UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

UCL

WORKING PAPERS SERIES

Paper 42 - Feb 02

Empiricism and Stochastics in Cellular Automaton Modeling of Urban Land Use Dynamics

ISSN 1467-1298

casa

Centre for Advanced Spatial Analysis University College London 1 - 19 Torrington Place Gower St London WC1E 7HB
Tel: +44 (0)20 7679 1782 casa@ucl.ac.uk www.casa.ucl.ac.uk

Empiricism and Stochastics in Cellular Automaton Modeling of Urban Land Use Dynamics

Cláudia Maria de Almeida[1], Antonio Miguel Vieira Monteiro[1],
Gilberto Câmara[1], Britaldo Silveira Soares-Filho[2],
Gustavo Coutinho Cerqueira[2], Cássio Lopes Pennachin[3],
& Michael Batty[4]

February 2, 2002

---

[1]National Institute for Space Research (INPE), Avenue dos Astronautas, 1758 – 12227-010, São José dos Campos, São Paulo, Brazil, [2]Department of Cartography, Federal University of Minas Gerais (UFMG) Avenue Antônio Carlos, 6627 – 31270-900, Belo Horizonte, Minas Gerais, Brazil, [3]Intelligenesis do Brasil Ltda, Avenue Brasil, 1438-1505, 30140-003, Belo Horizonte, Minas Gerais, Brazil, [4]Centre for Advanced Spatial Analysis, University College London, 1-19 Torrington Place, London WC1E 6BT, UK

{almeida@ltid.inpe.br, miguel@dpi.inpe.br, gilberto@dpi.inpe.br, britaldo@csr.ufmg.br, cerca@csr.ufmg.br, cassio@intelligenesis.net, mbatty@geog.ucl.ac.uk}

Abstract

An increasing number of models for predicting land use change in regions of rapid urbanization are being proposed and built using ideas from cellular automata (CA) theory. Calibrating such models to real situations is highly problematic and to date, serious attention has not been focused on the estimation problem. In this paper, we propose a structure for simulating urban change based on estimating land use transitions using elementary probabilistic methods which draw their inspiration from Bayes' theory and the related 'weights of evidence' approach. These land use change probabilities drive a CA model – DINAMICA – conceived at the Center for Remote Sensing of the Federal University of Minas Gerais (CSR-UFMG). This is based on an eight cell Moore neighborhood approach implemented through empirical land use allocation algorithms. The model framework has been applied to a medium-size town in the west of São Paulo State, Bauru. We show how various socio-economic and infrastructural factors can be combined using the weights of evidence approach which enables us to predict the probability of changes between land use types in different cells of the system. Different predictions for the town during the period 1979-1988 were generated, and statistical validation was then conducted using a multiple resolution fitting procedure. These modeling experiments support the essential logic of adopting Bayesian empirical methods which synthesize various information about spatial infrastructure as the driver of urban land use change. This indicates the relevance of the approach for generating forecasts of growth for Brazilian cities particularly and for world-wide cities in general.

Keywords

Land Use Change, Transition Probabilities, Bayesian Methods, 'Weights of Evidence', Cellular Automata, Urban Growth, Urban Planning

Acknowledgments

We wish to thank the Water Supply and Waste Water Disposal Department (DAE), and the Bauru Planning Secretariat (SEPLAN) for providing the city maps. We are also grateful for the help and co-operation of the technical and administrative staff of the Center for Remote Sensing of the Federal University of Minas Gerais (CSR-UFMG). Crucial financial support has been granted to this research by the Academic Coordination for Remote Sensing of the Brazilian National Institute for Space Research (PG-SER/INPE), the São Paulo State Foundation for Research Support (FAPESP) and the Brazilian National Foundation for the Undergraduates and Graduates Upgrade (CAPES). This paper was written when Cláudia Maria de Almeida was an academic visitor at University College London (CASA) from October 2001 to February 2002.

1 Introduction: Cell-Space Models of Land Use Change

There is now widespread recognition that most operational urban models, fashioned as far back as the 1960s for policy prescription, are strictly limited in their abilities to generate meaningful predictions in situations of rapid urban growth. These models are largely cross-sectional in structure and at best, can only be used for comparative analysis of long term change where it is assumed that a future system has adjusted to some temporal equilibrium (Wegener, 1994). One of the reasons why such models have fallen out of 'theoretical fashion', although they are still used in practice, is that they fail spectacularly to deal with the kinds of dynamics which characterize many urban situations; in developing countries, for example, where urbanization is still composed of substantial migration from the rural hinterland to the central city, and in the developed world where many cities are facing an explosion of growth through urban sprawl. In short, such models cannot handle rapid urban growth whether it be created by an increasing population and/or the demand for more living space.

To date, however, few models have been built which truly represent the dynamics of urban growth which is consistent with what we know about such change and the data that is available to measure it. As the dynamics is intrinsically complex and all but impossible to unravel in terms of the data we have, the models that have been built so far usually begin with the simplest of ideas. Cellular automata (CA) models have become popular largely because they are tractable, generate a dynamics which can replicate traditional processes of change through diffusion, but contain enough complexity to simulate surprising and novel change as reflected in emergent phenomena. CA models are flexible in that they provide a framework which is not overburdened with theoretical assumptions, and which is applicable to space represented as a raster or grid. These models can thus be directly connected to raster data surfaces used in proprietary GIS (geographic information systems). They are being used as much to implement map algebras on raster grids as in IDRISI, for example, as to simulate the intrinsic dynamics of systems that can be represented in this form.

Although early proposals for the use of CA in urban modeling tended to stress their pedagogic use in demonstrating how global patterns emerge from local actions (Tobler, 1979; Couclelis, 1985), increasingly models have been proposed which depart from the basic elements. Strict CA articulate the growth (or change) process in terms of highly localized neighborhoods where change takes place purely as a function of what happens in the immediate vicinity of any particular cell. Action-at-distance is forbidden for it is argued that the intrinsic dynamics which generates emergent phenomena at the global level, is entirely a product of local decisions which have no regard to what is happening outside their immediate neighborhood (Batty, 2000). Early models such as Tobler's (1979) model of Detroit and Couclelis's (1989) model of developer behavior in Los Angeles were pedagogic in this spirit, yet the attractiveness of the approach which grew alongside the enormous interest in GIS, has eventually led to a flurry of more practical applications to urban problems. In this, the strict adherence of CA to the most local of neighborhood is inevitably relaxed, and the models that have emerged are best called cell-space – CS models rather than CA (Albin, 1975).

There are cell-space models dating back to the 1960s such as that developed for Greensboro, NC by Chapin and Weiss (1968) but these models are more strictly econometric in their structure, and do not appeal to CA in any form. However since the early-1990s, there have been more than 20 significant practical applications although in all cases, local neighborhoods have been generalized to regions or fields, while the emphasis on fitting the actual land development process implied by these models to data has been weak (Schock, 2000). Nevertheless, there are now many varieties of cell-space model which involve matching their predictions to data, and at least three approaches to their estimation have emerged. The more traditional models such as those developed by White and Engelen (1993, 1997, 1998) for Cincinnati and other US cities, and for the island of St. Lucia, simply dimension the model's parameters with data taken from these applications. In contrast, various brute force approaches to searching the parameter space of such models has been tried by Clarke et al. (1997, 1998) in their various models of US metropolitan growth. The most promising methods of estimation, however, are largely data driven, and use contemporary pattern-fitting methods such as neural nets (Wu, 1998; Xia and Yeh,

2000) and evolutionary learning (Papini, et al., 1998). But in all these applications, it is assumed that the discrete dynamics of the growth process is unknown, hence untestable, and thus these models usually fall back on assuming their dynamics and testing predictions in a cross-sectional fashion at a single point in time.

In this paper, we will develop a model of land use change which is operationalized using a CA-like framework but whose locational structure is data-driven, in the same manner as that used by Chapin and Weiss (1968) over 40 years ago. In short, we will determine transition probabilities governing changes in land use as functions of a variety of socio-economic and infrastructural factors whose relationships with changes in different land use types are measured through spatial correspondences akin to the methods of map overlay. The formal framework used to determine the probabilities will be Bayesian, involving an updating of prior probabilities through the 'weighted evidence' provided by these factors. These probabilities are then used in various heuristic procedures which select particular cells to be developed according to ranking rules determined by the cellular operations (Soares-Filho, Cerqueira, and Pennachin, 2002). The results of fitting the model to a medium sized town in Brazil are then discussed where a fit statistic appropriate to map counting procedures is employed.

We will first state the model in terms of its aggregate, non-spatial structure, introducing the idea that the process of urban change is one of transition which can be easily envisaged as a first-order Markov process. We then unpack this model structure to deal with change at the level of the cell. The amount of change is an external input to the model for the focus here is on its spatial allocation. The 'weights of evidence' process is then outlined and the way this is embedded into the algorithmic structure of the cell-based operations introduced. The data available, the selection of factors determining location transition probabilities, and the method of choosing evidence from this data are then discussed. The model is used to make various predictions which are examined in terms of their fit to land use change in the town of Bauru between 1979 and 1988. The fit is deemed acceptable although several problems relevant to the structure of the model are identified and taken forward as a basis for further research.

2 The Model Framework

2.1 Transition Dynamics

The system that we define has N cells which can take on M different and mutually exclusive states. Each cell is identified by a location $i, j = 1, ..., N$ while each state is a land use (or related activity type) which falls in the range $k, l = 1, ..., M$. It is thus assumed that a typical cell *i* has only one land use *k* at time *t* which is thus defined as

$$N_i^k(t) = 1, \quad N_i^l(t) = 0, k \neq l, l = 1, ..., M, \quad \sum_k N_i^k(t) = 1 \quad . \tag{1}$$

From equation (1), it is easy to define aggregates in terms of cells or land uses. Then

$$N^k(t) = \sum_i N_i^k(t) \quad , \tag{2}$$

$$N(t) = \sum_k N^k(t) = \sum_k \sum_i N_i^k(t) \quad . \tag{3}$$

We also note that the total number of cells in the system is fixed through time, that is $N(t) = N, \forall t, t = 1, ..., T$ where *T* is the total number of time periods considered, and that although changes between the total number of distinct land uses in cells is what this representation allows, the total size of the system is conserved. In this way, growth or decline is conceived of as a transition from one state or land use *k* to another *l*. This formulation means that densities in each cell are the same, and it implies that cells must be the same size. Thus the total density of any land use in the system is the proportion of land use $\rho^k(t)$ defined as

$$\rho^k(t) = N^k(t) / N \quad . \tag{4}$$

All of this is consistent with a grid system which is at sufficiently fine level of spatial resolution to enable each cell to be associated with one and only one land use at any one time. Whether or not this is the case is an empirical matter in terms of the specific application.

