CALIBRATION AND VALIDATION OF A CA BASED MODEL FOR URBAN DEVELOPMENT SIMULATION USING AN EVOLUTIONARY ALGORITHM. A CASE STUDY IN MEXICO CITY.

CUPUM 05 LONDON

Mauricio SANTILLANA
Researcher
Centro de Investigación en Geografía y Geomática,
"Ing. Jorge L. Tamayo" A.C. CentroGeo.
Contoy No. 137 esq. Chemax,Col. Lomas de Padierna.
C.P. 14740, Mexico D.F.
Mexico.
Tel: +52 55 2615 2224
Fax: +52 55 2615 2403
E-mail: msantillana@centrogeo.org.mx

Fidel SERRANO
Graduate Student
Centro de Investigación en Geografía y Geomática,
"Ing. Jorge L. Tamayo" A.C. CentroGeo.
Contoy No. 137 esq. Chemax,Col. Lomas de Padierna.
C.P. 14740, Mexico D.F.
Mexico.
Tel: +52 55 2615 2224
Fax: +52 55 2615 2403
E-mail: fidel@centrogeo.org.mx

Abstract: Disordered and non-regulated urban growth in the outskirts of Mexico City represents a major problem that has led to the degradation of both natural and social environments. The following study was undertaken to provide decision-making agencies with a suitable tool to generate probable scenarios of urban development based on a given set of hypothetical conditions, such as the implementation of urban policies or urban infrastructure developments. Variations of Cellular Automata (CA) based models have been applied to study the evolution of urban systems in terms of land use, traffic, and intra-urban migration. Calibration of these models is a central need that requires special attention. In this study we use an evolutionary algorithm as a tool to calibrate a CA based model built specifically to mimic the dynamics of land development in vulnerable areas of Mexico City.

Keywords: Urban Modelling, Cellular Automata, Evolutionary Algorithms, GIS-based Simulation, Calibration in GIS.

CALIBRATION AND VALIDATION OF A CA BASED MODEL FOR URBAN DEVELOPMENT SIMULATION USING AN EVOLUTIONARY ALGORITHM. A CASE STUDY IN MEXICO CITY.

1 INTRODUCTION

The natural and social environments of Mexico City are at risk due to disordered and non-regulated urban growth occurring in the outskirts of the city. Aquifer recharge has been blocked in many critical regions, natural preserves have been urbanized, and areas with a high risk of flooding are commonly populated. These problems have become a major concern in recent years. The following study was undertaken to provide decision-making agencies with a suitable tool to generate probable scenarios of urban development based on a given set of hypothetical conditions, such as the implementation of urban policies or urban infrastructure developments. Such a tool must be sensitive to dynamic input from specialists in urban planning as well as decision-makers, and it should be able to incorporate new knowledge into the simulation concurrent with its use. Thus two key points relevant to creating a useful product are interaction with the user through a constant feedback process (Reyes, 2005) and a reliable simulation technique. In this presentation we will focus on the second aspect, emphasising the idea of modeling, not in order to make firm predictions, but rather as a systematic manner of learning about reality, by developing a model that can generate for itself the trajectory of the system in the past (Allen, 1997).

In this project, we developed an ad hoc mathematical model subject to the limitations and needs of a particular case study in which urbanization took place in a protected natural preserve. This is a common phenomenon in third world countries where existing regulatory or environmental laws are not an important factor in determining the evolution of urban settlements. The lack of relevant data played a major role in the development of a methodology unique to the particular phenomenon in consideration. Restrictions brought about by a lack of data caused us to develop a simple CA based approach that may have applications beyond the specific context in which this methodology was developed. In other words this calibration technique may prove to be illuminating in situations where, unlike in Mexico City, a large quantity of data is available. Thus, the main emphasis in our approach was the calibration process. Due to the nature of so called *black box* models, where the relationships among all participating variables and conditions are far from certain, evolutionary algorithms where found to be an efficient way to characterize our system in optimal computational time.

2 THE LAND DEVELOPMENT MODEL

We can understand Cellular Automata (CA) as discrete dynamical systems that consist of regular grid cells in a spatial domain, each cell being in a state that depends on the previous state of the neighboring cells via a transition rule. Variations to this formalism have been implemented to include relevant spatial factors such as terrain elevation and slope, connectivity, distance to

roads, land price, among many others, in the dynamics of the model. These variations have been called Cellular Automata based models and have been a natural choice in urban modeling since they are capable of mimicking urban global structures or patterns, from local interaction rules. An extensive presentation can be found in Batty M. *et al.* (1999) and O'Sullivan *et al.* (2000).

