Spatiotemporal Modeling of Urban Sprawl Using Machine Learning and Satellite Data

Alexander Troussov, Dmitri Botvich, and Sergey Maruev
International laboratory for mathematical modeling of social networks,
RANEPA,
Moscow, Russia
e-mails: {troussov, dbotvich, maruev}@gmail.com

***Abstract*—The paper discusses the issues related to the use of machine learning in designing and creating predictive models of the territorial development of cities. Special attention is paid to models that use satellite data, which is the most easily accessible type of data and allows for the use of machine learning. The proposed provisions are based on the authors' current experience in developing an information system for forecasting the territorial development of cities based on remote sensing data of the Earth. A novel approach to developing predictive models of urban growth is presented, which is based on machine learning methods. The approach uses satellite data from the Defense Meteorological Satellite Program/Operational Linescan System and Visible Infrared Imaging Radiometer Suite/Day-Night Band for training. The corresponding machine learning optimization problem is presented and discussed.**

Keywords*—*Earth remote sensing; urbanization; evolutionary models of urban development; cellular automata; machine learning.

I. Introduction

Existing computational models of urban evolution are too resource-consuming, and the validation of forecasting results is quite complicated. However, the growing flow of new data, primarily data from remote sensing of the Earth, enables the use of the 'Big Data' approach for both prediction and validation purposes. Remote sensing data have a high spatial resolution and global coverage and are collected according to a common methodology. To learn the model of a particular city (or agglomeration) development, one can use both historical data from a given city and data from other "similar" cities. The parameters of the computational model based on "Big Data" can be optimized by statistical and machine learning methods. The quality of the simulation can be evaluated by training the model on a piece of data. For example, by training a predictive model of evolution on time series until 2017, you can get the system forecasts for 2018, 2019, and subsequent years, and compare the forecast provided by the system with the real territorial development of the city in these years. A computing system based on such a model can be used to analyze and refine the city's development plans by comparing the development plan of an already developed city with the forecast of city development created based on previous development and the geographical characteristics of the earth's surface. A comparative analysis of these two spatiotemporal series of maps – "plans" and "forecasts" – can indicate possible planning shortcomings and suggest ways to eliminate them. The information system for forecasting the territorial development of cities, on which the provisions of this work will be illustrated, uses a computational model of the spatiotemporal evolution of cities, developed by the authors. Machine learning is very demanding on the quality of data, so in the course of the research, the authors paid great attention to data preprocessing (for methods of processing satellite data, see [1]).

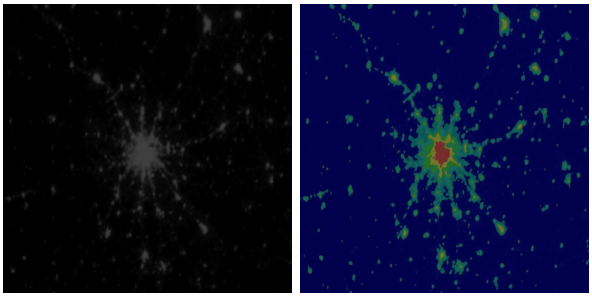

Figure 1: On the left – stable night-time lights of Moscow region in 2008. On the right – an urbanization map that was created based on these night-time lights.

Based on remote sensing data of the earth's night surface, the light footprints of cities are calculated. Night-time remote sensing data, in turn, allow calculating the so-called urbanization maps showing several levels of urbanization for a given city - the city center, nearby neighborhoods, and so on, a total of five levels. From these maps over several years, a time series of urbanization maps for a certain period is compiled.

Figure 1 features the stage of defining the boundaries of the compact residence of the population at a certain point in time. The time series of such maps allow for studying the territorial evolution of the city.

To model the dynamics of changes in this series, it is necessary to use physical characteristics of the earth's surface that hinder or contribute to the development in this area. These characteristics are extracted from the daytime multispectral satellite images. Machine learning can be useful at different stages of the system. For example, machine learning can be used to process space images to detect and classify the characteristics of the earth's surface and objects located on it (see, for example, [2]). In this work, the authors focus on the role of machine learning in optimizing model parameters and evaluating the quality of prediction results.

