An attraction-based cellular automaton model for generating spatio-temporal population maps in urban areas

Mehdi Khakpour, Jan Ketil Rød

Abstract In this paper, a cellular automaton model is developed to generate spatio-temporal population maps that estimate population distributions in an urban area in a random working day. The resulting population maps are at 50 by 50 meter spatial resolution and 5 minutes temporal resolution, showing clearly how the distribution of population varies throughout a 24-hour period. Places that are sparsely populated during night-time can be densely populated during day-time. The generated maps can be used to estimate population-at-risk in the wake of major disasters when they occur in an urban area at any time of a day. In addition to assessing exposure to hazards, the resulting maps also reveal the movement patterns, transition trends, peak hours, activity levels, etc. Possible applications, thus, range from public safety, disaster management, transport modeling and urban growth studies to strategic energy distribution planning. The developed cellular automata model assumes that the population transition trends follow similar dynamics and propagation patterns of a contagious disease. Thus the cellular automaton is designed to change the states of each grid cell (Stable/Dynamic) similarly as state changes of an individual being exposed to an infective disease (Susceptible/Infected). The modeling space is further informed by several geographic features, such as the transport routes, land use categories, population attraction points, etc. The model is geosimulated for the city of Trondheim in Norway, where the synthetic day population could be validated upon an estimated day-population map based on the registered work place addresses and employee statistics.

Keywords: Cellular automata, Attraction-based, Population-at-risk, Ambient Population Maps, Day population, Geosimulation

1 Introduction

Exposure assessment is an important aspect of risk-based disaster modeling (Modarres, 2006), which is nevertheless often either neglected or simply assumed to be embedded into the vulnerability indices (IPCC, 2012; Purdy, 2010; UNISDR, 2011). In any disaster management framework, saving human lives takes the highest priority. This implies that, the size of any formal evacuation planning, search and rescue operation, recovery and emergency housing,

etc., must necessarily be incorporated to a reliable estimate of the number of people, being exposed to the hazard.

The efficiency of such estimation is to a large extent, a matter of accuracy as well as the spatio-temporal resolution of the available population datasets (Dobson et al., 2000). For example, being able to accurately estimate the number of people at risk in the wake of a major hazard, such as a hurricane, can clearly determine the required capacity of evacuation routes, the number of logistical vehicles or the amount of food and medicines that must be stored when planning for an emergency response operation. The most commonly used method for locating people in a city is by using the population density maps, made from official censuses that are developed based on the registered residence addresses. This neglects the fact that many people may stay at home for only a few hours from late afternoon until the next morning, when they commute to work, study, shop or to where other urban daily activities take place. As a result, developing models for estimating the ambient population or the number of people actively moving within the built environment of a city during the day becomes vital. Population movements could also involve diurnal activities of sub-urban commuters travelling to main cities for work, who account for a significant difference between day and night population counts as well as part of the peak hour traffic congestions (National Research Council (US), 2007).

The conceptual model in this paper, adopts epidemiological concepts into the geosimulation of population movements. Spatially explicit elements of the simulated phenomenon are assumed to spread through time and space, following stochastic epidemic modeling approaches. In this sense, the modeling space has turned into a discrete modeling environment of autonomous geographical elements (e.g. gridded land cells).

This approach is based on the assumption that, the emergent behaviors of the population movements as a complex system would probably follow the dynamics of an epidemic disease. There are several mathematical formulations related to studying the dynamics of contagious diseases' epidemics, which could be used to mimic the population transition patterns. Comparing their characteristics and behaviors, population movements and the disease propagation have many similar features and characteristics in common. For example, if we consider people's movements as a contagious disease that can travel through the modeling space, there are some times of the day, during which the level of activity in a city peaks and drops again, such as the morning and the afternoon rush hours. This behavior is analogous to the epidemic behavior of the contagious diseases, when their break out and suppression periods varies in different months of their propagation season. Other

characteristics of these complex systems also motivate the use of epidemiologic predictions for modeling population movements. The way people are transported in a diffusive manner from core residential areas to commercial areas in the morning and the way back in the afternoon resembles the cyclic behavior of epidemic diseases with diffusive growth patterns. Another analogy can be seen based on the possible states that can be similarly assumed for the dynamic state of gridded land cells of a city. A land cell can become dynamic or stable the same way as the individuals that can be susceptible or become infected.

The main purpose of this paper, which should be considered more of a proof of concept rather than a fully operational predictive structure, is to take the first step to incorporate the epidemiological concepts into the modeling of population movements so that the resulting geosimulations would be able to generate synthetic population maps that have a temporal resolution. This modeling approach is expected to enhance the credibility and usefulness of previous models by reducing the size of required input data, with a concentration on validation and calibration issues. A two-dimensional evaluation strategy is therefore employed in order to improve the operational as well as the numerical reliability of the modeling results.

2 Literature review

The available models for mapping the ambient population are mainly limited to the Landscan Global and the GPW (Gridded Population of the World), which are developed based on the disaggregation of low resolution census counts to a higher resolution often using preferential weighing techniques. These probabilistic approaches commonly require several additional data inputs, such as satellite imagery. They are, therefore, heavily dependent on the availability of the required cutting-edge technology and resources. Both of these models employ diverse methodologies such as dasymetric techniques, spatial interpolation, disaggregation, and imagery analysis to estimate an average population value for the day population.

Despite the acceptable level of consistency delivered by these invaluable efforts to estimate tempo-spatial population maps, there are two main reasons why we might have to look for alternative methodologies. Firstly, the practical usefulness and usability of these types of population representations is curbed to a few countries, in which a large quantity and variety of spatial census data sets are available, which ironically is not usually the case when it comes to severe natural disasters. A significant part of the disasters occur in less-developed parts of the world, where poor nations suffer more death tolls than rich nations from natural

disasters (Kahn, 2005). Secondly, since the population distribution is based on local disaggregation or some other forms of extrapolation, the estimation results are heavily dependent on the preferential weighting of the estimation parameters such as land use, time of the day, slope gradient, etc. This approach is normally based on human inference approximations, or the so-called preferential weightings ("LandScan - Documentation," 2013), which could be hard to validate if used as an input for generating the model estimations rather than for calibration purposes. The temporal resolution of the mentioned models also may not satisfy the required needs for exposure assessment of several time-sensitive natural hazards, in which the number of casualties is heavily dependent on the time of their occurrence, such as earthquakes and tsunamis (Freire, 2010).

Furthermore, the intrinsic stochasticity of people's decision-making process in their choice of destination, route selection, etc. can not be well explained when models tend to represent the currently estimated data rather than actually predicting values for the population counts. This is a result of the weighting process that comes with the assumption of uniform distribution of disaggregated population assignments to all of the areas with the same land use category in every time interval of the day (Bhaduri et al., 2007).

