see from this section, the new data types and operations are needed to express these queries. This kind of queries cannot be supported by the existing models since the concept (the abstract data type) of spatial profile is not considered, nor the operations that allow handling spatial or temporal profiles of a measure.

IV. IMPLEMENTATION

In this section, we address some of the implementation issues of the presented model and language that we currently implement as an extension of a DBMS. The objective is to offer a general view regarding some implementation aspects, rather than a thorough, detailed presentation. Thus, Section IV-A presents the database system architecture. Section IV-B details the sliced representation of the abstract data types. Sections IV-C and IV-D deal with the optimization of the aggregate operations and the operators.

A. Database System Architecture

Currently, the support for spatio-temporal data in the existing DBMS is limited. However, most DBMS today offer possibilities for extensions to meet the needs of certain application domains. Rather than developing a prototype from scratch, we chose to implement the proposed model under such an existing system, i.e., the Oracle DBMS. Thus, all types are implemented as new object types in Oracle 11g. The operations are implemented in Java (Oracle DBMS integrates a Java Virtual Machine) and stored as a package in the database. These operations can be used in SQL queries along with the existing operations in the DBMS.

Finally, some filtering operations, i.e., operations used to identify the MOs that verify a certain spatial, temporal or on-value predicate, are indexed in order to accelerate the query response time and to provide a system scalable with the dataset size (see Section IV-D). To this end, we have proposed PARINET, a novel partitioned index for in-network trajectories [20]. We integrated the indexes by using the data cartridges in Oracle. The general architecture of the system is given in Figure 3. Notice that other DBMS systems that allow DBMS extensions can be used.

B. Data Type Representation

The model in [6][7] that we extended in Section III-B is an abstract model. A finite representation of this abstract model is needed in order to be able to implement it. For all *moving* types, the so-called *sliced representation* has been proposed in [6]. A *moving* object in the abstract model is a temporal partial function. The sliced representation represents the MO as a set of so-called *temporal units* or *slices*. Figure 4 shows a simple example of a temporal profile that is composed of four *units*.

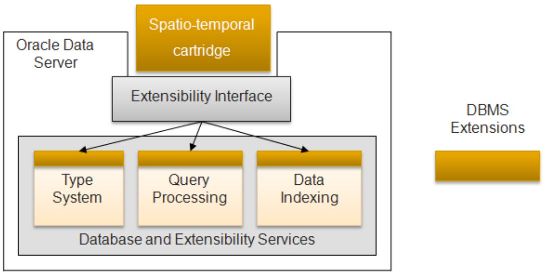


Figure 3. Database System Architecture.

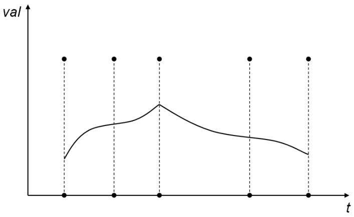


Figure 4. Example of sliced representation of a temporal profile.

A temporal unit for a moving data type α is a time interval where values taken by an instance of α can be described by a “simple” function. The “simple” functions used for the representations are the linear function or quadratic polynomials. The motivation for this choice is a trade-off between the richness of the representation and the simplicity of the representation of the discrete type and of its operations. For example, a unit for a moving(real) object is represented as a tuple *(a, b, c, t1, t2)*, where *a, b, c* are the coefficients of a quadratic polynomial and *t1, t2* is the unit time interval. The moving value at a time instant *t* inside the unit time interval is computed as $a \times t^2 + b \times t + c$. More complex function for unit representation can be imagined but are not considered in this paper. Also, for the sake of simplicity we ignore that the unit intervals are left-closed and/or right-closed. For all *gmoving* types that we introduced in Section III-B, we adopt the sliced representation as proposed in [6]. This is straightforward as the sole difference is to replace the unit's time interval, which is the support for temporal profiles, with a spatial interval, which is the support for spatial profiles. A spatial interval given a network space has the following elements: (*rid, pos1, pos2*), where *rid* is a road identifier and *pos1, pos2* are relative positions on the road. For example, a *gmoving(real)* object will contain a set of units with the following attributes: *(a, b, c, rid, pos1, pos2)*. To calculate a value for a given position, we first locate the corresponding unit, i.e., where the spatial interval includes the position, then we calculate the value as $a \times pos^2 + b \times pos + c$.

C. *Handling Aggregation Functions*

Unlike the functions on one or two profiles, aggregate functions operate on a set containing a (possibly) large number of profiles. This can lead to a fragmentation of the result profile in a large number of small units and a degradation of the query performance. Consider the example in Figure 5. The first graphic shows two profiles with their decomposition into units and the second one represent the maximum aggregate of these profiles. Notice that the units of the two profiles do not have the same spatial distribution. The units' space intervals of the two profiles overlap partially. This is common because the unit slicing is unique to each profile and depends on the variability of the observed measure at the observation time. Therefore, the result of an operation on a set of profiles is another profile that contains more units that the initial profiles. This process of fragmentation of the result is not disturbing when the calculation is done only on two profiles. However, in the case of the aggregation it can significantly slow down the computation time.

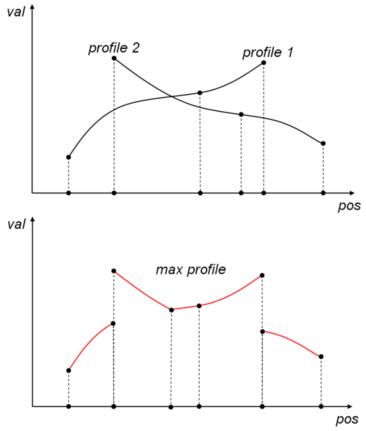


Figure 5. Example of fragmentation after using the max_agg (second graphic) on two profiles (first graphic).

To accelerate the aggregate operations, we propose a regular temporal or spatial slicing of profiles, independent of the initial slicing. This method offers a compromise between efficiency and the quality of the results. Thus, for aggregates on *gmoving* types for example, we uniformly divide the space, beginning with the start point of each road in intervals of a given length, e.g., 10 meters. Smaller intervals will produce higher quality results but at a cost of a slower performance, and vice versa.

Figure 6 presents an example of using uniform slicing for computing aggregates on two spatial profiles. The first graphic shows the profile decomposition by regular intervals (represented by the vertical dotted lines). For each interval, we compute or extrapolate first the values on the end limits of the interval. Thus, for the first profile (in red) we find the values v_1^1 and v_2^1 the first interval, v_2^1 and v_3^1 for the second interval, and v_3^1 and v_4^1 for the third interval. From these values computed for all profiles, we apply the corresponding scalar aggregate function (e.g., the aggregate max for **max_agg**) in order to generate the values of the resulting profile on the same limits of the intervals.

Overall, this approach to implement the aggregate functions produces approximate results, but in return it offers a good optimization of this costly type of operation. The analysis of the result quality depending on the granularity of slicing is left for future work.

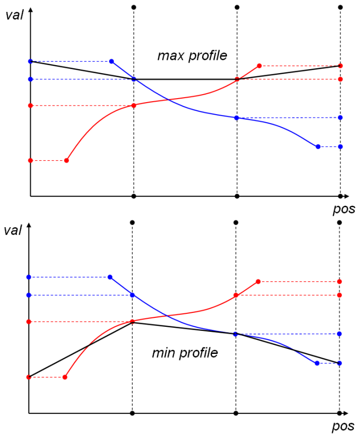


Figure 6. Example of calculating the max_agg (first graphic) and min_agg (second graphic) on the two profiles of Figure 5 using a regular slicing.

D. *Operators*

In the field of spatio-temporal databases, the indexing techniques that permit processing efficiently the spatial, temporal and on-value queries are complementary to modeling the moving objects. Our prototype uses PARINET [20] for querying trajectories. A discussion on the indexing methods is out of the scope of this paper. Instead, we present in this section the mechanism through which the indexes are linked to the algebra, i.e., object types and operations. This mechanism is based on the operators.

Operators are a subset of the algebra operations, mostly predicates such as **present** or **passes**, that benefit of an index based evaluation in addition to the basic function implementation. The functional implementation is used when the operator is invoked in the select list of a SELECT command or in the ORDER BY and GROUP BY clauses. However, when the operator appears in the condition of a WHERE clause, the DBMS optimizer chooses between the indexed implementation and the functional implementation, taking into account the selectivity and the cost when generating the query execution plan.

The operators that we implement are spatial, temporal and on-value predicates, or predicates that combine two of the three possible dimensions (i.e., spatio-temporal, on-value spatial and on-value temporal). For example, to select only the profiles that spatially intersect a given network region, one will use the **present** operator. In the same way, the **passes** operator is used to select only the objects for which a certain measure assumes a given value (e.g., acceleration is above 10m/s^2). Finally, the two-dimensional predicates verify that the conditions in each dimension are simultaneous verified. For example, a spatio-temporal operator ensures that the trajectory of a MO intersects (or is included) a (in) spatial network region at a given time interval. Similar reasoning can be applied for the rest of the operators.

V. CONCLUSION AND FUTURE WORK

The use of sensors embedded in vehicles leads to new applications, which give rise to new research problems. In this paper, we addressed the problem of modeling and querying mobile sensor data. In this context, the existing work in moving objects databases is limited. A DBMS capable of managing in a unified manner the moving object data and the (embedded) moving sensor data is needed for these applications.

The contribution of this paper is to propose a model for such a DBMS by extending an existing framework for MOs. We first analyzed the limitations of modeling mobile sensor data. Indeed, existing models can represent the data flows from a temporal point of view. We have shown that these measures are equally dependent of the object's position and a representation relative to the space is needed. Therefore, we have extended the existing type system with functions that describe the evolution of measures in space. We have also proposed a collection of operations in view of the enhanced system. We introduced the concept of *spatial lifting* inspired by the idea of the existing *temporal lifting*. We have redefined all the temporal operations and changed the semantics of some of them for the new data types. Finally, we proposed a collection of operations appropriated for analyzing moving sensor data. An illustration of use of the DBMS is given by query examples involving the new defined types and operations. The current prototype includes a partial implementation of the algebra as a data cartridge in Oracle DBMS.

This work is part of a Ph.D thesis. Further details could be found in the report [19]. As future work, we intend study proper indexing techniques for the new types. Although this is a similar to the query optimization problem in MOD, the distribution of sensor values may lead to specific optimizations in our system. We also investigate the problem of mining such databases [10]. Finally, adapting the data resolution to the application needs (some applications need data of all sensing points whereas others need just a summary) raises new challenges

REFERENCES

[1] Botts, M., Percivall, G., Reed, C., and Davidson, J.: OpenGIS Sensor Web Enablement: Overview and High Level Architecture. OpenGIS. White Paper. (OGC 07-165), 2007

[2] Desphande, A. and Madden, S.: MauveDB: supporting model-based user views in database systems. ACM SIGMOD 2006, pp. 73-84, Chicago, Illinois, June 2006

[3] Ehrlich, J., Marchi, M., Jarri, P., Salesse, L., Guichon, D., Dominois, D, and Leverger, C.: LAVIA, the French ISA project: Main issues and first results on technical tests, 10th World Congress & Exhibition on ITS, November 2003

[4] Forlizzi, L., Güting, R.H., Nardelli, E., and Schneider, M.: A Data Model and Data Structures for Moving Objects Databases. ACM SIGMOD 2000, pp. 319-330, Dallas, Texas, May 2000

[5] Grumbach, S., Rigaux, P., and Segoufin, L.: Spatio-Temporal Data Handling with Constraints, GeoInformatica, 5(1), pp. 95-115, 2001

[6] Güting, R.H., de Almeida, V.T., and Ding, Z.: Modeling and Querying Moving Objects in Networks. VLDB Journal 15(2), pp. 165-190, 2006

[7] Güting, R.H., Böhlen, M.H., Erwig, M., Jensen, C.S., Lorentzos, N.A., Schneider, M., and Vazirgiannis, M.: A Foundation for Representing and Querying Moving Objects. ACM Transactions on Database Systems, 25(1), pp. 1-42, 2000

[8] Güting, R.H. and Schneider, M.: Moving Objects Databases, Morgan Kaufmann, 2005

[9] ISO 19107:2003, Geographic Information – Spatial Schema, WG 2

[10] Kharrat, A., Sandu Popa, I., Zeitouni, K., and Faiz, S.: Clustering Algorithm for Network Constraint Trajectories, 13th International Symposium on Spatial Data Handling, SDH 2008, pp. 631-647, Montpellier, France, June 2008

[11] Koubarakis, M., Pernici, B., Schek, H.J., Scholl, M., Theodoulidis, B., Tryfona, N., Sellis, T., Frank, A.U., Grumbach, S., Güting, R.H., Jensen, C.S., Lorentzos, N., Manolopoulos, Y., and Nardelli, E. (Eds.): Spatio-Temporal Databases: The CHOROCHRONOS Approach. Springer-Verlag, Lecture Notes in Computer Science 2520, 2003

[12] NHTSA (2006). The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment. Report No. DOT HS 810 593

[13] Oracle Database Data Cartridge Developer's Guide, 11g Release 1, 2007

[14] Pelekis, N.: STAU: A Spatio-Temporal Extension to ORACLE DBMS, PhD Thesis UMIST, Department of Computation, 2002

[15] Pelekis, N., Frentzos, E., Giatrakos, N., and Theodoridis, Y.: HERMES: Aggregative LBS via a Trajectory DB Engine, SIGMOD 2008, pp. 1255-1258, Vancouver, Canada, June 2008

[16] Percivall, G. and Burggraf, D.: OGC Moving Object Snapshot: An application schema of the OGC Geography Markup Language, OGC™ Discussion Paper, 2010.