Figure 2 features an example of how the model calculating a forecast works. The top row shows urbanization maps for

 ISBN: 978-1-68558-049-0

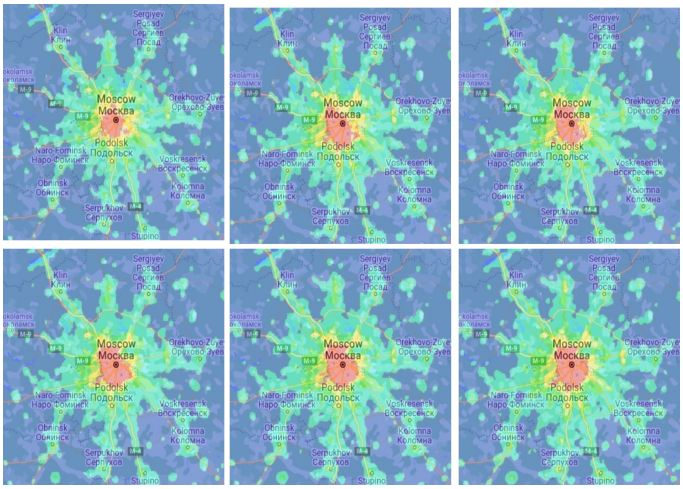

Figure 2: An example of calculating a forecast using semi-supervised learning based on data for Moscow.

2004, 2005, and 2006 (from left to right). The bottom row shows a forecast for 2007, 2008, and 2009 (from left to right). Such a forecast can be compared to real urban development to assess the quality of the forecast and adjust model parameters.

The urban growth model is an important tool for predicting and managing the expansion of cities. This paper discusses urban growth prediction using historical satellite data from the Defense Meteorological Satellite Program (DMSP) / Operational Linescan System (OLS) and Visible Infrared Imaging Radiometer Suite (VIIRS) / Day-Night Band (DNB). Nighttime DMSP and VIIRS data provide high-quality satellite images that capture the urbanization process over time. A machine learning optimization problem for the urban growth model based on the Probabilistic Cellular Automata (PCA) using historical satellite VIIRS data is also presented. The PCA is a popular method for modeling urban growth.

In particular, the well-known SLEUTH (slope, land-use, exclusion, urban extent, transportation, and hill shade) approach is also based on PCA (see [3]). The model was trained and its parameters were optimized with the help of historical VIIRS data. Different approaches to the solution of the optimization problem are also discussed in this paper.

The paper is organized as follows. Section II presents a short review of the urban growth predictive models based on machine learning. The authors' approach to the application of machine learning methods to urban growth is described in Section III. Section IV features the comparison of two predictive systems: the one developed by the authors and SLEUTH. Section V describes the advantages of the authors' system. Finally, Section VI concludes the paper.

II. Urban growth prediction and machine learning: outline

Urban growth models can be defined as spatiotemporal models, which can be generally grouped into three categories: (1) Cellular Automata (CA), (2) Agent-Based Models (ABM), and (3) Machine Learning models. Different machine learning algorithms have been used in the context of urban growth, including Logistic Regression (LR), Artificial Neural Networks (ANN), Linear Regression (LN), Decision Trees (DT), Random Forests (RFs), and so on.

In a systematic review of urban growth models in [10], their evolution, common frameworks, and applications were discussed. In particular, the review considered Cellular Automata models, Agent-based models, and Machine Learning models. In another review [7], it was demonstrated that RFs, Convolutional Neural Networks (CNN), and Support Vector Machines (SVM) are among the best algorithms for the classification and analysis of patterns in data obtained from earth observation as well as in urban growth problems. It was also emphasized that hybrid approaches combining machine learning and cellular automata have the potential for better performance in terms of accuracy, efficiency, and computational cost.

Machine learning techniques have been applied to different aspects of urban growth models, including calibration, sensitivity analysis, and validation. In [11], SVMs were applied to identify the most influential factors affecting urban growth in the federal state of North Rhine-Westphalia. The study found that land use change, population density, and road density were among the most important factors. In [12], an artificial neural network was used to develop an urban growth model for the five largest cities in Greece. The model incorporated various impact factors, such as social, economic, biophysical, neighboring-related, and political driving forces.

In [8], a novel technique combining supervised classification, prediction of urban growth, and machine learning was developed to predict urban growth boundaries and evaluate urban expansion. The technique showed that the expansion of Nasiriyah City was haphazard and unplanned, resulting in disastrous effects on urban and natural systems. In [9], Artificial Neural Network-Cellular Automata (ANN-CA) and Random Forest machine learning algorithms were used for urban growth modeling (2021–2041) in Islamabad, Pakistan. The study found that the city's urbanization has been unplanned and erratic, leading to dire consequences for the environment and urban systems.

Overall, machine learning techniques have shown promising results in improving the accuracy and efficiency of the SLEUTH model for urban growth prediction. However, more research is needed to further explore the potential of machine learning in this field.