The dynamics in the model are represented in a straightforward manner as transitions from one land use *k* to another *l*. At the cellular level, a transition in *i* where there is a land use *k* at time *t* to a land use *l* at time $t+1$ is defined as

$$\Delta N_i^{kl} = 1 \quad \text{where} \quad N_i^k(t) = 1 \quad \text{and} \quad N_i^l(t+1) = 1 \quad . \tag{5}$$

The aggregate transition in land use for the entire system is

$$\Delta N^{kl} = \sum_i N_i^l(t+1) - \sum_i N_i^k(t) = N^l(t+1) - N^k(t) = \sum_i \Delta N_i^{kl} \quad . \tag{6}$$

Actual change – growth or decline – in aggregate for each land use is thus

$$\Delta N^k = N^k(t+1) - N^k(t), \quad \text{where} \quad \sum_k \Delta N^k = 0 \quad , \tag{7}$$

which indicates the conservation imposed by having *N* cells with only one land use per cell.

Although the model works at the cellular level, it is useful to consider these transitions at the aggregate level for this enables the long term dynamics of the model to be articulated in a simple and direct manner. Total land use *l* at $t+1$ is calculated as

$$N^l(t+1) = \sum_k \Delta N^{kl} \quad , \tag{8}$$

which can be written in transition probability form as

$$N^l(t+1) = \sum_k P^{kl} N^k(t) \quad , \tag{9}$$

where the probability is defined as

$$P^{kl} = \frac{\Delta N^{kl}}{N^k(t)} = \frac{\Delta N^{kl}}{\sum_l \Delta N^{kl}}, \quad \text{and} \quad \sum_l P^{kl} = 1 \quad . \tag{10}$$

We can write the process implied in equation (8) in matrix-vector form as a first-order Markov process if we define the probability of any land use *k, l* at time $t, t+1$ as being

$$\pi^k(t) = \frac{N^k(t)}{N}, \quad \text{and} \quad \pi^l(t+1) = \frac{N^l(t+1)}{N} \quad . \tag{11}$$

Using these definitions, we write equation (8) as

$$\boldsymbol{\pi}(t+1) = \boldsymbol{\pi}(t)\ \mathbf{P} \quad , \tag{12}$$

where the limit of equation (12) for constant transition probabilities and relatively weak conditions of connectivity in the matrix $\mathbf{P}$, leads to

$$\boldsymbol{\pi}(t+T) = \boldsymbol{\pi}(t)\ \mathbf{P}^T \quad . \tag{13}$$

As T goes to the limit, then $\mathbf{P}^T \rightarrow \mathbf{Z}$, and equation (13) gives the steady state probabilities $\boldsymbol{\pi}$ as $\boldsymbol{\pi} = \boldsymbol{\pi}\ \mathbf{Z}$ (JRC, 1994). This is a very convenient form. It not only illustrates that the total flows are conserved to the *N* cells of the system which in a sense is trivial, but it also demonstrates that for constant transition probabilities, which might be assumed for sufficiently large time periods, then there is an implied long term equilibrium. In the case of the applications assumed here, then we can illustrate this for the simple 2 x 2 state case where the two land uses in question are undeveloped and developed land. Imagine a set of transition probabilities where there is substantial conversion of undeveloped to developed land but there is no land that changes in the other direction. The $\mathbf{P}$ matrix for this case can be assumed and then it is easy to show that the $\mathbf{Z}$ matrix is characteristic of an absorbing Markov chain where all the land moves to development in the limit of *T*. In fact in such a simple system, you can guess the limit as the following example shows:

$$\mathbf{P}=\begin{bmatrix}0.75 & 0.25\\ 0.00 & 1.00\end{bmatrix}\text{, guessing }\mathbf{Z}=\begin{bmatrix}0.00 & 1.00\\ 0.00 & 1.00\end{bmatrix}\text{ with }\mathbf{P}^{10}\approx\begin{bmatrix}0.06 & 0.94\\ 0.00 & 1.00\end{bmatrix}.$$

The transition to one state in this case is rather quick. In real applications, transitions are likely to be defined as being much slower – the time periods being much shorter – and of course transitions are not stable from time period to time period. However in the example we develop in this paper for the city of Bauru, this kind of analysis is useful because it points out the intrinsic trends in the system as well as the focusing on the speed at which transitions take place.

2.2 Locational Dynamics

In most cell-space models of this type, there is strong separability of locational from growth mechanisms in that invariably growth or decline is assumed at the aggregate level, and the focus of such models is on allocating land use of various kinds to cells. In short, it is not possible to write the aggregate dynamics in terms of the disaggregate locational processes in the simple form we have just shown. It is therefore not possible to produce a simple algebra for the model from the bottom up and this is due to the fact that the model's locational mechanisms are expressed as stochastic decision rules, often of binary variety. This means that the locational processes adopted by the model distribute quantities of the form ΔN^k. These totals are not calibrated in any way; they are exogenous to the system as the previous exposition assumes. In this section however it will be clear that the model's locational processes are subject to a calibration in that we will assume like many before us (for example, Chapin and Weiss, 1968 to Clarke et al., 1998) that we can define social and infrastructural factors that determine the best transition probabilities at the cellular level.

As we have stated, we define cellular change as $\Delta N_i^{kl}=1$ or 0. However, although this is in binary form, this change is associated with a transition probability $p_i^{kl}(t)$

relevant to predicting the transition from *k* to *l* but not observable. We are only able to observe the discrete change but we assume that there is a set of such probabilities at the cellular level which is used in making decisions about what land use change is most likely. We choose these probabilities to optimize the predicted transitions $(\Delta N_i^{kl})'$ to the observed ΔN_i^{kl} over all cells *i*. The generic form for these probabilities is

$$p_i^{kl}(t) = \Theta\left[f\left(X_i^1, X_i^2, \ldots X_i^E\right),\ \ g\left(N_i^k(t), N_j^l \in \Omega_i, l \in M\right)\right] \quad . \qquad (14)$$

This equation is in two parts. The first which is governed by the function $f(\bullet)$ combines a series of intervening factors for each cell $\{X_i^e, e = 1, 2, ..., E\}$ using a Bayesian updating procedure based on the 'weights of evidence' logit approach applied to map overlays (Bonham-Carter, 1994). The second function $g(\bullet)$ relates the existing land use type in cell *i* to all those land uses *l* in their Moore (8 cell) neighborhood Ω_i. Two operations – the *expander* and *patcher* – are used to effect this by choosing cells which are at the 'frontier' of development (expander) or in distinct but separable development nuclei (patcher). The cell probabilities determined from the first function $f(\bullet)$ are modified by these operations when the second function $g(\bullet)$ is applied. We will illustrate this in the next section.

Once the final set of probabilities $\{p_i^{kl}(t)\}$, normalized to $\sum_{ikl} p_i^{kl}(t) = 1$, is determined, then these are ranked across all cells and land uses as $p_i^{kl} > p_j^{mn} > .....$ and so on. Various methods can then be used to determine the land use conversion. In fact in the application here, a Monte Carlo algorithm is used to choose the actual transitions. This is based on sampling random numbers which are drawn until the total cells required for each land use transition in the entire system is met. Another way would simply be to define the minimum number of cells for each land use transition which would be composed of the maximum ranked probabilities, but in this case the algorithm would no longer keep its stochastic nature. The actual procedure is a little trickier than this in practice because the expander and patcher procedures are iterated in a certain order for reasons that will be explained below.

It is possible however to give some indication about the long term dynamics of the model from the cellular level using the same kind of first-order Markov process defined for the aggregate transitions in equations (8) to (13) above (Hobbs, 1993). We first define a land use transition for each cell *i* as

$$Q_i^{kl} = \frac{p_i^{kl}}{\sum_l p_i^{kl}} , \quad \sum_l Q_i^{kl} = 1 \quad , \tag{15}$$

from which we define the relevant Markov transitions as

$$\pi_i^l(t+1) = \sum_k \pi_i^k(t)\, Q_i^{kl} \quad , \tag{16}$$

with the appropriate matrix-vector format as

$$\boldsymbol{\pi}_i(t+1) = \boldsymbol{\pi}_i(t)\, \mathbf{Q} \quad . \tag{17}$$

The long term dynamics implicit at the cellular level is the same as the aggregate, that is $\mathbf{Q}_i^T \rightarrow \mathbf{Z}_i$, and the steady state equation for each cell *i* is $\boldsymbol{\pi}_i = \boldsymbol{\pi}_i\ \mathbf{Z}$. As the overall dynamics of the model is built from the bottom up, we would not expect the implied distributions from equation (17) to meet the actual distribution predicted by the full model. In short, if we choose the overall transition for each cell from equation (17) as

$$N_i^k(t+T) = 1 = \max_l \left\lfloor \pi_i^l(t+T) \right\rfloor \quad , \tag{18}$$

we can check to see how close $\sum_i N_i^k(t+T)$ is to $N^k(t+T)$ which is the aggregate prediction from equation (13). These would differ because of the way $f(\bullet)$ and $g(\bullet)$ are applied but we would not expect them to be very different. We have not done this, nor have we extended the mathematical formulation into multilevel form. This is possible using the kinds of accounting methods developed by Rees and Wilson (1977) for population modeling but in this context, it does not add anything to the model we have built. It is a direction for further research.

3 The Model Mechanisms

3.1 Locational Probabilities through Bayesian Updating

To introduce the analysis, we will drop the specific land use, location, and temporal notation for the time being, simply referring to a generic probability of land use change ΔN which is influenced by a factor X. We first assume that we have a prior probability of land use change from *k* to *l* for any cell *i* which we call $P(\Delta N)$ but that what we want to estimate is the posterior probability of such change which is influenced by the factor X in question. We call this posterior $P(\Delta N \mid X)$. We begin with the standard form for updating a prior probability to a posterior which incorporates Bayes' rule (Whittle, 1970) and this is

$$P(\Delta N \mid X) = P(\Delta N)\frac{P(X \mid \Delta N)}{P(X)} \quad . \qquad (19)$$

Equation (19) gives the probability when a change in land use takes place – is present, that is $\Delta N_i^{kl} = 1$, but in the case where there is no land use change – is absent, that is $\Delta N_i^{kl} = 0$, the probability must be written as

$$P(\Delta \overline{N} \mid X) = P(\Delta \overline{N})\frac{P(X \mid \Delta \overline{N})}{P(X)} = 1 - P(\Delta N \mid X) \quad . \qquad (20)$$

We have already introduced an important element to the analysis in that we assume the presence or absence of a factor in a cell rather than some continuous value to the factor which varies over all cells. This will enable us to deal with factors which are in binary form – presence or absence – and this is particularly suited to physical infrastructure with socio-economic implications such as the presence or absence of transport routes, utilities, social housing and so on in different cells.