[17] Pelekis, N., Theodoulidis, B., Kopanakis, I., and Theodoridis, Y.: Literature Review of Spatio-Temporal Database Models, The Knowledge Engineering Review Journal, 19(3), pp. 1-34, 2005

[18] Pfoser, D. and Jensen, C.S.: Indexing of Network-constrained Moving Objects. ACM-GIS 2003, pp. 25-32, New Orleans, Louisiana, November 2003

[19] Sandu Popa, I.: Modeling, Querying and Indexing Moving Objects with Sensors on Road Networks. Ph.D. Thesis, University of Versailles-Saint-Quentin, 2010

[20] Sandu Popa, I., Zeitouni, K., Vincent, O., Barth, D., and Vial, S.: PARINET: A Tunable Access Method for in-Network Trajectories, 26th IEEE International Conference on Data Engineering, ICDE 2010, pp. 177-188, Long Beach, California, March 2010

[21] Sistla, P., Wolfson, O., Chamberlain, S., and Dao, S.: Modeling and Querying Moving Objects. IEEE ICDE 1997, pp. 422-432, Birmingham, U.K., April 1997

[22] Project euroFOT: http://www.eurofot-ip.eu/ (retrieved on december, 6th 2011)

Land-use Mapping of Valencia City Area from Aerial Images and LiDAR Data

Txomin Hermosilla, Luis A. Ruiz, Jorge A. Recio
GeoEnvironmental Cartography and Remote Sensing
Research Group.
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia, Spain
txohergo@topo.upv.es; laruiz@cgf.upv.es;
jrecio@cgf.upv.es

José Balsa-Barreiro
Instituto Cartográfico Valenciano
Generalitat Valenciana
C. Santos Justo y Pastor 116, 46022Valencia, Spain
balsa_jos@gva.es

Abstract - **Land-use classification of urban environments is usually limited by the number and complexity of the considered classes and the capability of the selected methodology for the efficient discrimination of these classes. Thus, this paper analyses and assesses the performance of a contextual object-based classification methodology in urban environments considering a comprehensive land-use legend, including several complex urban land-uses –*historical buildings, urban buildings, open urban buildings, semi-detached houses, detached houses, industrial/warehouse buildings, religious buildings, commercial buildings, public buildings, gardens and parks*–, and agricultural classes –*arable lands, citrus orchards, irrigated crops, carob-trees orchards, rice crops, forest, greenhouses*–. Object-based approach was achieved by using cadastral plot limits for object definition. An exhaustive set of object-based descriptive features were computed informing about the spectral, texture, structural, geometrical, three-dimensional and contextual properties. Classification was performed by means of decision trees algorithm combined with boosting multi-classifier. The overall accuracy reached classifying the urban area of Valencia reached 84.8%, which is a significantly high value when considering a large number of complex urban classes.**

Keywords - Object-based classification; high spatial resolution imagery; LiDAR; urban areas; mapping

I. INTRODUCTION

Urban areas are dynamic and changing environments both in land covers and land uses. This entails that cartographical information referred to cities becomes rapidly obsolete. An efficient urban management requires an accurate and up-to-date knowledge about the land cover situation and evolution in urban and surrounding areas. This enables a wide range of applications including physical planning –viewshed analysis, impact assessment, environmental issues–; economic planning –accessibility, location analysis, transport studies –; social planning – population and other socio-demographic distributions, urban structures–; or forecasting models –diffusion and urban growth– [1].

Traditionally the process of creating land-use/land-cover (LU/LC) maps of urban areas involves field visits and classic photo-interpretation techniques. These methodologies are expensive, time consuming, and also subjective, requiring skilled operators. Remotely sensed data and digital image processing techniques help to reduce the volume of information that needs to be manually interpreted, satisfying current demands for continuously precise data for an automatic, systematic and efficient territorial and urban management.

Image classification processes to produce land-cover maps in urban areas can be considered straightforward when compared to the complex process of deriving information on urban land use [2], since the land use is an abstract concept that represents a socio-economic criterion referring to the dominant activity of a place, and may include category subdivisions with differing levels of detail [3]. The definition of an extensive land-use legend enables a deeper and better knowledge of the "*actuality*" of the urban scenario, but it also entails additional difficulties in the discrimination of classes, since a large number of complex land-use classes generally lead to reach limited results.

When considering high spatial resolution imagery, object-based approaches are generally used to classify land uses in urban areas, where objects can be defined using automatic segmentation methods or –most commonly– by means of urban blocks or plot limits derived from existing cartography. Moreover, plot-based image classification allows to directly relate the information extracted from the remotely sensed data to LU/LC geo-spatial database objects.

Reference [4] considered eleven complex land use/land cover classes, but without assessing the quality of the methodology. Reference [5] obtained discrete accuracy values when classifying eight different urban land uses; [6] obtained an accurate result considering five classes of urban development, and [7] differenced six land uses reaching medium accuracy values. The definition of a contextual framework through a multi-resolution analysis permits to increase the classification accuracy of urban environments considering several classes, as demonstrated in [8].

The addition of three-dimensional information using LiDAR (Light Detection and Ranging) data allows to increase the number of classes in the legend, and to reach higher accuracy values. Therefore, [9] obtained high accuracies when classifying a suburban area distinguishing between land uses and other additional land-cover classes to fully complete the area. Reference [10] considered nine different complex urban land uses reaching unbalanced accuracies due to the extreme differences in the number of per-class samples. In the same sense, [11] defined ten land use classes to completely classify the city of Brussels including different housing typologies, but no accuracy assessment was presented. Combining three-dimensional information and context-based descriptive features [12] attained accurate results distinguishing between five complex urban classes plus two agricultural classes.

The aim of this paper is to analyse and assess the performance of a contextual object-based classification methodology using high spatial resolution multispectral imagery and LiDAR data when classifying urban environments considering a comprehensive land-use legend containing a large number of classes, including several complex urban land-uses and agricultural classes. This paper is organized as follows. Section 2 describes the area where the study was performed and the data employed. Section 3 describes the object-based classification methodology and the accuracy assessment followed. Section 4 reports and analyses the results. Section 5 presents the conclusions.

II. Study Area and Data

This study was performed in the administrative area of the municipalities of Valencia and Paterna, located in the Mediterranean coast of Spain (see Figure 1.). Valencia is the largest city and capital of the Valencian Community, having 809,267 inhabitants in 2010 [13]. Valencia is a compact city composed by a central historical area surrounded by buildings of different typologies, depending on the date of construction. The northern area is covered by citrus orchards and horticulture crops, while the natural park of *l'Albufera* presents extensive rice crops and forests, and is located in the southern zone. Paterna is a contiguous municipality located in the metropolitan area of Valencia with a population of about 65,921 inhabitants [13], presenting large extensions of low-density suburban housing and several industrial and commercial areas.

The limits of the plots were provided by vectorial cadastral cartography in shapefile format, produced by the Spanish General Directorate for Cadastre (*Dirección General de Catastro*). This cartography presents a scale of 1:1,000 in urban areas and 1:2,000 in rural areas.

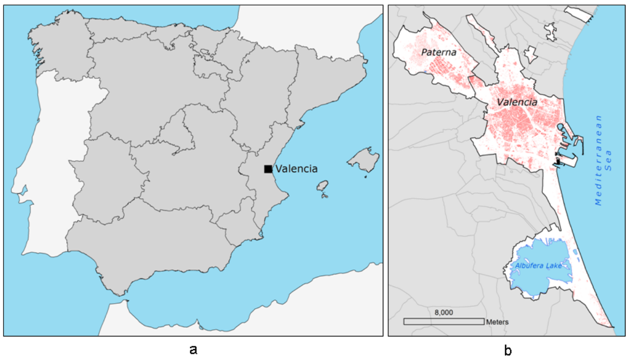


Figure 1. Location of Valencia in Spain (a) and representation of the two municipalities considered in this study: Valencia and Paterna (b).

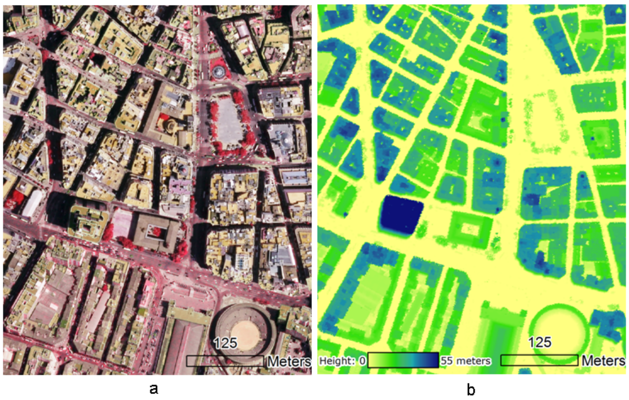


Figure 2. Detail of a high spatial resolution image in colour infrarred composition (a), and digital surface model created with the LiDAR data (b), depicting a zone in the urban centre of Valencia.

Imagery and LiDAR data were acquired in the frame of the Spanish National Plan of Aerial Orthophotography (PNOA). The images were collected in August 2008 and they have 0.5 m/pixel spatial resolution, 8 bits radiometric resolution and four spectral bands: red, green, blue and near infrared. The images are distributed orthorectified and georreferenced, panchromatic and multispectral bands fused, mosaicking and radiometric adjustments applied, as part of the PNOA project. An example of the multispectral imagery employed is show in Figure 2. a.

LiDAR data were acquired in September 2009 with a nominal density of 0.5 points/m^2 using a RIEGL LMS-Q680 laser scanner device. A normalised digital surface model (nDSM), i.e., the difference between the digital surface model (DSM) and the digital terrain model (DTM), representing the physical heights of the elements present over the terrain, was generated from LiDAR data. The DTM was computed by means of an algorithm that iteratively selects minimum elevation points and eliminates points belonging to any aboveground elements, such as vegetation or buildings [14]. Figure 2. b shows an example of the nDSM of the centre of Valencia.

III. Methodology

Land use classification was carried out in five steps: class definition; sample selection; descriptive feature extraction; object classification; and evaluation. Object definition was done using cadastral plot limits, and these objects were exhaustively described by different types of image derived features: three-dimensional features computed from LiDAR data, structural features derived from the semivariogram graph, geometrical features, and context-based features. Classification was performed by means of C5.0 decision tree algorithm combined with the boosting technique. The classification accuracy was assessed by analyzing the confusion matrix.

A. *Definition of classes*

Land-use class definition was performed based on the specifications of the Land Cover and Land Use Information

System of Spain (SIOSE) database. The legend was composed of seventeen classes, discriminating between ten urban land use classes and seven agricultural classes. The samples (plots) were collected by using a restricted randomization scheme [15], consisting on a random sampling selection, to ensure the spatial homogeneity of the samples, followed by a redistribution and addition of some samples in order to maintain the appropriate number of samples according to the variability into each class. The urban classes defined were: *historical buildings* (264 samples), *urban buildings* (225), *open urban buildings* (142), *semi-detached houses* (90), *detached houses* (153), *industrial/warehouse buildings* (139), *religious buildings* (30), *commercial buildings* (24), *public buildings* (173), including schools, universities, sport facilities and civic and governmental buildings, and *gardens and parks* (57). The agricultural classes defined were: *arable lands* (92), *citrus orchards* (141), *irrigated crops* (81), *carob-trees orchards* (63), *rice crops* (74), *forest* (39) and *greenhouses* (43). Examples of the defined classes in colour infrared composition are shown in Figure 3.

B. *Object-based descriptive feature extraction*

Object-based features describe each object as a single entity based on several aspects that reflect the variety of information used, and these were computed using the object-based image analysis software FETEX 2.0 [16]. The computed features provided information regarding spectral, texture, structural, geometrical, three-dimensional and context based properties.

Spectral features provide information about the intensity values of objects in the different spectral bands. Statistical descriptors were computed for each plot in the available bands and in the NDVI image. Texture features quantify the spatial distribution of the intensity values in the analysed objects. Texture was characterized by means of kurtosis and skewness, the descriptors derived from the grey level co-occurrence matrix proposed by [17], and the edgeness factor [18]. Structural features provide information of the spatial arrangement of different elements in the object, in terms of randomness, and these were derived from the semivariogram graph [19] [20]. Geometrical features describe the dimensions of the plots and their contour complexity. Three-dimensional features were derived from the nDSM computed using LiDAR data.

Context was described by characterizing the higher and lower aggregation levels of the plots. Thus, internal context features describe an object attending to the land cover types of the elements contained within the object (denoted as sub-objects). In this case, buildings and vegetation, which were extracted by applying a multiple-threshold based approach, as described in [21]. External context is defined characterizing each object by considering the common properties of adjacent objects that combined create an aggregation higher in hierarchy than plot level, such as urban blocks in urban areas. This context is described by means of specific building-based, vegetation-based, geometrical and adjacency features [12].

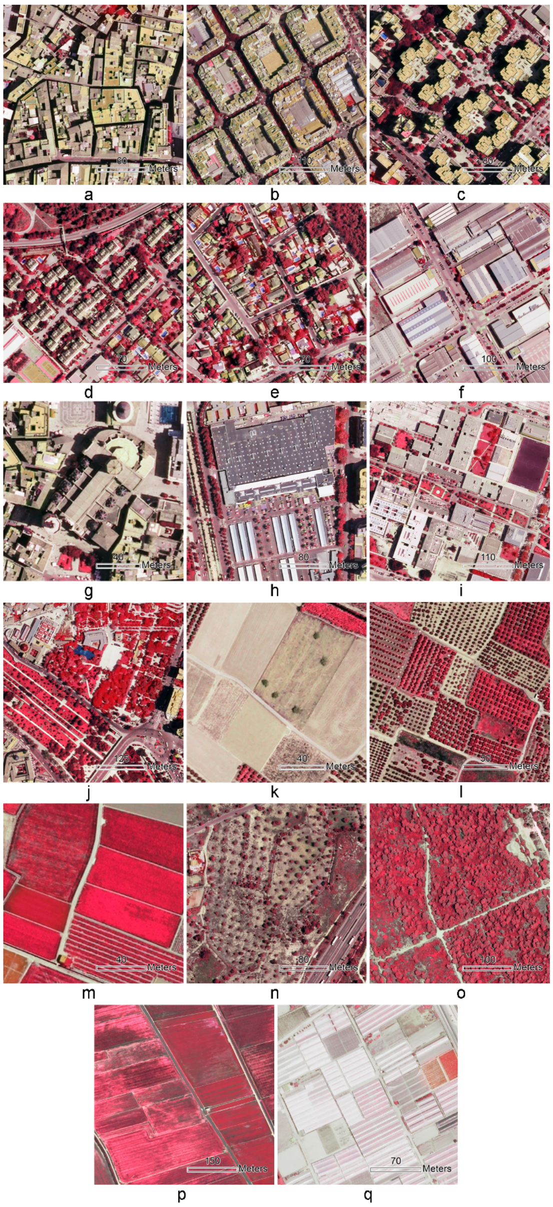


Figure 3. Examples of the considered classes: *historical buildings* (a), *urban buildings* (b), *open urban buildings* (c), *semi-detached houses* (d), *detached houses* (e), *industrial/warehouse buildings* (f), *religious buildings* (g), *commercial buildings* (h), *public buildings* (i), *gardens and parks* (j), *arable lands* (k), *citrus orchards* (l), *irrigated crops* (m), *carob-trees orchards* (n), *forest* (o), *rice crops* (p), and *greenhouses* (q).