III. Our application of machine learning methods to urban growth

The problem considered here is the prediction of the future growth of cities using PCA and historical satellite VIIRS data. Specifically, it is necessary to find the parameters of the PCA model that minimize the loss function, which measures the difference between the predicted urban growth and actual urban growth.

 ISBN: 978-1-68558-049-0

The optimization problem can be formulated as follows:

$$\underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}(\theta), \tag{1}$$

where θ represents the model parameters, Θ is the parameter space, and $\mathcal{L}(\theta)$ is the loss function that measures the discrepancy between the predicted urban growth and the actual urban growth observed in the VIIRS data.

The loss function $\mathcal{L}(\theta)$ can be defined as the sum of two terms: a reconstruction loss $\mathcal{L}_{rec}$ and a regularization loss $\mathcal{L}_{reg}$. The structure of the reconstraction loss function in the area $\mathcal{A}$ is based on the mean squared error (MSE) between the predicted urban growth and the actual urban growth at each time step t :

$$\mathcal{L}_{rec}(\theta) = \frac{1}{nT} \sum_{x \in \mathcal{A}} w_i(t) \sum_{t=1}^{T} (y_{x,t} - \hat{y}(i_{x,t}; \theta))^2 \tag{2}$$

where n is the number of cells in the study area $\mathcal{A}$, T is the number of time steps, $w_x(t)$ is the weight assigned to cell x at time t, $y_{x,t}$ is the observed urban growth level at cell x and time t, $d_{x,t}$ is the input data at cell x and time t, and $\hat{y}(d_{x,t}; \theta)$ is the predicted urban growth level at cell x and time t based on the model with parameters θ.

The regularization loss $\mathcal{L}_{reg}$ penalizes the complexity of the model, and generally helps prevent overfitting. A common choice for $\mathcal{L}_{reg}$ is the L_2 norm of the parameters:

$$\mathcal{L}_{reg} = \lambda \cdot ||\theta||_2^2. \tag{3}$$

The motivation behind such a formulation of the problem is to develop a model that can accurately predict urban growth and be quite stable and robust. Urbanization is a complex process that involves various factors such as population growth, economic development, and land use change. Accurately predicting urban growth can help urban planners and policymakers make informed decisions regarding infrastructure development and resource allocation. Satellite VIIRS data is a rich source of information about night-time lights that provide unique information about population density. By leveraging this data, machine learning models can learn to identify patterns and predict urban growth with high accuracy.

To solve the optimization problem (1) efficiently, it is possible to use the stochastic gradient descent method (SGD) or one of its variants to update the parameters iteratively based on a minibatch of data. In particular, the optimization problem can be solved using gradient descent, which iteratively updates the parameters θ in the direction of the negative gradient of the loss function:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta \mathcal{L}(\theta), \tag{4}$$

where α is the learning rate and $\nabla_\theta \mathcal{L}(\theta)$ is the gradient of the loss function with respect to the parameters θ.

The gradient of the loss function with respect to the parameters can be computed using backpropagation through the neural network. It is also possible to use such techniques as early stopping and learning rate scheduling to improve the performance of the model.

IV. Comparison of two predictive systems – ours and SLEUTH

The model developed by the authors of this paper uses a computational framework of probabilistic cellular automata: the earth's surface is divided into cells, the interactions of which are described by some rules. The evolution of cell states over time is modeled by discrete iterations of interactions according to these rules. The same scheme is used in the well-known SLEUTH model [3] [4].

SLEUTH is the most popular urban expansion model to date. The SLEUTH model reflects the patterns of urban expansion based on several input variables and is therefore relatively easy to understand. The interaction of different parameters of this model, as in any other cellular automaton, is based on explicitly described rules. It is important to note the already known shortcomings of SLEUTH (see [5]): adding new variables to the model and assessing their impact (e.g., beautification, land values, socio-economic distribution, access to utilities, etc.) are complex and time-consuming; the model considers only the fixed spatial neighborhoods of variables, although their spatial influence is not always the same; poor performance.

The authors expect that the use of their model will help to overcome the shortcomings of SLEUTH and will allow moving from "manual selection" of model parameters to their automated assessment, applying statistical and machine learning methods.

The SLEUTH system was developed and implemented in C programming language in the late 1990s. It is designed to model the spatial structure of land use changes. Since then, the system has practically not changed, having undergone only minor changes after 2001 (the last change was in 2017). It runs on a dedicated server and does not use parallelization technologies. The system is based on the use of cellular automata, which play a key role in it. The SLEUTH system uses cellular automata to model the spatial structure of land-use changes.

