International Journal of Remote Sensing
Vol. 31, No. 21, 10 November 2010, 5689–5707

Taylor & Francis
Taylor & Francis Group

Estimating small-area population growth using geographic-knowledge-guided cellular automata

F. BENJAMIN ZHAN*†‡§, FELIPE OMAR TAPIA SILVA†¶ and MAURICIO SANTILLANA¶|

†Texas Center for Geographic Information Science, Department of Geography, Texas State University, San Marcos Texas, TX 78666, USA

‡School of Resource and Environmental Science, Wuhan University, Wuhan 430079, China

§Laboratory for Earth and Space Information Technologies, Shenzhen Institute of Advanced Technologies, Shenzhen 518055, China

¶Research Center in Geography and Geomatics 'Ing. Jorge L. Tamayo' A.C., México D.F.

|Center for the Environment, Harvard University, Cambridge, MA 02138, USA

Estimation of small-area population counts in an intercensal year and in a future year is a challenging task. This paper presents preliminary results in the development of a geographic-knowledge-guided cellular automata (CA) for modelling growth in small geographic areas. Geographic knowledge contains rules dictating growth patterns that typically cannot be captured by a traditional CA model. Nighttime stable light images and census population counts in censal years are used to determine base-year population counts in each cell in the CA model, and these estimated base-year population counts are used to manually calibrate the model. We use census data in 1990 and 2000 in El Paso County of Texas as the base-year population data, develop a set of rules based on specific urban-growth situations in El Paso and use the model to estimate population counts in block groups in a future year in the study area. Preliminary results in El Paso County suggest that the model has the potential to produce reasonably accurate population counts in sub-county areas in a future year. Future work will include the development of computational procedures that can be used to automate the calibration of the CA model.

1. Introduction

In some applications, such as health research and water demand management, it is necessary to use estimated population counts in small geographic areas (e.g. sub-county areas). In the developed world, population data in small geographic areas are known in censal years, but estimation of population counts in small areas in an intercensal year and in a future year with acceptable accuracy has remained a challenging problem. When the base-population data in a censal year is known, the key for estimating population counts in a small area from the base population in a censal year is the determination of population growth/decline in the small area from the censal year. The problem then becomes how to model population growth/decline in a small area from that censal year and then use the results from the model to estimate population counts in the small area in a given year.

*Corresponding author. Email: zhan@txstate.edu

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals
DOI: 10.1080/01431161.2010.496802

Among a variety of methods, cellular-automata (CA) models have been widely used to simulate urban growth. One limitation of existing CA models is that they lack local geographic knowledge that typically dictates growth patterns in a specific region. We first argue that the inclusion of geographic knowledge in CA models is important for more accurately simulating urban-growth patterns and associated spatial distribution of population growth in a given area. We then discuss procedures for developing such a model and present preliminary results of a geographic-knowledge-guided CA model that can be used to estimate small-area population growth.

2. Related work

2.1 *Nighttime imagery (NTI) and its application in population estimation*

Nighttime imagery (NTI) is a product of the Operational Linescan System (OLS) of the US Defense Meteorological Satellite Program (DMSP). Some researchers have verified the usability of NTI to distinguish and characterize urban areas and their extensions. For example, Henderson *et al.* (2003) used DMSP stable lights and radiance-calibrated images to delineate the boundaries of urban areas in cities where the levels of urbanization and economic development were different. They compared these results with those obtained from high-resolution Landsat Thematic Mapper (TM) images and computed light thresholds that minimized the discrepancies between TM images and NTI. They then used the thresholds to calibrate NTIs to monitor the growth of cities with comparable levels of development and urbanization. Amaral *et al.* (2006) used NTI data to detect and estimate urban population in the Amazon region. These researchers recorded urbanized settlements that were larger than 2.5 km^2 in the study area. Kohiyama *et al.* (2004) used NTI to estimate areas damaged by natural disasters based on an index measuring the loss of city light in the impacted area.

Milesi *et al.* (2003) used a 1992 Landsat-based land-cover map, Moderate Resolution Imaging Spectroradiometer (MODIS) data and NTI derived from DMSP OLS to estimate the extent of urban development and its impact on net primary productivity. Their approach provides a means for rapidly assessing changes in urban land use and their impacts on ecosystem resources at a regional scale. Doll *et al.* (2000) considered the NTI lit area of a city and combined it with statistical information to estimate socio-economic parameters and greenhouse-gas emissions. Sutton (2003) used NTI as a proxy measure of urban areas and used a combination of census block-group level data from the 1990 US census and NTI data to estimate the population size of some urban areas. In this study, we use an approach similar to the one used by Sutton to estimate population size in urban areas, but in much smaller geographic areas, in cells at a resolution of 85 m.

2.2 *CA models for urban-growth modelling*

CA models can be understood as discrete and nonlinear dynamic systems that consist of regular grid cells in a spatial domain, each cell being in a state that depends on the previous states of the neighbouring cells via a transition rule. Variations to this formalism in the dynamics of the model have been implemented to include relevant spatial factors such as terrain elevation, slope, connectivity, distance to roads and land price, among many other factors. These variations have been called CA-based models and have become a natural choice in urban modelling because these models can be used to mimic urban global structures or patterns from local interaction rules.

CA models have been widely used to simulate urban growth (e.g. Batty *et al.* 1999, Wu and Webster 2000). In CA models, various approaches have been investigated to define cell sizes and transition rules. These approaches included Monte Carlo simulations (Clarke and Gaydos 1998) and neural networks (Li and Yeh 2002, Guan *et al.* 2005). O'Sullivan and Torrens (2000) discussed some of the limitations of using CA models as representations of human systems and indicated that various attempts had been made to improve CA models to model urban growth. These researchers suggested that theoretically motivated improvements in the formalism of CA models are necessary to understand how variations of the model would affect the behaviours of model dynamics.

Other researchers have observed serious problems of the CA technique as a 'bottom-up' simulation approach. These researchers have proposed some improvements in the model to address influences of certain processes associated with urban growth in large geographic areas. Ward *et al.* (2000) considered macro-scale economic, political and cultural driving forces in the model and studied how these factors would influence urban expansion and how the inclusion of these factors in the model would improve the performance of general growth rules defined normally in the context of a 'bottom-up' CA-based modelling approach. Using a similar approach employed by Ward *et al.* (2000), He *et al.* (2006) treated urban expansion as a complex process that is self-organizing at a local level, but constrained and modified by several broad-scale factors in a broader context. Examples of these factors include socio-economic and political systems, urban and regional planning policies, as well as environmental and natural resource constraints. He *et al.* (2006) coupled a CA-based model and one 'top-down' system dynamics (SD)-based model to accomplish this goal. The coupled model had the capacity of predicting complex system changes under different 'what-if' scenarios.

In this study, we include a geographic-knowledge layer in the CA model to identify regions where settlements are most likely to occur or grow, resembling human-expert knowledge about possible growth in a given area. By human-expert knowledge, we simply mean the knowledge that urban and regional planners and developers may have about a certain geographic region. We advocate that in future implementations of CA models, real expert knowledge should eventually be acquired and used in the models. In the following sections, we briefly describe the Strabo technique (in honour of the ancient Greek cartographer) as an example of a process that may be useful in synthesizing such knowledge into a geographic-knowledge layer. We then describe an extension of the methodology developed by Santillana and Serrano (2005) through an inclusion of the geographic-knowledge layer mentioned above to represent additional factors to be considered in the transition rules of the model.