If we compare equations (19) and (20), we can write their product in terms of 'odds': the probability of something happening divided by the probability of it not happening, where we call this odds $O(\bullet)$. Dividing equation (20) into (19) gives

$$\frac{P(\Delta \overline{N} \,|X)}{P(\Delta N \,|X)} = \frac{P(\Delta N)}{P(\Delta \overline{N})} \frac{P(X \,|\Delta N)}{P(X \,|\Delta \overline{N})} \quad , \tag{21}$$

which can be written in short form as

$$O(\Delta N \,|X) = O(\Delta N) \frac{P(X \,|\Delta N)}{P(X \,|\Delta \overline{N})} \quad . \tag{22}$$

The ratio $P(X \,|\Delta N) / P(X \,|\Delta \overline{N})$ is a likelihood which is called the sufficiency ratio which updates the odds of event ΔN taking place in the presence of the factor X, with the ratio being related to support for the event taking place. This equation is best represented in logit form as the 'positive weight of evidence' by taking the logarithms of equation (22) to give

$$\begin{aligned} \text{logit } (\Delta N \,|X) &= \text{logit } (\Delta N) + \log \frac{P(X \,|\Delta N)}{P(X \,|\Delta \overline{N})} \\ &= \text{logit } (\Delta N) + W^{+} \end{aligned} \quad , \tag{23}$$

where W^{+} is the positive weight of evidence associated with X.

An exactly symmetric analysis can be derived for the log odds associated with the absence of a factor $\overline{X}$. In analogy to equation (22), the odds for the absence is

$$O(\Delta N \left| \overline{X} \right.) = O(\Delta N) \frac{P(\overline{X} \left| \Delta N \right.)}{P(\overline{X} \left| \Delta \overline{N} \right.)} \quad , \tag{24}$$

which in logit form becomes

$$\begin{aligned} \text{logit } (\Delta N \left| \overline{X} \right.) &= \text{logit } (\Delta N) + \log \frac{P(\overline{X} \left| \Delta N \right.)}{P(\overline{X} \left| \Delta \overline{N} \right.)} \quad , \\ &= \text{logit } (\Delta N) + W^{-} \end{aligned} \tag{25}$$

where the ratio or likelihood in equation (24) is now called the necessity ratio and in its log form in equation (25) is the 'negative weight of evidence', W^{-}.

We are now in a position to generalize this to many different factors X^{e}. The probability equations that we use will depend strongly on the extent to which the multiple factors $\{ X^{e}, e = 1, 2, ..., E \}$ are independent of one another. This must be tested prior to using these equations, and if there is strong spatial dependence or association between the factors, then more complicated forms must be used with this kind of analysis being less suitable. In fact, independence from irrelevant alternatives is necessary in logit analysis for if the factors are associated with one another, then the probability estimates are biased. Assuming independence, we can write the conditional or posterior probability as $P(\Delta N \left| X^{1}, X^{2}, ..., X^{E} \right.)$. The generalized forms of equations (23) and (25) for positive and negative weights of evidence respectively can now be stated as

$$\left. \begin{aligned} \text{logit } (\Delta N \left| \, X^{1}, X^{2}, ..., X^{E} \right.) &= \text{logit } (\Delta N) + \sum_{e} W_{e}^{+} \\ \text{logit } (\Delta N \left| \, \overline{X}^{1}, \overline{X}^{2}, ..., \overline{X}^{E} \right.) &= \text{logit } (\Delta N) + \sum_{e} W_{e}^{-} \end{aligned} \right\} \quad . \tag{26}$$

We will apply equations (26) to each cell probability for the relevant land use change. Imagine that of the E factors affecting each cell, then some of these are present and some are absent and we now call these $\widetilde{X}_i^e$. We can now state a combined general form for equation (26), notating it with respect to a specific transition *kl* and a particular cell *i*, and assuming the weights W_{ie}^+ and W_{ie}^- associated with presence of absence of the relevant factor are written as $\widetilde{W}_{ie}$. We write this as

$$\text{logit}\,(\Delta N_i^{kl} \mid \widetilde{X}_i^1, \widetilde{X}_i^2, ..., \widetilde{X}_i^E) = \text{logit}\,(\Delta N_i^{kl}) + \sum_e \widetilde{W}_{ie} \quad . \tag{27}$$

Now we convert the log odds in equation (27) back to probability form, first as

$$\frac{P(\Delta N_i^{kl} \mid \widetilde{X}_i^1, \widetilde{X}_i^2, ..., \widetilde{X}_i^E)}{1 - P(\Delta N_i^{kl} \mid \widetilde{X}_i^1, \widetilde{X}_i^2, ..., \widetilde{X}_i^E)} = \left[\frac{P(\Delta N_i^{kl})}{1 - P(\Delta N_i^{kl})}\right] \exp \sum_e \widetilde{W}_{ie} \quad , \tag{28}$$

which simplifies to

$$P(\Delta N_i^{kl} \mid \widetilde{X}_i^1, \widetilde{X}_i^2, ..., \widetilde{X}_i^E) = \frac{\left[\dfrac{P(\Delta N_i^{kl})}{1 - P(\Delta N_i^{kl})}\right] \exp \sum_e \widetilde{W}_{ie}}{\left\{1 + \left[\dfrac{P(\Delta N_i^{kl})}{1 - P(\Delta N_i^{kl})}\right] \exp \sum_e \widetilde{W}_{ie}\right\}} \quad . \tag{29}$$

We finally generalize equation (29) to each cell transition probability specified in the previous section as $p_i^{kl}(t+1)$. This is

$$p_i^{kl}(t+1) = \Psi \left[\frac{P^{kl}(t+1)}{1 - P^{kl}(t+1)}\right] \exp \sum_e \widetilde{W}_{ie} \quad , \tag{30}$$

where $P^{kl}(t)$ is the prior probability of a transition taking place from land use category *k* to category *l* as observed in the entire system between the time period t and $t+1$, computed as $\Delta N^{kl} / N$. Ψ is a normalizing constant which is required to ensure that the probabilities sum to 1. This will clearly be close to the denominator on the RHS of equation (29). We will assess the suitability of the method when we develop and calibrate the model in a later section but to anticipate the outcome there, the factors chosen will be relatively independent of one another, and equation (30) thus appropriate.

3.2 Cellular Operations and Heuristics

It would be possible to simply run the model with the probabilities as specified and to determine land use transitions either using the deterministic or the Monte Carlo technique noted earlier. In fact, early versions of the model developed by Chapin and Weiss (1968) were based on estimating probabilities from a linear regression of land use change against independent variables of the kind used here. These probabilities were then used to drive a simulation based on sampling randomly as in the Monte Carlo method (Chapin, Weiss, and Donnelly, 1965). Here however, there are two features of urban growth which need to be reinforced and which we cannot ensure are directly picked up through the estimation. First, we observe that in the situations for which this model has been applied, particularly in urban growth, land use transitions to given uses take place on the periphery or 'frontier' of large agglomerations of such land uses. Second we observe that free-standing small agglomerations act as the 'seeds of growth' around much larger growing agglomerations and the possibility for such innovation needs to be recognized.

We thus modify the probabilities computed from equation (30) for each land use transition and cell using two procedures which incorporate these two agglomeration features (Soares-Filho,1998). First the model uses an 'expander' heuristic which identifies frontier cells, alters the posterior probabilities accordingly, and then chooses appropriate transitions from this reduced set. Each land use type l has associated with it a set of frontier cells which do not contain land use l, that is cells which have land use type $k \neq l$ These cells are those which are candidates for conversion to use l. This means that a cell with land use l that is entirely surrounded by cells of the same use, does not have a frontier set of cells and this implies that cells on the edge of growing land use types are those which form the frontier. When this reduced set of cells has been identified, the probabilities of transition from land use k to l are weighted according to the amount of land use of type l in the eight cell Moore neighborhood surrounding the cell in question. Then the cell probability transition becomes

$$\hat{p}_i^{kl}(t+1) = \gamma \left[\frac{\sum_{j \in \Omega_i} N_j^l(t)}{8} \right] p_i^{kl}(t+1) \quad , \tag{31}$$

where γ is a normalization factor to ensure the reduced set of probabilities remains within range. The algorithm that implements this expander procedure essentially first examines the entire set of probabilities with the number of frontier cells identified, which is normally much larger than the number required for transition. If this is so, then this set is then reduced by making transitions through random sampling which identifies those probabilities which are the largest in the set.

However the total number of cells chosen by this procedure depends on a second operation which is referred to as the 'patcher'. In this case, we identify those cells which are not of land use type l and we define all cells in their neighborhood which are not of land use type l but are potentially subject to conversion to land use type l. If a cell were of type l, and there were land uses other than type l around it, then this would have generated frontier type cells in the expander operation. Thus, this second procedure examines the complement of the frontier set but does not subject these to the same weighting procedure as that implied by equation (31). When these two operations are combined, it is clear that the set of original posterior probabilities will have been weighted according to the presence and absence of related land uses in their neighborhoods. When the land use transitions are finally chosen, these algorithms are operated in such a way as to ensure that the total number of cells associated with each overall transition given by ΔN^{kl} are met. Contained therein is competition between the expander and patcher operations in terms of what particular cells get chosen, and this is a way of introducing the effects of land use configuration into the probabilities of land use transition.

4 Applications: Calibrations and Simulations

4.1 The Urban Database

We have developed this model for the city of Bauru which is located in West Sao Paulo State, and in 2000 had a rapidly growing population of 309,640. The period for which the model is fitted and for which we have detailed data is from 1979 to 1988 when the population grew from 179,823 to 232,005 implying an annual growth rate of just under 3 percent. Although the simulation experiment in question only refers to a partial span of a longer time series modeling to be developed at a future stage, we are by now able to examine the long term dynamics of the system using the aggregate Markov model introduced earlier. The database that we have constructed identifies $M = 8$ distinct land uses from which we detected 5 observable land use changes from 1979 to 1988. We explain these changes using $E = 12$ independent physical and socio-economic factors which we consider logically determine observed locational change.