Among other methodologies that are somehow capable of incorporating human decision-making mechanisms into geosimulations are the Agent-based models. In these models, urban areas are considered as complex systems. The complexity of the system emerges global and structural behaviors from actions, each of which are simple in themselves, of relatively autonomous agents, interacting with each other and other system elements (Batty et al., 1998). Even though this type of modeling might preserve some of the intrinsic stochasticity of the people's/agents' behavior, they appear to be relatively limited to the modeling of local movement patterns, and behaviors in only small-scale built environments such as shopping malls or a neighborhood's pedestrian flows (Ali and Moulin, 2005). This is partly because, trying to increase the number of the agents to the size of a large city's population, would need much higher resolution of spatio-temporal data that could affect their usability, as well as imposing even more challenges when trying to calibrate and validate them. In addition, the expensive handling of the massive raw data will limit the flexibility of the model and may make alternative calibration techniques, such as replicative calibration, infeasible. Similar approaches have been developed previously in modeling population growth and land use change using an activity-based CA (White et al., 2012). Cellular automata (CA) that are known to be mathematical idealizations of complex systems in discrete space and time, were developed by Ulam in the 1940s and improved by Von

Neumann exploring the logical nature of self-reproducible systems (Adamatzky et al., 2008). CA can be viewed as a simple model of a spatially extended decentralized system made up of a number of individual components, called cells, each of which communicate between themselves following a set of simple rules. Each cell is in a specific state, which changes over time depending on the state of its local neighbors and various inputs from outside the automaton (Ganguly et al., 2003).

3 Methods and Materials

A third approach after the dasymetric and agent-based models, which we follow, is an alternative automaton-based modeling infrastructure, in which the macroscopic behavior of a complex urban system, the urban area, is assumed to be obtained from the microscopic behaviors of their spatial entities.

The proposed model is designed in form of an algorithm driven by an attraction-based cellular automaton that distributes the night population of an urban area from their origin at the nighttime, the residential cells, to other land use categories during different time-intervals of a 24-hour day. The map of a city (Trondheim) is therefore gridded into 50 by 50 meter cells. Each cell can be visualized by the estimated number of people staying in that cell, in every time-step of the simulation.

In this approach, unlike an agent-based model, people in a city are not modeled individually but quite similarly, the cellular automaton follows an individual's mind-set of traveling through the city both autonomously and stochastically. In other words, instead of modeling the individuals physically, their state of the mind in making decisions for moving across the city's context is parameterized. For this reason, every gridded land-cell on the map can be affected by its surrounding movement trends and a movement trend can lead to other movements locally and form a global movement pattern. Alternatively, adding the people's need for travel to their stochastic behavior, some of the cities main activity-gaining places are introduced into the model as attraction-centroids. This is handled by choosing the city's most populated areas as attraction-centroids in residential areas, and the main daily active places in the commercial areas. In the simplest words, the residential attraction-centroids are assumed to attract the population during the afternoon hours and the commercial attraction-centroids are assumed to attract people in the morning. The model simply divides an entire 24-hour day into five main time-intervals and weights the different land use categories' attraction values in each interval. Figure 1 shows a schematic illustration of the time intervals and the way people are generally believed to move between various land use categories. By the following

sections, we will see; based on what modeling principles, and how, the movement patterns and population counts could be estimated.

The urban area is first gridded into 50 by 50 meter cells, which hold one of the two states at a time, (1) Stable and (2) Dynamic. Every cell is assumed to have a balance of emigrant and immigrant population to and from the neighboring cells when it holds the Stable state. When it is more likely that the grid cell is either losing or gaining population counts, it shifts to the Dynamic state. In order to estimate this likelihood, a random value is assigned to every cell in every time step. Then the average estimated value for a local Moore neighborhood of a cell is compared to a global threshold probability, Beta. Any Static state cell can turn into a Dynamic cell if the threshold probability is passed. It is therefore assumed from the actual behavior of people's transition within a city that the Dynamic state is propagated through a neighborhood effect that is influenced passively by some suitability factors, i.e. road accessibility.

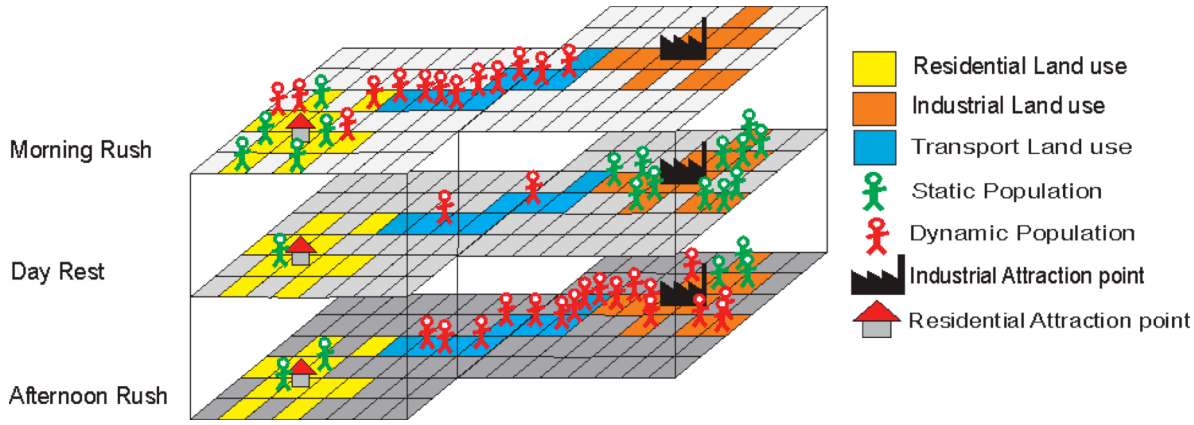

Figure 1: A schematic illustration of our basic understanding of urban movement patterns in three different time-intervals of a day.

The main idea is to determine, whether the complex population movement patterns could be drawn based on our basic understandings of their microscopic behaviors, such as their travel habits. The decision making process of the moving crowd in a city, if assumed to account for this microscopic behavior, can then be rationalized based on their obvious needs and the city's spatial characteristics. In other words, people can first be categorized based on their daily activities such as students, workers, etc., then assumed to move to their desired destinations based on their related land use category in different times of the day on a spatially explicit modeling platform of an urban area. For example, students go to institutional land use categories in the morning and commute back to the residential lands in the afternoon.

Another approach in setting destination for individuals, employed in this research, is to set highly populated commercial and institutional lands as attraction-centroids in the morning and highly populated residential areas in the afternoon, in a way that the autonomous land cells change their states based on the land use categories and day time intervals.

The CA's development direction is guided towards the main population attraction-centroids (similar to that of referred to as the ''city centroids'') that attract the majority of the population transition during different time intervals of the day. In simple words, work places and universities attract population counts during the morning rush hours of a city, while residential lands play the same role in the afternoon, when people are mostly heading back to their homes. This is analogous to the concepts of the conventional gravity-based transport modeling, where highly populated city-cores generate trips by attracting and repelling population during the 24 hours of a day. The algorithm chooses the largest population clusters as different attraction-centroids in different time intervals of the day. This attraction rule is simply implemented into the model, by estimating the cells' population probability, according to their Euclidean distance to the nearest attraction point and their associated attraction weight.

3.1 Attraction-based Cellular Automaton

A standard form of cellular automata can be generalized as follows:

$$ST^{t+1} = f(ST^t, N) \tag{1}$$

Where ST is a set of all possible states of the CA, N is a neighborhood of all cells providing input information, and f is a transition function that determines how the system states change from t to $t+1$.

