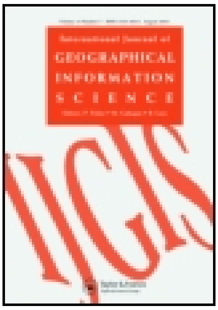


International Journal of Geographical Information Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

# Calibration of an urban cellular automaton model by using statistical techniques and a genetic algorithm. Application to a small urban settlement of NW Spain

A. M. García [a] , I. Santé [b] , M. Boullón [c] & R. Crecente [a]

[a] Land Laboratory, Department of Agroforestry Engineering, University of Santiago de Compostela, Lugo, La Coruña, Spain

[b] Agroforestry Engineering, University of Santiago de Compostela, Lugo, La Coruña, Spain

[c] SIT, edificio CACTUS, Santiago de Compostela University, Lugo, La Coruña, Spain

Version of record first published: 30 Jan 2013.

To cite this article: A. M. García , I. Santé , M. Boullón & R. Crecente (2013): Calibration of an urban cellular automaton model by using statistical techniques and a genetic algorithm. Application to a small urban settlement of NW Spain, International Journal of Geographical Information Science, DOI:10.1080/13658816.2012.762454

To link to this article: http://dx.doi.org/10.1080/13658816.2012.762454

*International Journal of Geographical Information Science*, 2013
http://dx.doi.org/10.1080/13658816.2012.762454

Taylor & Francis
Taylor & Francis Group



# Calibration of an urban cellular automaton model by using statistical techniques and a genetic algorithm. Application to a small urban settlement of NW Spain

A. M. García[a]*, I. Santé[b], M. Boullón[c] and R. Crecente[a]

[a]*Land Laboratory, Department of Agroforestry Engineering, University of Santiago de Compostela, Lugo, La Coruña, Spain;* [b]*Agroforestry Engineering, University of Santiago de Compostela, Lugo, La Coruña, Spain;* [c]*SIT, edificio CACTUS, Santiago de Compostela University, Lugo, La Coruña, Spain*

(*Received 18 June 2012; final version received 16 December 2012*)

Cellular automata (CA) stand out among the most commonly used urban models for the simulation and analysis of urban growth because of their ability to reproduce complex dynamics, similar to those found in real cities, from simple rules. However, CA models still have to overcome some shortcomings related to their flexibility and difficult calibration. This study combines various techniques to calibrate an urban CA that is based on one of the most widely used urban CA models. First, the number of calibration parameters is reduced by using various statistical techniques, and, second, the calibration procedure is automated through a genetic algorithm. The resulting model has been assessed by simulating the urban growth of Ribadeo, a small village of NW Spain, characterized by low, slow urban growth, which makes the identification of urban dynamics and consequently the calibration of the model more difficult. Simulation results have shown that, by automating the calibration procedure, the model can be more easily applied and adapted to urban areas with different characteristics and dynamics. In addition, the simulations obtained with the proposed model show better values of cell-to-cell correspondence between simulated and real maps, and the values for most spatial metrics are closer to real ones.

**Keywords:** urban growth; urban simulation; urban development

## 1. Introduction

Urban growth patterns arise from complex dynamics that are difficult to analyze because they stem from nonlinear and emergent processes caused by the interaction of several factors at a local scale (Allen 1997). Consequently, efficient tools are needed to scientifically study urban growth events, such that the problems derived from such events can be managed. Improving the knowledge and the analysis capability of urban dynamics allows for the design of more effective urban planning policies, the determination of the consequences of current planning instruments and the prediction of future problems derived from current decisions.

Cellular automata (CA) stand out among the most commonly used urban models for the simulation and analysis of urban growth because of their ability to reproduce complex

*Corresponding author. Email: andresmanuel.garcia@usc.es

dynamics, similar to those found in real cities, from simple rules. In addition, because CA models operate on a Euclidean space divided into an array of identical cells, they can easily incorporate raster data obtained from aerial photographs, satellite images, or GIS software. For the same reason, the results of the model can be visualized and analyzed directly within a GIS.

CA were developed in the 1940s, but their development as urban growth simulation tools was not attained until the 1990s (Berling-Wolff and Wu 2004a). There are many examples of the application of urban CA models to the simulation of growth in big cities such as Buffalo (Xie 1996), Cincinnati (White *et al*. 1997), San Francisco, the metropolitan area of Baltimore/Washington (Clarke and Gaydos 1998), Dublin (Barredo et al. 2003), Lagos (Barredo et al. 2004), Tokyo (Arai and Akiyama 2004) or San Diego (Kocabas and Dragicevic 2006).

Most of the existing models were developed in the field of scientific research to test urban theories or study the dynamics of real cities. However, CA models are yet to overcome some inherent problems such as lack of flexibility and adaptability to different urban dynamics, and the need for calibration methods that facilitate their implementation in urban areas with different characteristics (Santé *et al*. 2010).

The most traditional methods for the calibration of urban CA models are based on trial and error (Ward et al. 2000, Silva and Clarke 2002, Barredo *et al*. 2004; He *et al*. 2008) or on statistical techniques such as logistic regressions (Sui and Zeng 2001). In recent years, more elaborate empirical calibration methods have been developed (for example, see the article by Straatman *et al*. 2004). However, these methods are inapplicable to models of considerable size, which has given rise to the development of new techniques, which include neural networks (Yeh and Li 2003), data mining (Li and Yeh 2004) and, particularly, heuristic optimization techniques such as ant colony optimization (Liu *et al*. 2008), exhaustive search (Shan *et al*. 2008), or simulated annealing and genetic algorithms, which are the two most robust and widespread heuristic techniques used to solve optimization problems. A number of authors have used genetic algorithms (GAs) to calibrate CA models (Jenerette and Wu 2001, Goldstein 2003, D'Ambrosio *et al*. 2006, Li *et al*. 2007, 2008, Shan *et al*. 2008). Indeed, Al-Ahmadi *et al*. (2008) have demonstrated that GAs are more efficient than simulated annealing for CA calibration.

Most of the existing models were designed and evaluated for large cities, where urban growth is high and fast, which makes the identification of urban growth processes and drivers easier. For this reason, the application of these models to urban areas with different characteristics, such as areas with low and slow growth, is often difficult (García *et al*. 2012). This paper presents a method for the calibration of an urban CA model based on the combined use of statistical techniques and a GA. The proposed model and the calibration method presented here improve the adaptation of an urban CA model to the specific characteristics of the study area, which increases its flexibility. In addition, the calibration procedure has been automated to make it easier. The efficiency of the statistical and heuristic calibration methods used was evaluated by applying the model to Ribadeo, a small settlement of NW Spain characterized by slow urban dynamics and scattered growth along the major road (García *et al*. 2011b). The paper describes the structure of the model, based on the model of White *et al*. (1997), and explains the calibration procedure. Compared to the original model, the calibration procedure introduces two new features: the reduction of the number of calibration parameters using statistical techniques, and the calibration of the remaining parameters through a GA. Following the explanation of the calibration procedure, the results of the application of the model to the simulation of urban growth in Ribadeo are discussed and the conclusions are presented.