The initial data collection using various maps provided by the Bauru local authorities contained severe inconsistencies; illegal settlements are not usually shown while not all of the legally approved settlements are indicated. Some urban zones define areas which are not yet occupied, and there are inaccuracies as to the prevailing land use in some areas. As far as possible these difficulties can be resolved using appropriate satellite imagery which is able to resolve most of the omissions and misclassifications. The initial (1979) and final (1988) land use maps were subjected to a reclassification of zones according to their dominant and effectively existent use; residential zones of different densities were all reclassified to simply residential, and special use and social infrastructure were reclassified to institutional. Eight land use zone categories were thence defined, namely: residential, commercial, industrial, services, institutional, mixed use, leisure/recreation, and the all-embracing non-urban land use. Districts segregated from the main urban agglomeration by more than 10 km were judged outside the simulation area, and the traffic network was not considered to be at a fine enough scale to be represented as a land use. The land use maps for the two time slices are shown in Figures 1(a) and 1(b). All data used in this application was represented at 100m x 100m grid square, pre-processed using the SPRING GIS (from the Division for Image Processing of the Brazilian National Institute for Space Research – DPI-INPE), and was subjected to advanced cross-tabulation operations in IDRISI. The cells form a 487 x 649 grid, there being a total of 316,063 cells defining the region for simulation.

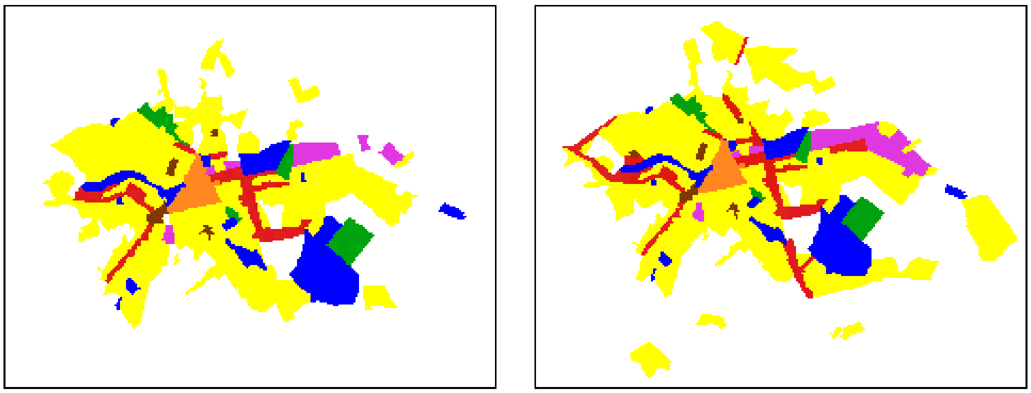


Figure 1: Land Use in Bauru in 1979 (left) and 1988 (right)

The changes between 1979 and 1988 are shown in Figure 2(a) with the most significant land use change – from non-urban to residential – shown in Figure 2(b). With M land uses, there are a possible $M(M-1)$ different transitions and in the case of 8 land uses, of the 56 possible transitions, there are only 5 which are observable. These comprise the critical variables $\{\Delta N_i^{kl}\}$ in the model which are to be explained by the 12 independent social and infrastructural factors, and we define them henceforth using the notation presented in Table 1. As we will show below, at the level of spatial resolution used, all transitions were asymmetric in that conversion from one land use to another implies that conversion the other way does not take place. If the level of resolution were made finer, then this would not hold.

Table 1: Significant Land Use Transitions

Notation	Land Use Transition
NU_RES	*Non-Urban to Residential*
NU_IND	*Non-Urban to Industrial*
NU_SERV	*Non-Urban to Services*
RES_SERV	*Residential to Services*
RES_MIX	*Residential to Mixed Use*

The selection of the variables used to explain the five land use transitions was determined by data availability as well as an appeal to the logic of the development process (Deadman, Brown and Gimblett, 1993; Batty and Xie, 1994). In other words, land use transitions are more likely to be determined by some factors than others in terms of the way developers and consumers of land consider the process of development and acquisition. For example, the location of services centers depends on very different kinds of accessibility to other land uses than say, residential. There is indeed a set of decisive factors for urban land use transitions, in the sense that they substantially account for the main driving forces governing such changes. These factors are suitable for the kinds of analysis we will use here and have effectively guided the modeling experiment at issue. These variables were subjected to a preliminary processing in SPRING which enabled vector editing, polygon identification, definition of distance links, and spatial statistical analysis such as smoothing using kernel density estimators. The 12 variables used in the statistical analysis of land use change to be reported below are listed in Table 2.

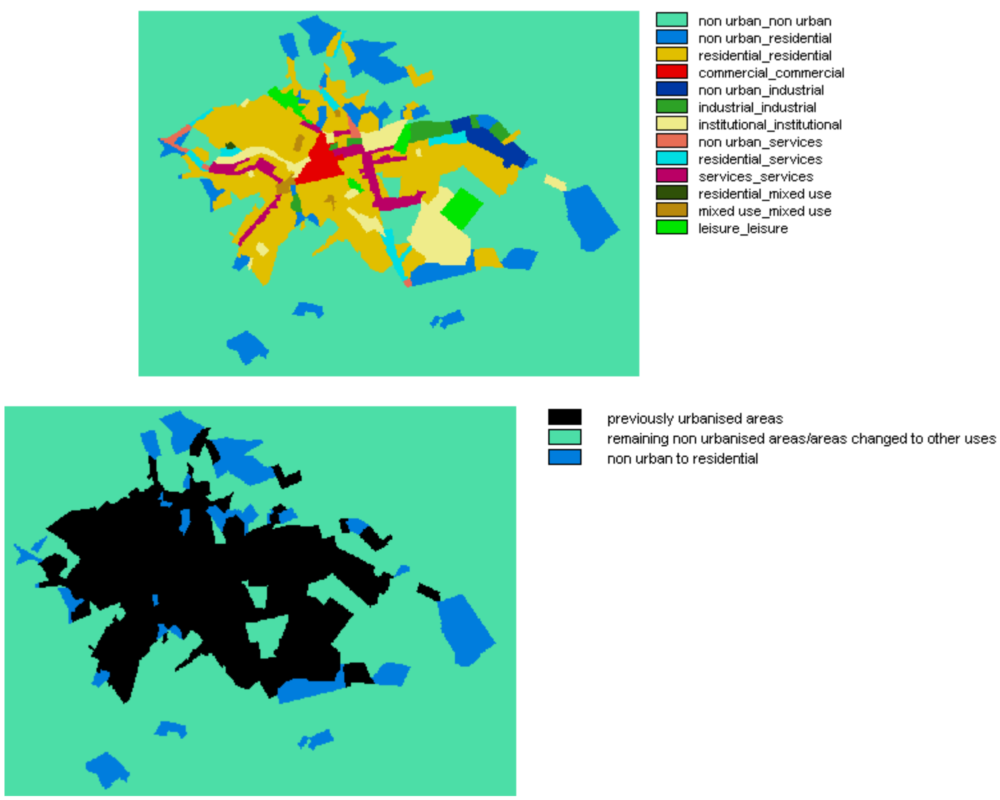


*Figure 2 (a): Land Use Change 1979 to 1988 (b): Non-Urban to Residential Land Use Change (**NU_RES**) 1979 to 1988*

Table 2: Definition of the 12 Independent Land Use Determinants

Notation	Physical or Socio-Economic Land Use Change Determinant/Factor
water	*Area served by water supply.*
mh_dens	*Medium-high density of occupation (25% to 40%).*
soc_hous	*Existence of social housing.*
com_kern	*Distances to different ranges of commercial activities concentration, defined by the Kernel estimator.*
dist_ind	*Distances to industrial zones.*
dist_res	*Distances to residential zones.*
per_res	*Distances to peripheral residential settlements, isolated from the urban concentration.*
dist_inst	*Distances to social infrastructure (institutional use), isolated from the urban concentration.*
exist_rds	*Distances to main existent roads.*
serv_axes	*Distances to the services and industrial axes.*
plan_rds	*Distances to planned roads.*

4.2 Temporal Dynamics

As previously mentioned, eight categories of land use zone were defined from the structures observed at 1979 and 1988 which we state again as residential, commercial, industrial, institutional, services, mixed use, leisure/recreation, and non-urban use. The mixed use land use basically comprises commercial, institutional, and services uses. The leisure/recreation use includes parks, the city zoo, and other public open spaces and green areas. To calculate the land use transition rates, the initial and final land use maps were converted to raster files at a resolution of 100m x 100m, and then exported to the IDRISI where a cross-tabulation operation was made between both land use maps shown in Figures 1(a) and 1(b). This is typical of the way the modeling process has been operated in that we use a variety of different packages and specialist code but with the main focus on GIS style map operations. It would be possible to accomplish this in purely numerical terms but as much of the data is in GIS format, then it has proved convenient to generate transition probabilities for the five types of land use change in the manner specified.

These land use transitions are shown in Table 3 where the asymmetry of the $\mathbf{P}$ matrix is immediately apparent. What is interesting about this matrix is that residential land use is not the state that captures all activity. In fact there is slow but persistent transition from residential into services and into the mixed land use category. In short we do not have the simplest absorbing state model that we illustrated earlier, although were we to define all urban land use as everything other than non-urban, then Table 3 would simplify to the following absorbing state form

$$\mathbf{P} = \begin{bmatrix} 0.9171 & 0.0829 \\ 0 & 1 \end{bmatrix}, \text{ with } \mathbf{P}^{10} \approx \begin{bmatrix} 0.3859 & 0.6141 \\ 0 & 1 \end{bmatrix}.$$

This speed of transition is quite fast although it would take nearly 500 years for 99 percent of the entire system to be converted from non-urban to urban. The disaggregate steady state matrix for Bauru 90 years on is shown in Table 4 and this reveals that there is still considerable non-urban land to be converted but services and residential strongly dominate. If we take this prediction forward to the steady state, then it is clear that services captures most of the activity although industrial is still significant. It is worth noting that the residential use in 500 years time would not be excluded from the town, but would simply lie beyond our 487 x 649 cell system. In fact the steady state is so distant in time (beyond 2000 years hence) that it is of no significant interest apart from the structural trend that this reveals. The system graph which is implicit in Tables 3 and 4 is also of interest in showing direct and indirect connectivities in transition.