As pointed out by O'Sullivan (2000), developers, firms, financers, regulatory authorities, landlords, tenants, and homebuyers maneuver, collaborate, and compete to change the city for their own purposes. The combination of their activities is what causes *state transitions.* Reducing human interactions to a simple or *solvable* set of transition rules in a dynamic model is clearly a major task. In our particular case study factors like migration of poor sectors of society in search of land to settle was observed to be a strong driving force. The factors included in the transition rules in this model are: closeness to roads, closeness to previous settlements, presence of forest, slope, major topographical accidents, and the sate of the second closest neighbours in the grid.

In any effort to produce a simulation, decisions have to be made to use a reasonable number of parameters to characterize the model in terms of computational time, completeness and goals to be met. Accuracy is something to be understood within the formulation of the simulation. A balance between over-sophistication and oversimplification has to be achieved. In our study we intend to produce a qualitatively acceptable scenario for environmental purposes.

2.1 Characteristics of this CA Based Model

In order to determine the factors to be included in the modeling process, we carefully studied the historical dynamics of land development observed in the areas surrounding the site of our case study in the south of Mexico City, paying particular attention to the geospatial information available in the region. Urban specialists have determined that urban settlements in the surroundings of Mexico City grow at different rates and according to different mechanisms depending on the location and connectivity, among many other factors, of the area (Bazant, 2001). For example, isolated suburban or rural areas are observed to grow at a slower rate compared to areas close enough to the city limits to be eventually absorbed by the city. Connectivity to the road network is noted to be of transcendental importance in the expansion of settlements. The mechanisms to be simulated in this region are basically diffusion either along roads, or due to the existence of previous settlements, and densification in a given settlement. Spontaneous growth was considered in the model but it was not relevant to this first case study.

We determined an appropriate cell size of 100m x 100m by inspecting the scales in which the relevant factors were available. Different approaches have been taken within cell states, such as binary non-urban to urban transition models (Wu, 2002), multi-state land use simulator (White *et al*, 2003), among others. In this paper, we explore the urban dynamics based on a four-state approximation. The basic variable being depicted by the cell state is the percentage of developed or constructed terrain occurring within each cell

(non-urban, low, medium and high). We selected this variable to match the needs of the project as well as to solve the feasibility aspect of data acquisition, since classified aerial photographs were used as an input.

In order to simulate urban development a probabilistic approach was undertaken. This was achieved through a spatial random number generator which determines whether a cell is to be considered for a state change or not according to a prescribed criterion.

3 CALIBRATION

The reliability of a given simulation in the urban context requires that the model should be validated in terms of whether the model can capture the structural similarity between simulated and actual land development. Thus, the need for calibration is great. Along this line of thought, recent CA models have used visual comparison to confirm the simulation results, as pointed out by Wu (2002).

Calibration can be understood as a way to find the best parameters that characterize a given model and, within our context, as a method to maximize the similarity between the output of the model and the actual situation. Thus, calibration can be thought of as an optimization process where the objective function is the distance index between images.

As described above, the variables involved in the dynamics of the current model are very limited and we can be certain that they will not capture the actual phenomenon. We should emphasize that we do not explicitly know the functional relationships among the variables and the objective function. And even if we could use a deterministic algorithm such as a sensibility analysis technique (Kocabas et al, 2004) to obtain estimates of the relationships, the physical meaning of parameters in the model are far from being close to reality. In fact, we could obtain similar outputs with different sets of parameters; thus, the "best parameters" may not exist in terms of a global minimum of the objective function. ***Nevertheless, we may be able to obtain a qualitatively reasonable output from existing data***. The probabilistic nature of our simulation, as well as the incompleteness inherent in this method, calls for a non-deterministic approach in the optimization technique to be used. In the following subsection we will explain our notion of *similarity measure* or *closeness* between two images from a visual perspective.

3.1 Optimization

The input of the CA based model consists of a set of j parameters $\{\theta_j\}$ (that characterize globally and locally the transition rules and the allowed number of cells to be processed) and the initial raster image I. The output of the model is an image $O(t_n, \theta_j)$ that depends on the set of parameters $\{\theta_j\}$ and the number of discrete time steps in years, t_n. See Figure 1.

We endeavored to find a set of parameters $\{\hat{\theta}_j\}$ such that the similarity between the real scenario image R and the output $O(t_n, \hat{\theta}_j)$ satisfies a certain reasonable condition, say for example:

$$i(R,O) \leq \delta\,, \tag{1}$$

where i represents a measure of the distance between both images, and δ represents a prescribed tolerance value. In other words, we look for a set of parameters that are able to qualitatively resemble the dynamics of the system. We will now define what we mean by distance between images and an effective method to satisfy the condition in Equation 1.

