Statistical Validation of Spatial Patterns in Agent-Based Models

William Rand, Daniel G. Brown, Scott E. Page, Rick Riolo, Luis E. Fernandez, and Moira Zellner [a]

[a]Center for the Study of Complex Systems
University of Michigan
Ann Arbor, MI, 48109-1120, USA
E-mail: sluce@umich.edu

KEYWORDS
land use, land change, urban sprawl, validation, spatial patterns, agent-based modeling, power laws

ABSTRACT

We present and evaluate an agent-based model (ABM) of land use change at the rural-urban fringe. This paper is part of a project that links the ABM to surveys of residential preferences and historical patterns of development. Validation is an important issue for such models and we discuss the use of distributional phenomena as a method of validation. We then highlight the ability of our ABM to generate two phenomena evident in empirical analysis of urban development patterns: a power law relationship between frequency and cluster size and a negative exponential relationship between density and distance from city center. We discuss these results in the light of validation of ABMs.

INTRODUCTION

Development at the urban-rural fringe has been linked to a variety of negative ecosystem impacts, including habitat and migration corridor destruction (Johnson, 2001). Understanding and predicting development patterns and learning how to construct policies that limit ecological damage requires knowledge of the processes that drive development.

One approach to modeling urban development is to borrow formalisms originally created to model physical processes that generate spatial patterns and dynamics which are similar to those seen in urban systems. To do this, the detailed mechanisms of social processes are mapped onto and interpreted as forces like those found in physical systems. In some models the analogies are very abstract, e.g., in diffusion limited aggregation (DLA) and correlated percolation models of urban growth (Makse et al., 1998), while in other cases there is an attempt to apply formalisms (e.g., Markovian Random Field Theory) in a way that allows a more direct mapping from model mechanisms to social processes (Andersson et al., 2002).

An alternative approach is to construct agent based models (ABMs) in which agents represent social actors (developers, residents, industries, etc.). An ABM approach provides a number of opportunities and challenges which sometimes complement and sometimes extend those available with other approaches. For example, including different agent types helps separate distinct entities and processes at play in the urban world, e.g., residents looking for satisfactory housing, developers trying to make a profit, and so on, while allowing the direct modeling of interactions between these entities. For instance, Otter et al. (2001) have shown that they can explore interesting questions regarding the interaction of households and firms in the creation of development patterns. In addition, the processes included within the ABM usually have direct analogs to processes in the real world, e.g., a resident purchasing a home can be modeled in a manner very similar to a housing search and decision process. Model agents have characteristics (e.g., preferences for views vs. short commutes) and capabilities (e.g., search heuristics) which may change over time (e.g., as a result lifecycle changes or of adaptation) in ways that reflect what is known about people, firms and so on. Note that heterogeneous agent characteristics, reflecting the diversity of characteristics found in social systems, can have important effects on the dynamics of complex systems (Holland, 1995; Rand et al., 2002). Further, policy tools are easy to incorporate in ABMs, e.g., to explore the effects of tax policy, zoning or other land-use control policies such as the creation of greenbelts (Brown et al., 2002).

Clearly ABMs can be made very complicated, to simulate the details of specific development histories (and futures) of particular places. On the other hand, ABMs can be very simple, e.g., by turning off mechanisms or setting parameters to special boundary conditions, the model can be compared to or calibrated with models with analytically demonstrable properties (Brown et al., 2002).

Here we describe and analyze an ABM that ultimately will be used to evaluate the ecological effects of development. The flexibility of the ABM approach suits our goal of building models at a level of detail and complexity appropriate to the questions we want to ask in each particular study. In general our goal is not to create models that generate the specific details of particular histories, but instead to create models that help us understand the social processes that lead to the patterns and dynamics common to a variety of situations. While ABMs do not have the formal simplicity of more abstract models (e.g., those based on physical systems), they do allow us to build models that can be more easily related to and used for policy decisions.

VALIDATION OF AGENT-BASED MODELS

A model is valid to the extent that it adequately represents the system being modeled, i.e. the model is able to correctly answer the questions it was designed to answer (Casti, 1997). There are a number of common approaches to validation (Sargent, 1988; Parker et al., 2003), including (1) matching model output to measured variables in the system being modeled, and (2) matching a model's component structures and processes to structures and processes in the system being modeled. In either case, a key modeling choice is how much detail the model is being designed to match. Sometimes the goal is to simulate specific cases, e.g., the detailed (micro) history of a particular urban area, while other times the goal is to explain general (macro) patterns that are observed across a wide variety of situations.