C. *Classification and accuracy assesment*

Classification was performed using decision trees constructed with the C5.0 algorithm [22] combined with the boosting technique. The process of building a decision tree begins by dividing the collection of training samples using mutually exclusive conditions. Each of these sample subgroups is iteratively divided by using the gain ratio as a splitting criterion until the newly generated subgroups are homogeneous, i.e., all the elements in a subgroup belong to the same class or a stopping condition is fulfilled. The gain ratio criterion employs information theory to estimate the size of the sub-trees for each possible attribute and selects the attribute with the highest expected information gain. The algorithm is based on searching partitions to obtain purer data subgroups, which are less mixed than the previous group from where they were derived. This is iterated until the original data set is divided into homogeneous subgroups.

The evaluation of the classification was based on the analysis of the confusion matrix [23], which compares the class assigned to each evaluation sample with the reference information, defined by photointerpretation. The overall accuracy of the classification and the kappa index were computed, as well as the producer's and user's accuracies for each class, that respectively expose the classification errors of omission and commission. In order to maximize the efficiency of the evaluation process, in terms of the number of samples, the *leave-one-out* cross-validation technique was employed. This method uses a single observation from the original sample set as validation data, using the remainder observations as training data. This is iterated 1590 times, until each observation in the sample set is used once as validation data.

IV. RESULTS AND DISCUSSION

The cartographic representation of the classification, depicting the centre of Valencia is shown in Figure 6. where the different urban structures are distinguished: historical centre, planned areas, industrial, civic and transportation facilities, parks, etc. The overall accuracy of the classification was **84.8**%, and the kappa coefficient **0.83**. These are sound results, especially considering the large number of classes defined (17) and the structural similarities between some classes, e.g., *semi-detached houses* and *detached houses*.

Analysing the per-class user's and producer's accuracies (see Figure 4.) it is remarkable the high performance achieved for agricultural classes, presenting values higher than 90% in the case of *arable lands*, *citrus orchards*, *carob-trees orchards*, *rice crops*, *forest*, and slightly lower for *irrigated crops* (with values around 88% for both accuracies) and the user's accuracy of the class *greenhouse* (84%). Among the urban classes, the lowest accuracies and the most unbalanced values were obtained for classes *commercial buildings* and *religious buildings*. The stu confusion matrix –graphically represented in Figure 5. – shows that *commercial buildings* had a poor performance and presented several misclassifications with *industrial/warehouse* and *public buildings* classes. *Religious buildings* class producer's accuracy reached a very low value (37%) due to the confusion with classes *urban* and *public buildings*. Medium user's and producer's accuracy values (70%) were achieved for *public buildings* and *semi-detached houses*. *Public buildings* presented a high degree of confusion with most of the building-related classes, due to their heterogeneity, and the fact that these buildings usually have significant morphological differences, producing misclassifications. Some particular public building plots containing covered sport facilities were erroneously assigned to *industrial/warehouse buildings* and viceversa. *Semi-detached houses* were especially confused with *detached houses*, due to their obvious structural similarities. *Gardens and parks* presented unbalanced accuracies, due to the misclassification with *citrus orchards* and *public buildings*. Other building-related urban classes achieved better classification performances with slight confusions between them, being especially significant for the pair of classes *historical* and *urban*, as shown in the confusion matrix (Figure 5.).

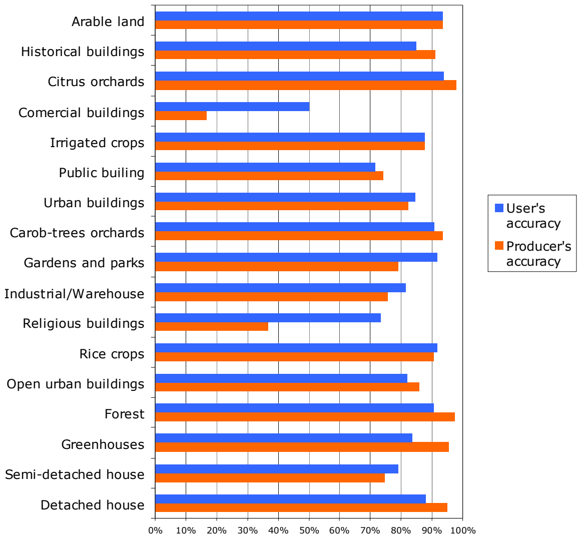


Figure 4. Classification user's and producer's accuracies for each defined class.

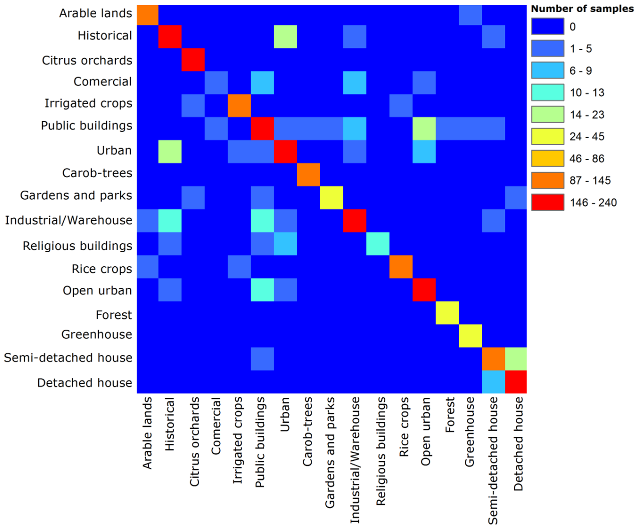


Figure 5. Graphic representation of the confusion matrix of the classification. Rows represent reference class and collums show classified data.

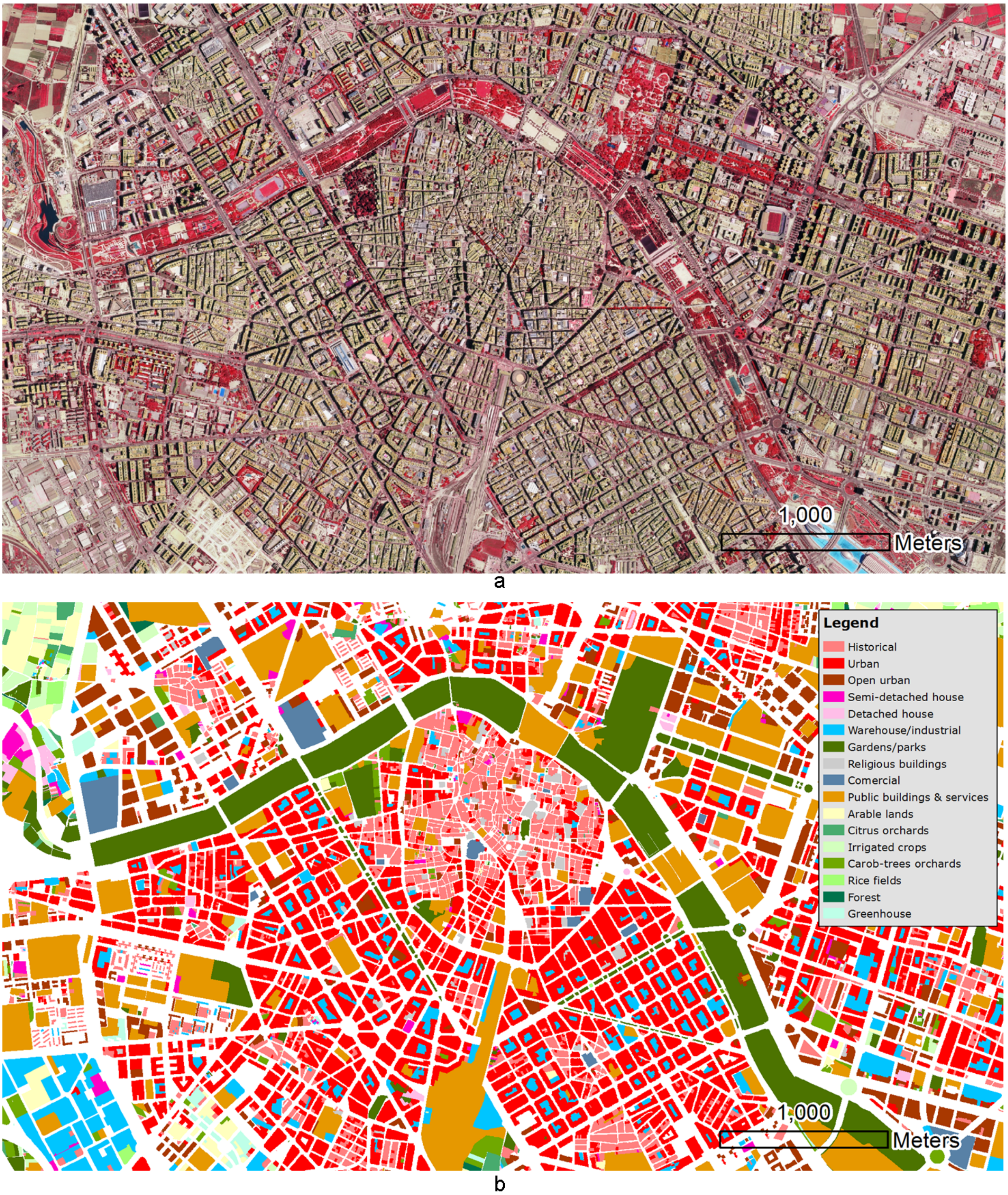


Figure 6. Thematic map composition showing the classes assigned to each plot of the urban centre of Valencia (a) and colour infrared composition of the same area (b).

V. CONCLUSIONS

In this paper, the performance of a contextual object-based classification methodology in urban environments was analysed and assessed, when considering an exhaustive land-use legend that includes several complex urban land-uses. A set of object descriptive features was extracted to characterise intrinsic properties of the plots –spectral, texture, geometrical, and three-dimensional–, and their context attending to two levels: internal –referred to internal covers in the plot–, and external –related to common properties of plots contained in the same urban block–.

The results showed the high potential of the proposed methodology to correctly and accurately discriminate and assign land use to a large number of different building typologies, and simultaneously a variety of agricultural land uses. Most of the agricultural classes were satisfactorily assigned. In general, urban classes were accurately classified. However, very heterogeneous building typologies concerning commercial, religious and public uses obtained a low performance, since the difficulty found to distinguish these classes from other urban building typologies. Additionally, due to the similarity of some classes, they presented minor mutual misclassifications, for example different typologies of suburban buildings, or planned urban areas and historical areas.

The proposed object-based classification methodology provides new tools that may increase the frequency, efficiency and detail level of urban studies, being useful for systematically mapping cities, urban landscape characterisation, automatic land-use change detection and updating LU/LC geospatial databases.

ACKNOWLEDGMENT

The authors appreciate the financial support provided by the Spanish *Ministerio de Ciencia e Innovación* and the FEDER in the framework of the Projects CGL2009-14220 and CGL2010-19591/BTE, and the Spanish *Instituto Geográfico Nacional* (IGN).

REFERENCES

[1] J. P. Donnay, M. J. Barnsley, and P. A. Longley, “Remote sensing and urban analysis”. In: J. P. Donnay, M. Barnsley, P. A. Longley, (Eds.), Remote Sensing and Urban Analysis, Taylor & Francis, London, UK, pp. 3–18, 2001.

[2] J. R. Eyton, “Urban land use classification and modeling using cover-type frequencies”, Appl. Geogr., vol. 13, pp. 111–121, April 1993.

[3] M. Barnsley and S. Barr, “A graph-based structural pattern recognition system to infer land use from fine spatial resolution land cover data”, Comput. Environ. Urban. Syst., vol. 21 (3-4), pp. 209–225, 1997.

[4] T. Bauer and K. Steinnocher, “Per-parcel land use classification in urban areas applying a rule-based technique”, GeoBIT/GIS vol. 6, pp. 24–27, 2001.

[5] J. Wijnant and T. Steenberghen, “Per-parcel classification of urban IKONOS imagery”, Proc. of 7th AGILE Conference on Geographic Information Science, Heraklion, Greece, pp. 447–455, April 2004.

[6] K. Zaremski, “Differentiation between forms of urban development using the object-oriented classification method with central Warsaw as the example”, Misc. Geogr., vol. 12, pp. 315–327, 2006.

[7] T. Novack, H. J. H. Kux, R. Q. Feitosa, and G. A. Costa, “Per block urban land use interpretation using optical VHR data and the knowledge-based system interimage”, Int. Arch. Photogram. Rem. Sens. Spatial. Inform. Sci., vol. 38 (4/C7), June 2010.

[8] A. Huck, S. Hese, and E. Banzhaf, “Delineating parameters for object-based urban structure mapping in Santiago de Chile using QuickBird data”, Int. Arch. Photogram. Rem. Sens. Spatial. Inform. Sci., vol. 38 (4/W19), June 2011.

[9] E. Hussain and J. Shan, “Rule inheritance in object-based image classification for urban land cover mapping”, ASPRS 2010 Annual Conference, San Diego, CA, April 2010.

[10] S. S. Wu, X. Qiu, E. L. Usery, and L. Wang, “Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use”. Ann. Assoc. Am. Geogr., vol. 99, pp. 76–98, January 2009.

[11] S. Vanderhaegen and F. Canters, “Developing urban metrics to describe the morphology of urban areas at block level”, Int. Arch. Photogram. Rem. Sens. Spatial. Inform. Sci., vol. 38 (4/C7), June 2010.

[12] T. Hermosilla, L. A. Ruiz, J. A. Recio, and M. Cambra-López, “Efficiency of context-based attributes or land-use classification of urban environments”, Int. Arch. Photogram. Rem. Sens. Spatial. Inform. Sci., vol. 38 (4/W19), June 2011.

[13] Insituto Nacional de Estadística. Revisión del Padrón municipal 2010 (Review of the census of the spanish municipalities), 2010. www.ine.es. Last access: 29 August 2011.

[14] J. Estornell, L. A. Ruiz, B. Velázquez-Martí, and T. Hermosilla, “Analysis of the factors affecting LiDAR DTM accuracy in a steep shrub area”, Int. J. Digit. Earth, vol. 4 (6), pp. 521–538, November 2011.