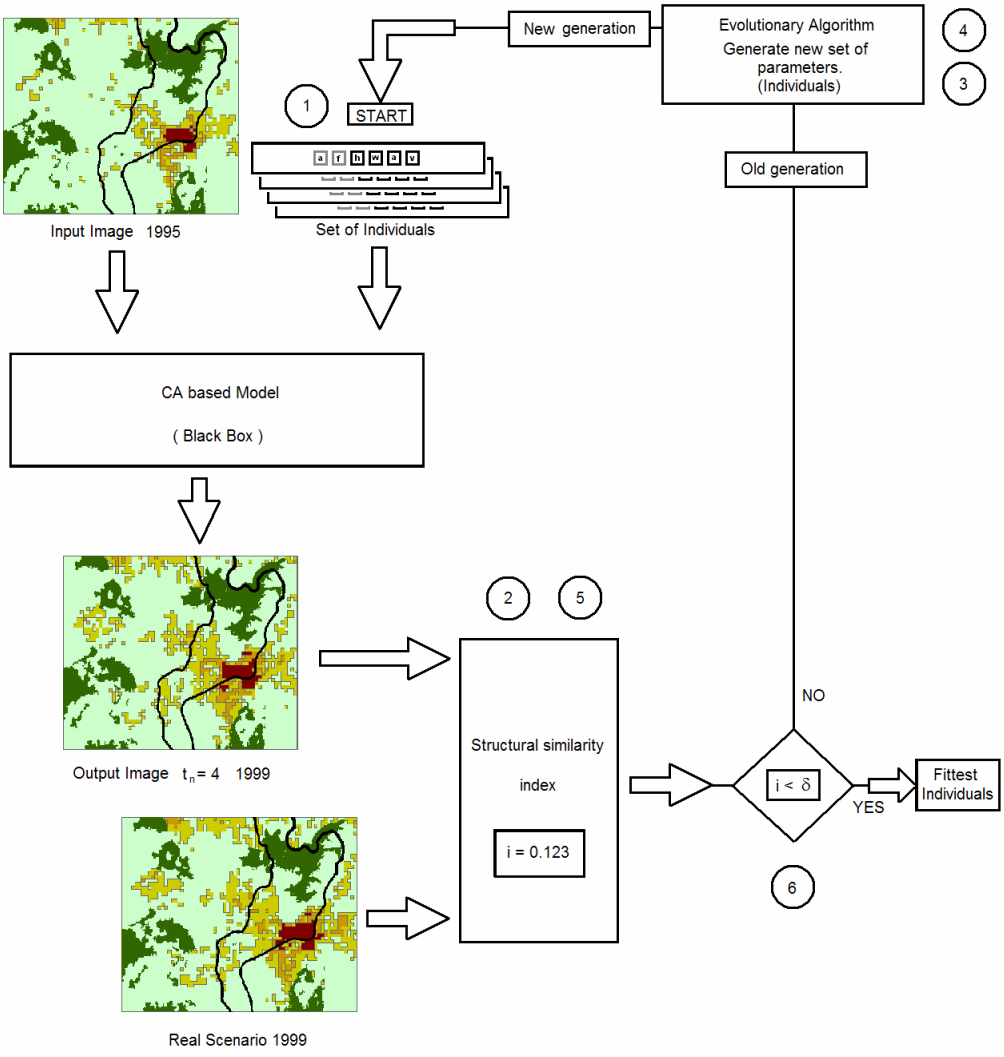

Figure 1 CA Based Model Calibration

3.2 A Measure of Structural Similarity

Image processing teams have developed many algorithms in order to measure how similar images are with respect to one another. A universal method to do this does not exist. Instead, different statistical techniques are built to meet specific targets depending on the application at hand.

We considered it appropriate for our purposes to construct a distance index i using binary (urban / non-urban) classified images in the following way: to find i between the real scenario image R, and the output image $O(t_n, \hat{\theta}_j)$ from the CA based model, we divided the spatial domain Ω in sub-domains Ω_j such that the union $\underset{j}{U}\Omega_j$ of all sub-domains covers the domain Ω totally, i.e. $\underset{j}{U}\Omega_j = \Omega$, and calculated the index i as:

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

where $N(\Omega_j)$ is the total number of sub-domains Ω_j, $c_k(R)$ and $c_k(O)$ represent the state of a given cell c_k in Ω_j, either 1 if urban or 0 if non-urban, in each image R or O. As an example, if we let $N(\Omega_j) = 1$, then $\Omega = \Omega_1$. The index i in this case, represents the square of the count of cells that are different from one image to the other. For such a choice, i does not provide any morphologic insight about the similarity of the images. See Figure 2 below.