Table 3 – The **P** *Matrix – Land Use Transition Rates for Bauru, 1979-1988*

Land Use	***NonU***	***Res***	***Comm***	***Indust***	***Inst***	***Serv***	***Mixed***	***Leis/ Rec***
NonU	**0.9171**	**0.0698**	0	**0.0095**	0	**0.0036**	0	0
Res	0	**0.9380**	0	0	0	**0.0597**	**0.0023**	0
Comm	0	0	1	0	0	0	0	0
Indust	0	0	0	1	0	0	0	0
Inst	0	0	0	0	1	0	0	0
Serv	0	0	0	0	0	1	0	0
Mixed	0	0	0	0	0	0	1	0
Leis/Rec	0	0	0	0	0	0	0	1

Table 4 – The Predicted $\mathbf{P}^{10}$ *Matrix for Bauru, 2069-2078*

Land Use	***NonU***	***Res***	***Comm***	***Indust***	***Inst***	***Serv***	***Mixed***	***Leis/ Rec***
NonU	**0.3859**	**0.3626**	0	**0.0703**	0	**0.1752**	**0.0057**	0
Res	0	**0.4945**	0	0	0	**0.4866**	**0.0187**	0
Comm	0	0	1	0	0	0	0	0
Indust	0	0	0	1	0	0	0	0
Inst	0	0	0	0	1	0	0	0
Serv	0	0	0	0	0	1	0	0
Mixed	0	0	0	0	0	0	1	0
Leis/Rec	0	0	0	0	0	0	0	1

4.3 Factors Determining Land Use Transitions: Testing for Independence

The first requirement for selecting factors which can be used in computing the locational probabilities is some measure of determining whether or not they are independent of one another for the particular land use changes in question. To do this with data which indicates presence p and absence a of any variable X_i^e in any cell or zone i, we need to define some measures of association or dependence. Two are used here: first a statistic due to Cramer (see Bonham-Carter, 1994) which is based on the chi-square (χ^2) and second, a statistic based on the joint uncertainty between any two distributions which is computed from entropies as used in information theory (Batty, 1976). Cramer's statistic is derived as follows. First we note that our variables X_i^e can be considered as showing presence p or absence a of some characteristic in a cell i and the two sets that these cells fall into are called Ω_p and Ω_a respectively, noting that each map is the union of these two sets $\Omega_p \cap \Omega_a = \Omega$. The numbers of cells in each of these sets can be defined as

$$X_p^e = \sum_{i \in \Omega_p} X_i^e , \qquad X_a^e = \sum_{i \in \Omega_a} X_i^e , \quad \text{where} \quad X_p^e + X_a^e = N \quad . \qquad (32)$$

Now the associations to be tested must be between two different distributions or maps X_i^r and X_i^s and we thus need to define the joint presence and absence of the map characteristics by the indices $u, v = 1$ (presence) and $u, v = 2$ (absence). We thus define any of the four combinations of presence and absence in these two distributions by the joint count

$$X_{uv}^{rs} = X_u^r \cap X_v^s , \qquad \sum_u \sum_v X_{uv}^{rs} = N \quad . \tag{33}$$

We first form the chi-square and this requires us to compare X_{uv}^{rs} with an expected value $(X_{uv}^{rs})'$ formed when the two map distributions are entirely independent. This expected distribution is defined as

$$(X_{uv}^{rs})' = \frac{\sum_v X_{uv}^{rs} \sum_u X_{uv}^{rs}}{N} \quad , \tag{34}$$

where the chi-square χ_{rs}^2 is computed as

$$\chi_{rs}^2 = \sum_u \sum_v \frac{[X_{uv}^{rs} - (X_{uv}^{rs})']^2}{(X_{uv}^{rs})'} \quad . \tag{35}$$

Cramer's statistic or coefficient is simply a normalized version of the chi-square and is defined as

$$V_{rs} = \sqrt{\chi_{rs}^2 / [N(N-1)]} \quad . \tag{36}$$

This statistic has a minimum value of 0 when the two distributions are completely independent of one another and a maximum value less than 1 depending upon the two map distributions.

The second statistic is based on joint entropies where a comparison is made between the actual entropy of the probability distribution formed from X_{uv}^{rs} and the marginal distributions based on variation across either the first or second variable. These are defined as

$$p_{uv}^{rs} = X_{uv}^{rs} \Big/ N, \qquad p_u^{rs} = \sum_v p_{uv}^{rs}, \quad \text{and} \quad p_v^{rs} = \sum_u p_{uv}^{rs} \qquad . \tag{37}$$

Relevant entropies are stated in the conventional manner as

$$H_{rs} = -\sum_u \sum_v p_{uv}^{rs} \log p_{uv}^{rs}, \; H_u^{rs} = -\sum_u p_u^{rs} \log p_u^{rs}, \; H_v^{rs} = -\sum_v p_v^{rs} \log p_v^{rs} \; . \tag{38}$$

The uncertainty is now defined as

$$U_{rs} = 2 \left\{ \frac{H_u^{rs} + H_v^{rs} - H_{uv}^{rs}}{H_u^{rs} + H_v^{rs}} \right\} \qquad , \tag{39}$$

and this statistic likewise has a minimum value of zero when the distributions are independent and a maximum value of 1 when the distributions are identical. The joint information uncertainty tends to be more robust than the Cramer's statistic, for the former works with percentage values for the overlapping areas between pairs of factors under analysis, whereas the latter uses absolute areas values.

We now need to introduce the twelve variables which are to be compared against the five land use transitions and first consider which of these variables is relevant to particular transitions. From our general understanding of the development process, we consider that there are limits on what factors should logically influence different types of land use change, and without going into specific details at this stage, we state these decisions in Table 5.

Table 5: Selection of Factors Determining Land Use Change

$\{\Delta N^{kl}\}$ $\{X_i^e\}$	***NU_RES***	***NU_IND***	***NU_SERV***	***RES_SERV***	***RES_MIX***
water				•	
mh_dens					•
soc_hous					•
com_kern	•		•		
dist_ind		•			
dist_res			•		
per_res	•				
dist_inst	•				
exist_rds	•				
serv_axes		•	•	•	
plan_rds					•
per_rds	•				•

From this table, we are able to see that there are 16 pairwise comparisons of variables to be made. You can read this table as follows. For changes from non-urban to residential use for example, 5 of the factors defined earlier in Table 2 as ***com_kern***, ***per_res***, ***dist_inst***, ***exist_rds***, and ***per_rds*** are relevant and thus there are 10 comparisons to be made – [5(5-1)/2] – to test for their independence from one another. The same kind of comparisons must be made with respect to factors affecting the four other land use changes, and in total, given that several variables are used to explain more than one transition, 21 such comparisons must be made. For each of these, we have computed Cramer's statistic from equation (36) and the joint information uncertainty from equation (39). We show these results in Table 6.

The criterion which is used to determine whether one factor is independent of another is to a large extent arbitrary as there is no large body of case results associated with the application of these methods. Where this particular variant of logit modeling has been used in the geosciences, Bonham-Carter (1994) reports that values less than 0.5 for Cramer's Coefficient and the Joint Information Uncertainty suggest less association rather than more. In all comparisons made here, these associations are less than this threshold. Indeed all values are less than 0.45 for V_{rs} and less than 0.35 for U_{rs}. As none of the association values surpassed these thresholds, no variables preliminary selected for the modeling experiment have been discarded from the analysis. In practice, the preliminary selection of these variables was also based on their visual correspondence when superimposed on the final land use map so that the meaning of these associations might be established. In other words, the fact that we associate variables which seem logically related and the fact that these have to be below the externally imposed thresholds for significant association also has to be tempered against the visual logic of their association. We can visualize these interactions for all cross-tabulations of the independent variable against land use change but we do so for only one of these – ***RES_MIX*** – which we show in Figure 3. If these factors still appear to have little connection to real land use change, then they must be discarded. This is a process that we used in the original selection of the variables and the land use transitions which we consider they determine.

Table 6: Associations Between the Independent Variables

Factor $\{X_i^r\}$	**Factor** $\{X_i^s\}$	**Cramer's Statistic** V_{rs}	**Uncertainty** U_{rs}
water	***serv_axes***	0.3257	0.0767
mh_dens	***soc_hous***	0.0460	0.0017
	plan_rds	0.2617	0.0701
	per_rds	0.0201	0.0003
soc_hous	***plan_rds***	0.1174	0.0188
	per_rds	0.0480	0.0047
com_kern	***dist_res***	0.4129	0.3447
	per_res	0.1142	0.0310
	dist_inst	0.1218	0.0520
	exist_rds	0.2685	0.1499
	serv_axes	0.2029	0.1099
	per_rds	0.0434	0.0064
dist_ind	***serv_axes***	0.1466	0.0477
dist_res	***serv_axes***	0.2142	0.1002
per_res	***dist_inst***	0.1487	0.0559
	exist_rds	0.0592	0.0078
	per_rds	0.1733	0.0553
dist_inst	***exist_rds***	0.0601	0.0108
	per_rds	0.0765	0.0238
exist_rds	***per_rds***	0.0239	0.0019
plan_rds	***per_rds***	0.0247	0.0029

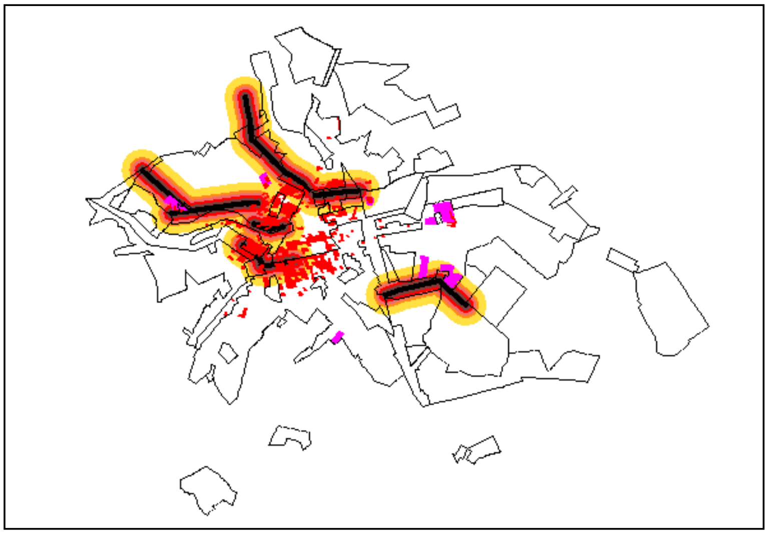


*Figure 3: Spatial Independence of Factors Determining the Transition from Residential to Mixed Use **(RES_MIX)***

The buffer bands are distance to planned roads (***plan_rds***), the darker diffused spots are areas of medium-high density of occupation (***mh_dens***), and the darker polygons correspond to social housing (***soc_hous***)

4.4 Estimating Locational Probabilities from Weights of Evidence

As previously presented, the 'weights of evidence' method, employed in the calculation of the cell transition probabilities, is based on 'Bayes rule of conditional probability'. The weights may act positively if a factor is present in the cell or negatively if the factor is not present. However of the 12 factors defined in Table 2, only 3 of these are in strict binary form. The other 9 are related to distances from various land uses and as such, there is always a value of distance for any cell. To code these factors then, we have divided them into distance bands which cover their spatial extent and as every cell is in a distance band, then it is not necessary to explicitly compute the evidence that a factor is not in any other band. In a general way, in cases where there is missing data, then the likelihood is set equal to 1 or the weight equal to 0. The evidences as estimated from the data are presented in Table 7. Note that positive evidence does not imply a positive coefficient or vice versa in these computations.