Validation is a critical issue for any modeling approach applied to any system, but it can be especially difficult when using ABM to model complex adaptive systems (CAS), like the social systems involved in urban development, because the behaviors and structures generated by the interacting, perhaps adaptive, agents often are very hard (or impossible) to predict (Holland, 1995; Bankes, 2002). In particular, the defining characteristics of CAS make matching micro-level details very difficult. First, CAS behavior can be dependent on detailed state, leading to extreme sensitivity to initial conditions. Also, CAS behavior often is path dependent, so that when re-run, CAS models can generate distributions of histories, sometimes resulting in multiple equilibria. One common cause of state and path dependence in CAS are the positive feedbacks commonly found in such systems (Arthur, 1994). This is all reflected in the literature on city formation, which suggests that luck and timing play large roles (Cronon, 1991). Further, agents may adapt to situations in unpredictable ways. The adaptability of agents can make it difficult or impossible to validate models of CAS by replicating all the micro-details of known histories, let alone by making detailed predictions.

In light of the above considerations, our goal is to create models that lead to a better understanding of the general processes and conditions that underlie urban development and, in particular, urban sprawl and its ecological effects. Thus our approach to model validation will include (1) matching model components and processes to real-world components and processes, and (2) matching macro-level, aggregate patterns, statistics and dynamics that are found across a variety of cases. While our goal is not to predict all the micro-details of histories of development in particular areas, we do expect the micro details of real-world histories to be consistent with the distributions of histories generated by our models.

Note that using an ABM approach leads directly to creating model components and processes that correspond to actors and processes in the social systems we are modeling. Andersson et al. (2002) call this building a model from "first principles." Ontologically, our model represents a system of decision making agents that interact with each other and their environment in ways that reflect theoretical notions of boundedly rational human behavior (Simon, 1955). The agents make, or can make, their decisions in ways that approximate various factors believed to affect human decision making, including a limited information set, foresight, interdependence of preferences, and non-rationality. Heterogeneity in the real world, observed through social surveys, can be incorporated in the model by characterizing the distributions of preferences, by grouping agents of like preferences, or by using the responses of actual survey respondents as inputs to individual agents in the model.

In this paper we focus on validating our model by matching the macro-level patterns generated by our model to those commonly found in the real world. Within urban development literature, two macro-level patterns have been found in many different locations and times.

First, it has long been known that there is a relationship between the size of developmental clusters and the frequency with which these clusters occur. Zipf (1949) first showed that there was a power law relationship between city populations and their rank. Several contemporary researchers have continued to explore this phenomenon and shown that it occurs in the relationship between developed clusters and the frequency with which they occur (Krugman, 1996; Batty and Longley, 1994; Caroll, 1982), as described by the following equation:

$$N(A) \sim A^r \tag{1}$$

where A is the size of an area, $N(A)$ is the number of areas of that size and r is some constant, empirically shown to be about 2 (Makse et al., 1998).

A second macro-pattern that is commonly observed is that the density of development is related to the distance from the urban center in an area. In particular, it was first shown by Clark (1951) that as the radius of a circle around a city center increases the density of development within that circle decreases exponentially, as described by:

$$y \sim Ae^{-bx} \tag{2}$$

where y is the density of the resident population, x is the miles from the city center, and b and A are constants. While the constants vary for different areas and over time in the same areas (Makse et al., 1998), the general exponential relationship remains.

In this paper we analyze the patterns generated by our ABM in an artificially constructed square lattice, to test whether the above macro-level patterns are created by the processes included in our model. While our overall project is aimed at studying urban development (and sprawl) in the Detroit Metropolitan Area (USA), starting with artificial landscapes makes it easier to study the behavior of our model, and it allows us to better understand which phenomena are idiosyncratic to a particular region and which are found across a variety of situations. For example, we have shown in previous work that our model is resilient to several parameter and structural changes within the model, such as whether utility functions are additively or multiplicatively separable (Rand et al., 2002).

EXPLANATION OF MODEL

The simplified model we present here was developed in Swarm using agents who have locational preferences. These agents exist on a heterogeneous two dimensional landscape, which can be defined using data stored in a geographic information system (GIS) or set to a hypothetical landscape. The model generates dozens of outcome variables of both a spatial and quantitative nature. We call our project SLUCE (**S**patial **L**and **U**se **C**hange and **E**cological effects) and this model, SOME (**S**luce's **O**riginal **M**odel for **E**xploration). The model is composed of three primary parts, the environment, the agents, and the agents interaction with the environment. We describe each in turn.