Even though Santillana and Serrano conceived their model subject to the limitations and needs of a particular case study in which urbanization took place in a protected natural preserve in the outskirts of Mexico City, the adjustable transition rules of their CA model allowed them to capture processes also observed in the dynamics of El Paso County, Texas, in which land development took place, for example, through diffusion along roads or due to the existence of previous settlements and through densification in a given settlement along roads or in surrounding areas. The list of spatial factors they used included: closeness to roads, closeness to previous settlements, slope, major topographical features and the state of the neighbourhood of a given cell in the grid. In this study, we use a modified version of their model and include a geographic-knowledge layer in the model. We keep the same list of factors that Santillana and Serrano used in their model. Details of the geographic-knowledge-guided CA model and its implementation are described in §4.

3. Estimation of past and current small-area population using census and NTI data

As mentioned in the previous section, the basic idea of using stable lights to estimate population counts is to determine the proportion of radiance of stable lights in a grid cell relative to the total radiance in a block-group polygon. Because the population counts in a block-group polygon are known, the population in a grid cell can be estimated using the proportion of radiance mentioned above. We used a four-step procedure summarized in figure 1 to estimate the population density in each grid cell.

In the first step, we aggregated the 1990 and 2000 point-population data to each block-group polygon and calculated the population counts in each polygon for 1990 and 2000. Although the boundaries of block-group polygons could change from 1990 to 2000, this change is not important for the purpose of estimating population counts in a grid cell, and the usage of the 2000 block-group polygons serve this purpose well.

Second, we constructed a grid covering the study area and obtain stable lights in each cell and in each block-group polygon for both 1990 and 2000. We then aggregated the stable lights to each cell and each block-group polygon. To minimize the boundary effect, we used a cell size of 85 m. Because of this fine resolution, the boundary effect became tolerable, although it cannot be completely eliminated. We conducted various experiments during the study and made sure that the boundary effect was within a tolerable range. For each cell that crossed the boundaries of different block-group polygons, we treated it as if the cell completely belonged to the block-group polygon that covered most of the cell among all polygons that covered part of the cell. This simple treatment avoids the problem of having to allocate population counts from different block-group polygons to the same cell.

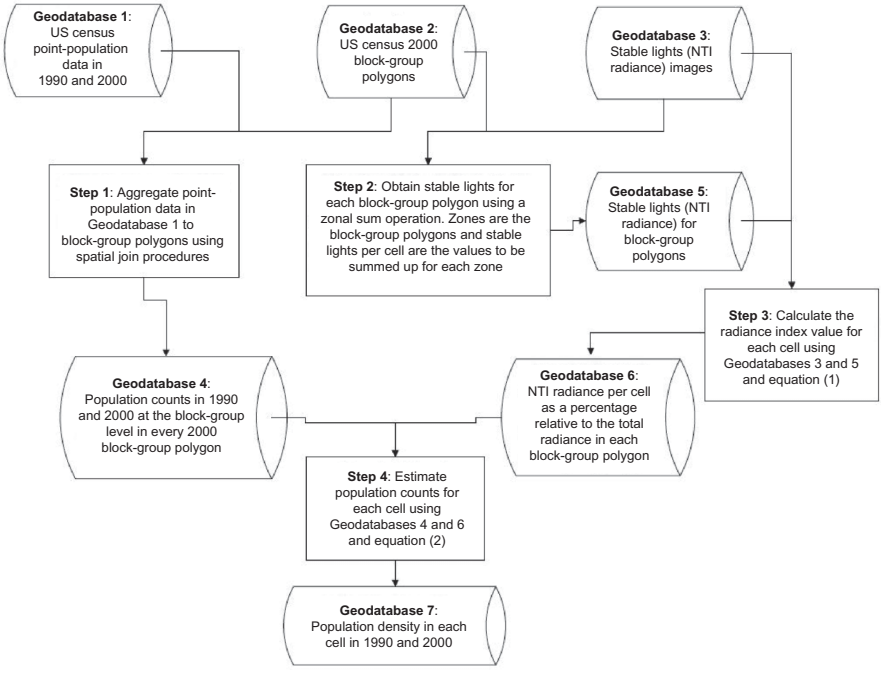

Figure 1. Procedures for estimating population density in a grid cell using census block-group data and nighttime imagery (NTI) data.

Third, we use equation (1) to calculate the radiance index associated with grid cell *i*:

$$P_{ik} = \frac{\mathrm{Rad}_i}{\sum_i^n \mathrm{Rad}_i} = \frac{\mathrm{Rad}_i}{\mathrm{Rad}^k}, \tag{1}$$

where P_{ik} is the proportion between the radiance in cell i and the radiance in block-group polygon k in which cell i is located, Rad_i is the radiance in cell i, n is the number of grid cells covered by block-group polygon k and Rad^k is the total radiance in block-group polygon k.

In the fourth and final step, we use equation (2) to estimate the population counts in each grid cell:

$$\mathrm{Pop}_{ik} = \mathrm{Pop}_k P_{ik}, \tag{2}$$

where Pop_{ik} is the population counts in cell i located in block-group polygon k and Pop_k is the population counts in block group k.

4. A knowledge-guided CA

4.1 *Basic components of the CA model*

Throughout the process of allocating population to grid cells using census data and NTI, we obtained a map layer with grid cells containing an estimated number of people in 1990 and 2000. We used population density as the state variable of a cell and defined four cell states based on the population density in each cell. These four cell states are: not-populated (no people in a cell), low density (from 1 to 5 people in a cell), medium density (from 5 to 10) and high density (more than 10). The reason that we decided to use population density as the main state variable in the CA model is based on the assumption that population density changes over time reflect urban growth, meaning areas with high growth would eventually have a higher population density and areas with less growth would have a lower population density.

As mentioned in §2, we extend the model developed and implemented by Santillana and Serrano (2005) and add a geographic-knowledge layer to enhance the model in this study. This additional layer provides the needed flexibility in the transition rules of the model to resemble human-expert knowledge. By expert knowledge, we mean knowledge about the *most probable tendencies* in the growth or decline of a given area. This type of knowledge may be understood as local rules governing the growth in a given area or may be viewed as the consensus from a group of experts who have significant insights about growth patterns in the area in question. The term *group of experts* refers to decision makers, city planners, policy makers, developers or other people who may have deep knowledge about how an area may grow or play an active role in policy issues that affect growth patterns in the area.

Even though the Strabo technique was conceived as a decision-making support tool to build geo-spatial consensus among a group of experts (Luscombe and Poiker 1983), we believe that it can serve as useful tool to define local-growth rules in the geographic-knowledge layer. The main idea behind the Strabo technique is to bring a group of expert individuals together so that they can dynamically solve a problem in a geo-spatial environment through the use of a geographic information system (GIS). For our purposes, the idea would be to bring urban planners and stakeholders together with the objective of building consensus about different growth patterns in an area of interest.

The transition rules in the model are defined according to certain observed processes that depend on the location, connectivity and the densification rate of the areas

represented as cells in the model. For a detailed discussion about these and other transition rules, see, for example, O'Sullivan and Torrens (2000) or White and Engelen (1993). Examples of these processes are:

- Expansion of a settlement as a result of the influence of its proximity to the road networks in an area, a process often called diffusion along roads.
- Growth of a settlement influenced by both the existing populated areas in areas surrounding the settlement and its connectivity to road networks in the area.
- Densification rate of a given settlement.