Despite their simplicity, CA has been used to model numerous physical and geographic phenomena, and ultimately, CA have been increasingly used to model spatial dynamics (Batty and Xie, 1994; Itami, 1994; Li and Yeh, 2000; Vliet et al., 2009; Wu, 1998). There have been numerous invaluable efforts made to alternate this uncomplicated format in order to enhance its suitability in different modeling environments, such as the development of constrained CA (White et al., 1997) and Geographical Automata Systems (Torrens and Benenson, 2005), by adding suitability or even geo-referencing factors into the right side of the equation. However, adding more variables and introducing much more input information into a model does not necessarily make it explain certain complex phenomena better (Wolfram, 2002).

Conforming to the preceding arguments, the cellular automata presented in this research is formulated as its classical structure, with the difference that the status change is influenced by a logical comparison of local averages probabilities to a global threshold. The CA's development direction is also guided in relation to its weighted Euclidean distance to the indicated urban attraction-centroids stochastically. In addition, the geographical system is transformed from its original geo-referenced format into unilateral grid cells so that it can be mathematically manipulated much easier in the form of matrices. Alternatively, the spatial information of every grid cell is converted into single-subscripted matrix indices so that it could be easily manipulated by the algorithm. The rest of the geo-referencing information, including the projection data, etc., is stored in the MATLAB text files during the conversion of GIS maps.

Hence, the attraction-based CA is formulated as:

$$ST^{t+1}\{i,j\} = f_{MN}(ST^{t}\{i,j\}, P_{ST}^{t}\{i,j\}) \qquad (2)$$

Where $ST^{t+1}\{i,j\}$ and $ST^{t}\{i,j\}$ represent the status (Dynamic or Stable) of cell $\{i,j\}$ at times t+1 and t respectively, $P_{ST}^{t}\{i,j\}$ is an estimated random variable representing the probability of cell $\{i,j\}$ transition to the state ST at time t, and f_{MN} is a transition function, averaging the cell probabilities in a Moore neighborhood of the cell $\{i,j\}$ including itself.

So,

$$f_{MN} : \begin{cases} ST = \begin{cases} Dynamic & \sigma \geq \beta \\ Stable & \sigma < \beta \end{cases} \\ \\ \sigma = \left(\sum_{M_1}^{M_9} P_{ST}^{t}\{idx + M\}\right)/9 \end{cases} \qquad (3)$$

Where σ is an average of the estimated probabilities assigned to the cell and all its eight Moore neighbors, idx is the Matrix index of cell $\{i,j\}$ in a $m \times n$ Matrix, and M is a set of offset values that indicate the nearest neighbors of the designated cell as illustrated here:

$$M = \{Cell, East, Southeast, South, Southwest, West, Northwest, North, Northeast\} \qquad (4)$$

$P_{ST}^{t}\{idx\}$ is estimated randomly based on $AtV\ \{idx\}$, the attraction value at cell $\{idx\}$, and $RAND^{t}\{idx\}$, which is a random variable that is generated in every iteration without a memory as following:

$$P_{ST}^{t}\{idx\} = (AtV\ \{idx\} * RAND^{t}\{idx\}) \qquad (5)$$

Note that, $P_{ST}^{t}\{idx\}$ is an estimate of P (ST) and is therefore not dependent on the ST itself. The attraction value is formulated as:

$$AtV^{t}\{idx\} = \left(\sum_{n=1}^{N} AtW_{n}\{idx\} * EUD\{idx\}\right) \quad (6)$$

in which, $AtV^{t}\{idx\}$ is calculated in every grid cell by summing the Euclidean distances of the cell to all the attraction-centroids multiplied by their associated attraction weight, denoted by AtW_n, given by a set of weights like:

$$AtW_n = \{AtW_1, AtW_2, AtW_3, \dots, AtW_N\} \quad (7)$$

These weights are generated by the algorithm based on the registered population size of the related attraction.

3.2 SIS-based Model

There are several classical models for explaining different diseases' dynamics that suit their characteristics and behaviors, with SIR and SIS being the most popular ones. In SIR models the disease leads to death or immunity, meaning that the individuals can hold three different states (Susceptible, Infectious, Removed) (Brauer and Castillo-Chávez, 2012). In our modeling environment, every gridded piece of land can only hold two possible states (Stable and Dynamic), which fits well in the SIS model representing no-mortality diseases.

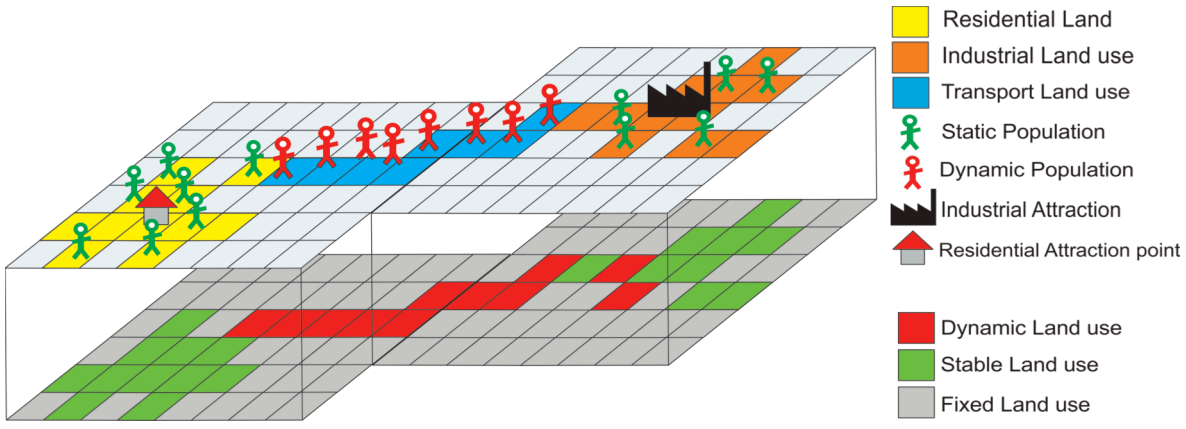

Figure 2: A schematic illustration of urban movement patterns between different land use categories and the epidemiological interpretation of the spread of the movement.

If the urban area would assumedly be divided into several unilateral grid-cells, each representing a population count point, every piece of land can then represent either a

susceptible or infectious individual that can change its status due to its contact with the neighboring cells. When people travel between the grid cells, every grid cell is assigned by an estimated number of people staying in that certain piece of land in a certain time of the day. Then in different times of the day, some parts of an urban area are more likely to attract or repel the population, while others are more likely to maintain equilibrium between the number of people travelling to and from these places. The infectious state is assumed to be similar to a condition, under which a grid cell is more likely to lose or gain population, rather than remaining stable. Similarly, the level of activity in a grid cell is assumed to be stable, when the number of people moving to and from that cell is almost the same. Figure 2 illustrates the urban movements between different land use categories and its epidemiologic interpretation.