## 2. Urban cellular automata model

A large number of urban CA models have been developed and used both at the theoretical level for the study of urban dynamics, and at a practical level for the simulation of urban growth in real cities. The first widespread empirical application of these models was developed by White *et al*. (1997) based on the previous models of White and Engelen (1993, 1997). In contrast to most urban CA found in the literature, the model of White *et al*. (1997) simulates the dynamics of various types of land uses, instead of only urban and nonurban land uses. By simulating the dynamics of various types of land uses, the model captures the differences between the dynamics of the various land uses present in the study area with a greater degree of detail (Berling-Wolff and Wu 2004b). Another advantage of this model is the flexibility provided by the neighborhood (Santé *et al*. 2010), which models the influence of the neighboring land uses according to the distance from each neighboring land use to the cell for which the transition potential is calculated. In addition, the transition rules of the model of White *et al*. (1997) are quite close to the formal rules of orthodox CA, which helps preserve much of the simplicity of these models and facilitates the analysis of the results obtained. Because of these advantages, many recent urban CA models have based their structure on the structure of the model of White *et al*. (for example, Barredo *et al*. 2003, 2004, Yüzer 2004, Kocabas and Dragicevic 2006). However, such models have inherited from the model of White a complex calibration derived from the large number of parameters involved. This paper proposes a model inspired by the original rules of the model of White that keeps the advantages of the original model while at the same time overcomes the calibration complexity by automating calibration using statistical techniques and a GA.

### *2.1. Model of White* et al.

The model developed by White *et al*. (1997) considers two types of land uses: fixed land uses, which influence the dynamics of other land uses but do not change their state, and active land uses, which influence the dynamics of other land uses and change their state according to the simulated growth demand for that land use. The model of White is an exogenously constrained CA in which land-use demand is determined exogenously. The transition rule of the model is based on Equation (1), which provides the transition potential of each cell from land use $h$ to each active land use $j$ ($P_{hj}$):

$$P_{hj} = vs_j \left(1 + N_j\right) + H_j \tag{1}$$

where $s_j$ is the cell suitability for land use $j$, $N_j$ is the neighborhood effect, and $H_j$ is an inertia parameter that models the resistance of land use $h$ to change to land use $j$. $v$ is a stochastic variable that introduces randomness in the model and is determined by Equation (2):

$$v = 1 + [-\ln(rand)]^{\alpha} \tag{2}$$

where *rand* is a random number between 0 and 1 and $\alpha$ is a coefficient that controls the degree of randomness introduced in the model. The effect of the neighborhood ($N_j$) is calculated from Equation (3):

$$N_j = \sum_d \sum_i m_{kjd} I_{id} \tag{3}$$

where $I_{id}$ is 1 if cell $i$ at distance $d$ is occupied by land use $k$ and 0 otherwise. In White *et al*. (1997), a circular neighborhood with a radius of six cells was used, where the influence of each cell was modeled by means of a coefficient $m_{kjd}$, whose value depended on land use $k$ in cell $i$ and on the distance $d$ from cell $i$ to the central cell. Cells transition at every iteration of the model to the land use for which the transition potential is higher until the demand for that land use is satisfied. In the unlikely event of cells with the same transition potential value for several land uses, a land use is assigned hierarchically by prioritizing residential over commercial and commercial over industrial uses. The flexibility of this model gave rise to the development of many other models that were based on it, such as the models described by Engelen *et al*. (1999), White and Engelen (2000), Barredo *et al*. (2003, 2004), Yüzer (2004) or Kocabas and Dragicevic (2006).

### *2.2. Proposed model*

In the proposed model, the stochastic component of the model of White *et al*. (1997) was modified based on the results of a previous study (García *et al*. 2011a) that showed that the use of an exponential function like the function used by Wu (2002) helped control the degree of randomness introduced in the model. Accordingly, the logarithmic function in Equation (2) was replaced by the exponential function in Equation (4):

$$v = \exp(-\alpha \times (1 - rand)) \tag{4}$$

In addition, the suitability value was scaled by using coefficient $\beta$ in order to model the relative importance of suitability $s_j$ with regard to the neighborhood. Coefficient $\beta$ can give less weight (low values of $\beta$) or more weight (high values of $\beta$) to suitability in the calculation of the transition potential, thus giving more or less importance, respectively, to the neighborhood effect in the discrimination between cells with higher or lower transition potentials. Therefore, the suitability factor is scaled in a similar way to the stochastic variable with the $\alpha$ coefficient in Equation (2).

Finally, a number of restrictions ($R_j$) have been introduced to consider the areas excluded for land use $j$ because of urban planning constraints or the presence of elements that prevent land-use change, such as cemeteries, churches or landfills, among others. $R_j$ takes a value of 0 if land use $j$ is excluded and a value of 1 if land use $j$ is allowed. These three modifications resulted in an equation for the calculation of the transition potential that was quite different from the equation of the original model (Equation (5)):

$$P_{hj} = R_j * v * s_j^{\beta} * \left(1 + N_j\right)_j + H_j \tag{5}$$

## 3. Model calibration

The main drawback of the model of White is the large number of parameters involved, which makes the calibration process extremely complex. In most of the models inspired by the model of White, calibration has been carried out by trial and error or expert knowledge (e.g. White *et al*. 1997, Barredo *et al*. 2003, 2004, Yüzer 2004, Kocabas and Dragicevic 2006). Even the most modern versions of the model such as Metronamica (http://www.riks.nl/products/metronamica) use trial and error in spite of making interface improvements that make the model simpler to calibrate. Nevertheless, trial and error calibration is time consuming. In addition, neither trial and error calibration nor expert knowledge calibration guarantees accurate results. For this reason, a method that overcomes the accuracy problem has been designed by reducing the number of calibration parameters and calibrating the remaining parameters using a GA.

### 3.1. Calculation of parameters using statistical techniques

In the original model, the suitability factor ($s_j$) is determined by expert knowledge, usually as the weighted summation of several suitability evaluation factors. In order to reduce the number of calibration parameters, the weights of the suitability evaluation factors have been determined using logistic regression.

Logistic regression allows for the analysis of the contribution of a number of independent variables to the probability of occurrence of a dependent binomial variable that takes a value of 1 when a specific land use is present and a value of 0 otherwise. The relationship between the independent variables and the dependent binomial variable is established by adjusting a linearized logistic function. A first logistic regression has been used to identify the most significant variables for the calculation of the probability of change to each active land use by determining the probability that variation is due to chance (Pr). If such a probability was low, the levels of significance of the variables for the prediction of change were high. Accordingly, the variables with Pr values below 0.001 were selected for use in the calculation of the suitability maps for each active land use using a second logistic regression. Because high-resolution maps were used, a sample of 10% of the data was selected by using stratified random sampling in order to remove spatial correlations.

### 3.2. Simplification of the neighborhood coefficients

In spite of having reduced the number of calibration parameters by using logistic regressions, the calibration of the model is still complex because of the large number of parameters required to calculate the neighborhood effect. The neighborhood effect models the influence that the land uses present in the neighborhood have on the transition potential according to the distance from each neighboring cell to the central cell. In the proposed model, a circular neighborhood with a radius of three cells has been used. The influence of neighboring land uses on the transition potential of the central cell can be modeled by using a distance function, which could take various shapes (linear, logarithmic, exponential, etc.). Accordingly, several types of functions (Xie 1996) should be considered in the calibration of the model. To simplify this, two linear functions have been used (Equations (6) and (7)), which has allowed us to model various types of distance-decay functions by using only four parameters ($a$, $b$, $c$, and $d$) for each pair of land uses (Figure 1):

$$f(x) = a + bx \tag{6}$$

$$g(x) = c + dx \tag{7}$$

where $x$ is the distance between the neighboring cell and the central cell, and $a$, $b$, $c$ and $d$ are the coefficients of the linear functions.