[15] C. Chatfield, “Avoiding statistical pitfalls”, Stat. Sci., vol 6 (3), pp. 240–252, August 1991.

[16] L. A. Ruiz, J. A. Recio, A. Fernández-Sarría, and T. Hermosilla, “A feature extraction software tool for agricultural object-based image analysis”, Comput. Electron. Agric., vol. 76, pp. 284–296, May 2011.

[17] R. M. Haralick, K. Shanmugan, and I. Dinstein, “Texture features for image classification”, IEEE T. Syst. Man Cyb., vol 3, pp. 610–621, November 1973.

[18] R. N. Sutton and E. L. Hall, “Texture measures for automatic classification of pulmonary disease”, IEEE Trans. Comput., vol 21 (7), pp. 667–676, July 1972.

[19] A. Balaguer, L. A. Ruiz, T. Hermosilla, and J. A. Recio, “Definition of a comprehensive set of texture semivariogram features and their evaluation for object-oriented image classification”, Comput. Geosci., vol. 36, pp. 231–240, February 2010.

[20] A. Balaguer-Besser, T. Hermosilla, J. A. Recio, and L. A. Ruiz, “Semivariogram calculation optimization for object-oriented image classification” Modelling Sci. Educ. Learn., vol. 4 (7), pp. 91–104, June 2011.

[21] T. Hermosilla, L. A. Ruiz, J. A. Recio, and J. Estornell, “Evaluation of automatic building detection approaches combining high resolution images and LiDAR data”, Remote Sens., vol. 3, pp. 1188–1210, June 2011.

[22] J. R. Quinlan, C4.5. Programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.

[23] R. Congalton, “A review of assessing the accuracy of classications of remotely sensed data”, Remote Sens. Environ., vol. 37, pp. 35–46, July 1991.

Generation and Validation of Workflows for On-demand Mapping

Nick Gould and Omair Chaudhry
Manchester Metropolitan University
Manchester, UK
emails: {nicholas.m.gould@stu.mmu.ac.uk, o.chaudhry@mmu.ac.uk}

***Abstract*— The paper presents a method to automatically select and sequence the tasks required to build maps according to user requirements. Workflows generated are analysed using Petri nets to assess their validity before execution. Although further work is required to select the optimal method for generating the workflow and to execute the workflow, the proposed method can be used on any workflow to assess its validity.**

Keywords-automated map generalisation; workflows; Internet mapping; Petri nets.

I. Introduction

The development of Google Maps and similar products has led to a vast number of ‘mashups’ where users can overlay their own data on Google Maps backgrounds and make the resultant map available to others. The problem with this approach is that the user is limited to the background maps supplied by Google; there is no, or very little, flexibility to vary the content depending on the context and there is no data integration [1]. This is highlighted in Fig. 1 where the street names are obscured by overlaid cycle routes. Further problems may occur when the scale changes. For instance, a minor road that may be part of a cycle route may disappear at smaller scales since the two datasets are independent.

What is required is a system to allow data from a variety of sources to be mapped at a variety of scales. Since, the possible combination of datasets and scales is too numerous to be pre-defined, on-demand generalisation (deriving smaller scale maps from larger scale maps) is necessary.

Cartographic generalisation is a complex process [2] and much effort has gone in to developing automation techniques that reduce or eliminate human involvement [3].

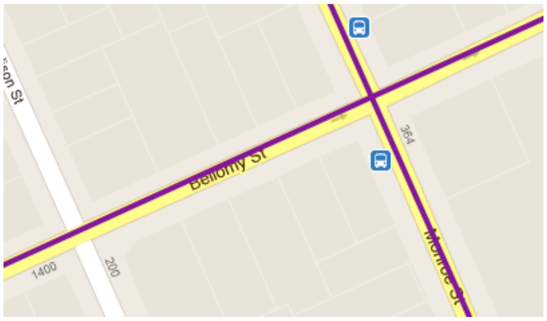


Figure 1. Google Maps with cycle routes overlaid

The focus has, until recently, been on allowing National Mapping Agencies (NMAs) to automate the production of maps at different scales from a single master source [4][5]. Automatic generalisation is applied to a pre-defined set of map features at pre-defined scales to produce a pre-defined set of products. However, the advance of neo-geography and Volunteered Geographic Information [6] means that on-demand generalisation is required allowing users to integrate their data with that of NMAs and other mapping resources. There have been attempts to generate online on-demand maps to user requirements, but such systems have been developed by applying a fixed sequence of generalisation operations to known datasets [7][8].

An on-demand mapping system will require a number of components including a means of taking high level user requirements (e.g., “I want a city-wide map of road accidents”) and producing a machine-readable specification of the map [9]. The system will also need a knowledge base to store cartographic rules of the type: “if the scale is greater than 1:30,000 omit minor roads”. A set of map generalisation services are then required to satisfy such rules or constraints. Traditionally the selection of map generalisation operators and their sequencing is done by cartographic experts, but for on-demand mapping, aimed at the non-expert user, a system is required that can automatically generate, validate, execute and monitor these operations; in other words a workflow needs to be generated and executed [9]. The focus of this research is on developing a workflow engine that, given the specification, using the rules, will automatically select, sequence, and execute the map generalisation services required to generate the map or spatial output.

This paper describes the initial attempts to automatically generate a workflow for building a map based on user requirements and suggests how to validate that workflow.

To illustrate the process, a use case involving the mapping of road accidents will be employed. Fig. 2 represents a detailed map of accidents at a road junction.

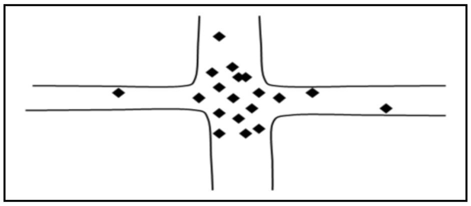


Figure 2. Accidents at a road junction

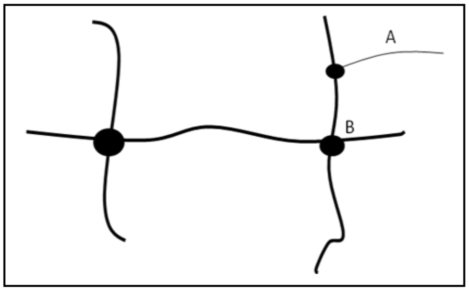


Figure 3. Generalised data at a small scale

To represent all of the data at a smaller scale the road network is generalised by *eliminating* any minor roads and *collapsing* (reducing to single lines) the major roads. To avoid information overload the accidents are *clustered* (Fig. 3). Elimination, collapse and clustering are common generalisation operations. The junction in Fig. 2 is represented by ‘B’.

The generalised map serves to highlight accident hot spots. However, by removing the minor road ‘A’, context is lost, since the cluster will appear to be on a straight section of road, so a step is required to reinstate those minor roads that intersect a cluster. What we have is a set of tasks that have to be carried out, some of which are in a particular order, i.e., we cannot reinstate the minor roads until we have created the clusters. Since there a number of tasks to execute, some of which have to be completed before others can start, a workflow is required to express these relationships. In addition that workflow has to be valid, i.e., all of the tasks must, at least, be called.

The method used was based on the premise that workflow definitions can be analysed using Petri nets for flaws that would stop the workflow from completing execution [10][11].

Firstly, a technique was implemented to generate the list of tasks and their dependencies based on applying user requirements to a set of applicable rules (described in Section II). From this a workflow definition could be created, represented by a directed graph (Section III). A method was developed to produce a Petri net from a given directed graph (Section IV). The Petri net could then be analysed for flaws in the workflow definition (Section V).

II. Generating a List of Tasks and Dependencies

There are a number of different techniques for automatically generating workflows including Case Based Reasoning [12] and product structures, where the map is treated as a product that has to be constructed from component parts [13].

The technique used in this research was one that employs user preferences to define the selection and execution of a set of rules [14]. This technique was selected because of its focus on the user’s needs, which is essential for on-demand mapping.

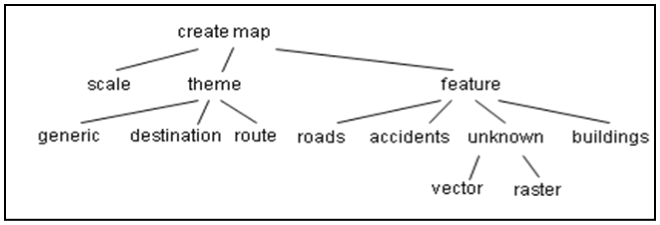


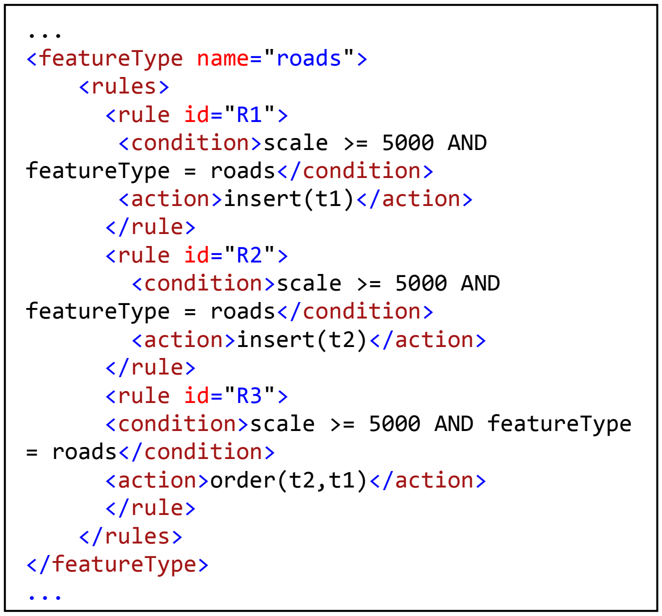
Figure 4. Knowledge hierarchy

The user preferences are gathered by navigating a knowledge/rule hierarchy (Fig. 4). If a particular branch is not selected by the user than that branch is closed off. For example, if the user does not select an ‘unknown’ feature type they are not prompted for ‘vector’ or ‘raster’. In the prototype the user is simply prompted for his or her preferences using text boxes and drop down boxes in a web page. Each leaf node in the hierarchy has one or more associated rules; these are added to a set of applicable rules as the user requirements are gathered.

If the user selects the ‘roads’ feature type then the rules associated with that feature type (R1, R2, R3) are added to the set of applicable rules. Rules consist of a condition and an action (e.g., *insert* a task to the workflow or *order* two tasks) and are stored in an XML file (Fig. 5). Using XML allows for the use of schemas to enforce correct structure.

The gathered user requirements are held in memory as ordered pairs, for example:

< scale,50000 >
< theme,generic >
< featureType,roads >
< featureType,accidents >



```
...
<featureType name="roads">
    <rules>
      <rule id="R1">
       <condition>scale >= 5000 AND
featureType = roads</condition>
       <action>insert(t1)</action>
      </rule>
      <rule id="R2">
        <condition>scale >= 5000 AND
featureType = roads</condition>
        <action>insert(t2)</action>
      </rule>
      <rule id="R3">
      <condition>scale >= 5000 AND featureType
= roads</condition>
      <action>order(t2,t1)</action>
      </rule>
    </rules>
</featureType>
...
```

Figure 5. Knowledge/rule hierarchy (partial) as XML

Once these have been collected, the applicable rules are then evaluated, checking rule conditions against user preferences to generate a set of tasks and a set of task dependencies. For example, if the user has selected the feature type “roads” and a map scale of greater than 1:5000 then the conditions for rules R1, R2 and R3 (Fig. 5) will be met and their actions triggered, e.g. task t1 is inserted into the workflow. The action ‘order(t2,t1)’ means task t1 is a dependency of task t2 and should only be run after t2 has been executed.

Using the above use case, the selected tasks may be:

t1: collapse roads
t2: delete minor roads
t9: add copyright notice for roads dataset
t3: cluster accidents
t8: reinstate minor roads on clusters

and the dependencies:

t2 ≺ t1
t1 ≺ t8
t3 ≺ t8

In this case, there are five tasks to perform and there are three dependencies (or precedence constraints). For example, we want to delete any minor roads (task t2) before we collapse the roads (task t1) as it is inefficient to collapse a subset of the road network that we are later going to delete. Task t9 is not involved in any dependency and is classified as an independent task.

The above output can be expressed as a directed graph where tasks are represented as nodes and dependencies as edges (Fig. 6).

However, the graph does not constitute a workflow. The next section describes why and then what is needed to construct a workflow definition.

III. Creating a workflow definition

The directed graph (Fig. 6) generated from the example set of task dependencies does not make up a workflow definition. This is because there is no place for any independent tasks (in our case task t9). In addition, the following rules must be satisfied for a workflow definition according to [15]:

- The workflow graph should have a single source node and a single sink node
- Every other node should have at least one parent and at least one child

This ensures that the workflow has a defined start and end and that there are no unnecessary tasks.

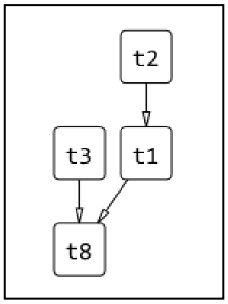


Figure 6. Directed graph based on dependencies

A workflow definition directed graph can be created by the following procedure:

1. Add start (A) and end (B) nodes
2. Create an edge for every dependency
3. For every node that has no children add an edge to the end node
4. For every node without a parent add an edge from the start node
5. For each independent tasks link the task directly to the start and end node.

The revised graph can be seen in Fig. 7.

The method so far has produced a workflow definition for the given case study but is it valid? For instance, it is relatively easy to ensure that there are no directly contradictory dependencies between the selected tasks so that both t1 ≺ t2 and t2 ≺ t1 did not appear in the same workflow. However, indirectly contradictory dependencies such as that seen in Fig. 8 would be harder to identify. In this example the dependency “t3 precedes t14” has introduced *deadlock* [15] into the workflow. Task t3 will not start until t8 starts; but t8 will not start until t14 starts, which will not start until t3 starts. So, tasks t14, t8, t5, t3 and subsequent tasks will never be executed.