In this metaphor, β is the probability by which the Dynamic state (Infectious) spreads through the Stable state (Susceptible) cells, while υ is the probability by which the Dynamic cells become Stable again. Then R becomes an efficiency factor being β/υ so that whenever R>1, the disease (Dynamic state) is dominating the modeling site, and if R≤1, then the Dynamic state is being suppressed, and the Stable state dominates. These two latter settings are mimicking the real situation within the peak traffic hours in the morning and afternoon. R is therefore randomly selected between 1 and 2 during the peak hours, and between 0 and 1 during other time intervals of a day using an exponential time decaying function. This is also consistent with the periodic emergence and decline of a disease, like flu during the cold seasons. β is also generated by a random variable generator with values chosen between 0 and 1, and υ is the dependent variable.

3.3 Geosimulation Infrastructure

Automaton-driven geosimulations, irrespective of their modeling engine, whether agent-based or CA, are computationally quite demanding, due to the integration of large spatially explicit datasets with necessary computation of updates for agents/cells in every time step of the simulation. This process can become unaffordably expensive when the simulation algorithm has to be repeated for several replicates for calibration purposes. It is therefore beneficial to build a robust simulation infrastructure that would be able to, not only handle the large number of replicates (10^4), but also to visualize the spatial data simultaneously, so that it keeps all the map projection information intact.

In this framework, where the gridded maps resemble a large matrix of land cells, MATLAB is chosen to perform the numerical computations. The use of matrix cells representing the gridded land cells could be particularly helpful in a discrete simulation

architecture, in which there are several data layers for the mapped area, namely the land use category, accessibility, attraction-centroids and the Day/Night population. These numerous data layers can easily be represented by several uniform matrices that interact with each other and get updated in every time step, without losing their spatial precision, through a MATLAB-scripted algorithm.

As the visualization interface, ArcGIS was integrated into the geosimulation platform, in which the MATLAB algorithm takes charge of data handling and CA-driven simulation as well as replicative calibration of the simulation. The geo-referenced data, in the form of text files, are automatically imported into the MATLAB environment, where the simulation algorithm builds the CA (using Image processing, Mapping and Statistical toolboxes), runs the computations and exports the results back to an environment, accessible by Python, that automatically applies the ArcGIS toolboxes and exports the maps. Figure 3 is a schematic illustration of the geosimulation platform used in this research.

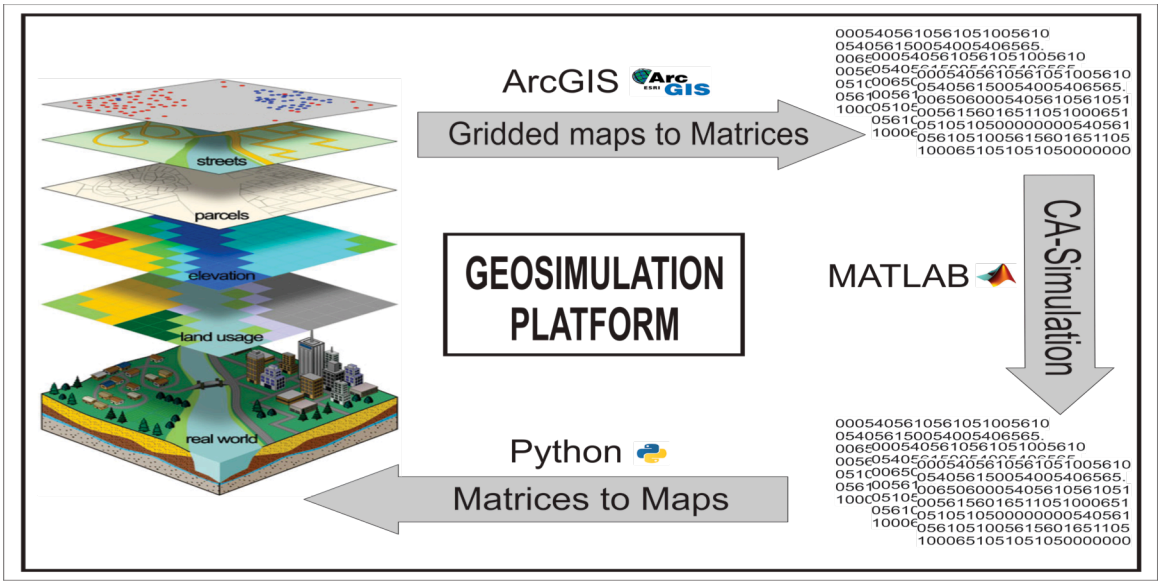

Figure 3: A schematic illustration of the geosimulation platform.

4 Study Area

In this research, the developed model was applied to the city of Trondheim for simulation and validation purposes. Trondheim has Norway's third most populous urban area with the population of over 176 000 ("Statistics Norway," 2013), a relatively medium size for a non-capital European city. The main activities in Trondheim are dominated by the Norwegian University of Science and Technology (NTNU), St. Olav's Hospital and a few other leading research firms and organizations such as SINTEF, Statoil, and the Norwegian Geological Survey, as well as a few leisure and shopping centers that form the city's main crowd-

transition attraction centroids. It is also recognized as a natural hazard-prone area in Norway, especially with regards to quick clay slides (Nadim et al., 2008). Figure 4 indicates the geographical location of Trondheim on the map of Norway.

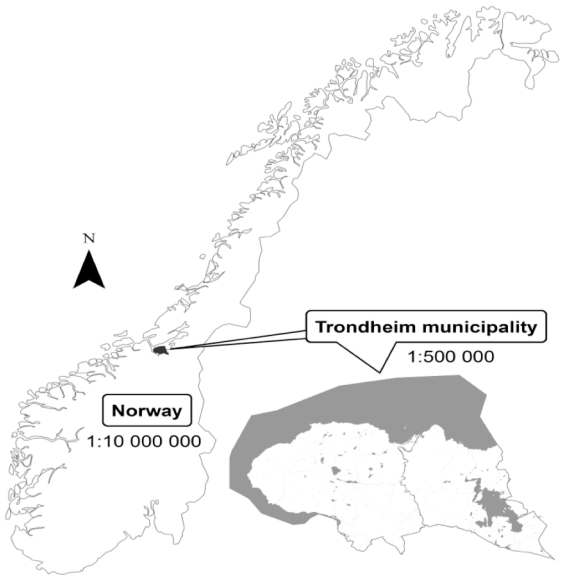

Figure 4: Location of Trondheim, the study area on the map of Norway.

5 Simulation Results

In every replication, the output of the geosimulation consists of 288 data sets corresponding to a 5-minute estimation of Trondheim's ambient population in a 50-meter spatial resolution. Figure 5 compares the synthetic ambient population map of Trondheim, representing day and night population respectively. It highlights dramatic shifts in population transition trends within the city's urban areas in daytime versus the nighttime. There are several places on the day population map, clearly indicated by the dark red color that show high-density population areas, which could cause a serious under-estimation in the process of calculating the number of people at risk.