By using these two linear functions, the need to calibrate one $m_{kjd}$ coefficient for every land use $k$ and cell equidistant from the central cell is avoided (for a circular neighborhood with a radius of three cells, seven coefficients $m_{kjd}$ would have to be calibrated for every pair of land uses, as shown in Figure 2). Actually, the four parameters that define the two

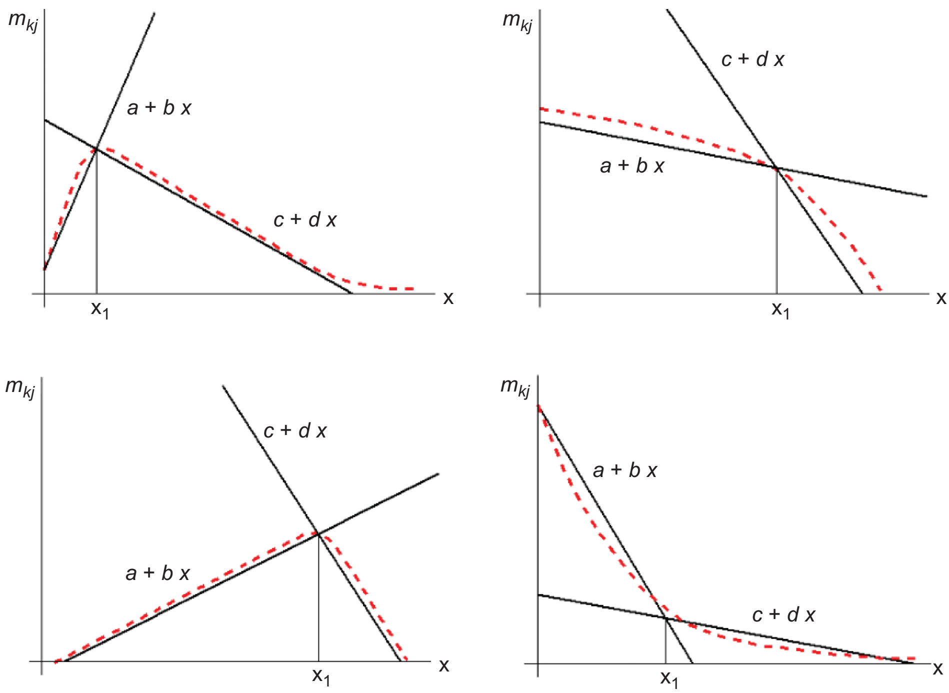


Figure 1. Examples of simplification of distance-decay functions using two linear functions.

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | 6 | | | |
| | 5 | 4 | 3 | 4 | 5 | |
| | 4 | 2 | 1 | 2 | 4 | |
| 6 | 3 | 1 | 0 | 1 | 3 | 6 |
| | 4 | 2 | 1 | 2 | 4 | |
| | 5 | 4 | 3 | 4 | 5 | |
| | | | 6 | | | |

Figure 2. Cells numbered according to distance from the central cell in a three-cell radius neighborhood.

lines ($a$, $b$, $c$, and $d$) are sufficient to model the distance-decay effect. This is achieved by determining the cross point between both lines ($x_1$), i.e. the point at which both lines have the same value (Equation (8)). To calculate coefficient $m_{kj}$, the $f(x)$ function is used for distance values between 0 and $x_1$ and the $g(x)$ function is used for distance values above $x_1$.

$$a + bx_1 = c + dx_1 \Rightarrow x_1 = \frac{a - c}{d - b} \tag{8}$$

### 3.3. Genetic algorithm

In spite of having reduced the number of calibration parameters significantly, there are still quite a few and, therefore, the calibration process remains complex. As aforementioned, GA and SA are the most robust heuristic techniques to solve optimization problems, both have been widely used for CA calibration and some authors have even proven that GA outperforms SA by the time of calibrating CA models. For this reason, a GA was designed to automate the calibration of the $\alpha$ coefficients of the stochastic variable, the $\beta$ coefficients of the suitability factor and the inertia coefficients $H_j$ for each active land use, as well as coefficients $a$, $b$, $c$, and $d$ of the linear functions that model the distance-decay influence of neighboring land uses on the transition potential for each active land use.

Genetic algorithms (Holland 1975) are inspired by the genetic evolution of populations in the search for the optimal solution. First, a random initial population of possible solutions is generated. Each individual in the population corresponds to a chromosome whose alleles are the calibration parameters. The goodness of each individual is evaluated through a fitness function. The best individuals of a generation are selected for generating an offspring population by means of crossover operators. In each generation, mutation operators are used to randomly modify the allele values in order to prevent the algorithm from being caught in a local optimum. After several generations a near optimal solution is reached.

A GA comprises the following phases: (i) initialization, during which an initial population of random individuals is generated; (ii) evaluation, during which the fitness value of each individual is calculated; (iii) selection, during which the best individuals are selected according to their fitness value; (iv) crossover, during which the selected individuals are used to create the offspring population; and (v) mutation, during which random variations are introduced in the offspring. Each of these phases can be implemented in many different ways (Goldstein 2003), an optimal method not existing for all the cases.

In the GA designed in this study, an initial population of 700 individuals with 117 alleles each (108 corresponding to the parameters of the linear functions, 3 to the $\alpha$ coefficients of the stochastic variable, 3 to the $\beta$ coefficients of the suitability factors and 3 to the inertia coefficients $H_j$) was created by generating random numbers. Based on the findings reported by Wu (2002), the $\alpha$ coefficients were forced to range from 0 to 10, whereas the $\beta$ coefficients were forced to range from 0 to 3, according to empirical results obtained in the first tests of the algorithm. For the inertia coefficients $H_j$, a broad value range (from 0 to 1,000,000) was defined in order to prevent land-use transitions. For the coefficients of the linear functions, the parameters that determine the slope of the line were calculated by generating a random angle and calculating the tangent of that angle. The other two parameters of the linear functions were generated within a range of values [$-100$, 100] in order to obtain neighborhood-effect values similar to those reported by White *et al*. (1997).

Once the initial population is generated, the parent population is evaluated and the best individuals are selected by using a tournament method, according to which two individuals are randomly selected and the individual with the highest fitness value becomes a parent. When two parents are obtained, a crossover operator is used to generate two child solutions. In the crossover operation, two recombination points in the chromosomes are randomly set, from which the alleles of each parent are exchanged to generate two sons. In the tournament process, all individuals in the population are forced to compete to be parents at least once. The individual of the parent population with the best fitness value survives in the offspring population. Once the offspring is obtained, a mutation rate of 0.008% is applied to all individuals, with the exception of the survivor of the parent population.