In general, working with the SLEUTH system is not an easy task, including compiling its source codes, configuring, preparing data, running diagnostics, analyzing calculation results, etc. Adding new modules to the SLEUTH system is also difficult, due to its monolithic architecture and underdeveloped diagnostics and I/O functions.

V. Advantages of our system

The prototype of the prediction module developed by the authors of this paper has been tested and has shown encouraging quality results.

The proposed system has demonstrated very good performance, thanks to the extensive use of parallelization of computing and cloud technology. Due to the use of modern architecture, this system is also easier to maintain and develop.

The performance of the system is very important, since the high speed of operation allows for conducting a large number

 ISBN: 978-1-68558-049-0

of experiments with different data and various model parameters, and potentially opens up the possibility of using machine learning, which will improve the quality of predictions.

Our approach has significant advantages over the SLEUTH approach in terms of architecture, data used, model parameter settings, and the implementation of the model itself. Using the unique features of satellite data and cloud technologies, our model can provide faster and potentially more accurate forecasts of future territorial development.

At the architecture level, our system works in the cloud and uses all its advantages, such as high scalability, parallelization, the convenience of data storage, etc. The SLEUTH system, in turn, runs on a single dedicated server with a limited set of service functions and does not support parallelization. Thus, to increase the performance of SLEUTH, you need to upgrade the server at the physical level.

The authors use satellite data over different years, including "night lights" (DMSP, VIIRS), which provides several additional opportunities for forecasting territorial development. In particular, "night lights" help to reliably determine the boundaries of the compact residence of the population, as well as to estimate the population density. In contrast, SLEUTH relies on land-use maps and urbanization data, which may be outdated and incomplete.

Historical data also provides a unique opportunity to assess the quality of the predictive model and adjust additional global parameters of the model. The use of more extensive and diverse data sets allows for capturing changes over time better and providing more accurate forecasts of future development.

The model parameter settings in the approach developed by the authors differ from those used in SLEUTH. In the authors' implementation of integer automata, the history of changes in local data on the satellite data set used is taken into account. By evaluating the parameters based on this story, including the relationship between neighboring cells, the proposed model can better account for the complex and dynamic nature of urban development.

Effective application of machine learning and statistical methods is possible only if there is a large amount of data. The results of the work of "prediction" algorithms should be easily interpreted and credible. One can recall the statement of the economist J. M. Keynes where he claimed that the theory of probability for many "has a taste of astrology or alchemy" [6] (p. 8)). The same could be probably applicable to the results of machine learning. Therefore, in the developed information system, the initial focus is aimed at assessing the accuracy of projected urban sprawl maps using validation methods as well as visualizing the results at various stages of work.

The input data has spatial dimensions (latitude and longitude) and a temporal component (time or time interval). Localization by space allows considering the local characteristics of specific areas of interest and automatically adjusting the local parameters of the algorithms using machine learning. The presence of a time component in the data, in turn, allows making additional adjustments to the parameters based on historical data to improve the quality of the forecast and, where it is possible, to compare the forecast and the real development.

In the proposed system (see Figure 2), the main manipulated object is a time series of maps, where each pixel of data (correspondent to some area of the surface) is classified, for example, as "urban" or "non-urban". The classification is not binary but has several values corresponding to different ranges of population density in the corresponding area.

The Map Web application allows simultaneously visualizing different data layers and modeling results. Visualization, in turn, provides an opportunity to obtain, study, understand, and transfer to others difficult-to-formalize knowledge about the process of urbanization.

The software created by the authors allows:

- Creating and processing time series of two-dimensional data;
- Calculating the likely continuation for a given time series of data;
- Making joint calculations of several time series (in particular, calculating a binary measure of the similarity of two series of data cards.

As applied to the problem of spatiotemporal modeling of cities, the functions mentioned above allow:

- Preparing data;
- Training and calculating the forecast;
- Visually comparing the forecast obtained after training on historical data with the real development by calculation;
- Calculating a measure of the quality of the forecast, which provides an opportunity to directly improve the parameters of the forecast calculation algorithm;
- Conducting a comparative analysis of "plans" made by policymakers and "forecasts" computed by the proposed system.

The authors believe that a new round of progress in the field of predictive computing systems for the territorial development of cities will occur primarily due to the improvement of nonfunctional characteristics, such as the simplicity of data preparation, system performance, visualization efficiency, and the ease of results interpretation at various stages of work. Assessing the quality of modeling by training a model on a piece of data at an early stage of system development helps to improve the model and quantify the usefulness of certain data. The use of the new model allows shifting from the manual selection of model parameters to the selection of real data on the territorial development of cities. The growing flow of new instrumental data and data from techno-social systems, such as social networks, will soon provide a substantial improvement in the quality of forecasting with the help of statistics and machine learning.