A more general index E can be generated as a weighted linear combination of indexes i_α :

$$E = E(R,O) = \sum_{\alpha} w_\alpha i_\alpha(R,O) \quad \text{thus} \quad 0 \le E \le 1 \tag{3}$$

where α represents a given choice of sub-domains Ω_j satisfying $\underset{j}{U}\Omega_j = \Omega$, i_α defines the previously defined index for such a choice, and w_α the importance or weight for the given scale. Note that $\sum_{\alpha} w_\alpha = 1$. This index E defines a natural way to take into account different scales within the method to measure structural similarity.

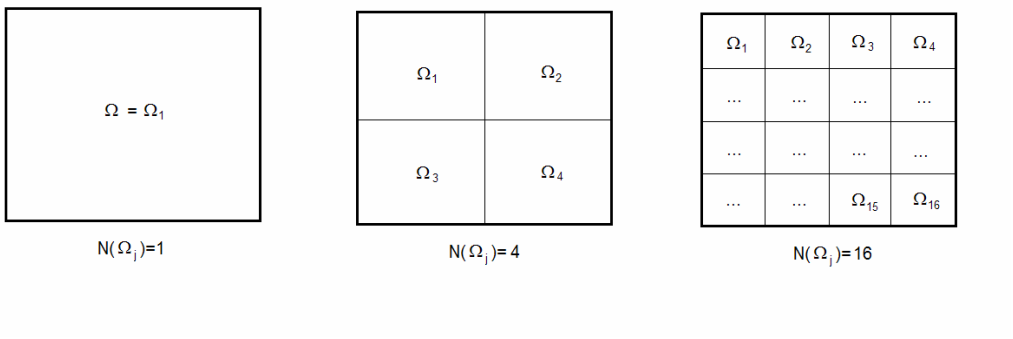

Figure 2 Examples of Domain Partition

In fact, if we consider images $\{I\}$ to be scalar functions whose domain is a subset $\Omega \in R^2$, i. e. $I : \Omega \to R$, and chose $N(\Omega_j)$ to coincide with the number of cells in the CA model, then calculating the index i_α for such a partition

$\alpha = \alpha_{\max}$ is actually calculating the square of the discrete $L^2 - norm$ normalized by the size of the domain $\mu(\Omega)$, of the difference between image R and O, i.e.

$$i_{\alpha_{\max}} = \frac{1}{\mu(\Omega)} \|R - O\|_{L^2}^2$$

From this perspective, it does make sense to think of E as some kind of distance function. For our purposes we chose E to be $E = w_1 i_1 + w_2 i_2 + w_3 i_3$, where $\alpha = 1,2,3$ correspond to $N(\Omega_j) = 1,4,16$ respectively. The choice of weights $\{w_\alpha\}$ was empirically determined after several numerical experiments to obtain best performance. It should be emphasized that the index E was introduced as a means to calibrate the model in visual terms and it should not be understood a precise way to measure how close an image is to another.

3.3 Evolutionary Algorithms

Evolutionary algorithms are stochastic search methods inspired by the work of Charles Darwin on the role of selection in the origin of species (Darwin, 1859). Evolutionary algorithms operate on a population of potential solutions (individuals) applying the principle of survival of the fittest to produce better and better approximations to a solution. Since they work on populations (generations) as opposed to single individuals, the search is performed in parallel. They have been widely used in the last two decades to solve scientific problems. Banzhaf (1998) characterized the classes of problems where this approach proved to be useful as follows: a) where the interrelation among variables is poorly understood, b) where finding the size and form of the solution is the most difficult aspect of the problem, c) where the traditional mathematical analysis does not or cannot provide an analytical solution, d) where an approximate solution is acceptable (or it is even the only possible result), e) where little improvements in the performance are highly appreciated, and f) where there is a great quantity of data, in computer readable format, that requires examination, classification, and integration (such as proteins and DNA sequences in molecular biology, astronomical data, data from satellite imagery, etc.). Clearly, our model falls within several categories where evolutionary algorithms have been successful. A detailed description of these algorithms can be found in Krzanowski (2001). Table 1 presents a equivalence between the languages that characterize CA models and evolutionary algorithms.