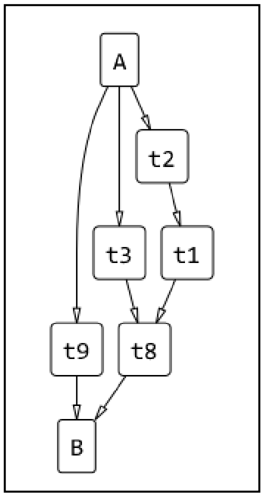


Figure 7. Workflow definition graph

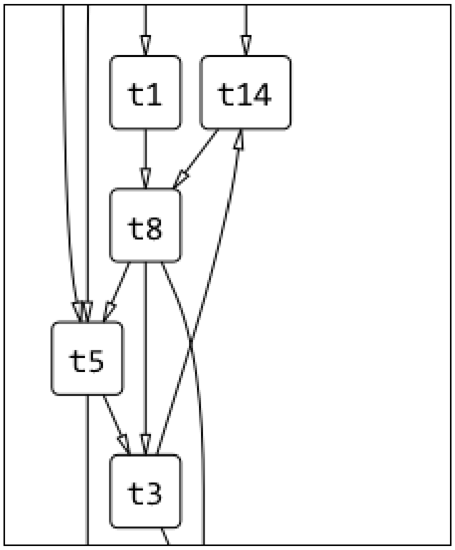


Figure 8. Deadlock in a workflow

A method of testing the soundness of a workflow before attempting execution is needed. The simplest way of checking for deadlock is by performing a topological sort on the graph. However, the application of Petri nets will allow for a more expressive form of graph and the ability to describe more complex workflow patterns [16] than that described above. In addition to describing workflows, the *mathematical foundations* of Petri nets [17] allow for the analysis of workflows and are applicable to more complex analysis than the deadlock problem [10][11]. Extensions to Petri nets, such as coloured Petri nets, which allow for the investigation of delays and throughput, have been defined formally [11]. Petri nets offer a number of advantages over PERT charts such as the ability to model nondeterministic behaviour and loops in the workflow [31]. The adoption of Petri nets at an early stage will allow the design to be scaled to more complex workflows. But, first, we need to generate the Petri net from the directed graph.

IV. Generating a Petri net

A Petri net is a particular class of directed graph, defined as a bipartite directed graph consisting of two types of nodes called transitions and places [17]. Nodes are linked by arcs such that arcs cannot link a place to a place or a transition to a transition. Transitions, denoted by rectangles, represent events or, in our case, workflow tasks. Places denoted by circles, represent states (Fig. 9).

Staines [18] describes the process for generating a directed graph from a Petri net, which can be reversed to generate a Petri net. The procedure used is as follows:

1. Nodes (tasks) are converted to transitions
2. Each edge generates arc-place-arc
3. Extra places are added preceding the start node (A) and following the end node (B).

The Petri net generated from the workflow shown in Fig. 7 can be seen in Fig. 9.

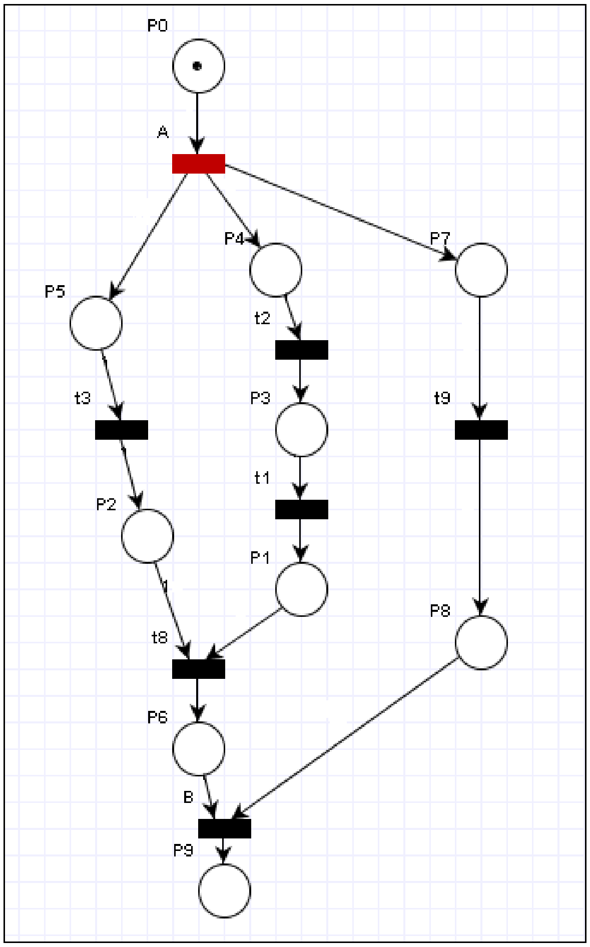


Figure 9. Petri net for valid workflow

The starting place, P0, contains a single *token*. Tokens can be used to model the workflow. A transition may be fired only if there are one or more tokens in all of its input places [17]. In this example, transition A can be fired. When a transition fires it takes a token from each of its input places and places a token in each of its output places. So after the firing of transition A, there will be a token in each of the places P5, P4, P7 (but no longer P0) thus enabling transitions t3, t2 and t9. Note that t8 will not be fired until both t3 and t1 have, which models the original dependencies.

Code was written to generate an XML file in a format that can be read by the PIPE software [19]. PIPE can then be used to visualise and animate the Petri net firings to ascertain whether the workflow is executable.

Deadlock can be identified visually or by using a Petri net analysis tool such as PIPE. It needs, however, to be identified as part of the on-demand mapping system. The following section describes how this was done.

V. Validating the workflow

Our system generates a *workflow net* [15], a particular type of Petri net such that:

- The net has a single source and a single sink node

- Every tasks lies on a directed path between the source and the sink nodes.

However, as has been shown, a workflow net containing anomalies such as deadlocks can still be generated. A *sound* process is defined as one that contains no unnecessary tasks and where every case submitted to the process must be completed in full and no reference to it remaining in the process, i.e., for every token that is in the start place there is one token in the end place and no others in the net [15].

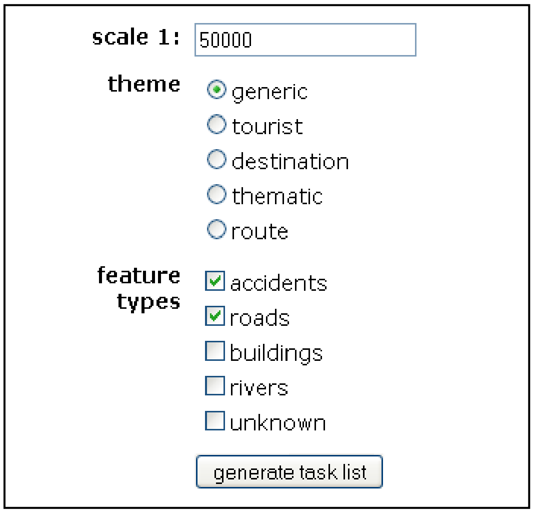
There is a number of, sometimes complex, techniques for checking the soundness of a workflow net. Fortunately the Petri nets derived from our workflow generation are a particular sub-class of Petri nets known as conflict free or T-systems where every place has no more than one input and one output transition [20]. In effect conflicts are ruled out; there are no logical ORs in the system. This makes them easier to analyse [21].

In the prototype the Petri net is checked for deadlock by looping through the set of places and firing any transitions that are enabled until no more transitions can be fired. If all the transitions are fired then the workflow is valid, if there are transitions that cannot be fired then these are listed.

In addition to the case study, the method was tested on a number of randomly generated task lists and dependencies. A demonstration version of the prototype can be seen at www.ondemandmapping.org.uk (Fig. 10).

VI. Conclusion and Future Work

The increasing availability of once inaccessible datasets and the explosion of crowd-sourced data, alongside the growth of web-based mapping, have led to the need for on-demand mapping. The requirement to integrate data from a number of disparate sources means that there is a need to automate the creation of the workflow required to generate such maps.



scale 1: 50000

theme
● generic
○ tourist
○ destination
○ thematic
○ route

feature types
☑ accidents
☑ roads
□ buildings
□ rivers
□ unknown

Figure 10. Prompting for user requirements

This paper has presented two aspects of automatic workflows; firstly the generation of the workflow from simple user specifications and secondly the generation of Petri nets from the workflow definitions to allow for their validation. In particular the work done so far has highlighted the potential problem of contradictory rules that can generate deadlocks in workflow definitions.

It was assumed that the generation of the workflow is a static scheduling problem, i.e., the workflow is deterministic, known in advance of execution [22]. This is likely to be a simplification of the on-demand mapping problem; it will be necessary to consider how the workflow may change during execution when, e.g., a particular generalisation service is not available at execution time. For this reason adaptive and autonomic workflow techniques [23][24][25] may need to be investigated. However, it could be argued that any replacement service or set of services would not affect the workflow if the replacement(s) could be represented as a sub-net with a single point of entry and a single point of exit to replace the failed service.

Further work is also required on the means for expressing the cartographic rules. For example, in the case study (Fig. 5) three rules had the same conditions but different actions. Could the rules be expressed more concisely? Also required is further investigation into how the rule base is to be populated and the knowledge hierarchy defined.

The execution of workflows will consist of calling a number of web services that provide generalisation operators. Web services are usually orchestrated using Business Process Execution Language (BPEL) [26]. Once sound workflow nets can be generated and validated using Petri nets it will be useful to investigate the process of generating BPEL from Petri nets [27][28].

Previous research into the orchestration of generalisation services in particular [29][30] will also need to be considered with a view to investigating how to integrate such services into the system.

A major problem with the work done so far is the lack of a data model. The method lacks the concept of tasks doing work on spatial datasets. Datasets have to be managed as they progress through the workflow and conflicts have to be handled when two different tasks attempt to work on the same dataset at the same time. One possibility may be to regard the presence of a dataset as a pre-condition to the firing of a transition. The transition would not fire until the dataset was available. The output from the transition would then be the processed dataset, e.g., a set of clustered accidents.

Whatever the eventual process is employed for generating the workflow, it is believed that the method described in this paper can be used to validate the workflow definition before an execution is attempted.

Acknowledgements

This project is funded by the Dalton Research Institute at Manchester Metropolitan University and the Ordnance Survey of Great Britain. Thanks to Martin Stanton of MMU for guidance on the use Petri nets.

REFERENCES

[1] J. Gaffuri, “Improving web mapping with generalization,” Cartographica: The International Journal for Geographic Information and Geovisualization, vol. 46, no. 2, January 2011, pp. 83-91.

[2] E.M. João, Causes and consequences of map generalisation. London: Taylor and Francis, 1998.

[3] Z. Li, “Digital map generalization at the age of enlightenment: A review of the first forty years,” Cartographic Journal, vol. 44, no. 1, February 2007, pp. 80-93.

[4] J. Stoter, et al., State-of-the-art of automated generalisation in commercial software. 2010. Available from: http://www.eurosdr.net/projects/generalisation/eurosdr_gen_final_report_mar2010.pdf 01.09.2011

[5] A. Ruas, and C. Duchêne, “A prototype generalisation system based on the multi-agent system paradigm,” in Generalisation of Geographic Information, A.M. William, R. Anne, and L.T. Sarjakoski, Eds., Amsterdam: Elsevier Science, 2007.

[6] M. Goodchild, “Citizens as sensors: the world of volunteered geography,” GeoJournal, vol. 69, no. 4, 2007, pp. 211-221.

[7] F. Grabler, M. Agrawala, R. W. Sumner and M. Pauly, “Automatic generation of tourist maps,” Proc. ACM SIGGRAPH 2008 papers, Los Angeles, California: ACM, 2008.

[8] K. Johannes, A. Maneesh, D. Bargeron, S. David and M. Cohen, “Automatic generation of destination maps,” ACM Trans. Graph., vol. 29, no. 6, December 2010, pp. 1-12.

[9] S. Balley and N. Regnauld, “Collaboration for better on-demand mapping,” in ICA Workshop on Generalisation, Paris, France, 2011.

[10] N. R. Adam, V. Atluri and W. K. Huang, “Modeling and analysis of workflows using Petri Nets,” Journal of Intelligent Information Systems, vol. 10, no. 2, March/April 1998, pp. 131-158.

[11] W. M. P. van der Aalst, “The application of Petri nets to workflow management,” Journal of Circuits, Systems and Computers, vol. 8, no. 1 , 1998, pp. 21-66.

[12] A. Aamodt and E. Plaza, “Case-based reasoning: foundational issues, methodological variations, and system approaches,” AI Communications, vol. 7, no. 1, 1994, pp. 39-59.

[13] W. M. P. van der Aalst, “On the automatic generation of workflow processes based on product structures,” Computers in Industry, vol. 39, no. 2, 1999, pp. 97-111.

[14] S. Chun, V. Atluri and N. Adam, “Domain knowledge-based automatic workflow generation,”, Proc. Database and Expert Systems Applications, 13th International Conference, Aix-en-Provence, France, Berlin: Springer, 2002, pp. 778-838.

[15] W. M. P. van der Aalst and K. M. van Hee, Workflow management models, methods, and systems, Cambridge, Mass.: MIT, 2004.

[16] N. Russell, A. ter Hofstede, W. van der Aalst, and N. Mulyar, “Workflow Control-Flow Patterns: A Revised View,” Technical Report BPM-06-22, 2006; Available from: http://www.workflowpatterns.com 01.09.2011

[17] T. Murata, “Petri nets: properties, analysis and applications,” Proceedings of the IEEE, vol. 77, no. 4, 1989, pp. 541-580.

[18] A. S. Staines, “Rewriting Petri Nets as Directed Graphs,” International Journal of Computers, vol. 5, no.2, 2011, pp. 289-297

[19] N. Akharware, PIPE - Platform Independent Petri net Editor 2. 2005; Available from: http://pipe2.sourceforge.net/. 01.09.2011

[20] J. Desel and J. Esparza, Free Choice Petri Nets, Cambridge: Cambridge University Press, 1995.

[21] P. Alimonti, E. Feuerstein, U. Nanni and I. Simon, “Linear time algorithms for liveness and boundedness in conflict-free Petri nets,” in LATIN '92 Lecture Notes in Computer Science, Berlin: Springer, 1992, pp. 1-14.

[22] J. W. Herrmann, C.-Y. Lee, and J. L. Snowdon, “A classification of static scheduling problems,” in Complexity Issues in Numerical Optimization, P. M. Pardalos, Ed, World Scientific Publishing Co.: Singapore, 1993, pp. 203-253.

[23] R. Muller, U. Greiner, and E. Rahm, “AgentWork: a workflow system supporting rule-based workflow adaptation,” Data & Knowledge Engineering, vol. 51, no. 2, 2004, pp. 223-256.