### *3.3.1. Fitness function*

In order to use a GA in the calibration of an urban CA, an objective function or fitness function that allows for the comparison of the simulated and real maps must be designed. Previous studies used only measurements of cell-to-cell correspondence (Li *et al*. 2007, 2008, Shan *et al*. 2008) or only spatial metrics (Jenerette and Wu 2001, Goldstein 2003) as the fitness function. In the design of the fitness function for the proposed GA, it was assumed that the indices that evaluate cell-to-cell correspondence do not consider that, if a simulated urban cell does not match the real urban cell but is located close to it, the result is better than if the simulated cell is located away from the real cell. To reduce this problem, the cell-to-cell correspondence was evaluated by using the index described in Pontius (2002), which compares maps at multiple resolutions. In addition, the simulated and real urban spatial patterns were compared by using three spatial metrics.

The index described in Pontius (2002) is calculated by running windows at various resolutions $g$ (1 cell per side, 2, 3, 4, . . . , $n$ cells per side) all over the real $R$ and simulated $S$ maps. In each window, the number of cells $n$ occupied by each land use $j$ in the real $Rn,j$ and simulated $Sn,j$ maps is calculated. The lowest value for each land use in both maps is chosen and the values for all the land uses are added. Then, each window is weighted by the number of cells covered by the window, $Wn$, and all the values of all the windows into which the maps are divided at every resolution are added $g(Ng)$. The resulting value is divided by the number of cells in the map. Equation 9 is used to calculate the value of the index for every window resolution ($Pg$). The index takes a value of 1 if the number of cells for each land use in each window coincides in both maps and a value of 0 if the number of cells does not coincide for any land use in any window.

$$Pg = \frac{\sum_{n=1}^{Ng} \left[ Wn \sum_{j}^{J} MIN(Rn,j,\ Sn,j) \right]}{\sum_{n=1}^{Ng} Wn} \tag{9}$$

In the fitness function, the global index $P$ (Equation (10)), which results from the weighted summation of the indices for each resolution, was used. The weighting coefficient $Vg$ for each map resolution $g$ was obtained by means of an exponential curve scaled by the coefficient $b$, which was assigned a value of $-1.2$.

$$P = \frac{\sum_{g}^{G} \exp^{b \times g} \times Pg}{\sum_{g}^{G} \exp^{b \times g}} \tag{10}$$

In addition, three spatial metrics were used in the fitness function, namely: number of patches (NP), mean patch area (AREA_MN) and edge density (ED). These metrics allow for the comparison of the shape and complexity of simulated and real spatial patterns. Only the values of these metrics for the active land uses were considered in the fitness function.

The values of the spatial metrics for the simulated map were subtracted from the values for the real map. The absolute values of the subtractions were used in the fitness function. To make these values vary within the same range, the maximum and minimum values for each index at every iteration of the GA were used to normalize the indices (Equation (11)).

Finally, the normalized values were added and divided by 3. The same process was used to normalize the values of the index proposed in Pontius (2002). The normalized value of this index was added to the normalized value of the three spatial metrics and the resulting value was divided by two. The inverse of the resulting value was used in the fitness function.

$$\text{Normalized value} = \frac{\text{Value} - \text{Minimum value}}{\text{Maximum value} - \text{Minimum value}} \tag{11}$$

## 4. Case study

The study area is located in the municipality of Ribadeo, NW Spain (Figure 3). Ribadeo is located at a junction of important routes connecting the regions of Asturias and Galicia and concentrates the commercial activities and services of the surrounding areas. Ribadeo has a population of 6000 and has gained 1000 inhabitants in the last decade. The study area is formed by the main urban core of Ribadeo and four surrounding parishes (a sub-municipal administrative division in the region of Galicia), toward which the urban core is expanding. The urban area of Ribadeo has experienced a slow urban growth process in the last three decades that has taken place in relatively small, scattered plots. In addition, the evolution of the urban spatial pattern in this area has been strongly conditioned by specific events or characteristic elements present in the area (García *et al*. 2012).

Land-use maps were obtained through photo interpretation of aerial photographs of 1978, 1995, and 2007, which were used as input data for the urban CA model. In addition to road maps obtained also by photo interpretation, a digital elevation model developed

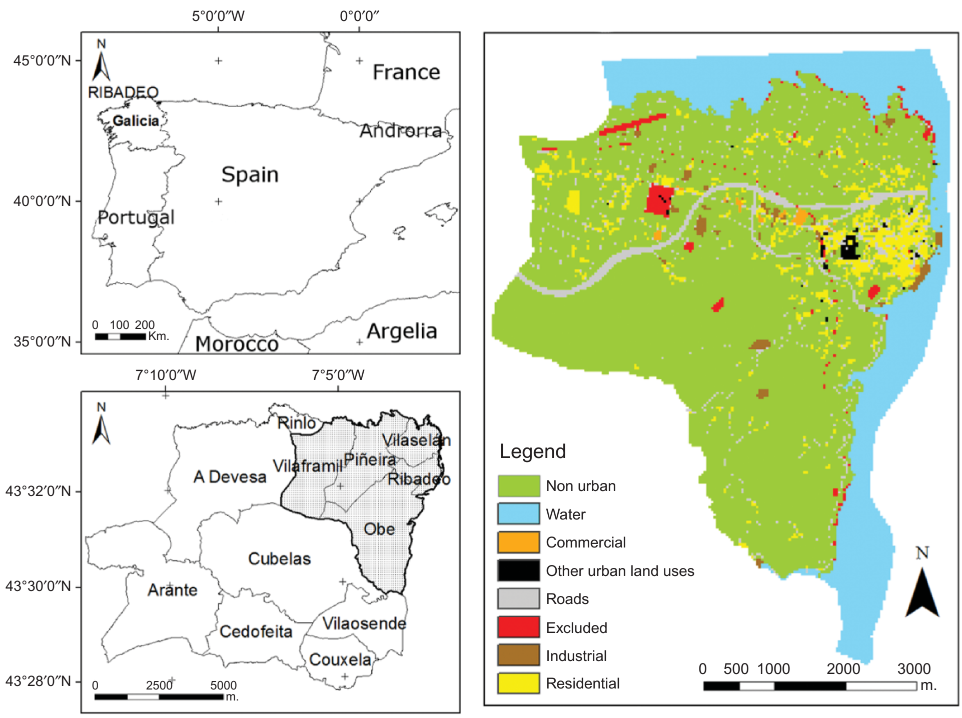


Figure 3. Location of the study area and land-use map of 2007.

from the national topography map and a cadastral plot map were used. All these maps were converted to raster format with 35 m resolution and processed to obtain the maps of the input variables.

Land uses were classified following the model of White *et al*. (1997). Water bodies, roads, institutional buildings, parks, recreational areas, and railways were classified as fixed land uses, whereas commercial, industrial, and residential land uses were classified as active land uses. Agriculture and forestry were classified as fixed land uses rather than as land reserve for land uses because they influence the dynamics of the other classes but do not take part in them, i.e. agriculture and forestry cells could transition to urban cells but the dynamics of the agricultural and forestry land uses were not simulated.

First, the suitability maps ($s_j$) were calculated by using the logistic regression techniques described in Section 2.1.Table 1 shows the independent variables used in the first logistic regression, which were identified as the main drivers of urban growth in the town of Ribadeo in previous studies (García *et al*. 2011b). In Table 1, the most significant variables, i.e. the variables with Pr values below 0.001, were identified with $^{***}$. The significant variables were used in the second logistic regression to calculate the suitability maps for the active land uses (Figure 4).