We have initially tried to present the raw numbers, i.e. the number of people in every gridded cell in every time step, but due to the very high resolution of our datasets (50 meters and 5 minutes) the resulting images and graphs were almost unreadable. Instead, we have aggregated the results into visual representations so that we can show the population trend changes in coarser time and space scales. Another reason for using visual representations instead of numbers is that the actual number of people in one gridded cell in a single time-step of the simulation is not of our major interest, because the same as the real world, they are just

some random events when looked at individually but form a clear trend when considered as a whole complex system.

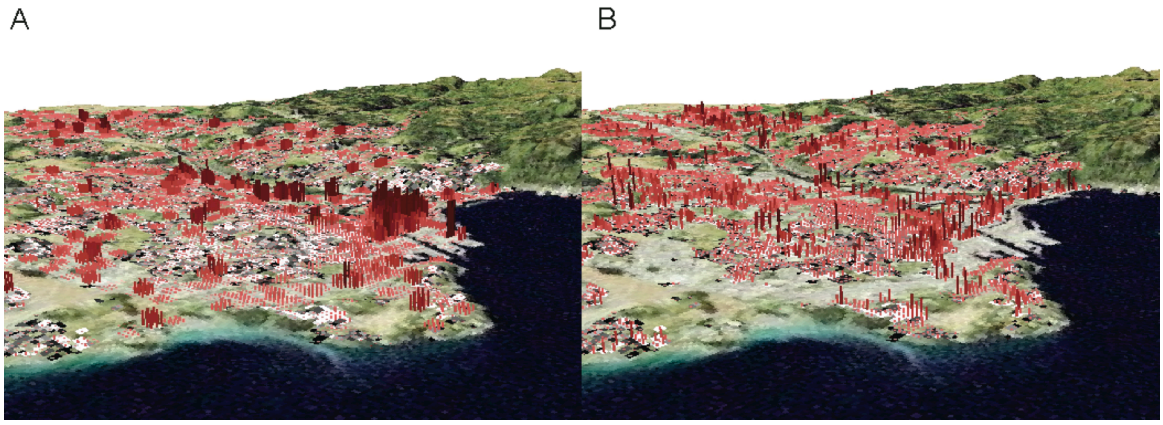

Figure 5: A Synthetic Ambient (Day) Population Map of Trondheim (Norway) based on an epidemiologically formulated geosimulation (A), compared to the so-called 'Night Population Map' based on the national census counts (B).

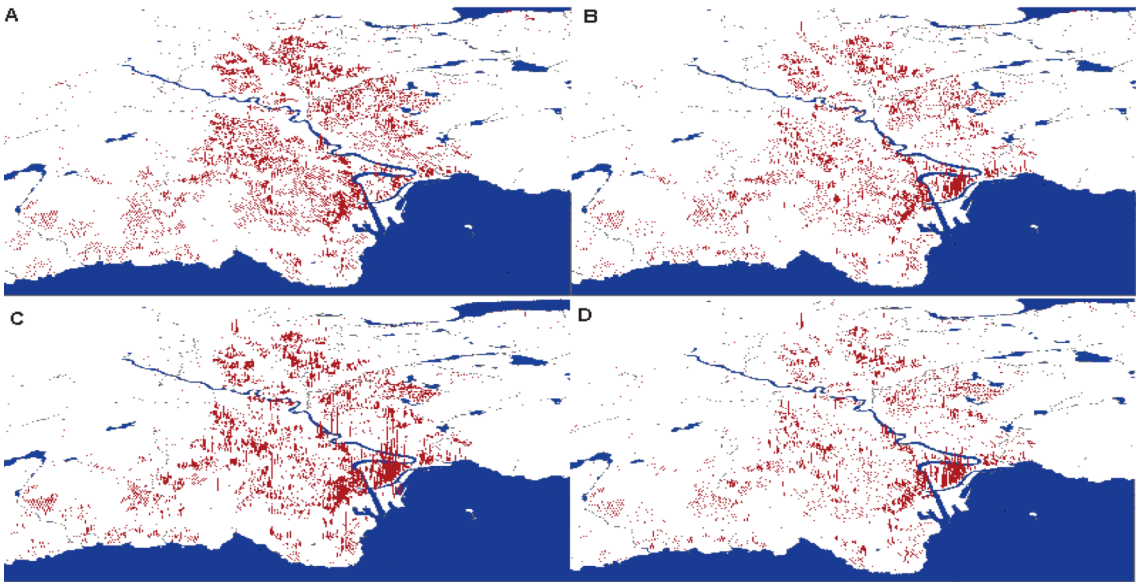

Figure 6: Geosimulation results for Night time, Morning-rush, Day-rest and the Afternoon-rush. (A to D)

Figure 6 shows the geosimulated ambient population distribution in Trondheim for four different time intervals: Nighttime, Morning-rush, Day-rest and the Afternoon-rush. Each elevated bar in every cell indicates the estimated value for the number of people locating in that part of the city. The blue areas indicate water bodies, such as rivers and lakes.

Figure 7 shows the CA output and the corresponding simulation for five selected times of the day: 09:15, 13:15, 17:30, 21:50 and 23:00. The red dots on the CA picture display cells

that hold the Dynamic state while black cells correspond to either the Stable state or some fixed cells (which are excluded from the simulation). The CA changes the states of the cells based on their initial state, neighborhood effect, proximity to attraction-centroids and roads, forming a local estimated value, which is compared to a global mean that functions as a threshold point in every iteration of the simulation. In parallel, the algorithm changes the estimated values for population counts in Dynamic state cells based on the prioritization of land use categories in different time-intervals and a random factor.

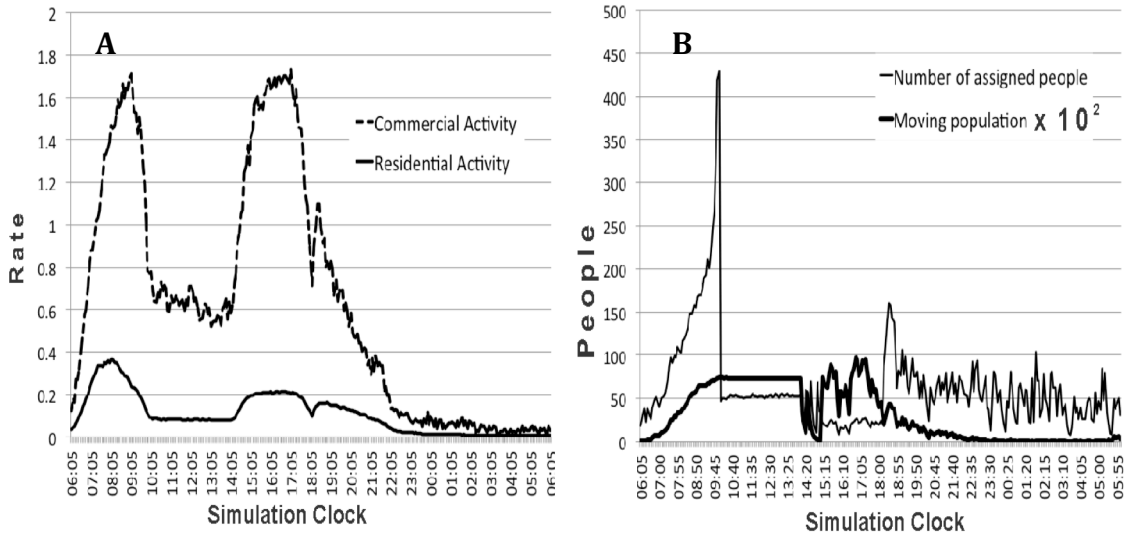

Figure 8: Plotting Commercial and Residential Activity values (A), and the Number of assigned people to different land uses, and the amount of moving population X 10^2 (B) against the simulation clock.

