Analysis of Clustering and Unsupervised Learning of Geospatial Demographic Data

Mikhail Kanevski and Jean Golay
Institute of Earth Surface Dynamics
Faculty of Geosciences and Environment
University of Lausanne, Switzerland
Mikhail.Kanevski@unil.ch, Jean.Golay@unil.ch

***Abstract*—The research discusses the methodological framework of an application of a newly developed spatial clustering algorithm – the functional multipoint Morisita index (fm-Morisita) - and an unsupervised learning algorithm – self-organizing Kohonen maps (SOM), for the comprehensive exploratory analysis and quantification of patterns in high resolution geospatial demographic data in Switzerland. fm-Morisita is used to analyse the complex clustering of the spatial distribution of the population. The SOM are used to reveal regional patterns of similarity using detailed information about ageing groups.**

Keywords - geodemography; unsupervised learning; spatial data clustering.

I. INTRODUCTION

The basic motivation for this research is to explore the structure and evolution of demographic and other geospatial data of the "SuisseMetropole" (SMP) using cutting-edge methods and algorithms from machine learning (mainly unsupervised learning and dimensionality reduction tools [1,2]), geocomputation (city clustering algorithms [3]) and spatial clustering using topological, statistical and fractal measures [4]. In this work, urban zones of Switzerland are considered as a "SuisseMetropole", i.e. a complex multivariate, multiscale (from hectometric intra-city to inter-city level), and high dimensional hierarchically organized system. One of the main intentions is to rely on real geodemographic data as much as possible, following the principle "let the data speak for themselves" in making as few prior hypotheses and assumptions as possible by following an unsupervised learning approach. Geocomputational approaches will help to carry out analysis at multiple selected scales and to better understand the structural and functional optimality of the "SuisseMetropole". A subsequent evaluation of the relevance of the results for the fields of urban, economic and social geography will be carried out.

Fundamentally, the first problem is to quantify and to understand the clustering of populated areas using a recently developed method – the functional multipoint Morisita index (fm-Morisita). The second one deals with the analyses of data in a high dimensional space (dimensionality >10-100) in order to understand the geodemographic patterns of "SuisseMetropole". Below, the data and the main methods proposed and how they will be applied are shortly discussed.

II. DATA

The Swiss Federal Statistical Office (OFS) and the Swiss Federal Office of Topography (Swisstopo) have provided the main data required by the research plan. OFS data are organized on a regular grid of high resolution (100m x 100m) and they can be subdivided into different sections. For the present research, the following data sets were mainly used: population census of years 1990, 2000 and 2010. They contain variables about the demographic and socio-economic characteristics of the population: resident population organized by language, place of residence, gender, age (19 classes), employment; information about education, socio-economic status, mean of transportation and households composition. The global distribution of the population in 2010 is given in Fig.1. The simulated data with known clustering structures (i.e. complete spatial randomness) were generated within the validity domains and the results were compared with the original geodemographic data (see Fig.2).

III. METHODS

Complex and challenging data demand the application of advanced data modelling tools, including machine learning algorithms.

In the present research, the two main methods applied for the study of Swiss complex geodemographic data are based on 1) machine learning – self-organizing maps (unsupervised learning) and 2) geospatial data clustering algorithms –the Morisita index and fractal measures. [1,2,4,5,6]. The basic idea is to find and to model geospatial patterns (structures in space) in the high resolution demographic data. The first approach - SOM is used to reveal patterns in the 19^{th} dimensional space of the Swiss demographic data composed of 19 classes of age groups (0-5, 5-10,…,>90 years) and to visualize them using Geographic Information System in a two dimensional space. SOM is a well know exploratory analysis and visualization tool making a topology preserving mapping (projection) from high dimensional space to two dimensional space. An important advantage of SOM is that it is an unsupervised algorithm and does not need a priori knowledge about the

 ISBN: 978-1-61208-326-1

number of similarity clusters in data. Here, the geographical information (i.e.the coordinates) were left aside and only the demographic data were used.

The second objective of the research takes into account the distribution of the population in the two dimensional geographical space using fm-Morisita and fractal measures of clustering in order to understand the spatial distribution of the populated areas [5]. In [5], fm-Morisita method was first time introduced for spatial data and it's the relationships of the multipoint Morisita index (on which fm-Morisita is based) with multifractality was demonstrated.

The multipoint Morisita index is calculated as follows: the region of interest in covered by non-overlapping cells of size l and the probability to find m points in a cell is calculated. Cell size and m are the free parameters. When taking into account the density of the population (not only the distribution in space) functional measures (depending on local density) based on the multipoint Morisita index can be implemented. The details can be found in [4,5]. In order to demonstrate that observed patterns are really structured, a comparison with point distributions generated generated by shuffling the original data was carried out.