In addition to the basic processes mentioned above, we need to develop the model in such a way that it can be used to account for a number of geographic factors that may affect growth in a given area. For example, cells whose slopes exceed a threshold value or cells located in areas where growth is not possible should be excluded from consideration of future growth. Some examples of these cells include areas covered by rivers, lakes, parks and roads.

Furthermore, a randomly defined map, referred to in the pseudo code as *RandomNum*, corresponding to a constant probability distribution throughout the study area, combined with a threshold value, was used to judge whether a specific cell would be processed at a specific time step in the simulation process using the CA model. This approach helps obtain more realistic growth patterns from the simulation. The state variable of a cell in the CA model may be assigned a value corresponding to one of the four cell states corresponding to different population densities as described above. During the simulation, a cell is only allowed to transition from its current state to the next state, meaning from not-populated to low density, from low to medium density, or from medium to high density.

The factors used to control a change of cell value from one state to the next include the minimum number of neighbours of a cell, the threshold distance from the cell to the closest road and the maximum distance to other urban settlements. The variable *neighbour value*, related to the minimum number of neighbours of a cell in a particular cell state, was obtained by adding the cell values of a circular neighbourhood of 3×3 cells. When computing the *neighbour value*, cells with category 1 population density (not-populated) receive a value 0, cells in category 2 (low density) are assigned a value 1, cells in category 3 (medium density) are given a value 10 and cells in category 4 (high density) get a value 100. These values are used to provide a one-to-one correspondence between the number of neighbours of a cell in a given state within the neighbourhood in question and the numeric value of the variable *neighbour value*. Other variables and the computation of their values can be understood similarly.

The growth in an area (change of cell states), as determined by the CA model, is controlled by the cell value calculated by the model and the values of variable *accelerate* from the geographic-knowledge layer of the model. A short pseudo code is shown below to illustrate a typical transition rule used in the CA model. In this example, not-populated cells are considered for a state change if they are located in a region where no geographic limitations exist and if the value associated with the cell by the randomly generated map is greater than the threshold value p, and if it satisfies the slope threshold criterion (related to the value of θ_1) and if it is located in an area where growth is most likely to occur as defined by the value of variable *accelerate* in the geographic-knowledge layer. Detailed discussions about the geographic-knowledge layer are provided in the next section.

The value of p is a number between 0 and 1 that the user provides in order to process more (closer to the value '0') or less (closer to the value '1') cells at each time step. In addition, the value of variable *accelerate* can be used to reflect the change of the state of a cell in two directions, an increase in population density or a decrease in population density. Once a cell satisfies the conditions mentioned above, the cell is considered for a state change provided that it also satisfies other appropriate conditions that account for its proximity to the road network, the state of its neighbours and its proximity to other settlements.

At a given time step:

IF (cell_state= not-populated) and (cell_factor= possible growth) and (RandomNum $\geq p$) and (slope $\leq \theta_1$) and (accelerate $\geq \gamma$)

THEN

IF (neighbour value $\geq \mu_2$) AND
IF (closeness_to_roads $\leq \mu_3$) and
IF (distance_to_centre_of_closest_settlement $\leq \mu_4$)

THEN (cell_state = low density).

4.2 *The multi-scale similarity index*

As stated above, we used a similarity index to quantitatively evaluate the results from the model against urban-growth situations on the ground in the study area when calibrating the model. The original similarity index defined by Santillana and Serrano (2005) can be used to assess similarity at multiple scales. However, we analysed the different domain partitions and decided to assess the similarity using 128 partitions only in the study area. More details will be provided in the next section. This approach allows us to objectively measure the effectiveness of the parameter values and the growth rules used in the model (i.e. the goodness-of-fit between results from the model and the actual data) and hence helps us choose the appropriate parameter values and local-growth rules.

The set of similarity indexes was previously defined by Santillana and Serrano (2005). The indexes produce a *multi-scale similarity metric*. To obtain these similarity indexes, we first divide the study area Ω into a set of sub-areas Ω_j such that the union of all sub-areas $U\Omega_j$ covers Ω completely, i.e. $U\Omega_j = \Omega$. Secondly, we use two binary images O and R (where O is the output image of the model and R is the actual image) to construct a similarity index i at a given scale using the formula below (Santillana and Serrano 2005):

$$i = i(R, O) = \frac{1}{N(\Omega_j)} \sum_{c_k \in \Omega_j} (c_k(R) - c_k(O))^2, \quad 0 \leq i \leq 1, \tag{3}$$

where $N(\Omega_j)$ is the total number of sub-areas Ω_j and c_k (R) and c_k (O) represent the state of a given cell in Ω_j, either 1 or 0, in the binary images R and O, respectively.

For each pair of images R or O, the value of i represents the square of the count of cells that are different from one image to the other, and it is normalized to 1. The higher the value of i is, the larger the difference between the two images. For the simplest case (given by the choice $N(\Omega_j) = 1$, where the only element of the partition $\Omega_1 = \Omega$), i does not provide any morphologic insight about the similarity of the

images. It only provides a normalized count of the number of cells that are different from one image to the other.

4.3 *Model calibration*

Once the model is initiated, the next step is to use an appropriate technique to calibrate the model. We used the following steps to calibrate the model. First, based on an initial state map with cells containing an estimated number of people in each cell in 1990, we manually changed the values of variables p and μ_i in the transition rules to conduct the simulations using the CA model without the geographic-knowledge layer and obtained population counts in each cell in 2000. As stated above, the value of p is a number between 0 and 1. The ranges of values of μ_i were set to vary as follows: closeness to roads, 200–300 m; and distance to centre of closest settlement, 200–1000 m. In addition, the status of a cell reflecting its closeness to human settlements is set to change from 0 to 1 when the sum of the cell values in a 3×3 window reaches 8; from 1 to 2 when the sum reaches 500; and from 2 to 3 when the sum reaches 1000.

We then compared the simulated population counts in the cells with the estimated population data in 2000. We initially compared the results from the model with the estimated data by manual visual inspections and realized that it was difficult to use manual visual inspection to distinguish improvements in the results of the model. Therefore, we used the similarity index proposed by Santillana and Serrano (2005) to quantitatively evaluate the results from the model against the estimated population data. We repeatedly ran the simulation a sufficient number of times until the best possible goodness-of-fit between the results of the model and the estimated population data in 2000 was obtained.

We observed that, even though the growth dynamics in the majority of the study area was captured by the model, some parts of the study area showed a strong level of disagreement, presumably due to the differences in the local-growth processes taking place in different parts of the study area. A geographic-knowledge layer was then added to the CA model to account for the effects of these local-growth rules. We provide a detailed discussion about the generation and implementation of the geographic-knowledge layer in the next subsection.

4.4 *Local-growth rules and the construction of a geographic-knowledge layer*

Local-growth rules in the geographic-knowledge layer can be defined in different ways. In this study, we examined growth patterns from 1990 to 2000 in the study area and identified five different categories of areas with atypical growth patterns as stated below. These five different categories of areas were used as local-growth rules in the geographic-knowledge layer.

- Areas that experienced population decline.
- Areas where new urban settlements were not likely to take place. For example, the presence of a military field (containing an airport and a big park in the middle of the city) in the study area prevented the establishment of new urban settlements. The presence of this military base, however, caused a higher rate of densification in the surrounding areas. This was taken into account in the model, through both imposing a physical barrier (such as a lake or infinite slope) in the model resembling the military area and setting a higher growth rate in the surrounding areas (through the value of variable *accelerate*).

