What Statistical Models Can Better Detect Land-Change Mechanisms?

Li An[1,4], Daniel G. Brown[1,4], William Rand[2,4], and Scott Page[3,4]

1. School of Natural Resources and Environment
The University of Michigan
Ann Arbor, MI 48109
Tel: +1 734-615-4897 (Li An),
+1 734-763-5803 (Daniel G. Brown)
FAX: +1 734-936-2195
Email: lian@umich.edu (Li An)
danbrown@umich.edu (Daniel G. Brown)

2. Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109
Tel: +1 734-763-3323
FAX: +1 734-763-9267
Email: wrand@umich.edu

3. Department of Political Science
The University of Michigan
Ann Arbor, MI 48109
Tel: +1 734-615-4191
FAX: +1 734-763-9267
Email: spage@umich.edu

4. Center for the Study of Complex Systems
The University of Michigan
Ann Arbor, MI 48109

Abstract

Statistical models are widely used to explore relationships between aggregate variables in a land-use system by inductively fitting empirical data, for example with regression models. Seldom do researchers know whether, and if so how much, the relationships thus obtained reflect the underlying mechanisms. This is partially due to the uncertainties in the system of interest, especially when the system is a complex adaptive system. This research addresses this issue by integrating a spatial agent-based model (ABM) and several regression models. Driven by a set of hypothetical residential location rules represented by associated parameters, the ABM generates a series of land-use patterns on a 2-dimentional virtual landscape over a number of time steps. For land pixels randomly sampled from the landscape, we recorded their emergent land-change dynamics, and calculated each pixel's characteristics (variables used in the ABM) of interest at selected time steps. Using these data, we applied survival analysis, logistic regression, and multivariate regression to discover the relationships between the land-change dynamics and the variables of interests. The findings suggest that these statistical models can capture the relationships encoded in the ABM in terms of some metrics (e.g., signs and significance status of regression coefficients) to varying degrees, and that survival analysis outperforms the other models. Our approach has substantial implications in calibrating and validating agent-based models, testing plausible scientific hypothesis, conducting sensitivity analysis, and many other model-based scientific endeavors.

1. Introduction

Statistical models have long been used to test hypotheses, detect relationships, and shed light on a system of interest in many fields (Fotheringham et al. 2000, p.1-14). Their power depends on many factors, such as the reliability of the data, the theory behind the model (or the researcher's conceptual understanding of the system), and specific model(s) that the researcher adopts. Unfortunately, much uncertainty arises from data noise (e.g., sampling biases, measurement errors), imperfect theory, poor understanding of the system, or inappropriate use of statistical models, especially when the systems under research are complex adaptive systems (CAS; see Axelrod and Cohen 2000, p. 32-38) which contain heterogeneity, non-linearity, and feedback. As such, researchers using statistical models find it difficult to assure that they have reached the best conclusions given the data available and the theory/theories at hand.

In land-change science, researchers have primarily adopted Markov Chain models (e.g., Baltzer 2000, Brown et al. 2000), logistic function models and variants (e.g., Mertens and Lambin 2000, Müller and Zeller 2002), and multivariate linear regression models (e.g., Mertens et al. 2000) to link land-change dynamics with exogenous biophysical and/or socio-demographic variables. These models are helpful in many situations, but rarely has their appropriateness been explored in detail: given the data and hypotheses to be tested,

are these models the best candidates that correctly use or make full use of the information in the data, and give reliable results? What other models may give different or better insights?" Furthermore, and perhaps most importantly, given the intervening effects of system interactions, feedbacks, and aggregations, what limits the ability of statistical models to accurately reveal the agent-level behaviors that gave rise to aggregate patterns of land-use? A generic exploration of the appropriateness (e.g., conditions, strengths, and weaknesses) and utility of several common statistical models seems very important. A pilot methodological study of statistical models in land-change science by An and Brown (in review) showed that survival analysis is very powerful in detecting temporal variations and can disclose some potential mechanisms that may be otherwise unavailable—the robustness of the conclusion is, however, sacrificed by uncertainties existing in the data quality and underlying mechanisms.

Fortunately, agent-based models (ABMs hereafter) show strong prospects in addressing this issue. As one type of simulation models, ABMs have an as-realistic-as-possible mapping of real-world entities and their relationships in terms of computer objects and the associated rules, and the emergent or macro-level phenomena arise from the interactions between these micro-level decision-making entities (agents) and/or between these entities and their environment(s). Generally represented through object-oriented programming, ABMs allows for representation of heterogeneous characteristics in agents, non-linear relationships and feedback among agents/objects and/or their environment(s), as well as cross-scale (both temporal and spatial) and cross-disciplinary integration of data, metrics, and models (e.g., Parker et al. 2003; An et al. 2005).