Table 7: The Weights of Evidence

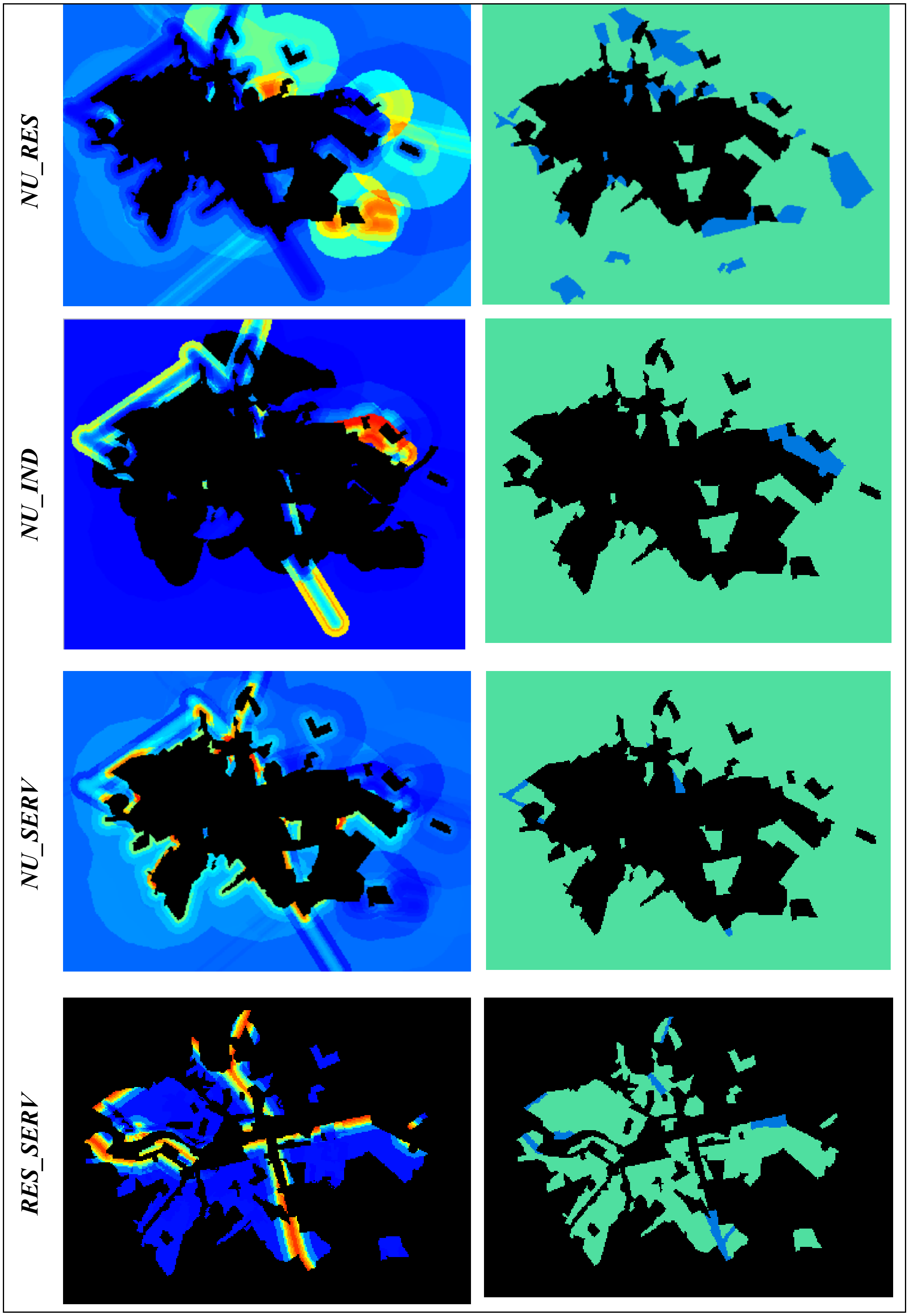
Land Use Transition	Factor $\{X_i^e\}$	Positive Weights of Evidence W_{ie}^{+}						
		1	*2*	*3*	*4*	*5*	*6*	*7*
NU_RES	***com_kern***[1]	3.749	2.106	1.864	0.491	-0.323	0	*na*
	per_res[3]	1.968	1.615	1.392	0.892	-0.626	-0.469	*na*
	dist_inst[4]	0.003	0.600	1.254	0.727	-0.359	-0.089	*na*
	exist_rds[5]	0.231	0.320	0.353	0.510	0.443	0.196	-0.085
	per_rds[6]	2.377	2.269	2.068	1.984	1.444	0.857	-0.127
NU_IND	***dist_ind***[2]	3.862	4.016	3.792	3.452	1.763	0	0
	serv_axes[5]	2.722	2.799	2.676	2.625	2.525	1.727	-3.832
NU_SERV	***com_kern***[1]	3.412	4.469	2.912	0.878	0	0	*na*
	dist_res[3]	2.144	1.523	0.621	-0.065	0	0	*na*
	serv_axes[5]	3.508	3.321	2.917	1.869	0.450	0	0
RES_SERV	***water***	*Presence* -0.6611				*Absence* 0.2883		
	serv_axes[5]	2.780	1.948	1.461	0.888	-0.297	-1.412	-3.284
RES_MIX	***mh_dens***	*Presence* 0.6452				*Absence* -0.0635		
	soc_hous	*Presence* 2.4678				*Absence* -0.3214		
	plan_rds[5]	3.506	1.863	0	0	0	0	0
	per_rds[6]	1.775	1.652	1.848	0.903	0	0	0

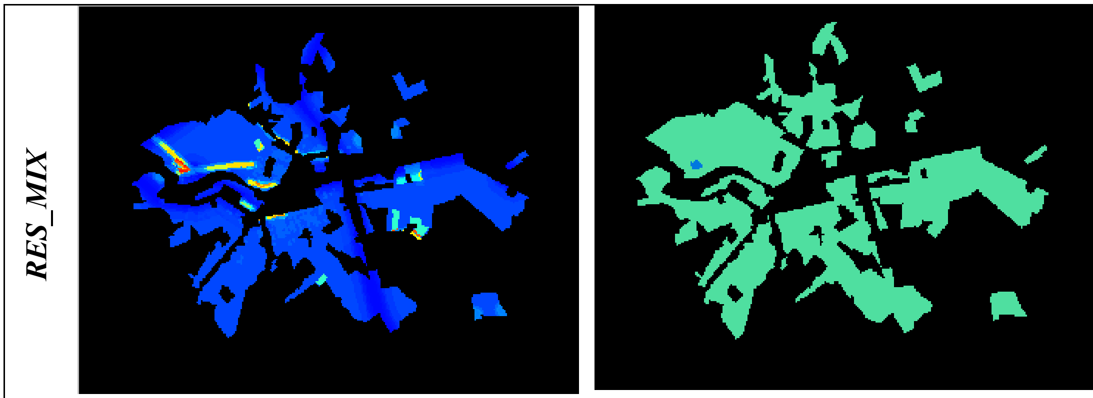
Note: Distance Bands in meters

[1] 1: 0 -500; 2: 500-1000; 3: 1000-1500; 4: 1500-10000; 5: 10000-30000; 6: > 30000
[2] 1: 0 -500; 2: 500-1000; 3: 1000-1500; 4: 1500-2000; 5: 2000-5000; 6: 5000-10000; 7: >10000
[3] 1: 0 -500; 2: 500-1000; 3: 1000-2000; 4: 2000-5000; 5: 5000-10000; 6: > 10000
[4] 1: 0 -500; 2: 500-1000; 3: 1000-3000; 4: 3000-8000; 5: 8000-15000; 6: > 15000
[5] 1: 0 -250; 2: 250-500; 3: 500-750; 4: 750-1000; 5: 1000-1250; 6: 1250-2000; 7: > 2000
[6] 1: 0 -250; 2: 250-500; 3: 500-750; 4: 750-1000; 5: 1000-1500; 6: 1500-2500; 7: > 2500

The weights to compute transition probabilities for each cell with respect to the five land use transitions are based on the values in Table 7. These are used in equation (30) with the prior probabilities set in proportion to the observed transition aggregated across all zones. The computed probabilities are shown in Figure 4 where these are compared with the actual land use transitions. These images are taken from ER Mapper, thus illustrating once again the range of software that is used to compute the various elements of this model. The range is from high probability of transition to low on a gray scale from light to dark. There are many points to be made about the definition of the variables used as evidence determining land use change. The majority of factors involve distance or accessibility and it is possible to examine the way the weight of evidence varies across the different distance bands. The trend lines produced in scatter plots relating factors and subcategories of factors (distances ranges in maps) with their respective positive weights of evidence vary in their regularity and produce mild support for including in the modeling analysis those factors whose plots present a good fit (Almeida et al., 2002). Although we have included all the prior factors due to their comparative independence from one another, the final choice towards the inclusion or exclusion of a given evidence must always rely upon a broader judgement as to the environmental importance of the evidence and its coherence concerning the land use transition being modeled (Soares-Filho, 1998).