Table 1: Cellular Automata and Evolutionary Algorithms

Cellular Automata Terminology	Evolutionary Algorithm Terminology
A parameter θ_j of the CA model	A gene
A set of parameters $\{\theta_j\}$ characterizing the CA model	A chromosome
The output $O(t_n, \theta_j)$ of the CA for a given set of parameters	A grown up individual
The distance E between the output image and the real one.	The adaptability index of a given individual

We used a simple evolutionary algorithm for our optimization process. Simple evolutionary algorithms can be presented as a series of 6 steps (Raper, 2000).

1. Randomly initialize a population of individuals.
2. Assign the adaptability index (Calculate the objective function: distance E).
3. Select the best individuals to produce offspring (Selection).
4. Parents are recombined (Cross-over, Mutation).
5. Calculate the adaptability index for new generation.
6. If optimization criteria are not met repeat from step 3. Stop otherwise.

See Figure 1 and Figure 3 to identify these steps within the context of the present project. The word *offspring* in this context means a recombination of two sets of parameters (chromosomes) from two individuals (the parents) to produce a new individual (the offspring). As presented in Figure 3, this process is called *cross-over* and it may be seen as a way to divide the parents´ chromosomes (sets of parameters) into subsections in order to mix them. The example presented in Figure 3 shows the generation of two new individuals from two parents. In order to encourage genetic variability, we add another mechanism called *mutation* which modifies one or more genes (single parameters) randomly.

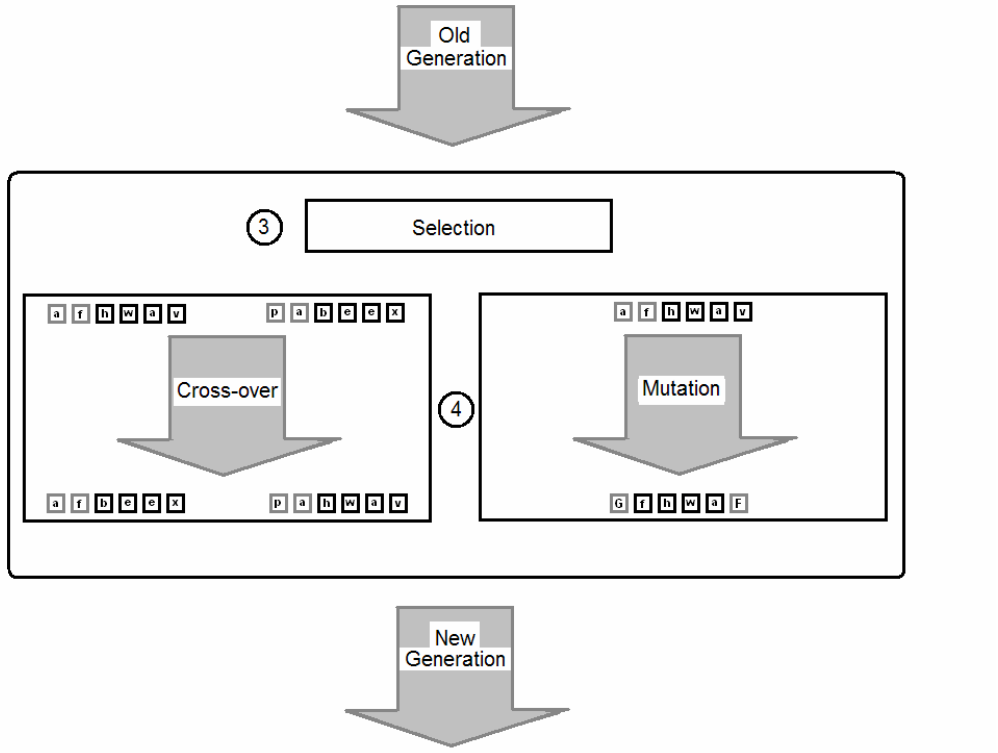

Figure 3 Evolutionary Algorithm

4 CASE STUDY. TOPILEJO

Our first case study was Topilejo, a town located in the south west of Mexico City where disordered urban growth, as described previously, has taken place. The speed of growth in Topilejo has increased considerably in recent years, causing the natural resources of the area to be at risk. The region where this town is located is surrounded by natural preserves which have attracted the attention of environmental agencies. In particular, the issues discussed previously are of great concern in this location.