Environment

We represent geographic space with a two-dimensional square lattice. The results presented below all take place on a 301 by 301 lattice. Each location on the landscape has two exogenous characteristics: a natural beauty score in the interval $[0, 1]$ (nb_{xy}) and the presence or absence of an initial service center (see 2.2). In the model presented here we place one initial service center in the center of the lattice to represent the initial city center.

We can then compute the distance to services for a cell at x, y, which we call sd_{xy}, by taking the sum of the inverse Euclidean distances (for simplicity) to the nearest eight service center locations from that cell. Thus, a cell that is surrounded by service centers would receive a score of 8. Because it seems reasonable that the residents of a cell would not receive additional benefit from more than about 2 immediately adjacent service centers, we set the service center score to a maximum of 2 and normalize the value. Thus,

$$sd_{xy} = max[1, \frac{(\frac{1}{\|sc_1\|} + \ldots + \frac{1}{\|sc_8\|})}{2}] \tag{3}$$

where $\|sc_i\|$ is the Euclidean distance to the $i-th$ nearest service center from x, y. Service center distance changes over time as new service centers arise.

The Agents

This simple model has two agent types: residents and service centers (e.g., retail firms). Residents and service centers enter the world at each time step, and each takes up one cell in the lattice. Both residents and service centers have the capacity to include heterogeneous attributes and behaviors, but at present service centers do not have any attributes themselves. However their presence greatly affects residential location decisions, as described below.

In the basic model, residents have two attributes: (1)

Beauty Preference ($\alpha_{nb} \in [0,1]$), the importance that an agent gives to the natural beauty of an area. The natural beauty of a cell (nb_{xy}) can be generated from a distribution or set to a particular value exogenously. The beauty value of a cell x, y to an agent r is $nb_{xy} \times \alpha_{nb,r}$; and (2) *Service Center Preference* ($\alpha_{sd} \in [0,1]$), the weight that an agent gives to the nearness of a cell to service centers. The service value of a cell x, y to an agent r is $sd_{xy} \times \alpha_{sd,r}$.

The distribution of attribute values across agents can be set to normal, uniform, homogeneous, or could be derived from empirical data.

Agent Behavior

In each step of the model, a group of new residents enters the map. The rate at which residents move into the landscape is determined exogenously. For the experiments below, we set the rate to 10 residents per step. Residents use a hedonic utility calculation to make their decisions about where to live, which takes into account the natural beauty and distance to service centers.

To select a cell, a new resident r looks at some number of randomly selected cells (10 for all runs presented here) and moves into the cell that has the highest utility for r (ties broken randomly). The utility function is a multiplicative model, which defines them to be dependent, e.g. being near a service is irrelevant if there is no natural beauty:

$$u_{xy} = nb_{xy}^{\alpha_{nb}} \times sd_{xy}^{\alpha_{sd}} \quad (4)$$

Every time some number of residents (arbitrarily set to 100) is created, a service center is created in an empty cell near the last resident to enter the model. This is an approximation of the idea that services follow settlements.

EXPERIMENTS AND RESULTS

We present the results of one set of experiments with homogeneous residents. In this experiment the preferences for natural beauty (α_{nb}) and service centers (α_{sd}) were set to 0.5 for all residents. The natural beauty of each cell was drawn from a normal distribution ($\mu = 0.5, \sigma^2 = 0.5$; if the value exceeded 1 or was less than 0 a new number was drawn), and then a local filter was applied to create spatial autocorrelation. We seeded the model with one service center located at the center of the world, and ran the model to time step 250 at which point we collected data (after 250 time steps, 2500 agents and 25 service centers have located in the world).