In terms of the five types of transition, the transition from non-urban to residential (***NU_RES***) appears to largely depend on the greater proximity of these areas to commercial activity clusters, on their general accessibility conditions, and on the previous existence of residential settlements in their surroundings, for this ensures the possibility of extending existing nearby infrastructure, if any. For non-urban areas to industrial use (***NU_IND***), there are two driving forces: the nearness of such areas to the existing industrial use – linkages between industries – and the availability of road access. For non-urban to services use (***NU_SERV***), three major factors are crucial: the proximity of these areas to clusters of commercial activities, their closeness to areas of residential use, and last but not least, their strategic location in relation to the N-S / E-W services axes of Bauru. The first factor accounts for the suppliers' market (and in some cases also the consumers' market) for services; the second factor represents the consumers' market itself; and the third and last factor corresponds to the accessibility for both markets related to the services use. The transition from 'residential to services use' (***RES_SERV***) involves the location of services into previously consolidated urban areas. This is determined in relation to the N-S / E-W services axes of Bauru, as well as the water supply. Finally, the last land use transition is the shift from residential to mixed use (***RES_MIX***). The mixed use zones cover the consolidation of secondary commercial centers, which also attract services and social infrastructure. New mixed use zones emerge in the more peripheral areas and are determined by the nearness to planned or peripheral roads, the existence of medium-high density of occupation, and the presence or proximity of social housing settlements, since these two latter factors imply the existence of a greater occupational gathering, and hence, economic sustainability for such zones.


NU_RES
NU_IND
NU_SERV
RES_SERV



Locational Transition Probabilities from Weights of Evidence | *Significant Land Use Transition 1979-1988*

Figure 4: Estimated Transition Probability Surfaces & Land Use Change 1979-1988

4.5 The Goodness of Fit: Alternative Simulations

We are now in a position to report the results of calibrating the full model. From the probabilities $\{ p_i^{kl} \}$ which are computed for each cell from the weights of evidence method in equation (30), the cellular heuristics which determine agglomeration and polynucleation are applied following the techniques associated with equation (31). These transform the probabilities into a new set $\{ \hat{p}_i^{kl} \}$ which in turn are subjected to continued iteration using the expander and patcher algorithms. At each stage of this iteration the probabilities are selected so that the total observed change for each land use ΔN^{kl} is approached although because this is a matter of trial and error, it takes five iterations of these algorithms to achieve the necessary balance. At each stage, the computed probabilities $\{ \hat{p}_i^{kl} \}$ are used to determine whether a transition takes place; that is, random numbers are drawn over a fixed range associated with the probabilities and if the number is less than the appropriately scaled probability a transition from *k* to *l* takes place. Formally

$$\text{if} \quad rand(\mathbf{num}) < \phi \; \hat{p}_i^{kl} \quad \text{then} \quad \Delta N_i^{kl} = 1, \quad \text{otherwise} \quad \Delta N_i^{kl} = 0 \quad , \; (40)$$

where ϕ is a scaling constant. The process implied by equation (40) does not ensure that the total number of transitions observed between time t and $t+1$ takes place and this necessitates the iteration. In this way, the proportion of cell probabilities modified by the expander and patcher operations is not known in advance. In the simulation reported here, for example, the relative balance of these for the five transitions is quite different; for ***NU_RES***, the proportion of cells modified by the expander was 0.65 and by the patcher 0.35; for ***NU_IND***, all the cells were modified by the expander; for ***NU_SERV***, the ratio was 0.5 to 0.5; for ***RES_SERV***, 0.1 to 0.9; and finally for ***RES_MIX***, all the cells were modified by the patcher.

Due to the randomness of allocation in the DINAMICA transition algorithms, even though the same DINAMICA parameter set drives the operation of the cell heuristics each time the model is run, distinctly different simulations results will be produced. Three such simulations which differ but are the best of many that we have run, are shown in Figure 5. The patcher algorithm proved to be greatly suited to simulating residential settlements disconnected from the main urban agglomeration. Nevertheless, the shapes of these settlements do not strictly coincide with those observed in the reality of 1988. This is partly because the mechanics of property subdivision and the kind of geometric constraints that govern development are not accounted for in the model. As Figure 5 implies, the services corridors were modeled very accurately in all simulations, while the zones of industrial use were also considerably well detected in all of the three simulation results. Shifts from non-urban areas to residential use represented the most challenging category of land use transition in these modeling experiments. The reasons for the difficulties in detecting their shapes have just been noted but it is worth remarking that 65% of these types of transition occur through the expander algorithm. An evident shortcoming of this algorithm which is being tackled at present by the CSR-UFMG team lies in the fact that after the random selection of a seed cell for transition, all neighboring cells are subject to transition and this is too blunt an instrument to accurately mirror the prioritization of development in real situations.

We can formally measure the fit of the model using a variety of correlation and chi-squared-like statistics but a particularly convenient form we adopt here is due to Constanza (1989) which involves examining the distribution of land uses at different levels of resolution. This also incorporates a moving-window filter and in this way ensures that spatial variation at all scales influences the ultimate evaluation of the goodness of fit. We define the fit F_w for a window of size $w \times w$ as

$$F_w = \left\{ \sum_{j=1}^{W} \left[1 - \left(\sum_{k=1}^{M} \left| \sum_{i \in w_j} \hat{N}_i^k(t+1) - \sum_{i \in w_j} N_i^k(t+1) \right| 2^{-1} w^{-2} \right) \right] \right\} \Big/ W \quad , \quad (41)$$

where W is the total number of windows of size $w \times w$ sampled in a scene, w_j is a window of size $w \times w$ anchored at cell j, $\hat{N}_i^k$ is the land use k of cell i in the simulated image at time $t+1$, and N_i^k is the land use k of cell i in the real image at time $t+1$. The summation over j for all windows is usually equivalent to the summation over all cells for if each cell is used to anchor a window, then $W = N$. To produce the total goodness of fit over all windows, then we simply sum the appropriate equation and take an average to get F. This can be weighted in the following way:

$$F_t = \frac{\sum_w F_w \exp[-\vartheta(w-1)]}{\sum_w \exp[-\vartheta(w-1)]} \quad , \quad (42)$$

where the summation of w is taken as an index over the different window sizes used and where ϑ is a tunable parameter.

The multiple resolution method was implemented for sampling window sizes of 3x3, 5x5 and 10x10 cells. The range of these fit statistics is in fact from 0 (no fit) to 1 (perfect fit) and for the three simulations shown in Figure 5, the fits were computed as 0.903, 0.986, and 0.901 respectively. These are particularly encouraging results and suggest that a very high proportion of the spatial variance in land use transition can be simulated by this model.

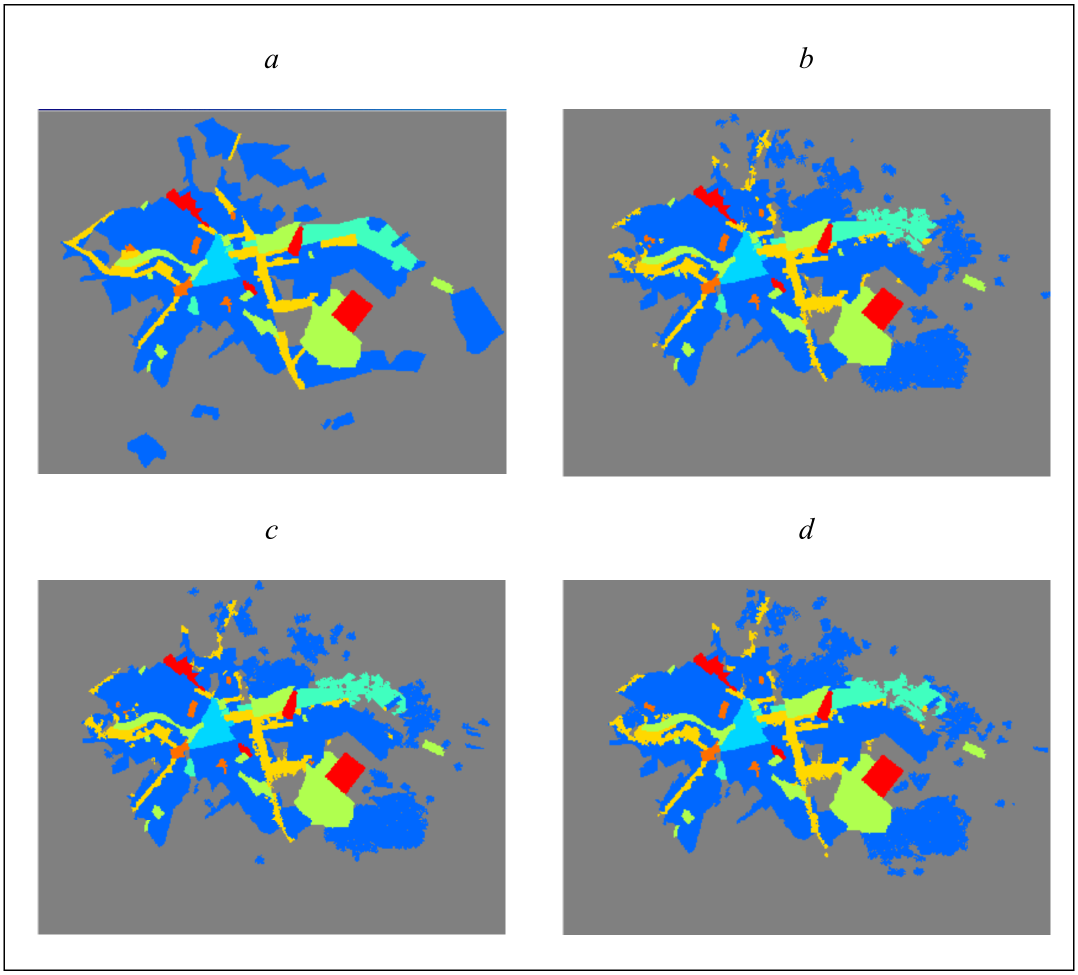


Figure 5: Actual Land Use in 1988 (a) compared to the Three Best Simulations (b), (c) and (d)

Leisure and recreation (light grey), institutional (dark grey), and the central commercial zone (mid grey) did not incur any transitions during the observed time period. The new mixed land use zone that emerged in the north-western part of the city was accurately modeled particularly in the first and third simulations.