The study area was limited to a rectangle of land measuring approximately 7 Km by 6 Km. It was divided into roughly 4000 cells (72 x 57). In order to provide the model with appropriate input, we proceeded to classify the state of each cell in time using aerial photographs from 1995, 1999, and 2002. The CA based model was implemented using ArcInfo (AML and GRID Module), coupled with the evolutionary algorithm coded in java. A set of randomly chosen parameters and the classified image of 1995 were used as the initial state of the evolutionary algorithm and CA based model, respectively. We calibrated the process by simulating a probable scenario for the area in 1999 and comparing the simulation with the actual situation as it occurred in 1999. For validation purposes, we used the set of parameters $\{\hat{\theta}_j\}$ obtained from the

calibration process and the actual scenario in 1999 as the initial image, and calculated a probable scenario for 2002 to be compared with the actual scenario in 2002. See Figure 4.

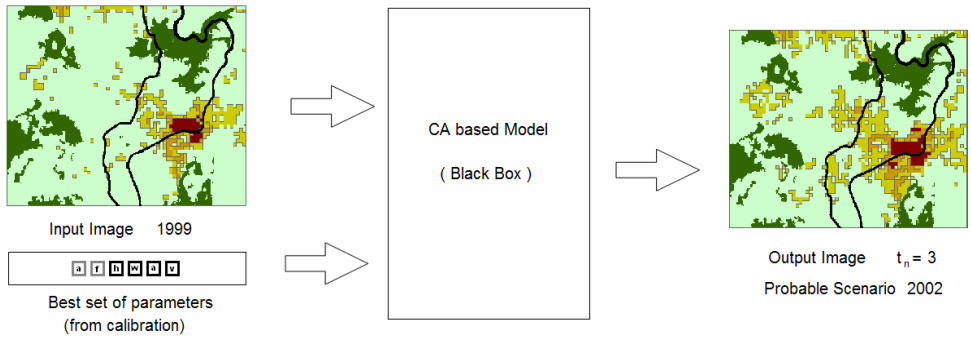

Figure 4 Probable Scenario Prediction

4.1 Results of the Case Study

In order to allow enough genetic diversity in the evolutionary algorithm, the number of individuals per generation was chosen to be 8. Four were generated from the process of cross-over between the parents, and the other four were the result of a mix of cross-over and mutation. See Figure 3. These numbers were determined empirically after several numerical tests. The performance of the evolutionary algorithm is shown in Figure 5.

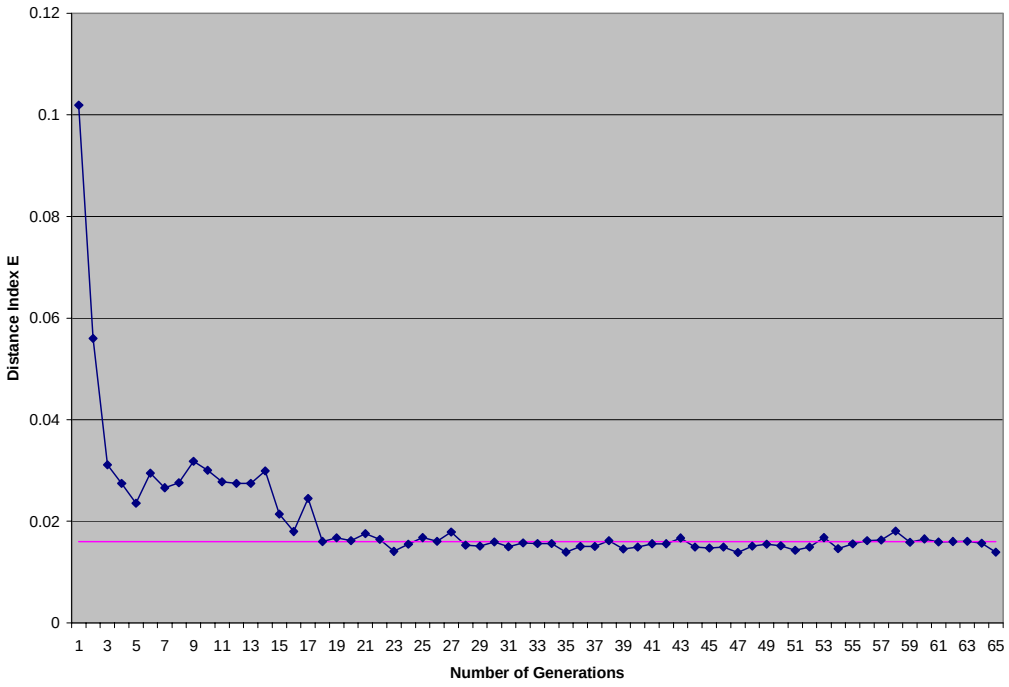

Figure 5 Calibration Performance

The graph in Figure 5 shows the adaptability index (Distance E between the real image in 1999 and the image produced by the model) of the fittest individual per generation. We can see that in approximately 19 generations, the algorithm reached a steady state in the index E of about 0.016. It has to be emphasised that as the algorithm progressed, the chromosomes or sets of fittest parameters, kept varying. In other words, similar images (with respect to the adaptability index) are produced with different parameters. This fact supports the use of evolutionary algorithms as opposed to deterministic optimization algorithms such as gradient based optimization, where the output of the model is expected to change continuously with respect to initial data. It is worth mentioning that the probabilistic quality of the CA based model also plays an important role. Mathematically we can understand the output $O(t_n, \theta_j)$ of the model as a sum of two terms