- Areas experienced either lower or higher rates of growth than other areas. Although these areas appeared to have the same conditions as other areas that exhibited average growth, growth rate in these areas was either lower or higher than that in other areas, presumably due to factors that are not or cannot be taken into account once the relevant variables in the model are chosen.
- Areas with close proximity to roads that exhibited more intensive densification than other areas. This growth situation was also reported by Silva and Clarke (2002).

These differences of growth patterns in these areas were represented in the model using different values of variable *accelerate* to help control the growth process in the CA simulations. These five categories of areas and their values of variable *accelerate* were: (1) areas that experienced population decline (*accelerate* = '–1'), (2) areas in which growth is not permitted (0), (3) areas that experienced low growth (0.001), (4) areas that experienced fast growth (10) and (5) areas with proximity to roads that experienced more growth (25). This geographic-knowledge layer resembles, in some sense, the expert knowledge layer that should eventually be generated *a priori* in order to simulate a possible future scenario.

5. Case study

We used El Paso County of Texas in the US as the case study area to test the CA model. El Paso is a county located at the southwest corner of the state of Texas. The county had a total population of 594 571 in 1990 and 685 508 in 2000 based on information from the US Census Bureau. The population in the county increased more than 15.29% in the 10-year period. The city of El Paso is located in El Paso County, and it is a border city between the US and Mexico. The city has experienced fast growth in the past decade and it has been projected to be a fast-growing city in the foreseeable future. To understand the growth patterns and future growth trajectories in the city, it is important to model and simulate population growth in different parts of El Paso County. We describe how we used the model developed in this study to simulate population growth in different parts of the county in the rest of this section.

5.1 *Data compilation*

Table 1 provides a summary of input data sources and operations applied to the data to generate the necessary layers as input to the CA model. The main input data for this research were the 1990 and 2000 census data and the NTIs over the years of this time period. Additional data used for the simulation were digital elevation data, road networks and water bodies in the study area. The census data were obtained from the Environmental Systems Research Institute (ESRI) Data & Maps Media Kit®. The data included layers of point population geo-referenced to the centroids of street blocks in 1990 and 2000, as well as population data at the block-group level in 2000.

NTI from 1990 and 2000 consisted of downloaded maps showing stable lighting activity (radiance) derived from the DMSP OLS website (NGDC 2006). According to information at this site, the DMSP currently operates satellites carrying the OLS in low-altitude polar orbits. The DMSP OLS has the capability to detect low levels of visible–near-infrared (VNIR) radiance at night. The nighttime lights of human settlements were separated from other classes of lights (e.g. fires) based on location, brightness/persistence and visual appearance.

Table 1. Overview of input data layers of the CA model.

Data layer	Data source	Description
Number of people per grid cell in 1990 and 2000	The 1990 and 2000 population data georeferenced to the centroids of blocks and block-group polygons from the ESRI Data & Maps Media Kit®; Nighttime imagery as maps of stable lighting activity from DMSP OLS sensor.	US census population data aggregated to centroids of blocks were re-assigned to block-group polygons. An index based on stable lighting activity (radiance) was generated for each cell at a resolution of 85 m and used to determine the population counts in each cell. Cells are classified into four categories based on population density: not-populated: <1 person, low density: 1–5 persons, medium density: 5–10, high density: >10. Cells corresponding to water bodies, rivers or roads were assigned a 'null' value.
Grid cells representing rivers, parks and roads	ESRI Data & Maps Media Kit®.	Rivers, parks and roads in the study area were extracted from ESRI Data & Maps Media Kit®.
Grid cells covering water bodies	ESRI Data & Maps Media Kit®; Map of water bodies from STRM.	The layer of water bodies from the ESRI Data & Maps Media Kit® was combined with the SRTM water bodies provided by the USGS EROS Data Center to obtain the complete layer of water bodies.
Grid cells containing slope in percentage	1 arc second (30 m) SRTM DTED® Level 2 'finished' data derived from SRTM IFSAR data.	DEM data were downloaded and used to calculate slope expressed in percentage. The DEM data were interpolated (bilinear) to obtain DEM data at the resolution of the cell size used in the CA model (85 m × 85 m).
Grid cells with distance to roads	Roads from the ESRI Data & Maps Media Kit®.	Euclidian distance from each cell to its nearest road was calculated using a Euclidian distance function in Arc/Info, and this distance was used to represent the minimum distance from the cell to its nearest road.

The NTIs are cloud-free composites produced from archived DMSP OLS smooth-resolution data for different calendar years. The products used in this study are the images corresponding to 30 arc second grids. The digital elevation model (DEM) data were obtained from the project Shuttle Radar Topography Mission (SRTM) operated by the US Geological Survey (USGS) Earth Resources Observation Systems (EROS) Data Center (2006). These DEM data were a part of the 1 arc second (30 m) SRTM Digital Terrain Elevation Data (DTED®) Level 2 'Finished' data derived from SRTM Interferometric Synthetic Aperture Radar (IFSAR) data. The layer of roads was also

obtained from this data source. The layer of rivers and other water bodies was created by combining the water bodies obtained from the SRTM and those from the ESRI map layer of geographic water bodies.

All geographic data were projected to the Universe Transverse Mercator (UTM) coordinate system with the parameters associated with WGS84 zone 13N. The downloaded SRTM DEM was converted to a grid-format data file and processed to fill out possible void areas. The slope (in percentage) map was derived from the DEM data. The roads were used to construct the map layer of minimum Euclidian distance from a cell to its nearest road. This distance was calculated using a function in the Grid module of Arc/Info. Cells corresponding to water bodies, rivers or roads were assigned the value 'null', meaning no growth was possible in these cells.

As stated above, we used a grid-cell resolution of 85 m in this study. This resolution was chosen after a number of experiments with the NTI data and the census population data. We found that when the original cell size of NTI (approximately 0.00833° or 850 m under the UTM projection) was used for estimating the population counts in each cell following the procedures discussed in §2 and summarized in figure 1, there was a loss of about 20% in the total population in El Paso County in 2000 when we compared the estimated population data against the 2000 census population data. When we changed the resolution to 85 m, the population loss no longer existed. Figure 2 shows the calculated radiance of grid cells in the study area in 1990.

When running the CA-based model, all input data layers were converted to a resolution of 85 m under the UTM projection. The SRTM DEM data were

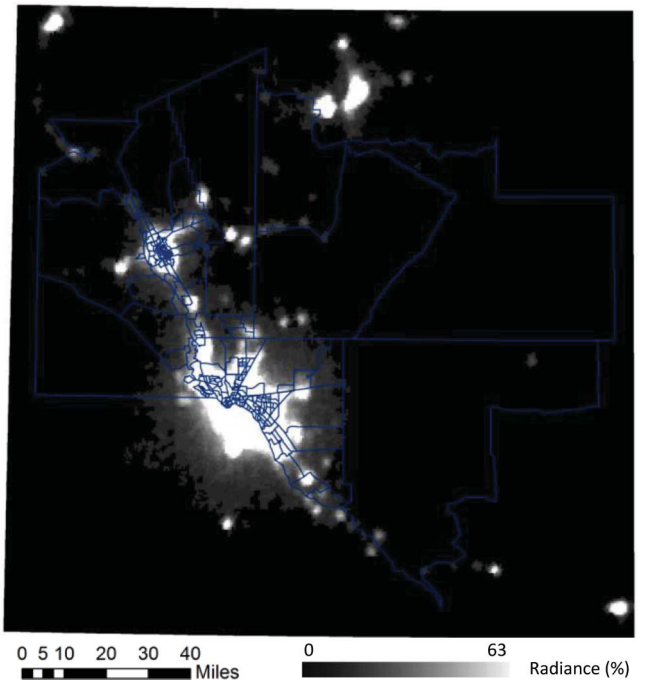

Figure 2. Radiance of grid cells in the study area for 1990 obtained from nighttime imagery (NTI) derived from the Operational Linescan System (OLS) of the US Defense Meteorological Satellite Program (DMSP).

re-sampled from its original resolution of 30 m. This is a normal procedure used by the SRTM team to give global access to the SRTM DEM for all countries in the world. They re-sampled the 30 m DEMs to obtain the 90 m DEMs that can be downloaded by the general public.