The model was calibrated using the land-use maps of 1978 and 1995. The GA was run until the best fitness value did not increase during several iterations (Figure 5). The coefficients obtained with the GA (Tables 2–4) were used to simulate land-use evolution between 1995 and 2007. Results were validated through their comparison with the land-use map for the year 2007. The amount of growth of each active land use at every iteration was determined by dividing the real growth in the simulated period by the number of model iterations. At every iteration, the cells transitioned to the land use with the highest transition potential until the growth demand calculated for that land use was achieved. Once all cells were allocated to the land use with the highest transition potential, and if the demand calculated for a specific land use was not reached, that land use was allocated to cells whose second highest potential corresponded to that land use.

The coefficients shown in Tables 2–4 are consistent with the land-use dynamics observed in the study area. For example, $\alpha$ is high for the residential land use because this use shows the most dispersed spatial pattern. For the slopes of the linear functions used for the calibration of the neighborhood effect (coefficients $b$ and $d$), positive coefficients correspond to a repulsion influence and negative coefficients correspond to an attraction influence. For example, the logic of these coefficients can be observed for the residential land use, which is attracted by residential, park, institutional, commercial, and agricultural land uses and repelled by water, roads, forestry, and industrial land uses in the nearest neighborhood. Contrastingly, the residential land use is attracted by water, parks, institutional, and residential land uses and repelled by roads, agricultural, commercial, forestry, and industrial land uses in the most distant neighborhood. These coefficients, together with $a$ and $c$, allow us to draw the distance-decay-effect function for each land use in the neighborhood.

The simulation maps of the proposed model were compared with those obtained using the original model of White. The neighborhood parameters used in the model of White were the same as those used in the application of this model to Cincinnati (White *et al*. 1997) since, according to the authors, these parameters should not vary too much between different cities. The stochastic variable and accessibility were calibrated by trial and error, whereas suitability was calibrated using a logistic regression and the suitability evaluation factors considered in the proposed model.

Table 1. Results of the first logistic regression.

| | Commercial coefficient | Pr | | Industrial coefficient | Pr | | Residential coefficient | Pr | |
|---|---|---|---|---|---|---|---|---|---|
| Plot area | −1.88E−05 | 0.05 | | −8.05E−06 | 0.009 | ** | −1.39E−05 | 2.78E–12 | *** |
| Distance to water bodies | 0.003 | 0.0002 | *** | 0.001 | 2.32E − 09 | *** | 0.0008 | 7.27E–13 | *** |
| Distance to Ribadeo | −7.28E−05 | 0.7 | | −0.0002 | 0.04 | | −0.0004 | 7.21E-23 | *** |
| Distance to municipal roads | 0.004 | 0.02 | | −0.0007 | 0.1 | | −0.002 | 1.26E-11 | *** |
| Distance to main roads | −0.02 | 2.14E−07 | *** | −0.002 | 3.19E−11 | *** | −0.002 | 2.18E-18 | *** |
| Distance to secondary roads | −0.0009 | 0.4 | | −0.002 | 1.33E−06 | *** | −0.002 | 5.01E-18 | *** |
| Distance to railways | 0.002, 825, 93 | 0.04 | | 0.001 | 0.0004 | *** | 0.0001 | 0.4 | |
| Plot shape index | 0.2 | 0.09 | | 0.1 | 0.0009 | *** | 0.2 | 8.65E-11 | *** |
| Slope | 0.01 | 0.7 | | −0.006 | 0.7 | | −0.04 | 1.28E–05 | *** |

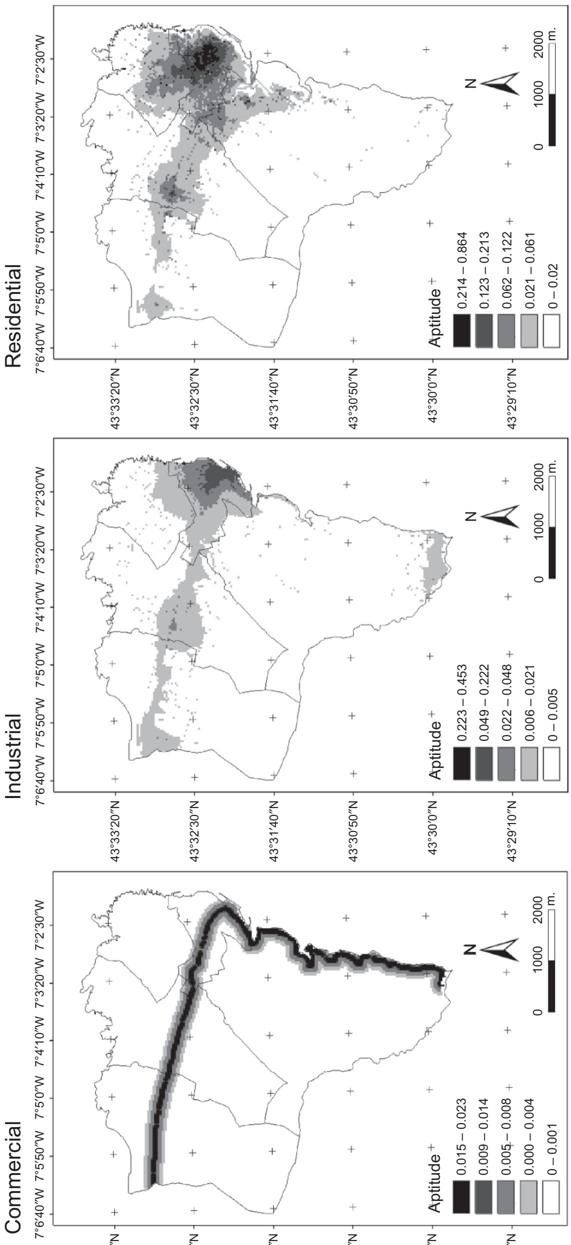


Figure 4. Suitability maps for active land uses.