Though ABMs have seen increasing applications in modeling human-environment interactions on real landscapes in a formal scientific context (e.g., Lim et al. 2002, An et al. 2005), their capacity to allow for hypothesis testing and exploratory modeling (sometimes for pedagogical use as well) in virtual landscapes still turns out to be one of their dominant strengths (e.g., Parker and Meretsky 2004, Brown et al. 2004), and some researchers even argue that this might be the only role of ABMs (Epstein 1999). Whatever the arguments, ABMs can undoubtedly provide researchers a platform that (1) encapsulates a small set of relevant agents and assumed (maybe highly simplified) relationships (rules), with other entities or relationships of no interests screened out, (2) performs simulation experiments under different model parameters corresponding to varying hypothetical conditions, and (3) tests hypotheses that are otherwise difficult to test. This utility of ABMs has found wide application in many disciplines such as economics (e.g., Arthur 1999), political science (e.g., Epstein 2002, Kollman et al. 2003, p. 1-12), and complexity theory (e.g., Axelrod and Cohen 2000).

In the research presented in this paper, we used ABMs to rule out the uncertainties arising from data quality and/or those usually (if not always) unknown underlying mechanisms, and explore the utility and conditions of statistical models in land-change science. Therefore, we pursued the following objectives: (1) establish a set of simplified hypothetical mechanisms that drive the land-use changes in an ABM; (2) use different statistical models to analyze the data obtained from such an ABM; and (3) explore the

situations in which these statistical methods reflect the true/hypothetical mechanisms or produce significant bias.

2. Methods

2.1 The SOME Model

Our project on Simulating Land-Use Change and Ecological Effects (SLUCE; see http://www.cscs.umich.edu/sluce) at the urban-rural fringe developed an agent-based spatial model called SOME (SLUCE's Original Model for Exploration), which studies how individual residential location decisions interact with environmental factors (Rand et al. 2003; Brown et al. 2005). Two types of agents, homebuyers and service centers, enter a virtual landscape of n × n lattice based on a set of assumptive, but empirically and theoretically reasonable, rules. The landscape has a set of characteristics upon which the homebuyers base their residence decisions, such as aesthetic quality, distance to service centers, and nearby density. The utility of each potential location (x, y) is calculated by this equation:

$$u_{x,y}(t) = \prod_{i=1}^{n} (1 - |g_i - z_{i,x,y}(t)|)^{a_i} \quad (1)$$

Where $u_{x,y}$(t) is the utility calculated at location (x, y) at time t, g_i is the ideal value for factor i (i = 1, 2,…, n; we assume that all individual homebuyers have the same ideal value for one specific factor over time), and $z_{i,x,y}$ (t) is the observed value for factor i at location (x, y) at time t. Both g_i and $z_{i,x,y}$ (t) are standardized to be between 0 and 1. a_i is the preference weight that the homebuyer places on factor i (time-invariant for the time being). To simplify the utility calculation and modeling process, we assume that all the g_i's are equal to 1, and only examine the factors of aesthetic quality and distance to service centers. According to this assumption and some relevant processes in the SOME model, equation (1) reduces to:

$$u_{x,y}(t) = (A_{x,y})^{a_A} \times (1/SC_{x,y}(t))^{a_{SC}} \quad (2)$$

where $A_{x,y}$ (standardized between 0 and 1) and $SC_{x,y}$(t) are the observed values of aesthetic quality and distance to service center (at time t) at location (x, y), and a_A and a_{SC} are the preference weights placed on these two variables. The SOME model can choose the values of these two parameters based on some common distributions such as normal and uniform distributions. To keep the a's meaningful in equation (2), negative values and values greater than 1 that arise from the sampling process are replaced with positive values re-sampled from the same distribution.

When the model is started, a service center is placed at the geometric center of the virtual landscape. At each time step, a total of n (n = 10 in our case) homebuyers enter into the landscape. Each homebuyer randomly samples a total of m (m = 10 in our case) unoccupied places (cells), and the one that provides the highest utility value (calculated

from equation (2) based on the observed values at (x, y), i.e, $A_{(x,y)}$ and $SC_{(x,y)}(t)$) will be the choice of that homebuyer. As the process continues, a new service center will be added to the landscape as soon as the existing service serves more than 100 homeowners; the location of the service center is near the location chosen by the last homebuyer. This process continues until a certain proportion of the landscape is occupied or a time specified by the researcher has been reached.