5 Conclusions: Potential Developments and Next Steps

Models of land use transition based on the GIS-map overlay paradigm which represents data in raster form inevitably appeal to ideas from cellular automata. However as we have argued, strict CA models are only appropriate as a pedagogic perspective on such transitions and as soon as an effort is made to calibrate such models to data, the CA paradigm becomes less significant. In this paper, we have presented a model which operates in a CA environment more akin to a cell-space (CS) than strict CA approach and although the model has been represented formally in the manner of CA, the main emphasis is on the way the transition probabilities are estimated. What we have done is to introduce an approach more widely used in ecology and the geosciences rather than urban modeling which builds a robust and parsimonious structure from the ground up. In fact the 'weights of evidence' approach provides a particularly simple and useful way of illustrating how less conventional map data in raster and in binary form can be used in a multivariate framework, thus linking CA and raster GIS to more conventional and well established methods of statistical estimation. We consider this approach to be one of the most promising ways of calibrating such raster models to data. What is particularly appealing about the model is the way in which factors which are directly considered by local municipalities and developers who have the greatest control over development are used as drivers of the growth process.

As we have implied above, we are extending the model in several directions which will improve its applicability and performance. Our group at DPI-INPE is currently committed to the development of a 2D and 3D land use CA simulation module which is to be integrated with the SPRING Geographical Information System developed at the Institute. This module shall be conceived as a flexible integrated multi-scale and multi-purpose device, which encompasses both deterministic and stochastic transition algorithms. In a more practical context, we intend to explore the solution space of the model in the manner indicated in the presentation of the three alternative simulations in Figure 5 as we consider that running the model under different conditions and different scenarios is the best way of beginning to understand its potential and its limits.

From the perspective of GIS, there have been many pleas for extending the functionality of standard systems to embrace the kinds of model introduced here. This has been partly achieved in some packages such as IDRISI but in general, spatial models remain largely disconnected from such systems. In fact, dynamic modeling within GIS constitutes an even greater and more immediate challenge. According to Burrough (1998), methods of open systems modeling of which CA is one of the best examples and which meet many of the general requirements for simulating dynamic processes quickly and efficiently, are rarely implemented in GIS. As a result, GIS remains surprisingly narrowly focused (Openshaw, 2000). All these opinions find support in our work and that of our colleagues (Câmara et. al., 2001), for whom the current computational paradigms of knowledge representation are essentially static and unable to appropriately model the temporal dimension and the dynamic context-based relationships amongst entities and their attributes.

Finally we prefer the integration between models and GIS being through a loose or tight coupling, rather than through embedding models directly within the GIS (Bivand and Lucas, 2000). This is largely because so much effort has been put into proprietary GIS to date that the connections now exist for making such couplings in an effective and efficient manner. Our application here is an example par excellence of the use of many kinds of software where the actual model is programmed in conventional terms once the inputs and estimation have been organized through separate statistical and GIS software. In fact, even the visualization of the inputs and outputs from our models is accomplished using different GIS packages ranging from our own SPRING to packages such as IDRISI and ER Mapper. Nevertheless, the need for integration remains strong. Parks (1993) gives three sound reasons: firstly, spatial representation is critical to environmental problem solving, but GIS usually lack the predictive and analytic capabilities necessary for representing and analyzing complex problems. Secondly, modeling tools lack sufficiently flexible GIS-like spatial analytical components and are often inaccessible to the non-specialist. Finally, modeling and GIS can both be made more robust through their linkage and co-evolution. It is this that we feel represents the long term relevance of this project and we consider this is best developed when models and GIS are applied to problems of spatial-temporal change.

6 References

Albin, P. S. (1975) *The Analysis of Complex Socio-Economic Systems*, Lexington, MA: D. C. Heath.

Almeida, C. M., Monteiro, A. M. V., Câmara, G., Soares-Filho, B. S., Cerqueira, G. C. and Pennachin, C. L. (2002) Modelling Urban Land Use Dynamics through

Bayesian Probabilistic Methods in a Cellular Automaton Environment, *Proceedings of the 29th International Symposium on Remote Sensing of the Environment*, Buenos Aires, Argentina, 8th-12th April.

Batty, M. (1976) Entropy in Spatial Aggregation, *Geographical Analysis*, 8, 1-21.

Batty, M. (2000) Geocomputation using Cellular Automata, in Openshaw, S. and Abrahart, R. J. (Editors) *Geocomputation*, London: Taylor & Francis, 95-126.

Batty, M. and Xie Y. (1994) From Cells to Cities, *Environment and Planning B*, 21, 31-48.

Bivand, R. and Lucas, A. (2000) Integrating Models and Geographical Information Systems, in Openshaw, S. and Abrahart, R. J. (Editors) *Geocomputation*, London: Taylor & Francis, 331-364.

Bonham-Carter, G. F. (1994) *Geographic Information Systems for Geoscientists: Modelling with GIS*, New York: Pergamon.

Burrough, P. A. (1998) Dynamic Modelling and Geocomputation, in Longley, P. A., Brooks, S. M., McDonnell, R. and MacMillan, B. (Editors) *Geocomputation: A Primer*, Chichester, England: John Wiley & Sons, 165-192.

Câmara, G., Monteiro, A. M. V., Medeiros, J. S. (2001) *Representações Computacionais do Espaço: Um Diálogo entre a Geografia e a Ciência da Geoinformação*, http://www.dpi.inpe.br/geopro/trabalhos/epistemologia.pdf 29/01/02, [online].

Chapin, S. F., Weiss, S. F., and Donnelly, T. G. (1965) *Some Input Refinements to a Residential Model*, Institute for Research in Social Science, University of North Carolina, Chapel Hill, NC.

Chapin, S. F., and Weiss, S. F. (1968) A Probabilistic Model for Residential Growth, *Transportation Research*, 2, 375-390.

Clarke, K. C., Hoppen, S. and Gaydos, L. (1997) A Self-Modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area, *Environment and Planning B*, 24, 247-261.

Clarke, K. C., and Gaydos, L. J. (1998) Loose-Coupling a Cellular Automaton Model and GIS: Long-Term Urban Growth Predictions for San Francisco and Baltimore, *International Journal of Geographic Information Science*, 12, 699-714.

Constanza, R. (1989) Model Goodness of Fit: A Multiple Resolution Procedure, *Ecological Modelling*, 47, 199-215.

Couclelis, H. (1985) Cellular Worlds: A Framework for Modeling Micro-Macro Dynamics, *Environment and Planning A*, 17, 585-596.

Couclelis, H. (1989) Macrostructure and Microbehavior in a Metropolitan Area, *Environment and Planning B*, 16, 141-154.

Deadman, P., Brown, R. D. and Gimblett, P. (1993) Modelling Rural Residential Settlement Patterns with Cellular Automata, *Journal of Environmental Management*, *37*, 147-160.

Hobbs, R. J. (1983) Markov Models in the Study of Post-Fire Succession in Heathland Communities, *Vegetation*, 56, 17-30.

JRC (1994) *Modelling Deforestation Processes – A Review*, JRC (Joint Research Centre, European Commission/Institute for Remote Sensing Applications) and ESA (European Space Agency/ESRIN – Earthnet Programme Office), Trees Series B, Research Report No.1, Luxembourg: ECSC-EC-EAEC.

Openshaw, S. (2000) GeoComputation, in Openshaw, S. and Abrahart, R. J. (Editors) *Geocomputation*, London: Taylor & Francis, 1-32.

Papini, L. Rabino, G. A., Colonna, A., Di Stefano, V., and Lombardo, S. (1998) Learning Cellular Automata in a Real World: The Case Study of the Rome Metropolitan Area, in S. Bandini, R. Serra, and F. Suggi Liverani (Editors) *Cellular Automata: Research Towards Industry: ACRI'96: Proceedings of the Third Conference on Cellular Automata for Research and Industry*, Springer-Verlag, London, 165-183.

Parks, B. O. (1993) The Need for Integration, in Goodchild, M. J., Parks, B. O. and Steyaert, L. T. (Editors) *Environmental Modelling with GIS*, Oxford: Oxford University Press, 31-34.

Rees, P. H. and Wilson, A. G. (1977) *Spatial Population Analysis*, London: Edward Arnold.

Schock, S. (Editor)(2000) *Projecting Land Use Change: A Summary of Models for Assessing the Effects of Growth and Change on Land Use Patterns*, EPA/600/R-00/098, National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Washington DC.

Soares-Filho, B. S. (1998) *Modelagem da Dinâmica de Paisagem de uma Região de Fronteira de Colonização Amazônica*, unpublished PhD dissertation, EPUSP, Universidade de São Paulo, São Paulo, Brazil

Soares-Filho, B. S., Cerqueira, G. C. and Pennachin, C. L. (2002) DINAMICA: A New Model to Simulate and Study Landscape Dynamics, *Ecological Modelling*, in press.

Tobler, W. R. (1970) A Computer Movie Simulating Population Growth in the Detroit Region, *Economic Geography*, 42, 234-240.

Tobler, W. R. (1979) Cellular Geography, in S. Gale and G. Olsson (Editors) *Philosophy in Geography*, D. Reidel, Dordrecht, The Netherlands, pp. 279-386.

Wegener, M. (1994) Operational Urban Models: State of the Art, *Journal of the American Planning Association*, 60, 17-30.

White, R. W., and Engelen, G. (1993) Cellular Automata and Fractal Urban Form: A Cellular Modelling Approach to the Evolution of Urban Land Use Patterns, *Environment and Planning A*, 25, 1175-1199.

White, R. W. and Engelen, G. (1997) Cellular Automaton as the Basis of Integrated Dynamic Regional Modelling, *Environment and Planning B*, 24, 235-246.

White, R. W., Engelen, G. and Uljee I. (1998) Vulnerability Assessment of Low-Lying Coastal Areas and Small Islands to Climate Change and Sea Level Rise – Phase 2: Case Study St. Lucia, Report to the United Nations Environment Programme, Caribbean Regional Co-ordinating Unit, Kingston, Jamaica: RIKS Publication.

Whittle, P. (1970) *Probability*, Harmondsworth, Middlesex, UK: Penguin Books.

Wu, F. (1998) Simland: A Prototype to Simulate Land Conversion through the Integrated GIS and CA with AHP-Derived Transition Rules, *International Journal of Geographic Information Science*, 12, 63-82.

Xia, L. and Yeh, A. G. (2000) Modelling Sustainable Urban Development by the Integration of Constrained Cellular Automata and GIS, *International Journal of Geographic Information Science*, 14, 131-152.