5.2 *Model calibration*

The model was used to simulate the population density in 2000, using the estimated 1990 population density from the census and NTI data as the initial state. Then, we manually calibrated the model, by changing the values of the parameters, in order to obtain a simulated population density as close as possible to the one estimated by the census and NTI data for 2000. We used 1 year as the time step in the calibration and the simulations. We ran the model more than 250 times when calibrating the model. The goodness-of-fit between the simulated population density in 2000 and the estimated population density in 2000 at each trial run was measured by the similarity index described in the last section. After a number of different trial runs, we observed that the domain partition defined by $N(\Omega_j) = 128$ was suitable for evaluating the performance of the model. We thus used this domain partition to compute the similarity index.

The goodness-of-fit of the CA model can be illustrated in different ways. Figure 3 gives the goodness-of-fit of the calibrated model when we classify the study area into not-populated and populated areas. Figure 3(*a*) is an image showing the not-populated and populated areas in 2000 based on estimated population density from census and NTI data, figure 3(*b*) is an image depicting the simulated not-populated and populated areas in 2000 using the CA model without local-growth rules and figure 3(*c*) is an image showing the simulated not-populated and populated areas in 2000 using the CA model with local-growth rules. A similarity index was constructed between the estimated 2000 population density and the simulated population density without local-growth rules (figure 3(*d*)), and another similarity index was constructed between the estimated 2000 population density and the simulated 2000 population density with local-growth rules (figure 3(*e*)). Figure 3(*f*) gives a graphical comparison of the two similarity indices.

For the majority of the domain partitions, the values of i in the image obtained by the CA model without local-growth rules are much greater than those of their counterparts in the image obtained by the CA model with local-growth rules. This situation indicates that the CA model with local-growth rules is more suitable to mimic the real growth situation on the ground from 1990 to 2000. In the images shown in figures 2(*d*) and 2(*e*), partitions (5,6), (8,7) and (10,10) have the greatest differences between their corresponding i values in the two images, suggesting that the inclusion of an 'expert' knowledge layer, or local-growth rules, in the model is an effective way to improve the power of the CA model.

In a similar manner, we can divide the study area into high-population-density areas and other areas, construct the images showing the two categories of areas and determine the similarity indices. Figure 4 illustrates the images showing the two categories of population density in the study area. These images can be understood in a way similar to those of figure 3 described above.

5.3 *Simulation results and evaluation*

Figure 5 shows the input population density in 1990 and 2000 in the study area estimated from census and NIT data (figure 5(*a*) and 5(*b*)), the simulated population density in 2000 using the CA model without local-growth rules (figure 5(*c*)), the

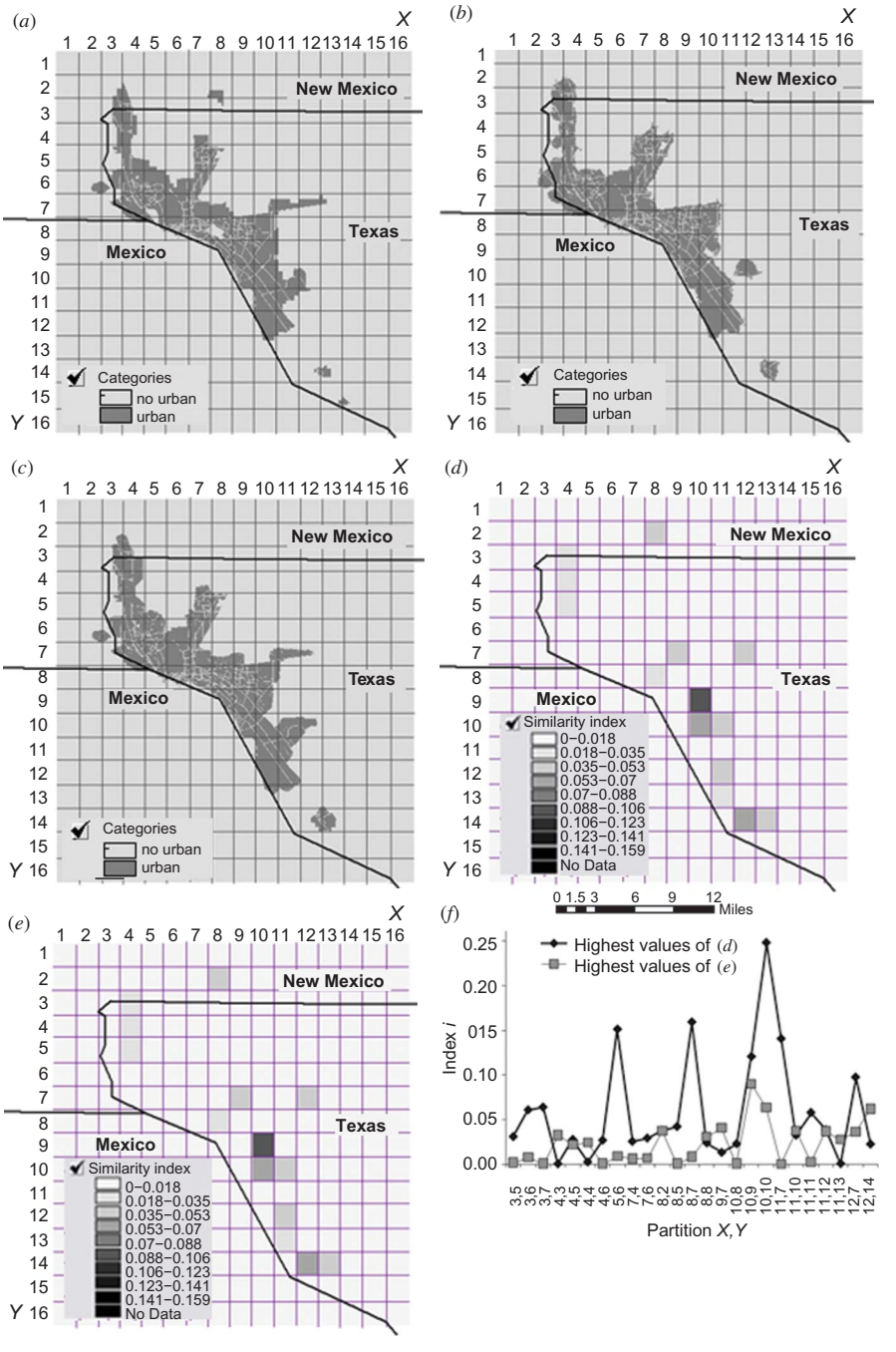

Figure 3. Goodness-of-fit images of estimated and simulated population density in not-populated areas (less than 1 person cell^{-1} at a resolution of 85m) and urban areas (1 or more people cell^{-1}) in 2000 as measured by the similarity index i for an image partition of $N(\Omega_j) = 128$. (*a*) Not-populated areas in 2000 as estimated from census and NTI data, (*b*) simulated not-populated and populated areas in 2000 without local growth rules, (*c*) simulated not-populated and populated areas in 2000 with local growth rules, (*d*) similarity index i based on 128 partitions (x,y) calculated with $R = (a)$ and $O = (b)$ using equation (3), (*e*) similarity index i based on 128 partitions (x,y) calculated with $R = (a)$ and $O = (c)$ using equation (3) and (*f*) a graphic comparison of (*d*) and (*e*).