2.2 Experiment Design

The SOME model served as a platform to perform experiments on a virtual landscape. We created a 121 by 121 random map of aesthetic quality with spatial autocorrelation in consideration (Figure 1), each aesthetic value ranging from 0 to 8191 (integers were used to ease computational complexity, but were standardized to be between 0 and 1 before calculating the utility in equation (2) or entering regression analysis later). The report frequency was 20 steps (one time-unit; each step could be understood as one year), and we reported the simulation results at steps 20, 40, 60, 80, and 100 (or at time-units 1, 2, 3, 4, and 5), and output them as ASCII files for statistical analysis and/or spatial analysis in ArcGIS.

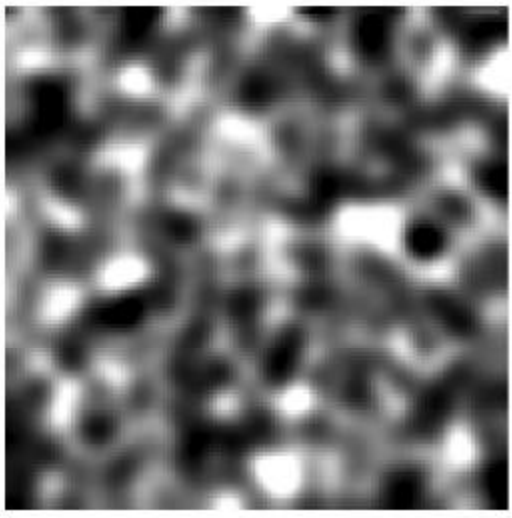

Figure 1. The 121×121 aesthetic quality map generated in ArcGIS with spatial autocorrelation considered. The light to dark areas represent low to high aesthetic values on the virtual landscape.

Our experiments focused on different levels of preferences towards the two environmental factors, the aesthetic quality and the distance to service center (for similar explorations of the effect of diversity in preferences, see Rand et al.2002). First, we chose average preferences for aesthetic quality and distance to service centers (a_A and a_{SC}, respectively) to be 0 and 1, respectively, with variable levels of variance increasing from 0.0 to 1.0, representing a situation/hypothesis that the homebuyers only consider distance to service centers with aesthetic quality unaccounted—but as the variance rises, this trend may be blurred. Experiments #1-4 were designed in this regard (Table 1). Next, we switched values of a_A and a_{SC} (a_A =1 and a_{SC} =0) and still keep their variances

increasing from 0 to 1 (Experiments #5-8 in Table 1), reflecting our hypothesis that if aesthetic quality is the only factor in the decision, the emergent spatial trajectory would vary and lead to quite different regression results. Aesthetic quality (represented as a random map with spatial autocorrelation) is free of temporal variations (we keep it so, though the SOME model has an option to change the aesthetic quality as new developments come in), while distance to service centers is time-dependent (addition of new service centers may change the distances for some cells) and path-dependent (i.e., the location of upcoming service center depends upon the distribution of current residence and service centers, which in turn depend upon locations of earlier ones; all these distributions are more or less related to the first service center located at the center of the virtual landscape). Lastly, we were interested in the scenarios where variances remain zero, but the mean preferences (a_A and a_{SC}) change. These experiments were complementary to the above two sets of experiments: aesthetic quality values of the landscape are random, while the distances to service centers are both time-dependent and path-dependent. Therefore, they may behave differently in the regression analysis even if the preferences to them rise at the same rate and are equal.

Table 1. Model parameter specification

Experiment #	a_A		a_{SC}	
	Mean	Variance	Mean	Variance
1	0	0.00	1	0.00
2		0.25		0.25
3		0.50		0.50
4		1.00		1.00
5	1	0.00	0	0.00
6		0.25		0.25
7		0.50		0.50
8		1.00		1.00
9	0.25	0	0.75	0
10	0.50		0.50	
11	0.75		0.25	

The parameter settings (Table 1), together with the processes described above, serve as the hypothetical mechanisms that drive the land-use dynamics, and the resultant emergent land-use patterns were sampled and analyzed using statistical models.