Figure 8a compares the percentage of Dynamic cells in Residential versus Commercial land use categories in different time intervals of the day. Percentage of Dynamic cells acknowledges the level of activity in the city in each time interval of a random day. When this percentage is increasing for the residential cells, it means that the residential cells are getting more active in sense of attracting or repelling population. It is obvious from the graph that the Dynamicity of the city goes to a long rest in the afternoon until the next morning in both residential and commercial land uses.

Figure 8b shows estimated values for the number of population counts assigned to different land use categories and the number of people on the move within the 24 hours of the day. Numbers on the vertical axis correspond to the estimated number of people who are assigned to different land use categories in every time step and the estimated number of people still being on an inter-city travel (x 10^2 for the number of people on the move).

A comparison between the trends of these graphs and our basic understanding of population transitions within different times of the day, clearly shows that the model is capable of mimicking the peak hours as well as the calm hours of the day. Studying these estimated values could especially be useful for transport modeling and strategic energy planning.

6. Model Evaluation

Adequate calibration and validation, which is commonly referred to as model evaluation, is critical for ensuring the credibility of any geosimulation structure (Marceau and Benenson, 2011). In this regard, the model's randomly selected parameters were calibrated first by a replicative mapping technique. Then, the numerical validation is applied in the form of statistical testing and deviance measurements, by comparing the synthetic maps and the available census datasets, generated by statistical surveys. As there are obviously no actual real-world ambient population datasets in a fine scale spatio-temporal resolution, the decline in the amount of standard errors (RMSE and MAE) and an increasing trend in the R^2 values is considered as a state of goodness-of-fit in this model.

6.1 Replicative Calibration

The idea of replicative calibration emerges in response to a major challenge in validating certain types of models, for which there are scarcely available 'observed' datasets and a set of parameters are generated stochastically. The results of such models could be considered as being sensitively dependent on the initial conditions. Replicative calibration, in this research, refers to an evaluation approach, in which the algorithm repeats the simulation for a significant number of replicates with different initial conditions every time. Then, the varying parameters are mapped against a form of measured error term so that the modeler can easily spot the criterion, for which the model has a better validity. This approach is conceptually similar to the "variant-invariant" method developed by (Brown et al., 2005) that shows areas, where the model has a better fit. The calibration method presented here has been previously used in similar works for running simulations with several initial conditions (Belcher et al., 2010; Gregory and Smith, 1991). Even though this method is not fully conforming to the formal calibration techniques, which are driven by error functions and adjust the model in every iterating step (Straatman et al., 2004), it has the advantage of preventing the error from propagating into the model results through the step-by-step corrections. Particularly, in

modeling environments where a large amount of variability is introduced into the model and the uncertainty levels are high, risk of over-controlling the model will significantly rise when using iterative correction or automatic calibration techniques. Over-control is a statistical term, generally used for situations, in which the modeler inserts more variation into the modeling process by imposing corrections (as random disturbances) after each estimation step, rather than informing the model (Montgomery and Runger, 2010). In turn, studying the model results with a significant number of different initial conditions could lead us to a better understanding of the sensitivity of the model to its parameters as well.

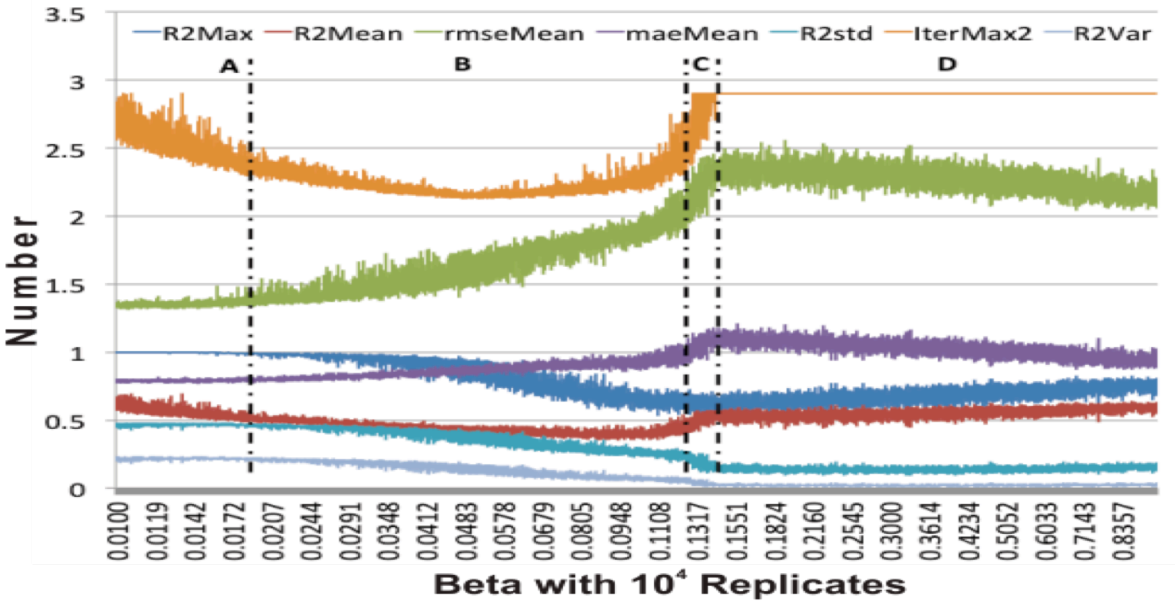

Figure 9: Plotting Simulation fitting values against the model parameter Beta with 10^4 replicates.

In our approach, the simulation is performed for 10^4 replicates with various values for Beta, Sigma and R, first for β in (0,1) and then for a suitable criterion (0.1310, 0.1551) as indicated by section C in figure 9. In every simulation, the estimated values were compared to a collection of samples from real data and the minimum RMSEs and MAEs were recorded in accordance to the recorded values for Beta, Sigma and R in that simulation. This way, efficient values for these parameters could be calibrated based on the model's goodness-of-fit. Figure 9 shows the plotted goodness-of-fit-values against randomly generated values of the model parameter β for 10^4 replicates.

6.2 Numerical Validation

Among common numerical validation techniques, statistical testing and deviance measurements are used to examine the goodness-of-fit of the model in this research, because of their suitability to the simulations' spatio-temporal attributes. Therefore, in every time-step

of every replication of the simulation, the coefficient of determination, denoted by R^2, was calculated as:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \tag{8}$$

in which, SS_{err} is the sum of squares of errors (residual sum of squares) and SS_{tot} is the total sum of squares. R^2 is often interpreted as the proportion of variation explained by the fitted regression line. In this statistical testing, R^2 values are recorded in every time step and plotted against time in a way that its variations in time shows the trend in the model's goodness-of-fit. As clearly shown in Figure 12, R^2 is solidly increasing from the beginning of the simulation and approaches toward 1.0 in every time-interval of the day. It should be mentioned that, because we are not comparing models of different parameter sizes in every time-step, it is not necessary to calculate the adjusted version of R^2, which is less sensitive than R^2, to additional parameters in model selection approaches.