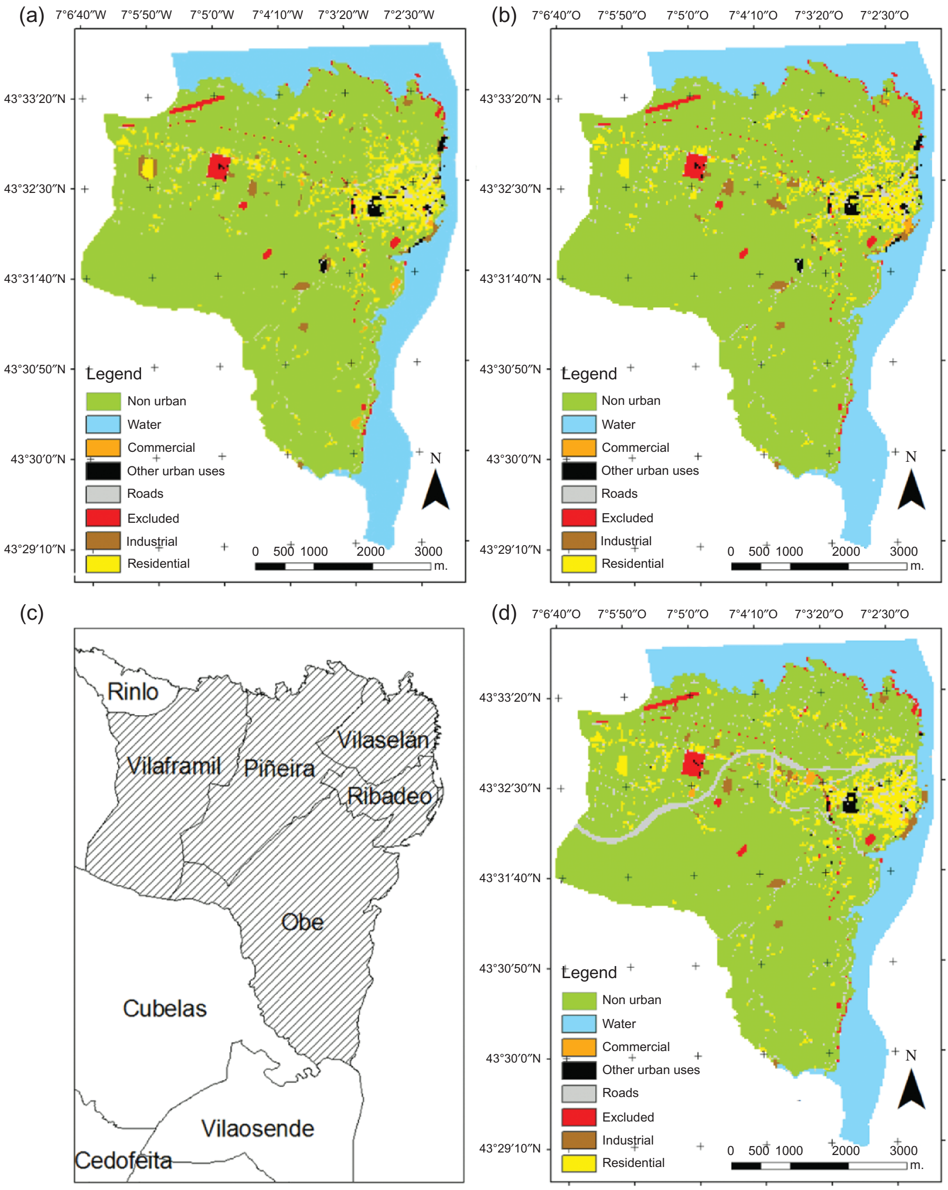


Figure 5. Evolution of the average value of the fitness function of the GA.



Table 5 shows the values of the metrics used to calculate the fitness function for the proposed model and the model of White *et al*. (1997). The cell-to-cell correspondence is higher for the proposed model, and the values of the spatial metrics obtained in the simulation performed with the proposed model are closer to the actual values than those obtained with the original model in all cases, except for the mean patch area (AREA_AM) of the residential land use. This is because the patches generated by the model of White are smaller than the patches generated by the proposed model, which are more similar in size to the actual patches (shown in map in Figure 5). However, the distribution of

Table 2. Coefficients calibrated with the GA for the calculation of the transition potential for industrial land use.

| Neighborhood land uses | *a* | *b* | *c* | *d* |
|---|---|---|---|---|
| Agriculture | −15.73 | −1.94 | 74.42 | −11.94 |
| Water | 18.35 | 1.42 | −82.68 | 7.03 |
| Commercial | 15.27 | 0.06 | 2.76 | −0.50 |
| Roads | −43.73 | −1.07 | 96.80 | 0.97 |
| Forestry | −19.81 | 2.11 | 55.90 | 0.78 |
| Industrial | −69.44 | −0.15 | 1.19 | −0.47 |
| Institutional | 89.94 | −0.14 | 15.98 | 0.00 |
| Parks | 34.18 | 0.34 | −14.99 | −0.51 |
| Residential | −98.98 | 0.74 | −85.63 | −2.51 |
| | | $\alpha$ | | 4.53 |
| | | $\beta$ | | 5.50 |
| | | $H$ | | 33,837.97 |

Table 3. Coefficients calibrated with the GA for the calculation of the transition potential for commercial land use.

| Neighborhood land uses | *a* | *b* | *c* | *d* |
|---|---|---|---|---|
| Agriculture | −1.05 | 0.61 | 33.10 | −5.96 |
| Water | −25.26 | −0.57 | −71.82 | 2.14 |
| Commercial | −69.79 | −2.34 | −18.75 | 0.07 |
| Roads | 41.08 | −1.60 | 78.94 | 0.02 |
| Forestry | 6.00 | 2.06 | −99.54 | −1.02 |
| Industrial | 82.46 | 0.64 | −11.04 | 0.11 |
| Institutional | −80.02 | 0.00 | 41.64 | −125.34 |
| Parks | −30.96 | 1.96 | −30.85 | 17.19 |
| Residential | 12.17 | −0.49 | 90.18 | −0.65 |
| | $\alpha$ | | 2.78 | |
| | $\beta$ | | 1.82 | |
| | $H$ | | | 265,816.4 |

Table 4. Coefficients calibrated with the GA for the calculation of the transition potential for residential land use.

| Neighborhood land uses | *a* | *b* | *c* | *d* |
|---|---|---|---|---|
| Agriculture | −4.58 | −1.64 | 63.85 | 0.95 |
| Water | 89.51 | 2.64 | 32.25 | −2.68 |
| Commercial | −39.35 | −0.93 | −0.55 | 0.86 |
| Roads | 48.06 | 0.21 | −12.97 | 13.97 |
| Forestry | −49.20 | 1.80 | −6.74 | 0.20 |
| Industrial | 35.41 | 1.98 | −97.94 | 0.11 |
| Institutional | 92.93 | −0.53 | 89.63 | −1.23 |
| Parks | 86.09 | −0.41 | 61.85 | −5.28 |
| Residential | −67.89 | −1.57 | 69.27 | −4.70 |
| | $\alpha$ | 5.93 | | |
| | $\beta$ | 3.44 | | |
| | $H$ | | | 113,284.80 |

Table 5. Values of the metrics used in the fitness function for the proposed and original models.

| | | Proposed model | Model of White *et al*. (1997) | Land-use map 2007 |
|---|---|---|---|---|
| Index described in Pontius (2002) | | 0.9201 | 0.9195 | |
| NP | Residential | 233 | 234 | 224 |
| | Industrial | 44 | 46 | 45 |
| | Commercial | 15 | 29 | 13 |
| AREA_AM | Residential | 26.12 | 19.52 | 18.73 |
| | Industrial | 1.96 | 3.24 | 2.07 |
| | Commercial | 1.52 | 1.35 | 1.76 |
| ED | Residential | 18.68 | 20.46 | 18.25 |
| | Industrial | 2.90 | 2.4 | 2.66 |
| | Commercial | 0.98 | 1.06 | 0.84 |

these patches is better in the simulation of the proposed model, as the simulation obtained from the original model produces an excessive concentration of these patches along the roads. This can be observed in the map, particularly in the patches of residential land use located in the north, at the top, which correspond to growth along a secondary road. This confirms that a single spatial metric is not sufficient to capture all the complexity of spatial patterns, as reported by Visser and Nijs (2006). Although the spatial pattern is quite similar to the real one, fully optimized values cannot be obtained for the spatial metrics because the fitness function includes both spatial pattern criteria and cell-to-cell correspondence criteria. Consequently, the algorithm searches for the compromise solution that provides the best trade-off between both criteria.