2.3 Sampling and data collection

To avoid the influences of spatial autocorrelation on regression and to speed up the computation, we sampled a total of 2175 cells (approximately 15% of all the cells in the virtual landscape) for each of the experiments in Table 1. For each of the sampled cells, we calculated the associated value of the aesthetic quality (homogeneous over time), and the distances to the nearest service center at each of the five time-units. Finally, we recorded the land-use trajectory for each cell based on the output pattern at each of these five time-units (see Figure 2): 0 for non-development, 1 for residential land-use, and 2 for service center. In our research, we were primarily interested in the drivers of the transitions from non-development to residential land-use, so we do not model the transitions from non-development to service center—actually, the placement of service center in the SOME model only depends on the needs from the homebuyers (each 100 residential units need a service center) rather than environmental factors.

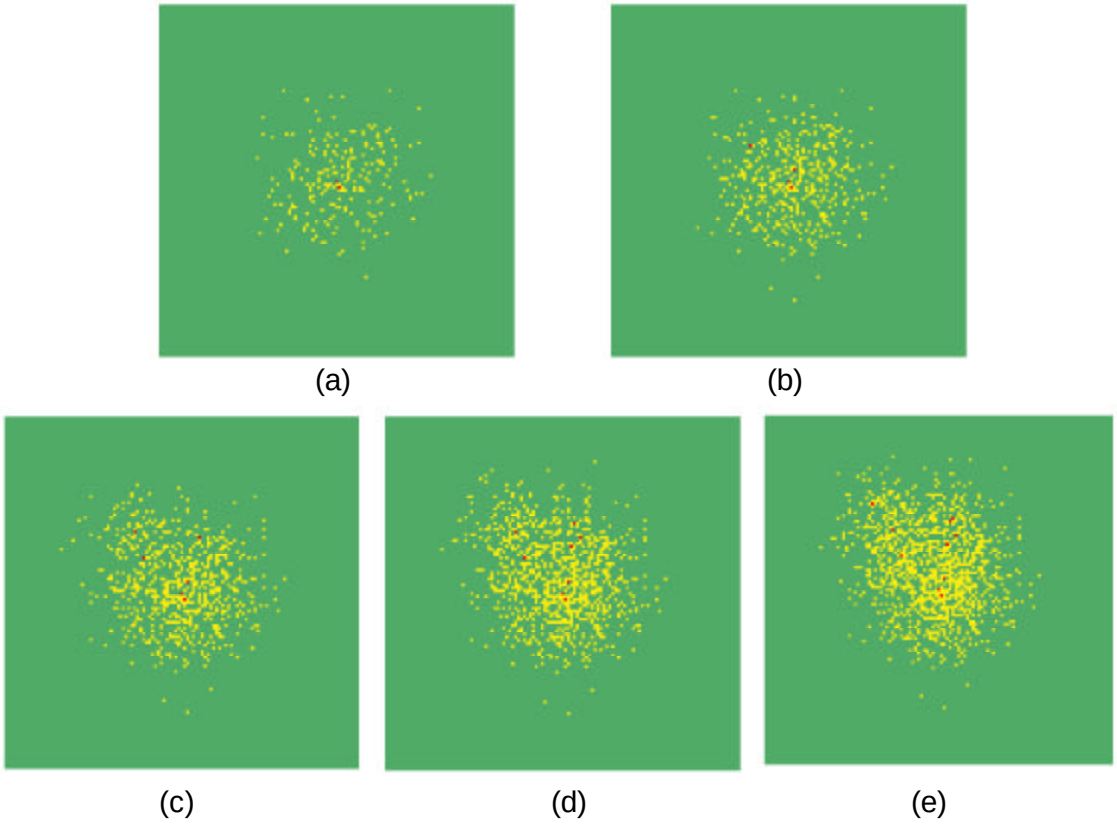

Figure 2. Snapshots of development dynamics in five time steps when $\alpha_a = 0$, $\alpha_{SC} = 1$, and variance = 0 (Experiment 1 in Table 1): (a) step 20, (b) step 40, (c) step 60, (d) step 80, and (e) step 100. Green (grey if printed in b/w) is for non-development, yellow (white if printed in b/w) for residential land-use, and red (black if printed in b/w) for service center.