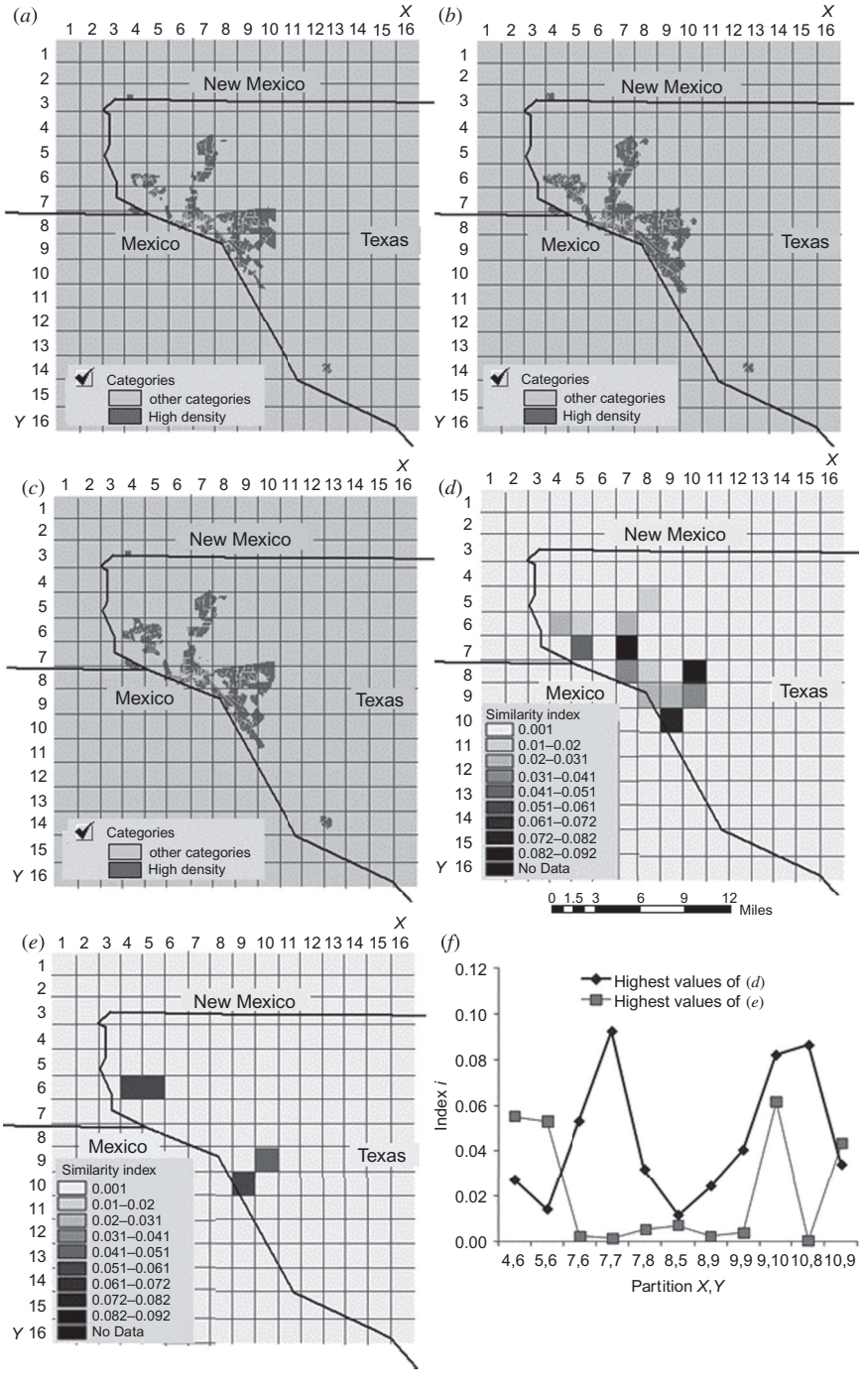

Figure 4. Goodness-of-fit images of estimated and simulated population density in areas with high population density (more than 10 people cell^{-1}) as measured in 2000 by the similarity index i for an image partition of $N(\Omega_j) = 128$. (*a*) High-population-density areas in 2000 as estimated from census and NTI data, (*b*) simulated high-population-density areas in 2000 without local growth rules, (*c*) simulated high-population-density areas in 2000 with local growth rules, (*d*) similarity index i based on 128 partitions (x,y) calculated with $R = (a)$ and $O = (b)$ using equation (3), (*e*) similarity index i based on 128 partitions (x,y) calculated with $R = (a)$ and $O = (c)$ using equation (3) and (*f*) a graphic comparison of (*d*) and (*e*).