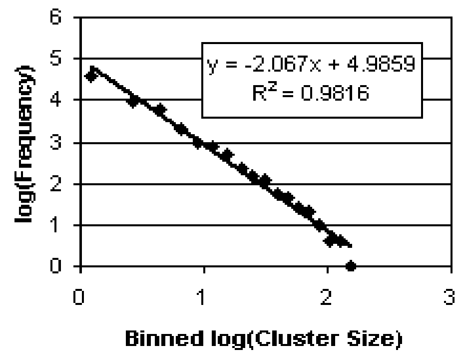

Figure 1. Frequency vs. Cluster Size

As mentioned earlier we are concerned with two sets of data. The first records the size of each developed cluster in the map for each model run. A developed cluster is any contiguous cluster using a Moore neighborhood. We then aggregated these data across as many as 50 runs and plotted the logarithm of the cluster size versus the logarithm of the frequency of the cluster size. The data were placed in logarithmic bins of size 1.2^k (where $k = 0, 1, 2 \ldots$). The plot is the logarithm of the frequency (aggregated over 50 runs) within these bins versus the logarithm of the upper bound of the bin. The results are presented in Figure 1.

The data appear to fall on a fairly straight line. A regression analysis shows the linear fit has an r^2 value of 0.981 and a slope of -2.07. Our model generates data which fits a power law for frequency of clusters versus size, with slope that falls within the range of published empirical data. For example, one value of the slope throughout the developed world is reported as -2.03 ± 0.05 (Zanette and Manrubia, 1997). Note that these results were obtained without any tuning of the model and were simply the first data that we generated.

The second measure we chose to analyze was of the density within a circle drawn around the center of our model. We calculated this value for radii from 1 to 150 which is the outer edge of the world. We averaged these values over all 50 runs. We then took the logarithm of the average density and plotted it against the radius. Here we fit a negative exponential instead of a power law. The results are presented in Figure 2.

Our model generates a result similar to the findings of Clark (1951). Using a semi-log plot we again have a fairly straight line. A regression analysis shows that the r^2 value is 0.998 and the

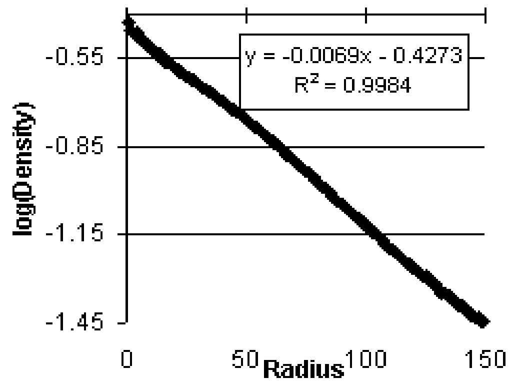

Figure 2. Density vs. Radius

slope is -0.006. Comparisons to previous work are difficult, due to a variety of reporting techniques. For instance, Clark (1951) uses natural logarithms whereas our results here are in base 10, if we use natural logarithms instead the slope of our regression is comparable to Clark's lower bounds. Moreover, Makse et al. (1998) shows that this value changes in time and in at least one observed data set (Berlin, 1945) the slope is -0.009 which is close to our observed value.

To summarize, our most basic model generates distributions of cluster sizes and declining density that, in structural form if not in precise parameters, aligns with data from a handful sample cities in the real world. This by no means proves that our model is correct, but it does suggest that our model has captured some aspects of the process of development. But at present we can make no claim that we have explained the real world settlement patterns.

DISCUSSION

Aggregate statistics of the sort discussed here can be useful in policy-making settings. For example, it is possible to imagine that a city planner cares less about where development occurs than that it occur in tight clusters. In these cases measures like those mentioned above may very well be useful in applying a model. The policy analysts could try out possible policy alternatives, e.g. zoning or the placement of greenbelts and see if they affect these aggregate statistics in favorable ways.

We have discussed previously the problems with trying to validate models that are naturally unpredictable. For instance, in our model if by chance the first additional service center locates north of the city instead of to the south then there will likely be more development to the north. However within our model (and often within the real system) this development might just as well have formed in the south. If we say that only those models that always give development in the north are correct, then we rule out the most realistic models. But all is not lost. The fact that a development occurred extending out from the city is important and may very well be useful. Therefore some mix of aggregate statistics and similarity to the actual settlement pattern may be the best way to perform validation.

Practically speaking, if urban growth models are to be useful to policy makers it is nearly as important that the mechanisms be something that can easily be understood as it is that the results are realistic. ABMs facilitate policy analysis because they can include representations of the restrictions, incentives, and disincentives that might be put in place to affect agent decision making. Our model can be used to evaluate policies like zoning, green belts, or incentives for clustered development by running versions of the model with these represented and evaluating the outcomes. The impacts that these policies have on patterns can be used to evaluate the effectiveness of policies toward specific outcomes in terms of development patterns.