[24] M. Polese, G. Tretola and E. Zimeo. “Self-adaptive management of Web processes,” Proc. Web Systems Evolution (WSE), 12th IEEE International Symposium, 2010.

[25] G. Tretola, and E. Zimeo, “Autonomic internet-scale workflows,” Proc. of the 3rd International Workshop on Monitoring, Adaptation and Beyond, Ayia Napa, Cyprus, New York: ACM, 2010.

[26] OASIS. Web Services Business Process Execution Language v2.0. 2007; Available from: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf 01.09.2011

[27] P. Sun, C. Jiang and M. Zhou, “Interactive Web service composition based on Petri net,” Transactions of the Institute of Measurement and Control, vol. 33, no. 1, 2011, pp. 116-132.

[28] W. M. P. van der Aalst and K.B. Lassen, “Translating unstructured workflow processes to readable BPEL: Theory and implementation,” Journal of Information Software Technology, vol. 50, no. 3, 2008, pp. 131-159.

[29] T. Foerster, L. Lehto, T. Sarjakoski, L. T. Sarjakoski, and J. Stoter, “Map generalization and schema transformation of geospatial data combined in a Web Service context,” Computers, Environment and Urban Systems, vol. 34, no. 1, 2010, pp. 79-88.

[30] G. Touya, C. Duchêne, and A. Ruas, “Collaborative generalisation: formalisation of generalisation knowledge to orchestrate different cartographic generalisation processes,” Proc. of the 6th international conference on Geographic Information Science, Zurich, Berlin: Springer-Verlag, 2010.

[31] D. Dubois, K. Stecke, “Using Petri nets to represent production processes,” Proc. of the 22nd IEEE Conference on Decision and Control, 1983.

Methodology for the Collection and Handling of Geological Data

Rodrigo Ávila Cipullo
UnB – University of Brasília
Institute of Geosciences
Brasilia – DF, Brazil
racshalom@gmail.com

Henrique Llacer Roig
UnB – University of Brasilia
Institute of Geosciences
Brasilia – DF, Brazil
roig@unb.br

Abstract— **Universities and other geological information producers have been suffering the negative effects of a lack of data organization and standardization. The purpose of this work is to create a new spatial geology database and to present a new methodology concept for internal procedures of data acquisition. This proposal follows the precepts of Spatial Data and vector Structure (EDGV), which were approved in 2008 by the National Commission of Cartography (CONCAR), created by the Brazilian government in order to standardize the structure of spatial data, facilitating data sharing, interoperability and rationalization of resources between producers and users of data and cartographic information. This is an important step for developing a geological model database that can be embedded in the context of the National Spatial Data Infrastructure (NSDI/ INDE).**

Keywords— geology data; managing data; web services; web mapping; geotool.

I. Introduction

Geoscience is a branch of science that requires the collection and processing of large databases, especially regarding spatial and geological phenomena. However, this development is overdue for improvement in safe storage,allowing a more secure use of databases in collaborative environments.

Many steps are being taken, in Brazil and internationally,to create an interoperability culture and standard of spatial data. In Brazil, the Decree No. 6666, from 2008, instituted the NSDI (National Spatial Data Infrastructure), which aims at establishing metadata and interoperability standards for basic cartography in Brazil.

These types of initiatives attempt to carry out the interoperability of spatial data. In this context, the OGC (Open Geospatial Consortium) must be highlighted. The OGC is defined on its own website [9] as a non-governmental, nonprofit organization, formed by volunteers from around the world, who intend to establish standards for spatial data and services.

Despite many efforts to set standards, the amount of geological information could not match the pace of the evolution of its standardization, and one of the only projects with global prominence is the GeoSciML [10]. This project has the goal of providing a data interoperability architecture that allows institutions with the most different types of databases to exchange information without changing their structures. The GeoSciML is an initiative of major geological surveys in the world, such as the IUGS [11] and the British Geological Survey [12].

A. Objective

The aim of this paper is to offer a new model of spatial geological databases, a new methodology for geological project management and the development of a tool that enables the implementation of the proposed model and methodologies.

Not every goal will be achieved at this first stage of the work, and the products presented in this article are:

- A conceptual database model,
- A proposed methodology for acquisition and management of geological data, and
- Easy use of web application for data visualization.

B. Structure of the Paper

This paper begins with an abstract and an introduction (I) that includes this topic. The other content topics are structured as follows: Proposed Architecture (II), Proposed Methodology (III), Data Acquisition (IV), Conclusion and Future Work (V) and References (VI).

II. Proposed Architecture

The discussion about the ideal architecture for the provision of GIS has already been discussed by several authors. Harrower [5] defended that the use of the Internet would allow new services to GIS. According to Harrower, the Internet has revolutionized the way we work with cartography, featuring some important points that led to this new paradigm of GIS distribution:

- Ease of availability and distribution of cartographic products,
- Universal access to map data,
- Significant increase in demand for mapping services,
- Emergence of tools that allow the development of an "on-demand" application.

Another important concept presented by Harrower [5] is the “on-demand maps." According to the author, one of the main benefits of modern GIS is that it allows the user to manipulate and organize his or her own data, in which case the maps are not ready and static, but are constructed according to the needs of the user.

While we consider the globalization of access to GIS tools, we must also rethink the concepts of usability of these tools.

We have a growing public that is hungry for information, trained in a digital world, but does not necessarily have the adequate training to properly handle a GIS. The development of new GIS must consider this problem, during all the process from architectural definition to manufacturing the final product layout.

Another way that technology has taken hold of late is with cell phones and tablets. Increasingly powerful and with characteristics very similar to computers, these devices enable the technology to be part of every moment of everyday life.

Considering that a geologist uses several tools in a field study, it is possible that mobile devices provide the missing link between the data acquisition and storage/ distribution of information.

The use of mobile data acquisition not only represents a huge reduction in operational costs, but mainly simplifies the publishing process and data interpretation.

According to this work, this type of technology contributes greatly to the successful implementation of the new GIS, which allows the technology to spread with greater speed and makes possible the creation of a technological culture that is conducive to the advancement of real security and interoperability of geological information.

All points shown indicate the following architectural features:

- Centralized storage and web presentation,
- Centralized data processing,
- Acquisition of data using mobile devices.

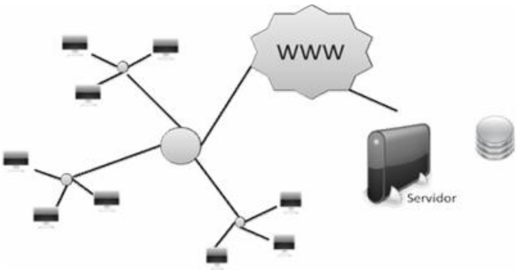


Figure 1. Architecture of centralized storage and processing type Legend: Servidor (Server)

A key to the success of the proposed methodology is to ensure that data is always available, and ensure that the database is always updated. For this, the tool uses web servers and the update will be done directly from mobile devices.

In order to provide this kind of service on the web, we need some essential tools, which are a web server and a map database with support for spatial data.

The web server application is responsible for providing a website or application hosted on one computer to another within a network, the global network known as the World Wide Web. The application server used is the Apache Web Server [14].

Map Server is the software that allows us to publish geographic data on the web. Through this application, we can provide a spatial database on the web through a series of specifications established by the OGC (Open Geospatial Consortium). Because they are published based in international open standards, the information can be reached by a wide variety of web and desktop software.

The application Geoserver [13] maintained by the Open Planning Project, was chosen in this project to serve the geological data, as shown in Figure 2.

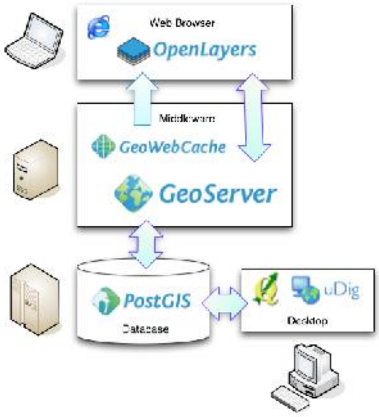


Figure 2. Basic publishing architecture of spatial data based on open source software [15].

III. PROPOSED METHODOLOGY

The proposed methodology is based on three steps:

- The standard remote data acquisition,
- Topologically-consistent database,
- Decentralized data management.

A. Data Acquisition

The data acquisition step represents the most important stage of the process, and the quality of data acquired following minimum standards is crucial to the success of the other steps. Thus, it is important that the data collection tool triggers a simple and standardized environment, ensuring flexibility in the collection that results in high quality data.

As shown in Figure 3, all acquired data are stored in a small SQLite database, exported by the system in CSV format and then imported by the Web tool.

This process will change in the next version of the mobile tool, when it will automatically synchronize the data with the server.

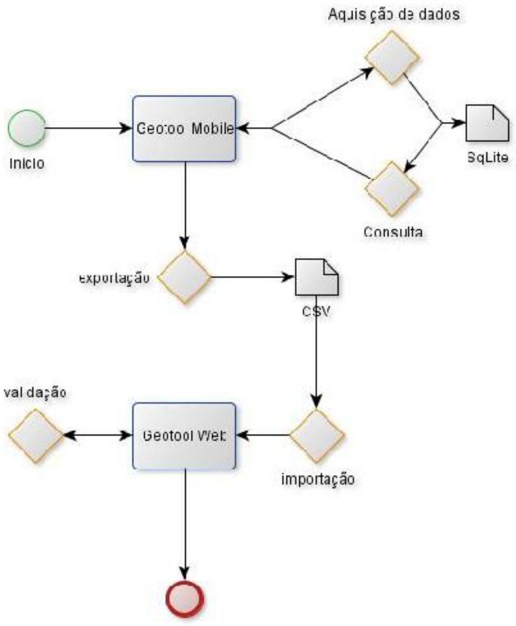


Figure 3. Flow data acquisition methodology. Legend: Início (Start); Aquisição da dados (Data Acquisition); Consulta (Query); Exportação (Export); Importação (Import); Validação (Validation).

B. Database

The second step of the methodology, the spatial database, was modeled using a methodology known as OMT-G.

The methodology OMT-G (Object Modeling Technique for Geographic Applications) was designed by Borges [4] based on one of the most popular models for modeling conventional databases, the OMT (Object Modeling Technique), which has the characteristic of representing the semantic aspects of data using an object-oriented approach. Thus, the OMT-G model, revised and extended by Borges [4], presents an object-based model that is also capable of representing objects and spatial relationships.

In addition to the OMT-G model, other proposals for modeling spatial data were created and must be remembered, such as the GeoOOA [1], MADS [2] and UML-GeoFrame [3].

The OMT-G brings together lots of geographic primitives proposed by various authors, as well as introduces new primitives that supply some deficiencies, such as the representation of multiple views from geographic entities. The OMT-G model is based on three main concepts: classes, relationships and spatial relationships [4]. This model works on the conceptual level as spatial classes featuring both conventional classes (Figure 8).

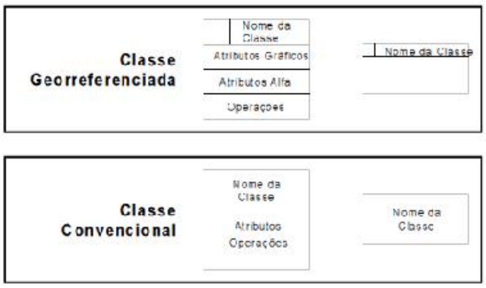


Figure 4. Types of classes and their representations in the OMT-G model. Borges et al. (2001). Legend: Classe georreferenciada (Spatial Class); Classe convencional (Conventional Class); Nome da Classe (Class Name); Atributos (Attributes); Operações (Operations).

Seeking an adequate representation of the types of spatial objects, spatial classes receive information from their geometry type. This information is known in the model as subclasses. The description of the main subclasses can be found in Figure 5.

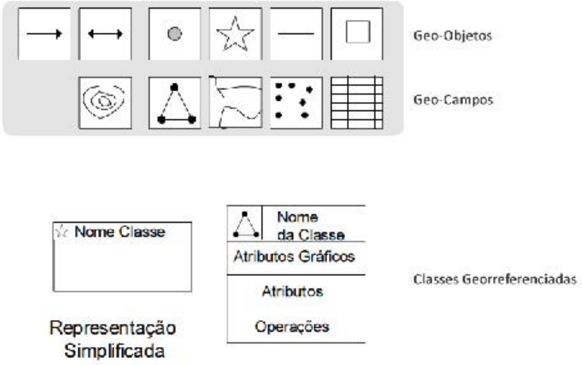


Figure 5. Georreferenced Classes model and its subclasses. Legend: Classe georreferenciada (Spatial Class); Nome da Classe (Class Name); Atributos (Attributes); Operações (Operations).

To generate the geographic database conceptual model, it was necessary to define some steps in the process of geological mapping. The following steps were also highlighted:

- Field survey (outcrops),
- Lamination,
- Geochronology,
- Geochemistry.

These steps were defined as superclasses of the model. These super classes do not have a real implementation, serves only as aggregators towards a better understanding of classes.

Each of the major classes that would be needed to correctly represent the stage of mapping emerge from the super classes. These major classes are presented in Table 1.

TABLE I. MAIN CATEGORIES

Superclass	Class	Description
Afloramentos (Outcrops)	Descriptive data	General data of the outcrop, such as name and description
	Toponymy	Information on the location of the outcrop that help visually its correct positioning.
	Geographical location	Information on latitude and longitude of the outcrop. It can be understood as coordinates west and south in the case of planar coordinates.
	Structural measures	Measures related to tectonic structures or primary structures.
	Photos	Storing pictures of the outcrops in question.
Amostras (Sample)	Samples	Samples related to an outcrop.
Laminação (lamination)	Lamination	General data such as name of the blade, which is owned and outcrop description.
	Paragenesis	Data on the composition of the blade modal.
	Photos	Storing pictures of the blade.
Geocronologia (Geochronology)	Geochronology	Basic information such as responsible for analysis, sample origin and age.
	Dating method	Method used in the analysis
Geoquímica (Geochemistry)	Geochemistry	Information on sample origin, responsible for analysis and consideration.
	Analysis	Types of analysis performed on the sample
	Results	Results of sample analysis. Load the type of analysis, the sample, the result, the laboratory and all considerations.

The next step is to define all the spatial classes. Each one of these classes will be implemented in the physical model. Following are the identified spatial classes.