Deviance measurements are often useful when observed and simulated data can be paired according to time, location, treatment, etc. and are normally calculated as the difference between the observed values and the predicted (simulated) values (Mayer and Butler, 1993). In this validation approach, the simulated values for population counts in every grid-cell of the map were compared with the observed values by calculating the Route Mean Square Error (RMSE) and Mean Absolute Error (MAE) values.

The best available population census maps for Trondheim were found only in 500-meter resolution, therefore the simulated values were geographically aggregated by averaging the number of synthetic population in every 10 grid-cell and the deviance measurements were performed consistently.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}} \tag{9}$$

$$MAE = \frac{\sum_{i=1}^{N}|P_i - O_i|^2}{N} \tag{10}$$

As shown bellow, the RMSE measures $(P_i - O_i)^2$, the distance between the Predicted and the Observed value, in a quadratic sense; therefore it is rather sensitive to outliers. Hence, MAE is alternatively calculated as another measure of deviance so that a few well-fitted points on the geosimulated map would not affect the validation.

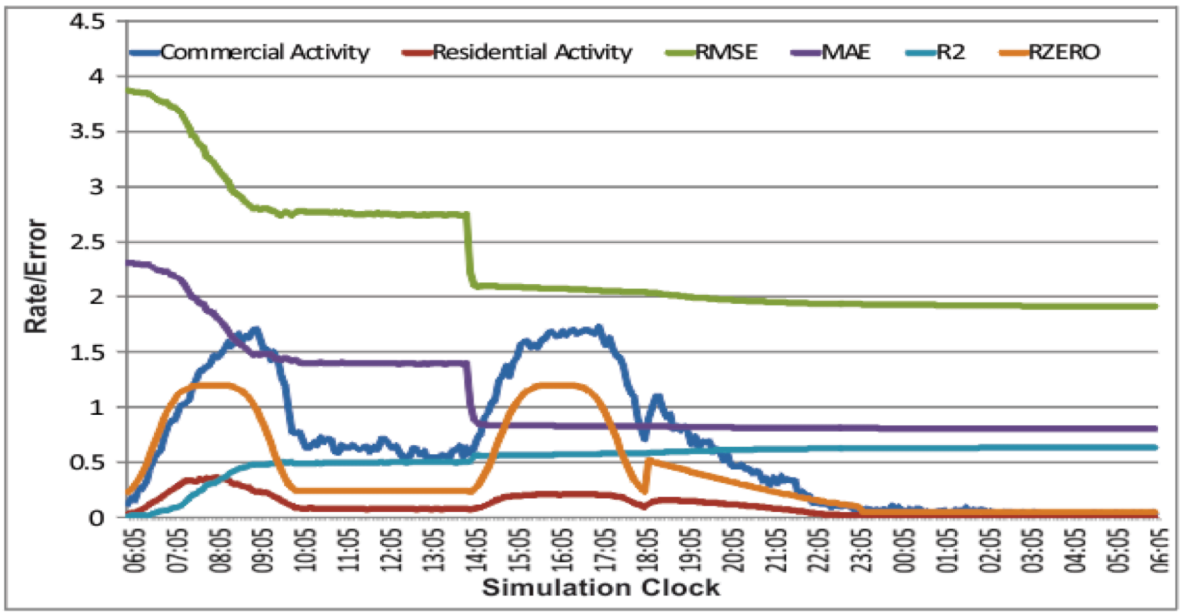

Figure 10: Plotting Goodness-of fit, Error measurements and activity values against the simulation clock. (RMSE=Root Mean Squared Error, MAE=Mean Absolute Error, R2= Coefficient of determination, RZERO= β/υ Basic reproduction number.

As shown in figure 10, the RMSE and MAE are both declining by time, which indicates that the values on the synthetic day population are gradually merging with the real-world day population maps. In other words, the predicted values for cell population counts are getting closer to their observed values, which are taken from the employment address points in Trondheim.

7 Conclusions

The goal of this work was to introduce a novel approach in modeling urban population movements in order to improve the existing methodologies in the field of automaton-based modeling of geographical systems.

A classical epidemiologic model was used as the driver of the cellular automata model that incorporated several advantages to this type of modeling. It enhanced the predictive capabilities, as well as the efficiency, of conventional automaton-based geosimulation platforms. The possibility of parameterizing the model by a notably few number of dependent and independent variables allowed for more efficient validation options, as well as less noise in the system. The implication of the model could have been exceedingly computationally expensive if the replicative calibration were to be performed on numerous independent variables, since the number of required simulation in every replicate grows geometrically, by increasing the number of initial conditions.

The appropriately small size of the required input data in this model, compared to other similar platforms that utilize a wide range of available input data from survey-based population census counts to high-resolution satellite imagery, opens the window for its further usability in other parts of the world. This would especially be more sensible in less developed countries, where the available technology and resources does not allow for the development of high resolution data collection. Instead, synthetic spatio-temporal population maps can be used to estimate the number of people being at risk in different times of a day in the wake of a major hazard. Studying the movement patterns can be used in planning for evacuation routes or even testing their efficiency by simulating the interruptions caused by different hazard scenarios. The range of hazards that the method can be used for is almost unlimited since incorporating a numerical exposure assessment could significantly reduce the damage caused by any natural as well as man-made hazard. Exposure assessment is a crucial part of the conventional Hazard-Vulnerability-Exposure trio in assessing the level or quantity of the threatening risk. Miscalculating the exposure, especially when estimating the endangered lives can disastrously affect the calculated risk, even if we have improved the vulnerability or predicted the hazard well in advance. This implies that there is an increasing need for more comprehensive models of exposure assessments. On the other hand, the growing boom in the availability of the so-called ‘’geo-located big data’’ about almost any urban service in modern cities has opened new grounds for modelers to think about finer scales of spatio-temporal models. Models similar to the one presented in this paper can benefit a lot from using these types of large data sets for calibration as well as validation of exposure assessment models.

Besides, the relative consistency of this type of modeling with the formal theories of complex systems’ dynamics, which is simulated through the stochastic development of the attraction-based CA, improves its credibility. The inter-dependency of the system components such as the cell states, the transition rule, the population probability estimation, the emergent behavior of the cell states, and the travel trends may all account for such consistency. Epidemiologically formulated geosimulation is therefore, considered as providing a new platform of modeling for future approaches in modeling spatially explicit complex systems.

It is inferable that it may be possible to build advanced models of urban movement patterns based on our basic understanding of the peoples’ inter-city travel habits. In other terms, a broadly accepted understanding of the microscopic decision making processes of the travellers could form a basis for predicting the emergent macroscopic behaviors of the system as a whole, which is useful for future conceptualization of modeling other urban systems. As discussed in this paper, even the obvious fact that various land use categories attract or repel

people differently in different time-intervals of the day could serve as a foundation for the development of high-resolution numerical estimation models.