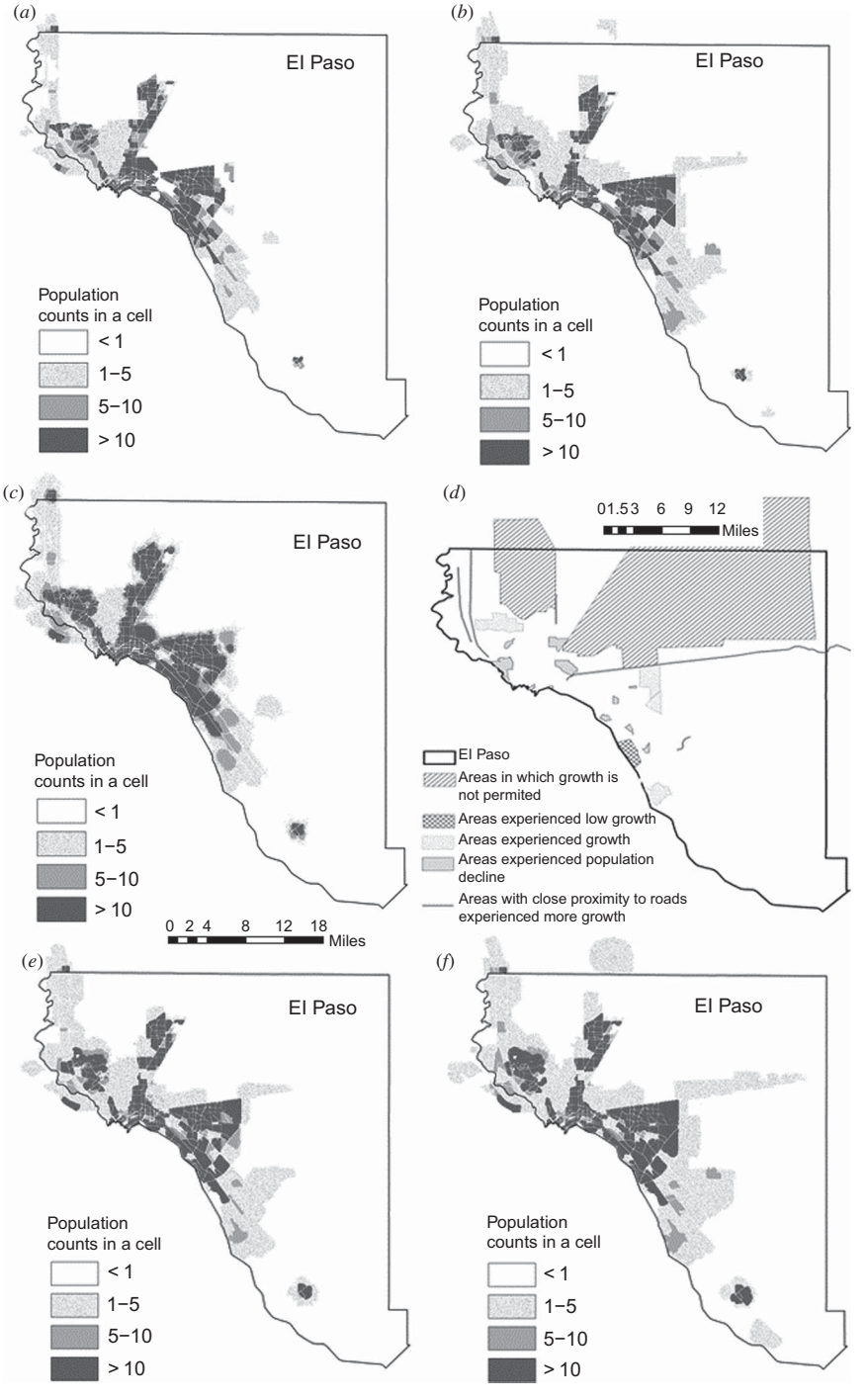

Figure 5. Estimated and simulated population density in El Paso County, Texas, in 1990, 2000 and 2011 (grid cell resolution: 85 m). (*a*) Population density in 1990 as estimated from census and NTI data, (*b*) population density in 2000 as estimated from census and NTI data, (*c*) simulated 2000 population density without local growth rules, (*d*) selected areas for defining local growth rules in the geographic knowledge layer, (*e*) simulated 2000 population density with local growth rules and (*f*) simulated 2011 population density with local growth rules.

selected area for defining and calibrating the local-growth rules (i.e. the geographic-knowledge layer) (figure 5(*d*)), the simulated population density in 2000 with local-growth rules (figure 5(*e*)) and the simulated population density in 2011 using the calibrated CA model with local-growth rules (figure 5(*f*)). The total number of people in El Paso County will be 803 408 in 2011 based on simulation results from the CA model. This 2011 total population obtained from the CA model is remarkably close to the total population projected by the Texas State Data Center, who projected that the total population in El Paso County could reach 804 349 in 2010 (Texas State Data Center 2007). Given that population in the county is assumed to continue to grow, the CA model only slightly underestimated the total population in the county.

To evaluate the simulation results, we also compared the simulated population counts in 2011 against the estimated population counts in 2011 from Claritas (2006). Based on data from Claritas, the estimated population counts in El Paso County in 2011 will be 742 687. This number is 60 721 (8.18%) less than the number estimated by the CA model, and it is also noticeably lower than the number projected by the Texas State Data Center. Nevertheless, this was the only projected population data in 2011 at the census block-group level that we had access to. Although there is no way to tell that the estimated population data at the census block-group level from Claritas are accurate, the Claritas data serve as one source of reference.

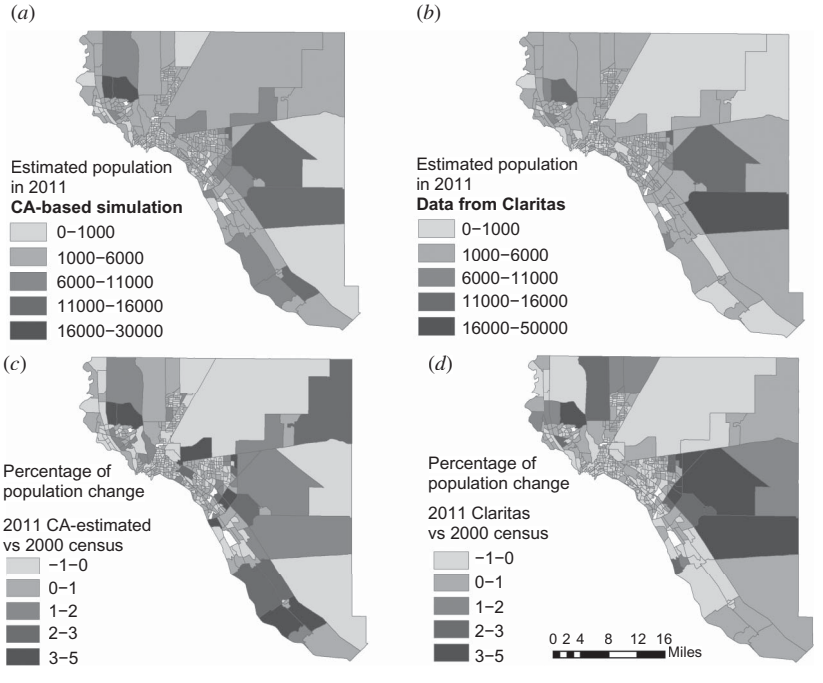

Figure 6. A comparison of simulated population counts and population change at the block group level from 2000 to 2011 between simulated data using the CA model and data from Claritas. (Note: percentage of population change is calculated as the population change in a block-group polygon from 2000 to 2011 divided by the population change in the whole county during the same time period.)

Because of the difference in the total population counts at the county level between the estimated data from the CA model and the Claritas data, we compared the percentage of population change in each block-group polygon from 2000 to 2011 between the estimated data from the CA model and the data from Claritas, instead of using the absolute data at the block-group level. Figure 6 shows the map of population counts as well as the percentage of population change at the census block-group level. The percentages of population change (P_{ci}) shown in figure 6(*c*) and figure 6(*d*) are calculated using:

$$P_{ci} = \frac{\Delta P^i}{\Delta P}\%, \quad \Delta P^i = P^i_{2011} - P^i_{2000}, \tag{4}$$

where P_{ci} is the percentage of population change from 2000 to 2011 in block-group polygon i, ΔP^i is the population change from 2000 to 2011 in block-group polygon i based on either the simulated data of the CA model or the estimated data from Claritas, ΔP is the population change from 2000 to 2011 in the whole county based on either the simulated data or the estimated data from Claritas, ΔP^i_{2011} is the population counts in 2011 in block-group polygon i based on either the simulated data or the estimated data from Claritas and ΔP^i_{2000} is the census population counts in 2000 in block-group polygon i.