TABLE II. SPATIAL CLASSES

Spatial Class	Description	Geometry type
Afloramento	Outcrops	Point
Estruturas (structures)	Tectonic structures	Linestring
Unidade Geológica (Geological units)	Differential unit of the crust for their compositional characteristics, age and physical boundaries	Polygon
Projeto (Projects)	Representation of the project area.	Polygon
Grupos (Groups)	Represents the work area of each work group.	Polygon
Contatos (Contacts)	Geological contacts	Linestring

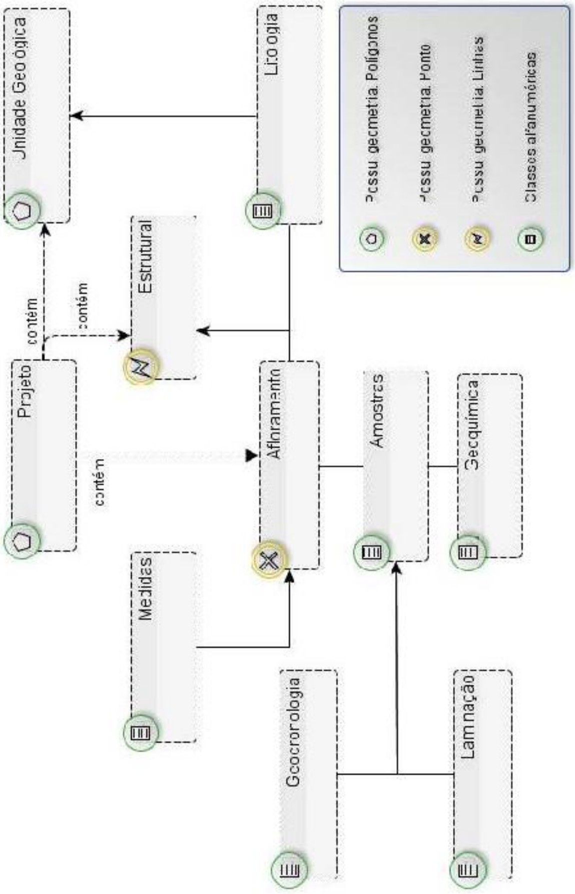


Figure 6. Model class indicating geometry
Legend: Table 1 and Table 2.

From the mapping of classes, a known model diagram class (Figure 6) was generated. This model was adapted to allow viewing the spatial types for each class. This adaptation allows us to clearly differentiate the non-spatial classes from others and display a preview of the spatial relationships that are present in the database.

Mainly to create the conceptual model of the geological database, each of the super classes have worked individually, especially in the quest for spatial delimitation of its features and possible relationships with other classes.

The first analyzed superclass, the spatial class project, aggregates all the others, functioning as the parent class in the model. It necessarily contains all other geometric classes of the database and represents the spatial delimitation of the mapping project. The class project represents a polygon geo-object, as already shown in Table 2.

The mapping project is formed from their groups. Groups are small areas that together, in the end, will aggregate to the final project area. Thus, all the mapping work is performed within a group. In practice, all classes are tied to the bank group, which is in turn aggregated by the project.

C. Data Management

The third step deals with the proposed availability of data. A tool able to manage its own projects and data was developed, allowing the publication of these data in other systems capable of utilizing WMS services.

1) Functional Requirements

- Access Control
- Management of users and permissions
- Project Management
- Creating projects and groups (subprojects)
- Definition of the project area and groups (subprojects)
- Outcrop Management
- Insert, edit and query outcrops
- Insert structural measures
- Photos
- Insert photo samples
- Entering, editing and deleting contacts in the sphere of geological group (subproject)
- Entering, editing and deleting slides related to the samples
- Management Isotope geochemistry samples held in
- The system must provide dynamic queries (spatial or not) and reports.

2) Non-functional Requirements

- User-friendly interface and simplified access to user data in a project,
- Light and simple software to compensate for the large volume of information.

3) Business Rules

- The user should be able to insert outcrops and contacts only within the boundaries of the area of his/her group,
- The user should be able to insert outcrops and contacts visually by map or via a form.

4) Mapping and Projects Management System

The main product developed, the management system of mapping projects, is an application that runs in a web environment and is able to manage since the creation of a new design until the closing of the geological map.

To expedite the development of this application and ensure easy maintenance in the code, we used the Zend framework [6]. The choice of this framework is given for the following reasons:

- This design pattern isolates the application logic from presentation logic. It allows them to be tested and/ or modified separately, reducing development time and enabling better reuse of code;
- Object-oriented library that allows easy extension and reuse of code;
- Abstraction of the database, allows for the automatation of a common operations database as well as decrease the impact of changes in the model;
- Automating operations AJAX and JSON, facilitating the task of integrating with the API for manipulating spatial data OpenLayers [7];
- Framework is maintained by the same company responsible for the PHP language, which makes it very popular, facilitating future maintenance.

In addition to the framework used in the development of the PHP application, some other open source APIs were used in order to improve user navigation. They are the following:

Openlayers [7]: Set of open source tools, available as a JavaScript API for viewing and manipulating spatial data based on OGC standards.

JQuery [8]: This is one of the most powerful JavaScript API today. The main objective of JQuery[8] is to provide a web browsing experience and practice based on modern concepts of AJAX (Asynchronous JavaScript and XML)

5) Restricted Access and Security

To ensure data security and proper control of projects, we implemented a system of user verification. The user is identified by his/her e-mail address and password (Figure 11). Immediately upon entering the system, the user will be prompted to choose which project to work on. Thus, he/she can work on any project that is registered.

Figure 7. Screen to access the system.

The definition of access permissions, registration of users, projects and subprojects in the system itself is made by an admin user. The system allows a user to be registered on several projects, however, the user must belong to a single subproject.

Once registered, the data remain in the system. This means that even if someone has access to a user account and makes lots of changes in it, or even delete some data, these data can be retrieved by the system administrator through a restitution version.

It is important to note that all project data is nested in its subprojects and always related to a registered user in a subproject. If you do not have a subproject it should be created so that the work can be started.

6) *Features*

The system is designed in modules, that is, it is possible that some features are only available to a user according to his/her profile. Modularity also facilitates system maintenance and code reuse for other future projects.

The modules developed so far include the major classes of database:

- Outcrops: Responsible for managing user outcrops in a particular subproject. The basic function is to list all the outcrops of the user on a given project and allow him/her to add, modify or delete outcrops. From the list of outcrops, the user can view all information related to an outcrop, including its visualization on the map.

  In addition to the basic features of the outcrop, several sub-modules that add functionality to the system were developed. They are:
  - Measurements: Stores all measurements taken at the outcrop. This module not only allows the listing and input measures, but measures of export in text format to be opened in structural geology software.
  - Samples: Relates physical samples to the outcrop.
  - Petrography: Stores blades from the outcrop samples.
  - Geochronology: Responsible for keeping the data for isotopic analysis performed on samples.
- Map: Module that aggregates spatial data and allows viewing and editing data.
- Structures: Displays all the structures of the related group and allows creating and editing new structures and contacts.
- Topology and closing contact: Validation responsible for implementing topological rules and close contacts between groups. Such procedures are necessary before generating the geometry of geological units, which are automatically created through spatial operations.

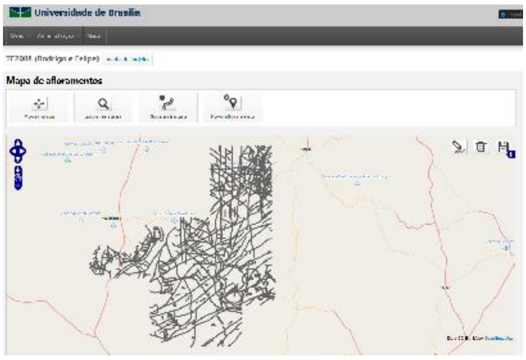


Figure 8. Module map showing some cartographic tools

IV. Conclusion and Future Work

The application used for deploying PostgreSQL is sufficiently mature and stable to include large databases like the one here proposed. Its spatial component, PostGIS, responds very well to spatial analysis, demanding, however, even more studies and tests on their topological functions in order to overcome some difficulties that persist.

It was hoped an algorithm that would allow the automatic generation, at runtime, of geological units. However, none of the generated algorithms were able to properly turn contacts into areas and receive individual attributes of geological units correctly. This feature was only partially implemented and needs further studies.

Applications developed for the management of geological data responded positively to the tests in which they were submitted. However, there are security-related points that remain to be discussed further, before using these tools on a large scale.

References

[1] G. Kösters, “GIS-Application development with GeoOOA”. Int. Journal of GIS, 11, 1997

[2] C. Parent, “Spatio-temporal conceptual models: data structures + space + time”. In Proc.7th ACM GIS, Kansas City, 1999.

[3] J. L. Filho, A. C. Costa, and C. Iochpe, "Projeto de banco de dados geográficos: mapeando esquemas GeoFrame para o SIG Spring". In Proc. GEOINFO – 1st Brazilian Workshop on geoinformatics, Campinas, 1999.

[4] K. A. V. Borges, C. D. Davis Jr and A.H.F. Laender. "OMT-G: an object-oriented data model for geographic applications". GeoInformatica, 5, 2001.

[5] M. Harrower, "A look at the history and future of animated maps". In: Cartographica n. 39: 2004. pp. 33-42.

[6] http://framework.zend.com (retrieved: August, 2011)

[7] http://openlayers.org (retrieved: November, 2011)

[8] http://jquery.org (retrieved: December, 2011)

[9] http://ogc.org (retrieved: December, 2011)

[10] https://eegrid.csiro.au (retrieved: November, 2011)

[11] http://iugs.org (retrieved: November, 2011)

[12] http://bgs.ac.uk (retrieved: November, 2011)

[13] http://geoserver.org (retrieved: December, 2011)

[14] http://apache.org (retrieved: December, 2011)

[15] http://opengeo.org (retrieved: December, 2011)

Comparison Between Generalization Processes at Large and Small Scales

Jacqueleen Joubran Abu Daoud
Faculty of Civil and Environmental Engineering,
Department of Transportation and Geo-Information
Engineering, Technion- Israel Institute of Technology,
32000 Haifa, Israel
e-mail: jacquele@tx.technion.ac.il

Izabela Karsznia
Faculty of Geography and Regional Studies,
Department of Cartography, University of Warsaw
Krakowskie Przedmiescie St., 00-927 Warsaw, Poland
e-mail: i.karsznia@uw.edu.pl

***Abstract*—The presented research touches on the questions of digital cartography, particularly on automated map generalization in advanced geographic information systems. The aim of this article is the investigation and comparison of generalization processes of settlements at different detail levels. Two generalization models are described and verified at two different detail levels. On the basis of this the directions of future research are outlined.**

Keywords-* *generalization; small-scales; large-scales; generalization models.

I. Introduction

Although automation of cartographic generalization has been an extensive field of research [3][7][11[14][15], there still remains a lack of a usable holistic generalization method. A holistic process that makes it possible to generalize the whole map including all the layers, to take into account the connections between the layers and to deal with levels of detail in the small and large scales at once. More recently, the demand for automated map generalization, which has been longstanding in the context of conventional GIS (Geographical Information Systems), has been reinforced by the prevalence of geographical information access on the Internet, that make it more complicated. There are several types of public access map-based Web sites that allow a user to zoom in and zoom out of a particular region, but at presently, this is usually based on stepping between independent pre-processed generalized datasets which may differ markedly in their degree of generalization. It would be desirable to be able to change the level of detail on such systems in a smooth and progressive manner rather than the quantum-leap changes that often characterize the current approach [12].

The aim of this article is to investigate and compare generalization processes on different detail levels. The author's intention is to describe and compare two generalization processes of one thematic layer, which are settlements. The first process touches large-scale generalization, from 1:10 000 to 1:50 000 and the second one concerns small-scale generalization, from 1:250 000 to 1:500 000. Two different models of generalization processes are proposed: one based on electric fields theory implemented in MATLAB (by MathWorks) and the second based on mathematical morphology implemented in Clarity (by 1Spatial). The main point of interest is to show the specifics of the generalization process at large- and small-scale elaborations in terms of characteristic of the generalized thematic layer and also characteristic algorithms and tools used at different level of details. The authors want to investigate whether it is possible to build one comprehensive model to manage the on-line generalization process at different levels of detail.

With respect to the goals of the article in section II the authors present the different problems, which a cartographer has to deal with, on different detail levels. In Section III the authors concentrate on some significant differences, as well as some similarities, between generalization processes at large and small scales. In Section IV two generalization models, for both detail levels, are presented and discussed. Finally in Section V the authors conclude their research and they point at the future research directions.

II. Different problems on different levels of detail

Cartographic generalization is a decision-making process aimed at reflecting the purpose of a map or database and emphasizing characteristics and relations of generalized objects. Due to its holistic nature, the generalization can hardly be transferred into a process of sequences of tasks which might be applied in computer environment. The necessary condition of de-composing the generalization process into tasks sequence is a formalization of cartographic knowledge [13].

In many countries, there are specifications of map redaction with additional remarks on the generalization process for topographic maps. Based on them, a cartographer is able to collect and formalize knowledge about the generalization process at the large scales. Those map specifications are the source of important constraints like: threshold values, minimum or maximum values of distances.

Unfortunately, a dominating part of existing elaborations touches both the maps and spatial data generalization expressed in large scales [1], [2]. [10]. By large-scales elaboration, we understand maps and databases at the scales from 1:500 up to 1:50 000 while as a small-scale elaborations we consider maps and databases from 1:200 000. The reason for that can be placed in the wide practical application of such kind of data. Basic spatial databases from country levels have been expressed just in the scales of 1: 10 000, 1: 25 000 and 1: 50 000 and hence the need for their automated generalization appears.

On the contrary, the generalization at small scales depends, in general, on the experience and knowledge of a cartographer. The result of the process depends on, not always consequent, decisions of its author. As a result, maps in particular scales or a spatial data of the same level of detail may differ from each other both in a range, as well as in a level of generalization. Due to the subjective character of the small-scale generalization process, none of the precise instructions of its redaction were elaborated on until now: what makes it significantly difficult to collect knowledge about the process, its formalization, and implementation in GIS systems.

III. Large *versus* small generalization

Generalization methods and processes have been changed and improved alongside development in the science and art of cartography and have been surely influenced by progress in computer science [8].

In the process of cartographic generalization at every level of detail we can point at four main stages:

- Selection of categories of objects (object classes) presented on a map and their classification.
- Selection of objects within particular categories.
- Change in a cartographic method of representation – replacing an outline of area feature with a signature [9].
- Simplification of built-up area’s outlines.