The outcome development trajectories take the form of, e.g., 0-0-0-0-0 (the cell remains undeveloped until the end of time-unit 5, or step 100), 0-0-1-1-1 (the cell gets developed between time-units 2 and 3, or steps 40 and 60), or 1-1-1-1-1 (the cell gets developed

between the start-up time and step 20, or between 0 and time-unit 1). To facilitate the OLS multivariate regression, we assigned its survival times (the length of time that the cell remains undeveloped) as t = 5.5, 2.5, and 0.5 time-unit. Note: assigning the 0-0-0-0-0 type (the cell remains undeveloped until the end of step 100) a fixed survival time (5.5 time-units, or 110 steps here) is required in OLS regressions, but may cause bias (that cell may survive longer than 5.5 time-units or 110 steps). This is one of the strengths of survival analysis, which will be introduced later.

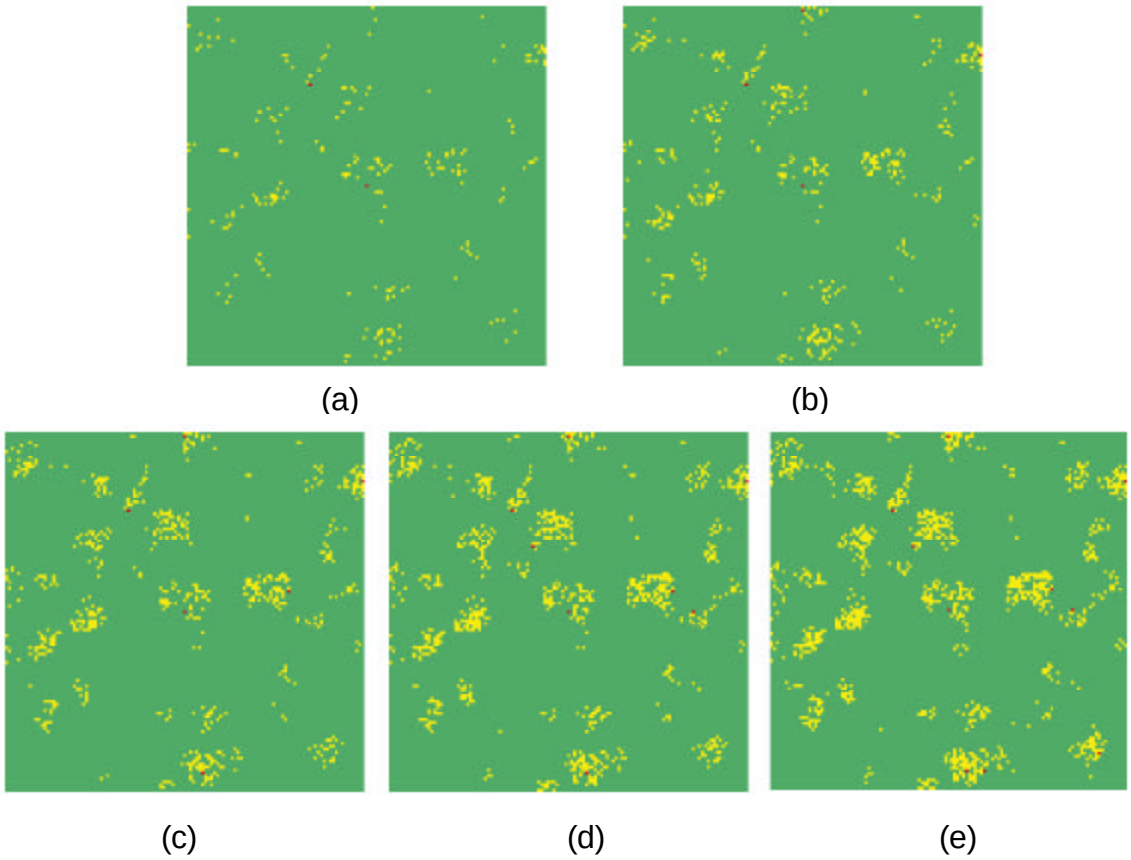

Figure 3. Snapshots of development dynamics on the landscape in five time steps when α_a = 1, α_{SC} =0, and variance = 0 (Experiment 5 in Table 1): (a) step 20, (b) step 40, (c) step 60, (d) step 80, and (e) step 100. See Figure 2 for the legend.

2.4 Statistical Analysis

Using the data thus generated, we conducted statistical analysis using three models, and recorded the estimated coefficients, χ^2, p value, and model-fit indices (- 2 log likelihood or R^2). The three models include a survival analysis model, a logistic model, and a regular ordinary least square (OLS) multivariate model.