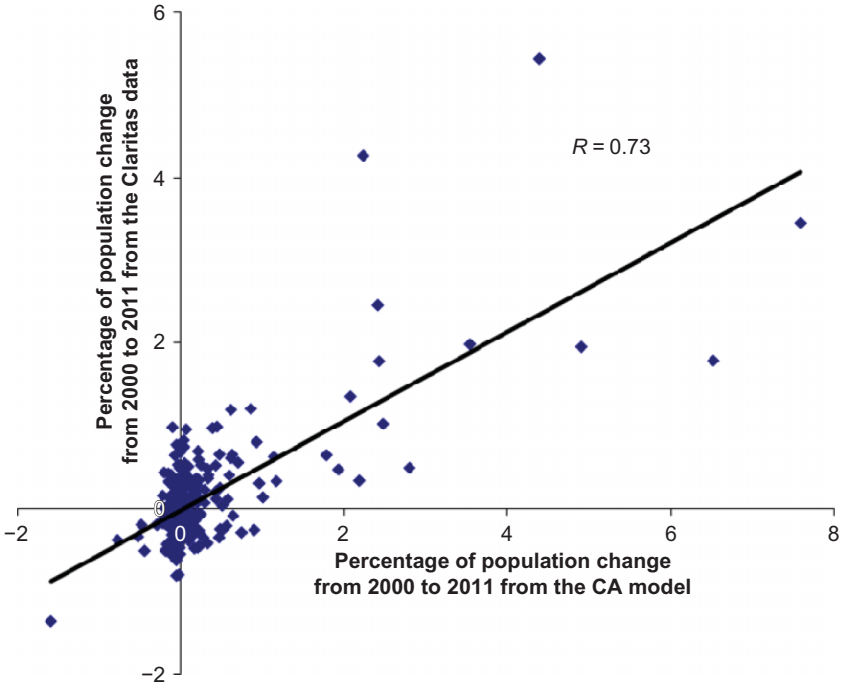

Figure 7. Correlation between percentages of population change from 2000 to 2011 at the block-group level between simulated results of the CA model and estimated data from Claritas (shown in figures 6(*c*) and 6(*d*)). (Note: percentage of population change was calculated as the population change in a block-group polygon from 2000 to 2011 divided by the population change in the whole county during the same time period.)

Figure 7 gives the Pearson correlation coefficient between the percentage of population change at the block-group level from 2000 to 2011 between the simulated population data from the CA model and the estimated population data from Claritas. A correlation coefficient of $R = 0.65$ was obtained when all 414 block groups in the study area were used in the analysis. When we excluded seven block groups (from a total of 414) that appeared to be outliers in the analysis, the correlation coefficient became 0.73 (figure 7).

6. Concluding remarks

We presented a geographic-knowledge-guided CA model and demonstrated how the model can be used in estimating small-area population growth in this paper. There are three essential components in this model: (1) a procedure for estimating past and current small-area population data using census and nighttime imagery (NTI) data, (2) a geographic-knowledge layer that can be used to represent local-growth rules governing the growth process in the area of interest and (3) a manual model-calibration process that helps fine-tune the model to more accurately reflect local-growth patterns. We used El Paso County in Texas, US, as a case-study area to test the model. Results from the case study suggest that the total population in the county in a future year (2011) aggregated from the estimated population counts at the census block-group level through the simulations matches that of the projected population reasonably well. In addition, the population counts at the block-group level obtained from the simulation are comparable to the data obtained from a commonly used data provider.

Acknowledgements

This study was in part supported by a grant from the US Department of Agriculture (USDA/CSREES award no: 2004-38 899-02 181). Part of this paper was written while Benjamin Zhan was visiting Wuhan University in China as a Chang Jiang Scholar Guest Chair Professor. Support from the USDA and the Chang Jiang Scholar Awards Program is greatly appreciated. The final completion of this manuscript was possible due to the generous Henson Environmental Fellowship awarded to Mauricio Santillana at the Harvard University Center for the Environment.

References

Amaral, S., Monteiro, A.M.V., Camara, G. and Quintanilha, J.A., 2006, DMSP/OLS nighttime light imagery for urban population estimates in the Brazilian Amazon. *International Journal of Remote Sensing*, **27**, pp. 855–870.

Batty, M., Xie Y. and Sun Z., 1999, Modeling urban dynamics through GIS-based cellular automata. *Computers, Environment and Urban System*, **23**, pp. 205–233.

Claritas, 2006, *Claritas Demographic Update Methodology*. Technical report (San Diego, CA: Nielsen Claritas). Available online at: http://www.claritas.com/collateral/methodology/2006_american_demographics_methodology.pdf (last accessed 7 October 2010).

Clarke, K.C. and Gaydos, L.J., 1998, Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. *International Journal of Geographical Information Science*, **12**, pp. 699–714.

Doll, C.N.H., Muller, J.P. and Elvidge, C.D., 2000, Nighttime imagery as a tool for global mapping of socio-economic parameters and greenhouse gas emissions. *Ambio*, **29**, pp. 157–162.

GUAN, Q., WANG, L. and CLARKE, K.C., 2005, An artificial-neural-network-based, constrained CA model for simulating urban growth. *Cartography and Geographic Information Science*, **32**, pp. 369–380.

HE, C., OKADA, N., ZHANG, Q., SHI, P. and ZHANG, J., 2006, Modeling urban expansion scenarios by coupling cellular automata model and system dynamic model in Beijing. *China Applied Geography*, **26**, pp. 323–345.

HENDERSON, M., YEH, E.T., GONG, P., ELVIDGE, C. and BAUGH, K., 2003, Validation of urban boundaries derived from global night-time satellite imagery. *International Journal of Remote Sensing*, **25**, pp. 595–609.

KOHIYAMA, M., HAYASHI, H., MAKI, N., HIGASHIDA, M., KROEHL, H.W., ELVIDGE, C.D. and HOBSON V.R., 2004, Early damaged area estimation system using DMSP-OLS night-time imagery. *International Journal of Remote Sensing*, **25**, pp. 2015–2036.

LI, X. and YEH, A.G.O., 2002, Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, **16**, pp. 323–343.

LUSCOMBE, B.W. and POIKER, T.K., 1983, STRABO: an alternative GIS approach to decision making for planning applications in data scarce environments. In *Proceedings of 6th International Symposium on Automated Cartography. American Congress on Surveying and Mapping*, Bethesda, MD, vol. 1, pp. 264–269.

MILESI, C., ELVIDGE, C.D., NEMANI, R.R., and RUNNING, S.W., 2003. Assessing the impact of urban land development on net primary productivity in the southeastern United States. *Remote Sensing of Environment*, **86**, pp. 401–410.

NATIONAL GEOPHYSICAL DATA CENTER (NGDC), 2006, US NOAA satellite and information service. Available online at http://www.ngdc.noaa.gov/dmsp/index.html (accessed 16 April 2006).

O'SULLIVAN, D. and TORRENS, P.M., 2000, Cellular models of urban systems. In *4th International Conference on Cellular Automata for Research and Industry*, September, Karlsruhe University, Karlsruhe, Germany.

SANTILLANA, M. and SERRANO, F., 2005, Calibration and validation of a CA based model for urban development simulation using an evolutionary algorithm: a case study in Mexico City. In *Proceedings of 9th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2005)*, University College London, London, UK.

SILVA, E.A. and CLARKE, K.C., 2002, Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. *Computers, Environment and Urban Systems*, **26**, pp. 525–552.

SUTTON, P.C., 2003, A scale-adjusted measure of 'urban sprawl' using nighttime satellite imagery. *Remote Sensing of Environment*, **86**, pp. 353–369.

TEXAS STATE DATA CENTER, 2007, Projected Texas population by county 2010. Available online at http://www.dshs.state.tx.us/chs/popdat/ST2010.shtm (accessed 10 March 2007).

USGS EROS DATA CENTER, 2006, Shuttle Radar Topography Mission. Available online at http://edc.usgs.gov/ (accessed 25 May 2006).

WARD, D., MURRAY, A. and PHINN, S., 2000, A stochastically constrained cellular model of urban growth. *Computer, Environment and Urban System*, **24**, pp. 539–558.

WHITE, R. and ENGELEN, G., 1993, Cellular automata and fractal urban form: a cellular modeling approach to the evolution of urban land use patterns. *Environmental Planning*, **25**, pp. 1175–1199.

WU, F. and WEBSTER, C., 2000, Simulating artificial cities in a GIS environment: urban growth under alternative regulation regimes. *International Journal of Geographical Information Science*, **14**, pp. 625–648.