With respect to the holistic character of a generalization process, in this article, we concentrate on one thematic layer - settlements. The reason for that is that most cartographers consider this thematic layer as the most important and at the same time the most difficult one.

The above mentioned four generalization stages will be applied differently at large and small scales elaborations. Both in large and small-scale generalization processes, the context of geographical information and topologic relationships are to taken into consideration.

A. Characteristics of the generalization processes at large- and small-scales

There are significant differences between large- scale and small-scale generalization processes. First of all at large scales (in this article we treat 1:10 000 scale as a source) settlements are presented as separate buildings, while at small-scale elaborations, all settlements are presented as signatures, and additionally the ones of them which are highly-populated are presented also as outlines (these are built-up areas). So, in the small-scale data model settlements are placed in two thematic layers.

In the small-scale generalization process in order to achieve desirable cartographic results, the following operators need to be applied. First and the key generalization stage is a selection of information which concerns both settlements presented by signatures and presented by outlines. A selection operator is based on attribute information concerning the importance of particular settlements, like population, administrative meaning, and area for built-up areas. It is also often connected to the spatial characteristic of a whole settlements’ network like density of settlements.

On the other hand, while generalizing at large scales, the goal of the process is to show the presented data with more details without spatial conflicts that destroy the correctness and the reality of the data. A selection operator is based on attribute information concerning the importance of a particular map object which is affected by its geometric and attributes properties and its topologic relation with other objects.

Another important operator during the small-scale generalization process of settlements presented by outlines is aggregation. At the small scales the goal is to put together all parts of the same locality (city) presented by outlines. The parts of a locality are usually aggregated based on two conditions: the distance between them and the name of the locality. In this article, we use aggregation operators originated from mathematical morphology (erosion/dilation) to keep the characteristics of shape of the outlines after the generalization process. At large scales this operator also plays a very important role as we aggregate the relevant buildings, based on its functions and the distance among them, in order to create built-up areas. At the same time those two aggregation processes differ in kind. At large scales, buildings will be aggregated by moving them to each other in order to decrease the presented area and increase the free area between them, while at such a scale the movement distances are small relatively. Another important difference, especially according to its relation to other thematic layers like roads is that in the large-scale generalization process, during building aggregation it is not allowed to aggregate together the buildings lying on the other side of the road. While at the small scales aggregation we are allowed to aggregate built-up areas (parts of one city) even if they are shared by the roads.

One of the last-used operators in the generalization of settlements, in both large- and small-scales processes, is a simplification operator. It is used in the generalization process of the outlines of built-up areas at small-scales but also for building simplification. The difference here lies in the application of a relevant algorithm of simplification. At small-scales where the goal is to simplify the irregular outline of built-up area we need to apply a simplification algorithm which lets us do the shape simplification while at the large scales we can, for example, use both simplification and squaring algorithm for buildings generalization.

IV. Limitations of proposed approaches

The main issue in the automated generalization process is the formalization of rules and cartographic constraints definition, which makes it possible to obtain correct cartographic results.

A. Small-scale generalization model

The concept proposed by Karsznia [5], [6] comprises the collecting of cartographic data, its formalization and implementation in a form of knowledge base in the Clarity system.

In the first stage, cartographic knowledge concerning small-scale generalization of settlements was collected and formalized in a form of a rules sequence for 1:500 000 level of detail. As a result, a knowledge base concerning generalization process of settlements was elaborated.

The knowledge base in the Clarity system [5][6] consists of generalization activities together with their implementation in a form of either algorithms or respectively, generalization tools. For that purpose, the available system functions have been used as well as new algorithms (by Java programming) and spatial analyzes tools have been proposed.

An important element of a knowledge base in this system was the development of the algorithms *cluster settlements* and the application of *action polygon erode* algorithm (supplied by Ordnance Survey), made it possible for more correct results (from the cartographic point of view) to be obtained from the aggregation of part of the built-up areas presented as outlines. In Clarity it is possible to aggregate built-up areas on the basis of distances between them, by using the algorithm *clustering*. Unfortunately, this may lead to connection of even a few different localities. The modification of the built-up areas aggregation algorithm in Clarity made it possible to connect selected parts of localities under the condition that the same name be used, and the limitation of an assumed distance between them applied. As for the algorithm *action polygon erode* deriving from mathematical morphology, its application allows the proper shapes of objects to be kept, after generalization (Fig. 1).

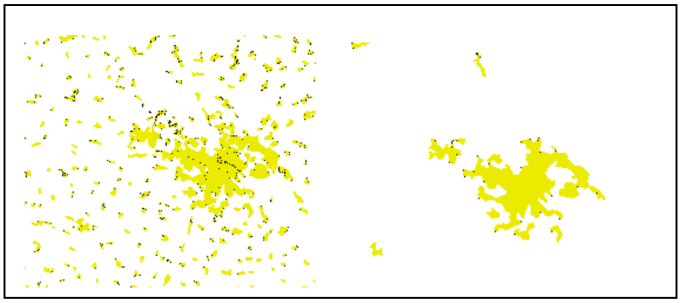


Figure 1. Aggregation results of settlements. Southern Poland

B. Large-scale generalization model

In large-scale elaborations where more details are needed and showed, more spatial conflicts arise and need to be solved.

`In [4], an automated process was modeled based on a sub model of a neural network to set relative importance values of the maps' objects taking into account the object's properties, its surrounding area and the map's target. Electric theories were implemented in the MATLAB environment in order to formalize the dynamic maps' object behaviors during the process of generalization.

Each object during the automated generalization process is treated individually; spatial analysis is implemented to define the object’s cartographic characteristics and topological relationships with its nearby objects that should remain. The electric model set powers for each object that expressed the relative importance of the cartographic objects in the treated map according to the object properties and the map target and required scale. The interaction between the objects' powers produces forces that act in order to solve spatial conflicts and insure clear presentation of the cartographic data in the required map. The translation of the forces into generalization operators is done with the respect of the cartographic rules and constraints dealing with the objects details and connections. The cartographic rules and constraints were set and formalized by setting a sub-model of neural network that learned previous cartographers and users decisions from the input training datasets. The process of the forces action and the presentation preparing contains of three main stages: 1). simplification of each map's object and deletion of minor objects, 2).clustering close objects of the same type according to the cartographic layer properties, taking into account topologic connections (Fig 2) and 3). movement and reshaping objects in order to solve spatial conflicts, results demonstrated in Fig 3.

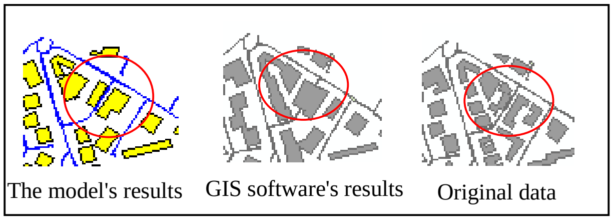


Figure 2. The aggregation results considering the roads layer and compared with GIS software results

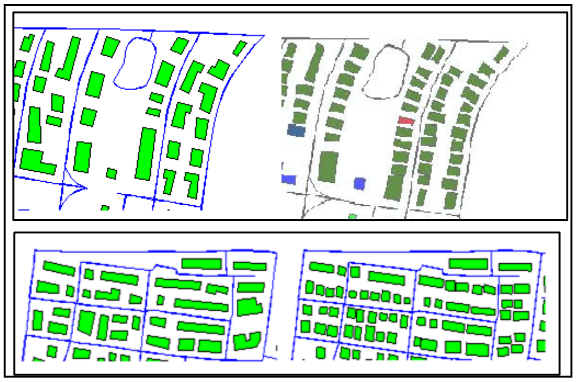


Figure 3. The results of conflicts solutions by MATLAB in the lef

V. Conclusion and Future Work

Successive and satisfying results were produced especially for large scale maps (Fig. 2), where the model elaborated in MATLAB succeeded to generalize the data taking into account the building’s properties and clustered buildings of the same kind, the model as demonstrated in the same picture for the two cases preserve the characteristics of

ISBN: 978-1-61208-178-6

the original data. What is more settlements presented by outlines at small scale were also aggregated and simplified successfully There were thematic attributes of the settlements taken into account (aggregate settlements from the same city). Fig 4, in the left side demonstrates the original data of middle Poland, while the right side shows the results of selection stage by MATLAB according to threshold of minimum area 1000000 m, and illustrates the results of aggregation taking into account the city name and the allowable distance 250 m with the respect to the knowledge base built by Karsznia [5], [6] at 1:500000 scale.

The comparison of small-scale generalization results obtained within MATLAB and Clarity makes it possible to formulate few interesting conclusions.

- Aggregation algorithms implemented in Clarity (Fig 4), based on mathematical morphology operations: erosion and dilation, made it possible to obtain more proper results in terms of keeping shape characteristic than algorithms implemented in MATLAB system (Fig. 5).

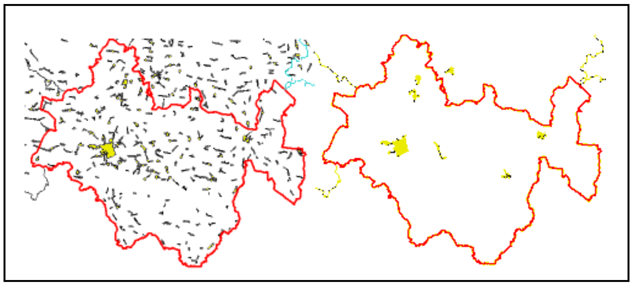


Figure 4. The original data of middle Poland at the left, and generalized ones in Clarity

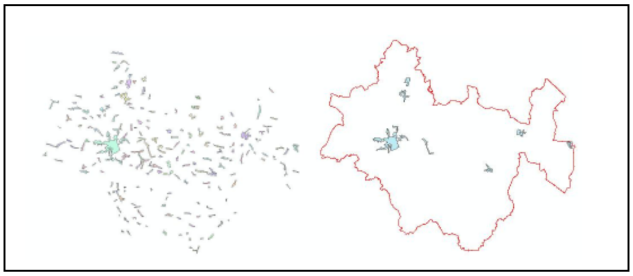


Figure 5. The original data of middle Poland at the left, and generalized ones in MATLAB

- Both aggregation algorithms, implemented in Clarity, based on mathematical morphology and also in MATLAB, made it possible to obtain more proper results in terms of keeping shape characteristic than other GIS software
- Automated generalization process by MATLAB succeeded to take into account the features properties and topological connections and produce satisfying results according to other GIS software.
- The specific character of both generalization process of small-scale and large-scale maps demands in many cases different solutions of the same problem depending on the context and objects' surrounding. In this context, important constraint is a lack of algorithms and tools having a context-like character which would make it possible to implement generalization steps on a higher conceptual level.

The conducted experiments proved that building one comprehensive generalization model to manage on-line generalization process at different levels of detail is a difficult but at the same time, challenging task.

In order to do more detailed comparison between large and small scale generalization on the way of building comprehensive generalization model, another experiment in being carried out. The cartographic knowledge of large scale generalization process collected in polish map specifications is being formalized at the moment and it will be implemented within Clarity based on Israeli data at 1:10 000 scale.

ACKNOWLEDGMENT

The authors want to thank to the Foundation for Warsaw University for financial support of presented research and to 1Spatial for Radius Clarity licensing as well as supporting during the research work.

REFERENCES

[1] Bildirici O., 2004, “Building and road generalization with the Change generalization software using turkish topographic base map data”. Cartography and Geographic Information Science, Vol. 31, No 1, pp. 43–54.

[2] Hardy P., and Lee D., and Van Smaalen J., 2008, “Practical research in generalization of european national framework data from 1:10K to 1:50K, exercising and extending an industrystandard GIS”. ICA Workshop on Generalization and Multiple Representation, Montpellier. http://aci.ign.fr/montpellier2008/papers/19_Hardy_et_al.pdf

[3] Harrie, L., 2003. Weight – setting and Quality Assessment in simultaneous Graphic Generalization. The Cartographic Journal, Vol. 40, No. 3, pp. 221-233.

[4] Joubran, A.J. and Doytsher, Y., 2008. “An Automated Cartographic Generalization Process: A Pseudo-Physical Model”. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2. Beijing, China, pp. 419-424.

[5] Karsznia I., 2011a, “Methodical principles of automation in the generalization of selected *general geographic databases elements”*. Proceedings of the XXV International Cartographic Conference, Paris.

[6] Karsznia I., 2011b, “On automatic generalization of General Geographic Database in the applications of geographic information systems”. Geographical Journal. (in press).

[7] Kilpelainen, T., 2000. “Knowledge Acquisition for Generalization__Rules”. Cartography and Geographical Information System, Vol. 27, No. 1, pp. 41-50.

[8] Mackaness W. A., and Ruas A., and Sarjakoski L. T., 2007, “Observations and research challenges in map generalization and multiple representation”. In: W.A. Mackaness, A. Ruas, L.T. Sarjakoski (Eds.), Generalisation of geographic

information: cartographic modelling and applications. Oxford: Elsevier, pp. 315-323.

[9] Ratajski L., 1967, *Phenomenes des points de generalisation*. „Intern. Yearb. Of Cartography" , Vol. 7, pp. 143–151.

[10] Revell P., 2005, "Seeing the wood from the trees: generalising OS MasterMap tree coverage polygons to woodland at 1:50 000 scale". ICA Workshop on Generalization and Multiple Representation, A Coruna, Hiszpania. http://aci.ign.fr/Acoruna/Papers/Revell.pdf

[11] Sester, M., 2005. "Optimization Approaches for Generalization and data Abstraction". International Journal of Geographical Information Science, Vol. 19, No. 8-9, pp. 871-897.

[12] Sester, M. and Burrener, C., 2005. "Continuous Generalization for Visualization on Small Mobile Devices". Proc. Conference on Spatial Data Handling –Springer, pp. 355-368.

[13] Steiniger S., 2007, "Enabling pattern – aware automated map generalization". PhD thesis, University of Zurich.

[14] Stieniger, S. and Weibel, R., 2007. "Relations among Map Objects in Cartographic Generalization". Cartography and Geographic Information Science, Issue 34 (3), pp. 175-179.

[15] Weibel, R. and Jones, C. B., 1998. "Computational Perspective on Map Generalization.",GeoInformatica, Vol. 2, No. 4, pp: 307-314.