Even though the predicted values were aggregated for numerical validation due to the lack of real-world data, we argue that it had little effect on the credibility of the output data since the model results have fairly satisfied the primary objectives of this modeling experiment. We believe that automaton-driven geosimulations, which have been previously validated mostly by visual techniques, must necessarily proceed toward methodologies that involve a combination of numerical as well as operational evaluation techniques. It might be the case, as it is here, that the unavailability of fine-scale real-world data does not let the formal numerical validation tests be comprehensively performed. In turn, using alternative numerical evaluation techniques such as replicative calibration and deviance measurements could be efficiently used in combination with operational tests so that an acceptable level of consistency, as well as validity, is maintained.

In this paper, the population was estimated for just a random working day in order to reduce the amount of calculations by the software so that we can replicate the simulation for several times. It could be a better practice if the time sensitivity of the algorithm is improved in a way that recognizes the difference between seasons, public holidays, weekends or even the days with extremely bad or desirable weather conditions. It is expected that population transition patterns are very much sensitive to these variations.

The evaluation of the data could have also been improved by comparing the datasets with other synthetic population models if found in similarly high resolutions, since the use of real-world input data such as mobile phone geo-locations are purposely avoided to extend the usability of the model.

There are broader potential applications to the presented work. Generating fine-scale spatio-temporal synthetic population maps, which are critically beneficial to the accuracy as well as the efficiency of disaster management or any similar contingency plans, where initially targeted in this research, while several other applications of this model could still be experimented. Specifically, the presented model is believed to be able to be used as a powerful decision-support tool if integrated to any other spatial modeling infrastructure. It could either act as an exposure assessment tool or a model for pattern detection and trend prediction applications with a wide variety of implications, ranging from transport modeling and urban growth analyses to business development location analysis, strategic energy planning, and climate-change adaptation strategy testing.

References

Adamatzky, A., Alonso-Sanz, R., Lawniczak, A., 2008. Automata-2008: Theory and Applications of Cellular Automata. Luniver Press.

Ali, W., Moulin, B., 2005. 2D-3D multiagent geosimulation with knowledge-based agents of customers' shopping behavior in a shopping mall, in: Spatial Information Theory. Springer, pp. 445–458.

Batty, M., Jiang, B., Thurstain-Goodwin, M., 1998. Local movement: agent-based models of pedestrian flows [WWW Document]. URL http://discovery.ucl.ac.uk/225/ (accessed 10.14.12).

Batty, M., Xie, Y., 1994. From cells to cities. Environ. Plan. B 21, s31–s31.

Belcher, C.M., Yearsley, J.M., Hadden, R.M., McElwain, J.C., Rein, G., 2010. Baseline intrinsic flammability of Earth's ecosystems estimated from paleoatmospheric oxygen over the past 350 million years. Proc. Natl. Acad. Sci. 107, 22448–22453.

Bhaduri, B., Bright, E., Coleman, P., Urban, M., 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal 69, 103–117.

Brauer, F., Castillo-Chávez, C., 2012. Mathematical Models in Population Biology and Epidemiology. Springer.

Brown, D.G., Page, S., Riolo, R., Zellner, M., Rand, W., 2005. Path dependence and the validation of agent-based spatial models of land use. Int. J. Geogr. Inf. Sci. 19, 153–174.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. Photogramm. Eng. Remote Sens. 66, 849–857.

Freire, S., 2010. Modeling of Spatiotemporal Distribution of Urban Population at High Resolution – Value for Risk Assessment and Emergency Management, in: Konecny, M., Bandrova, T.L., Zlatanova, S., Cartwright, W., Gartner, G., Meng, L., Peterson, M.P. (Eds.), Geographic Information and Cartography for Risk and Crisis Management, Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, pp. 53–67.

Ganguly, N., Sikdar, B.K., Deutsch, A., Canright, G., Chaudhuri, P.P., 2003. A Survey on Cellular Automata.

Gregory, A.W., Smith, G.W., 1991. Calibration as Testing: Inference in Simulated Macroeconomic Models. J. Bus. Econ. Stat. 9, 297–303.

IPCC, 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor, and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp.

Itami, R.M., 1994. Simulating spatial dynamics: cellular automata theory. Landsc. Urban Plan. 30, 27–47.

Kahn, M.E., 2005. The Death Toll from Natural Disasters: The Role of Income, Geography, and Institutions. Rev. Econ. Stat. 87, 271–284.

LandScan - Documentation [WWW Document], 2013. URL http://web.ornl.gov/sci/landscan/landscan_documentation.shtml (accessed 9.27.13).

Li, X., Yeh, A.G.-O., 2000. Modelling sustainable urban development by the integration of constrained cellular automata and GIS. Int. J. Geogr. Inf. Sci. 14, 131–152.

Marceau, D.J., Benenson, I., 2011. Advanced Geo-Simulation Models. Bentham Science Publishers.

Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecol. Model. 68, 21–32.

Modarres, M., 2006. Risk analysis in engineering: techniques, tools, and trends. CRC Press.

Montgomery, D.C., Runger, G.C., 2010. Applied Statistics and Probability for Engineers. John Wiley & Sons.

Nadim, F., Schack Pedersen, S.A., Schmidt-Thomé, P., Sigmundsson, F., Engdahl, M., 2008. Natural hazards in Nordic countries. Episodes 31, 176.

National Research Council (US), 2007. Tools and methods for estimating populations at risk from natural disasters and complex humanitarian crises. Natl Academy Pr.

Purdy, G., 2010. ISO 31000:2009—Setting a New Standard for Risk Management. Risk Anal. 30, 881–886.

Statistics Norway [WWW Document], 2013. . Stat. Nor. URL www.ssb.no

Straatman, B., White, R., Engelen, G., 2004. Towards an automatic calibration procedure for constrained cellular automata. Comput. Environ. Urban Syst. 28, 149–170.

Torrens, P.M., Benenson, I., 2005. Geographic Automata Systems. Int. J. Geogr. Inf. Sci. 19, 385–412.

UNISDR, 2011. Global Assessment Report on Disaster Risk Reduction: Revealing Risk, Redefining Development. United Nations International Strategy for Disaster

Reduction, Geneva, Switzerland, www.preventionweb.net/english/hyogo/gar/2011/en/home/index.html.

Vliet, J. van, White, R., Dragicevic, S., 2009. Modeling urban growth using a variable grid cellular automaton. Comput. Environ. Urban Syst. 33, 35–43.

White, R., Engelen, G., Uljee, I., 1997. The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. Environ. Plan. B 24, 323–344.

White, R., Uljee, I., Engelen, G., 2012. Integrated modelling of population, employment and land-use change with a multiple activity-based variable grid cellular automaton. Int. J. Geogr. Inf. Sci. 26, 1251–1280.

Wolfram, S., 2002. A new kind of science. Wolfram media Champaign.

Wu, F., 1998. SimLand: a prototype to simulate land conversion through the integrated GIS and CA with AHP-derived transition rules. Int. J. Geogr. Inf. Sci. 12, 63–82.