With regard to cell-to-cell correspondence, the results are also good with indices different from the index used in the fitness function. For example, the Kappa index has a value of 0.9195 for the proposed model and 0.9158 for the model of White *et al*. (1997). Small variations in this index are quite significant, since the proportion of cells whose land-use changes is quite low and the unchanged remaining cells increase cell-to-cell correspondence. Furthermore, the Kappa index does not allow for an in-depth analysis of cell-to-cell correspondence (Pontius and Millones 2011, van Vliet *et al*. 2011). For this reason, other metrics such as the figure of merit have been used. The figure of merit (Pontius *et al*. 2008) is obtained from the number of hits (observed change predicted as change) divided by the summation of hits, partial hits (observed change predicted as change but for a wrong land-use category), misses (observed change predicted as persistence) and false alarms (observed persistence predicted as change). As shown in Table 6, the figure of merit is higher for the results of the proposed model than for the results of the original model. The improvement of partial hits, i.e., the decrease of the errors caused by the simulation of a wrong category of urban land use, can be attributed to the improvement of the calibration procedure, which captures the specific dynamics of each land use more accurately.

Table 6. Figure of merit of the results obtained with the proposed and original models.

| | Figure of merit | Hits | Partial hits | False alarms | Misses |
|---|---|---|---|---|---|
| Proposed model | 8.16% | 77 | 6 | 457 | 404 |
| White *et al*. (1997) | 7.37% | 68 | 36 | 436 | 383 |

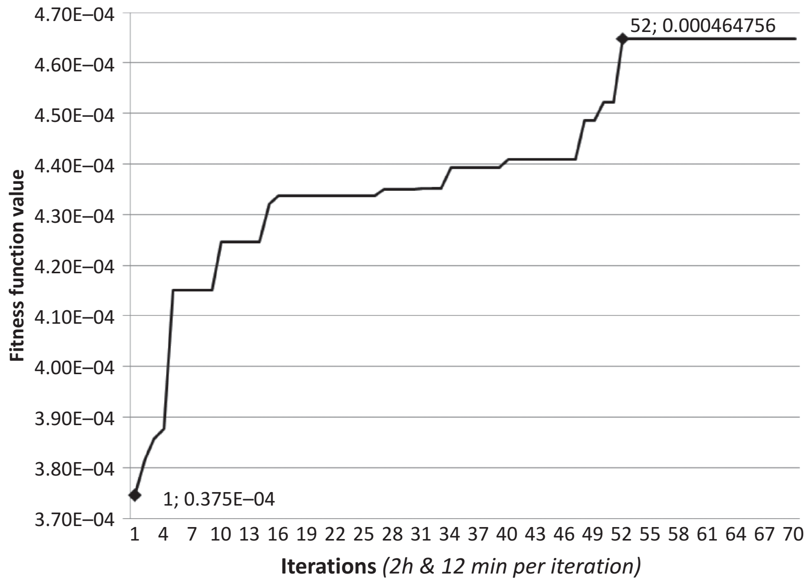


Figure 6. (a) Map simulated with the proposed model for 2007, (b) map simulated with the model of White for 2007, (c) parishes of the study area, (d) real land-use map of 2007.

As shown in the simulation maps (Figure 6), the urban growth patterns obtained with the proposed model are quite similar to the actual ones. Overall, the growth at the north of the settlement and along the provincial road was correctly simulated. In addition, the patterns obtained for the small patches located at the area in the south and along the main road are quite similar to the real ones. The urban growth simulated with the model of White is more concentrated along roads, and is distributed into patches that are smaller and closer to one another than in the real map. In the visual analysis, the most easily identifiable errors in both simulated maps correspond to industrial and commercial land uses. Because the growth of both land uses was very low in the simulated period, the calibration procedure could not accurately capture the dynamics and drivers of growth.

## 5. Conclusions

The model developed by White *et al*. (1997) has the advantage of using a flexible neighborhood that allows for the simulation of the influence of proximity to different land uses on the evolution of a specific land use. This characteristic provides the model with the ability to conform to different types of urban areas, which can be characterized by specific factors and dynamics, as in the small urban settlement analyzed in this paper. However, such a flexibility is possible due to the large number of parameters required to calibrate the neighborhood effect, which makes the calibration process complex. This paper has presented a method that allows to automate the calibration of the model without losing flexibility and analysis capability. This has been achieved, first, by modifying the original model to incorporate certain restrictions and to improve the control of stochasticity and of the influence of the suitability factor. Second, the calibration of the model has been simplified by the design of a method that comprises two procedures: the reduction of the number of calibration parameters and the calibration of the remaining parameters through a GA.

The application of the model to a small village of NW Spain characterized by a low, slow urban growth that is distributed in scattered small patches and strongly conditioned by specific events, has shown that the modifications to the original model simplified the model without involving any loss of analysis capability or simulation accuracy, which has enabled the use of a heuristic method to systematize calibration tasks. The automation of the calibration process allows for a more simple application and adaptation of the model to urban areas with different dynamics and processes.

The simulations of urban growth in the study area has shown that the results obtained with the proposed model are better than those obtained by calibrating the original model using trial and error and expert knowledge. The cell-to-cell correspondence between simulated and real maps is higher with the proposed model for all the analyzed indices, particularly for the figure of merit, which distinguishes correctly predicted change from correctly predicted persistence and considers partial hits. With regard to spatial metrics, the values obtained with the proposed model are better than the results of the original model in all cases, except for the mean patch area of the residential land use. Because the fitness function includes spatial pattern criteria (such as number of patches, mean patch area, and edge density) and cell-to-cell criteria, the GA searches for a compromise solution that provides the best trade-off between criteria. These results confirm that GA is a tool with great potentiality for urban CA calibration, in that it not only produces better results than expert knowledge calibration, but it also automates the calculation of the calibration parameters and avoids the knowledge of urban local dynamics.

The design of the proposed calibration method required knowledge of statistics and modeling, but expert knowledge is not needed once the procedure has been implemented in the software. Running the calibration procedure with the developed software is easier than determining the value of the calibration parameters by expert knowledge or by trial and error. As an example, 189 calibration parameters would be needed for the case study with the original model. The model and the calibration procedures proposed in this paper reduce the number of parameters and automatically determine their value, thus avoiding the need to tune a huge amount of parameters manually. This makes the application of the model to study areas with different characteristics easier and overcomes the issues identified in Santé *et al*. (2010) related to the need of making urban CA more flexible while keeping their simplicity by developing better calibration methods. Moreover, automated calibration improves the model results.

Future research should focus on the development of better validation methods that allow us to capture all the complexity of urban growth patterns and to assess cell-to-cell correspondence. Another research line is the use and testing of new techniques for the calibration of urban CA models. In fact, some authors have already used artificial intelligence techniques such as neural networks or data mining to define the transition rules of urban CA models. This field can be explored to find new calibration methods that can better capture urban dynamics and generate simulations that are closer to the real world. Finally, the viability of the proposed model as a decision support tool for planning could be evaluated by simulating future land-use scenarios under different assumptions to test the consequences of the planning decisions.