VI. Conclusion and future work

This paper is aimed at considering the issues related to the application of machine learning in designing and creating predictive models for the territorial development of cities. The use of satellite data in developing machine learning predictive

 ISBN: 978-1-68558-049-0

models is also discussed. The role and importance of remote sensing data, particularly night lights VIIRS/DNB satellite data, in the development of these models, is highlighted. Additionally, the authors share their experience in developing an information system for forecasting the territorial development of cities based on remote sensing satellite data.

Furthermore, a novel approach to developing a mathematical model for urban growth prediction is presented. This approach is based on machine learning using VIIRS/DNB satellite data. The corresponding optimization problem for finding optimal parameters of the urban growth predictive model is also raised and discussed.

In the future, the authors plan to carry out practical experiments involving urban growth predictive models in real-world urban planning scenarios. To evaluate the performance of the urban growth predictive model, quantitative metrics will be employed to compare the predicted values with the actual observed values using historical satellite data. These metrics will provide a clear indication of how well the model aligns with the observed urban growth in diverse real-world scenarios. In particular, the Root Mean Square Error (RMSE) metric will be used. By analyzing and comparing the RMSE values across different regions and time periods, a comprehensive assessment of the model's accuracy can be made, allowing for the identification of potential variations in its performance. Furthermore, the authors intend to employ validation techniques, such as cross-validation, to thoroughly assess the robustness and generalizability of the model.

REFERENCES

[1] M. Zhihin and A. Troussov, "Applied tasks of remote sensing of the Earth's night surface," Publishing House 'Delo', 2021, 159 pages., ISBN: 978-5-85006-311-5. (In Russian).

[2] A. Troussov, D. Botvich, and S. Vinogradov, "Detection of Gas Flares Using Satellite Imagery," PATTERNS 2021, The Thirteenth International Conference on Pervasive Patterns and Applications (Porto, Portugal, April 18-22, 2021), 2021, pp. 45-49.

[3] K. C. Clarke, S. Hoppen, and L. Gaydos, "A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area," Environ. Plan. B Plan. Des., 24, 1997, pp.247–261.

[4] K. C. Clarke and L. J. Gaydos, "Loose-coupling a cellular automaton model and GIS: Long-term urban growth prediction for San Francisco and Washington/Baltimore," Int. J. Geogr. Inform. Sci. 1998, 12, 1998, pp. 699–714.

[5] J. A. Gomez, J. E. Patino, J. C. Duque, and S. Passos, "Spatiotemporal Modeling of Urban Growth Using Machine Learning," Remote Sensing, vol. 12, no. 1, 2020, pp. 109-119.

[6] H. Cramer, "Special invited paper. Half a century with probability theory: some personal recollections," The Annals of Probability, Vol. 4, No. 4, 1976, pp.509–546.

[7] V. Chaturvedi and W. T. de Vries , "Machine Learning Algorithms for Urban Land Use Planning: A Review," Urban Science, vol. 5, no. 3, article no. 68, 2021.

[8] S. K. Hanoon, A. F. Abdullah, H. Z. M. Shafri, and A. Wayayok, "Urban Growth Forecast Using Machine Learning Algorithms and GIS- Based Novel Techniques: A Case Study Focusing on Nasiriyah City, Southern Iraq," ISPRS International Journal of Geo-Information, 12(2), 76, 2023.

[9] A. Khan and M. Sudheer, "Machine learning-based monitoring and modeling for spatio-temporal urban growth of Islamabad," The Egyptian Journal of Remote Sensing and Space Science, vol. 25, no. 2, 2022, pp. 541-550.

[10] X. Li, Y. Chen, X. Chen, and Y. Wang, "Urban growth models: progress and perspective," Science Bulletin, vol. 61, no. 21, 2016, pp. 1637-1650.

[11] A. Rienow and R. Goetzke, "Supporting SLEUTH – Enhancing a cellular automaton with support vector machines for urban growth modeling," in Computers, Environment and Urban Systems, vol. 49, 2015, pp. 66-81.

[12] P. Tsagkis, E. Bakogiannis, and A. Nikitas, "Analysing urban growth using machine learning and open data: An artificial neural network modelled case study of five Greek cities," Sustainable Cities and Society, vol. 89, 2023, p. 104337.

 ISBN: 978-1-68558-049-0