$$O(t_n, \theta_j) = \bar{O}(t_n, \theta_j) + \varepsilon \quad (3)$$

where $\bar{O}(t_n, \theta_j)$ represents the morphological or average aspect of the image, and ε represents the probabilistic noise inherent to the model. The dependence of ε with respect to t_n and θ_j is not explicitly written so as to simplify the notation. Our purpose is to get an image $\bar{O}(t_n, \theta_j)$ that resembles R in average, even though our efforts were made to minimize the distance between the real scenario R and $O(t_n, \theta_j)$ through the index E. This implies that different images are generated with the same set of parameters. These images are far from one another by ε, the noise.

For our best performance, the index E between the average output of the model and the real scenario image R_{1999} in the calibration process was calculated as the average of indexes $E\acute{}s$ between 50 outputs generated with the fittest parameters $\{\hat{\theta}_j\}$ and R_{1999}, obtaining that

$$E(R_{1999}, \bar{O}_{t_n=4}(\hat{\theta}_j)) = .016 \quad (4)$$

and the noise $\varepsilon = .0018$, which was identified as the standard deviation of such $E\acute{}s$ values. The index E between the average output of the model and the real scenario image in 2002, was calculated similarly, obtaining that

$$E(R_{2002}, \bar{O}_{t_n=3}(\hat{\theta}_j)) = .026 \quad (5)$$

and $\varepsilon = .0025$.

As a means to have an idea of the performance of the model, the indexes E between the real scenario images from 1995 and 1999,

$$E(R_{1995}, R_{1999}) = .052 \quad (6)$$

and between R_{1999} and R_{2002} ,

$$E(R_{1999}, R_{2002}) = .079 \quad (7)$$

were calculated. These latter quantities (6) and (7) are useful in order to normalize the results. In other words, if we considered the worst performance ($E = 1$) of a model to be such that it would not add anything to the initial image and the best to actually obtain an index E closer to zero then the model could be characterized as within 30% accuracy, since the ration between (5) and (7)

is 0.33 approximately. This assumption is probably too strict for a qualitative modelling approach, since a meaning to the value of E is not clearly defined, but it gives a sense of its qualitative use within this context. On the other hand, from the visual point of view the match between simulated and real images in 2002 is high. See Figure 5.

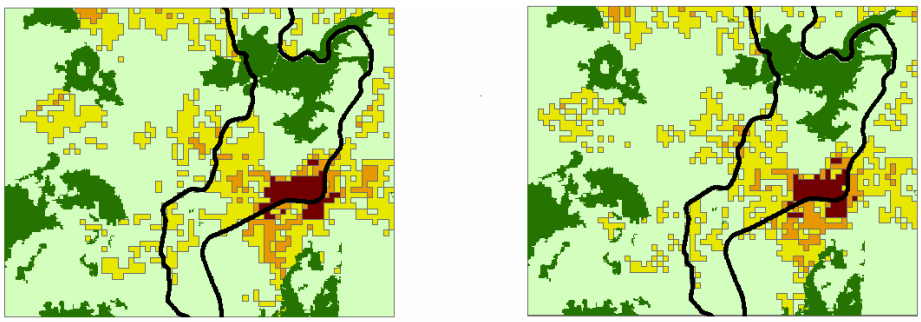

Figure 6 Results of the Case Study

5 Conclusions

Within the GIS environment, a validation of the model should be evaluated with respect to its ability to produce a visually reasonable output. From this point of view, we considered the performance of this methodology successful.

Due to the highly probabilistic features involved along the generation of a calibrated model in this approach, the index E should be understood as a useful tool to produce a properly calibrated model but not as a precise way to quantify its performance.

We would like to mention that within the black box models context, the meaning of fittest individuals (sets of parameters) is not related to the real physical variables. In other words, the values of the parameters only lead to obtaining qualitative relationships among variables but are not to be interpreted as describing the actual dynamics of the system.