**Acknowledgements**

This research was funded by the Spanish Government under Contract TIN2011-24326 ‘Agent-based model and GIS-web system for the development, evaluation and implementation of urban plans’.

## References

Al-Ahmadi, K., *et al*., 2008. Calibration of a fuzzy cellular automata model of urban dynamics in Saudi Arabia. *Ecological Complexity*, 6 (2), 80–101.

Allen, P.M., 1997. *Cities and regions as sell-organizing systems: models of complexity*. Amsterdam: Gordon and Breach Science.

Arai, T. and Akiyama, T., 2004. Empirical analysis for estimating land use transition potential functions – case in the Tokyo metropolitan region. *Computers, Environment and Urban Systems*, 28 (1–2), 65–84.

Barredo, J.I., *et al*., 2003. Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata. *Landscape and Urban Planning*, 64 (3), 145–160.

Barredo, J.I., *et al*., 2004. Modelling future urban scenarios in developing countries: an application case study in Lagos, Nigeria. *Environment and Planning B: Planning and Design*, 31 (1), 65–84.

Berling-Wolff, S. and Wu, J., 2004a. Modeling urban landscape dynamics: a review. *Ecological Research*, 19, 119–129.

Berling-Wolff, S. and Wu, J., 2004b. Modeling urban landscape dynamics: a case study in Phoenix, USA. *Urban Ecosystems*, 7 (3), 215–240.

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 (7), 699–714.

D'Amborsio, D., Spataro, W., and Iovine, G., 2006. Parallel genetic algorithms for optimising cellular automata models of natural complex phenomena: an application to debris flows. *Computers and Geosciences*, 32 (7), 861–875.

Engelen, G., *et al*., 1999. Dynamic GIS and strategic physical planning support: a practical application to the IJmond/Zuid-Kennemerland region. *In*: S. Stillwell, S. Geertman, and S. Openshaw, eds. *Geographical information and planning*. Berlin: Springer-Verlag, 87–111.

García, A.M., Santé, I., and Crecente, R., 2011a. An analysis of the effect of the stochastic component of urban cellular automata models. *Computers Environment and Urban Systems*, 35, 289–296.

García, A.M., *et al*., 2011b. Land development dynamics by morphological areas: a case study of Ribadeo, northwest Spain. *Environment and Planning B: Planning and Design*, 38, 1032–1051.

García, A.M., *et al*., 2012. A comparative analysis of cellular automata models for simulation of small urban areas in Galicia, NW Spain. *Computers, Environment and Urban Systems*, 36, 291–301.

Goldstein, N.C., 2004. Brains vs. Brawn: Comparative strategies for the calibration of a cellular automata-based urban growth model. *In*: P. Atkinson, G. Foody, S. Darby and F. Wu, eds., *GeoDynamics*. Boca Raton, FL: CRC Press.

He, C., *et al*., 2008. Modelling dynamic urban expansion processes incorporating a potential model with cellular automata. *Landscape and Urban Planning*, 86, 79–91.

Holland, J.H., 1975. *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.

Jenerette, G.D. and Wu, J., 2001. Analysis and simulation of land-use change in the central Arizona-Phoenix region, USA. *Landscape Ecology*, 16, 611–626.

Kocabas, V. and Dragicevic, S., 2006. Assessing a cellular automata model behavior using a sensitivity analysis approach. *Computers, Environment and Urban Systems*, 30 (6), 921–953.

Li, X., Yang, Q.S., and Liu, X.P., 2007. Genetic algorithms for determining the parameters of cellular automata in urban simulation. *Science in China, Series D: Earth Science*, 50 (12), 1857–1866.

Li, X., Yang, Q., and Liu, X., 2008. Discovering and evaluating urban signatures for simulating compact development using cellular automata. *Landscape and Urban Planning*, 86, 177–186.

Li, X. and Yeh, A.G.-O., 2004. Data mining of cellular automata's transition rules. *International Journal of Geographical Information Science*, 18 (8), 723–744.

Liu, X., *et al*., 2008. A bottom-up approach to discover transition rules of cellular automata using ant intelligence. *International Journal of Geographical Information Science*, 22 (11–12), 1247–1269.

Pontius, R.G., 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering and Remote Sensing*, 68 (10), 1041–1049.

Pontius, R.G., *et al*., 2008. Comparing the input, output, and validation maps for several models of land change. *Annals of Regional Science*, 42 (1), 11–47.

Pontius, R.G. and Millones, M., 2011. Death of Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32 (15), 4407–4429.

Santé, I., *et al*., 2010. Cellular automata models for the simulation of real-world urban processes: a review and analysis. *Landscape and Urban Planning*, 96 (2), 108–122.

Shan, J., Alkheder, S., and Wang, J., 2008. Genetic algorithms for the calibration of cellular automata urban growth modeling. *Photogrammetric Engineering and Remote Sensing*, 74 (10), 1267–1277.

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, 525–552.

Straatman, B., White, R., and Engelen, G., 2004. Towards an automatic calibration procedure for constrained cellular automata. *Computers, Environment and Urban Systems*, 28, 149–170.

Sui, D.Z. and Zeng, H., 2001. Modeling the dynamics of landscape structure in Asia's emerging desakota regions: a case study in Shenzhen. *Landscape and Urban Planning*, 53, 37–52.

van Vliet, J., Bregt, A.K., and Hagen-Zanker, A., 2011. Revisiting Kappa to account for change in the accuracy assessment of land-use change models. *Ecological Modelling*, 222 (8), 1367–1375.

Visser, H. and de Nijs, T., 2006. The map comparison kit. *Environmental Modelling and Software*, 21, 346–358.

Ward, D.P., Murray, A.T., and Phinn, S.R., 2000. A stochastically constrained cellular model of urban growth. *Computers, Environment and Urban Systems*, 24, 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. *Environment and Planning A*, 25, 1175–1199.

White, R. and Engelen, G., 1997. Cellular automata as the basis of integrated dynamic regional modeling. *Environment and Planning B. Planning and Design*, 24, 235–246.

White, R. and Engelen, G., 2000. High resolution modeling of the spatial dynamics of urban and regional systems. *Computers, Environment and Urban Systems*, 24 (5), 383–400.

White, R., Engelen, G., and Uljee, I., 1997. The use of constrained cellular automata for high-resolution modeling of urban land-use dynamics. *Environment and Planning B: Planning and Design*, 24 (3), 323–343.

Wu, F.L., 2002. Calibration of stochastic cellular automata: the application to rural-urban land conversions. *International Journal of Geographical Information Science*, 16 (8), 795–818.

Xie, Y., 1996. A generalized model for cellular urban dynamics. *Geographical Analysis*, 28, 350–373.

Yeh, A.G.-O. and Li, X., 2003. Simulation of development alternatives using neural networks, cellular automata, and GIS for urban planning. *Photogrammetric Engineering and Remote Sensing*, 69 (9), 1043–1052.

Yüzer, M.A., 2004. Growth estimations in settlement planning using a land use cellular automata model (LUCAM). *European Planning Studies*, 12 (4), 551–561.