FUTURE WORK

We plan to further validate the spatial patterns of our model using modifications to agent characteristics and behaviors. In doing so we hope to find that changes in our model influence parameters of the output distributions we see but maintain distributions similar to those we see in the real world. For instance, some parameter(s) of our model may allow us to influence the slope of the power law. This might in turn be used as a tool to influence how policy makers attempt to control patterns of development.

Second, it is important for us to examine questions of validation, and how to better connect our model with reality. As argued above recreating micro-details of development is not our goal, but we do want to validate our model against the real world. We have begun to investigate improved methods, for example dividing the world into variant and invariant regions and examining the relationship between these regions.

CONCLUSION

We have shown that a simple ABM can generate aggregate patterns that match those seen in empirical data. We argue that the macro measures that we are using to make this comparison provide useful information about the spatial patterns of development that occur in the real world. Moreover it seems clear that if these models are to be useful in real-world applications they must be understandable by those who use them. We should be able to construct models at a level of detail appropriate with the questions we want to ask about development patterns. Validation of models is a difficult issue and probably one of the most important ones facing the agent based modeling community. A renewed effort to explore and confront these issues is necessary. This paper represents our first step to address these issues within the realm of our particular model.

ACKNOWLEDGMENTS

This work is funded by the US National Science Foundation under the Biocomplexity and the Environment program, grant BCS-0119804.

References

Andersson, C., K. Lindgren, S. Rasmussen, and R. White, 2002. Urban growth simulation from "first principles". *Physical Review E*, 66(026204).

Arthur, W. B., 1994. *Increasing Returns and Path Dependence in the Economy*. Economics, Cognition, and Society. University of Michigan Press, Ann Arbor, MI.

Bankes, S. C., 2002. Tools and techniques for developing policies for complex and uncertain systems. *Proc. Natl. Acad. Sci.*, 99(Supp. 3):7263–6.

Batty, M. and P. Longley, 1994. *Fractal Cities*. Academic, San Diego.

Brown, D. G., S. E. Page, R. Riolo, and W. Rand, 2002. Modelling the effects of greenbelts at urban-rural fringe. In *Proceedings of IEMSS 2002*.

Caroll, G. R., 1982. National city-size distributions: what do we know after 67 years of research? *Progress in Human Geography*, 6:1–43.

Casti, J., 1997. *Would-Be Worlds: How Simulation is Changing the Frontiers of Science*. Wiley.

Clark, C., 1951. Urban population densities. *Journal of the Royal Statistical Society (Series A)*, 114:490–496.

Cronon, W., 1991. *Nature's metropolis: Chicago and the Great West*. W. W. Norton, New York.

Holland, J. H., 1995. *Hidden Order*. Perseus Books, Reading, Massachusetts.

Johnson, M. P., 2001. Environmental impacts of urban sprawl: A survey of the literature and proposed research agenda. *Environment and Planning A*, 33:717–735.

Krugman, P., 1996. *The Self-Organizing Economy*. Blackwell, Cambridge, MA.

Makse, H. A., J. S. A. Jr., M. Batty, H. Shlomo, and H. E. Stanley, 1998. Modelling urban growth patterns with correlated percolation. *Physical Review E*, 58(6): 7054–62.

Otter, H. S., A. van der Veen, and H. J. Vriend, 2001. Abloom: Location behaviour, spatial patterns, and agent-based modelling. *Journal of Artificial Societies and Social Simulation*, 4(4).

Parker, D. C., S.M.Manson, M.A.Janssen, M.J.Hoffman, and P.Deadman, 2003. Multi-agent systems for the simulation of land-use and land-cover change: A review. *Ann. Assoc. Amer. Geographers*, page Forthcoming.

Rand, W., M. Zellner, S. E. Page, R. Riolo, D. G. Brown, and L. E. Fernandez, 2002. The complex interaction of agents and environments: An example in urban sprawl. In *Proceedings of Agent 2002*.

Sargent, R. G., 1988. A tutorial on validation and verification of simulation models. In Abrams, M. et al., editors, *Proc. of 1988 Winter Simulation Conference*, pages 33–39.

Simon, H. A., February 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69: 99–118.

Zanette, D. H. and S. C. Manrubia, July 21 1997. Role of intermittency in urban development: A model of large-scale city formation. *Physical Review Letters*, 79(3): 523–526.

Zipf, G. K., 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.