In order to construct a useful tool to simulate probable scenarios, we previously identified two key points, the interaction with the user through a constant feedback process and a reliable simulation technique. We focused on the second one in our project and found weaknesses in our predictions, which, we believe, could be overcome by directing more effort toward the first point known as cybercartographic process. A cybernetic process is one that allows interactivity between the user and the computer, the user and modeller, and the modeller and computer. Cybercartography addresses such qualities in a geospatial context (Reyes, 2005). This could improve the interaction between decision-makers and experts and the modelling teams, by developing tools to incorporate current knowledge into the models. Such tools should include geospatial consensus techniques such as the STRABO methodology (Luscombe *et al*, 1983). This would bring a substantial

improvement in the model performance, for example to identify spontaneous growth seeds in a given area.

6 Ongoing and future work

Applications of this model include deforestation growth (which can be obtained as the inverse process of urban development), model coupling with a water recharge simulation, the use of dynamic street grids to improve performance. In order to study larger areas, a multi-domain, multi-scale approach is being considered. This would imply dividing a large study area into sub-domains and calibrating different sub-domains separately. Adequate matching criteria in the overlapping boundaries are to be carefully studied.

ACKNOWLEDGEMENTS

This study was supported by the Centro de Investigación en Geografía y Geomática, “Ing. Jorge L. Tamayo” A.C. CentroGeo. Authors are grateful for the enthusiastic support received from the institution.

REFERENCES

Allen, P. (1997) **Cities and Regions as Self-Organizing Systems: Models of Complexity (Environmental Problems & Social Dynamics).** Gordon & Breach, Amsterdam.

Banzhaf, W. (1998) **Genetic Programming. An Introduction.** Morgan Kaufman, San Francisco.

Batty, M., Xie Y., Sun Z. (1999) Modeling urban dynamics through GIS-based cellular automata, **Computers, Environment and Urban Systems, Vol. 23,** 205-233.

Bazant, J. (2001) **Periferias Urbanas. Expansión urbana incontrolada de bajos ingresos y su impacto sobre el medio ambiente**, Trillas, Mexico.

Darwin, C. (1859) **On the Origin of Species.** Murray, London.

Kocabas, V., Dragicevic S. (2004). Sensitivity Analysis of A Gis-based Cellular Automata Model. **XXth ISPRS Congress,** Istanbul, Turkey, July 2004.

Krzanowski, Roman (2001) **Spatial Evolutionary Modeling (Spatial Information Systems)**, Oxford University Press, Oxford.

Luscombe, B.W., Poiker T.K. (1983) Strabo—An alternative GIS approach to decision making for planning applications in data scarce environments. P. 264–269. **Proceedings of the Sixth International Symposium On Automated Cartography. Vol. 1**. American Congr. on Surveying and Mapping, Bethesda, MD.

O´Sullivan, D., Torrens, P.M., (2000) Cellular Models of Urban Systems, **Fourth International Conference on Cellular Automata for Research and Industry,** Karlsruhe University, Karlsruhe, Germany, September 2000.

Raper, F.J. (2000) **Multidimensional Geographic Information Science,** Taylor & Francis, London.

Reyes, C. (2005) **Cibercartography: Theory and Practice**. Chap 4. Elsevier Scientific, Amsterdam. In press.

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 A. Vol. 25**, 1175-1199.

White, R. and Engelen, G. (1994) Cellular dynamics and GIS: modeling spatial complexity. **Geographical Systems. Vol. 1**, 237-253.

Wu, F., Webster C. (2000) Simulating artificial cities in a GIS environment: urban growth under alternative regulation regimes, **Int. J. Geographical Information Science, Vol. 14, No. 7,** 625-648.

Wu, F. (2002) Calibration of stochastic cellular automata: the application to rural-urban land conversions, **Int. J. Geographical Information Science, Vol. 16, No. 8,** 795-818.

Filename: F_cupum05_article_Ref42
Directory: C:\Documents and Settings\Monica\Desktop
Template: C:\Documents and Settings\Monica\Application Data\Microsoft\Templates\Normal.dot
Title: PRESENTATION GUIDELINES FOR FINAL PAPER - CUPUM'05 LONDON
Subject:
Author: Sonja Curtis
Keywords:
Comments:
Creation Date: 5/1/2005 6:28 PM
Change Number: 3
Last Saved On: 5/1/2005 11:00 PM
Last Saved By: Monica Anne Brown
Total Editing Time: 7 Minutes
Last Printed On: 5/1/2005 11:06 PM
As of Last Complete Printing
Number of Pages: 14
Number of Words: 4,454 (approx.)
Number of Characters: 25,389 (approx.)