In the present research, both methods were applied for the first time for high resolution demographic geospatial data.

IV. Results and Conclusions

The preliminary results of the first analysis (SOM) are quite interesting and demonstrate that several demographic spatially structured patterns can be observed in Switzerland: dense urban zones, interurban regions and country patterns. More detailed analyses will also take into account also the level of education and other relevant socio-economic parameters.

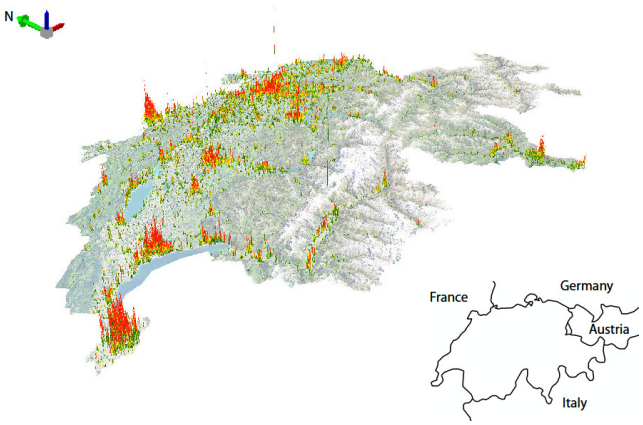

Figure 1: Distribution of the Swiss population in Year 2010.

The application of fm-Morisita has confirmed that the clustering of population in Switzerland (SuisseMetropole) is of multifractal nature.

In future research, the methods will be applied to several temporal datasets, which will help to understand the evolution of the detected geodemographic patterns in time. Moreover, multivariate measures – joint multifractals and multivariate fm-Morisita-will be analysed to relate demographic and socio-economic data.

Acknowledgment

The preliminary research was partly supported by the Faculty of Geosciences and Environment, University of Lausanne within the framework of the FINV "SuisseMetropole" project.

References

[1] T. Kohonen. "Self-Organizing Maps". Springer, 2001.

[2] M. Kanevski, A. Pozdnoukhov and V. Timonin. "Machine Learning for Spatial Environmental Data". EPFL Press, 2009.

[3] H. Rozenfeld, Rybski D., Andrade J, Batty M., H. Eugene Stanley H. E., and Makse H. "Laws of population growth". PNAS, vol. 105, No. 48, 2008.

[4] M. Kanevski (Editor). "Advanced Mapping of Environmnetal Data". iSTE and Wiley, 2008.

[5] J. Golay, M. Kanevski., C. D. Vega Orozco, and M. Leuenberger. "The Multipoint Morisita Index for the Analysis of Spatial Patterns". Available on arXiv (arxiv.org):1307.3756, 2013 [accessed March 2014].

[6] C. D. Vega Orozco, J. Golay, and M. Kanevski "Multifractal portrayal of the Swiss population" Available on arXiv (arxiv.org): 1308.4038, 2013 [accessed March 2014].

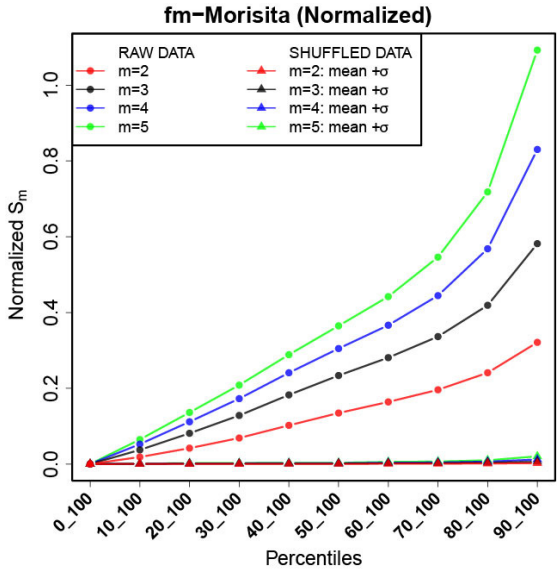

Figure 2: Application of fm-Morisita [3] to the distribution of the Swiss population. The X-axis corresponds to the percentage of the population above the mentioned deciles (e.g., "20_100" means that all the hectares associated with a value lower than the second decile are rejected) and the Y-axis is a normalized index derived from the m-Morisita index. The results for shuffled data are given as well. They partially overlap.

 ISBN: 978-1-61208-326-1