Survival Analysis is a collection of statistical methods designed to handle the occurrence and timing of events, which originated from the study of deaths in medicine, public health, and epidemiology (thus the name survival analysis; Allison 1995; Klein and Moeschberger 1997). Powerful in handling data censoring (if the event is known to have

occurred in an interval $(-\infty, t_2)$, (t_1, t_2), or (t_1, ∞), we say that the survival time is left, interval, and right censored, respectively) and time-dependent covariates (explanatory variables taking varying values over time), it has been a more powerful tool for explaining time-varying phenomena (Klein and Moeschberger 1997; Cantor 2002). We regressed the logarithm of hazards against a linear combination of the two explanatory variables, aesthetic quality and distance to service center:

$$h_{(x,y)}(t) = h_0(t) \exp (b_A A_{(x,y)} + b_{SC} SC_{(x,y)}(t)) \quad (3)$$

where $h_{(x,y)}(t)$ is the hazard that cell (x, y) is developed to residence at time t (not directly observed, but derived indirectly; see Allison 1995 for more details), $A_{(x, y)}$ and $SC_{(x,y)}(t)$ are the aesthetic value and distance to service center at location (x, y) at time t ($A_{(x,y)}$ is time invariant, though). Due to a significant contribution called "partial likelihood function" by Cox (1975), $h_0(t)$ can be eliminated during the estimation of the coefficients b's, which explains the fact that the estimated results do not contain an intercept term. For details about survival analysis and its substantive potential in land change science, see An and Brown (in review).

Logistic modeling is quite popular in land-change science (thus we do not provide an introduction here), and we used it as one alternative to capture the mechanisms as specified earlier in this article. The logistic model is estimated as below:

$$\text{Log} (P_{(x,y)}(t)/(1-P_{(x,y)}(t))) = h + b_A A_{(x,y)} + b_{SC} SC_{(x,y)}(t) \quad (4)$$

Where $P_{(x,y)}(t)$ is the probability that location (x,y) under consideration is developed to residence at time t (thus each location has five records corresponding to the five time-units in the regression model), h is an intercept term (we do not use α because it is used in the utility equation above), and $A_{(x,y)}$ and $SC_{(x,y)}$ are the aesthetic value and distance to service center at location (x,y). To account for the possible effects of time-dependent variables (distance to service center in our case), we spread out the data into the "location-time" format in estimating model (4), i.e., each sampled cell has five records corresponding to its five time steps. To address the issue of possible correlation between the five records of one cell, we treated them as records in one cluster using the surveylogistic procedure in SAS as suggested in An and Brown (in review).

OLS multivariate regression is one of the most-often used methods in land-change science and many other disciplines. We regressed the survival time (s_time) against the two explanatory variables $A_{(x,y)}$ and $SC_{(x,y)}$:

$$s\_time = h + b_A A_{(x,y)} + b_{SC} SC_{(x,y)}\ (t) \quad (5)$$

where $SC_{(x,y)}$ (t) is the distance to service center prior to the development, and other variables are similarly defined as in the logistic model.

3. Results

In general, the regression coefficients had the expected signs when they were significant (we chose $p < 0.05$ as the default significance level): aesthetic quality had a positive sign from survival analysis and logistic regression for Experiment 5 (Table 2), implying the larger the aesthetic quality, the higher the chance to be developed, which reflected what exists in the SOME model in Equation (2). In the OLS model, aesthetic quality had a negative sign because the response variable is the survival time, or the time that the cell remains undeveloped, implying that the larger the aesthetic quality, the shorter the survival time, which translates to higher probability or hazard of being developed. Distance to service center manifests exact the opposite signs compared to aesthetic quality in the models because it is the invert of distance to service center that is used to compute the utility function (Equation (2)). Similarly, in Experiment 1 the negative/positive signs on distance to service centers/aesthetic quality from survival analysis and logistic regression, and positive/negative sign for OLS regression, were consistent with the higher/lower likelihood of development closer to service centers/in places with higher aesthetic quality that was built into this model.

Table 2. Regression results in two extreme situations*

	Methods	-2log L	R^2	β's	Estimates	χ^2(t value in OLS)	p-value
Experiment 1 $a_A=0, a_{SC}=1$ (Var=0)	Survival analysis	911.81	0.20	β_A	-0.3909	0.29	0.676
				β_{SC}	-14.1199	215.18	0.000
	Logistic regression	2049.78	0.11	β_A	-0.0860	0.01	0.9279
				β_{SC}	-14.3151	226.45	<0.0001
	OLS regression		0.06	β_A	-0.0811	-0.97	0.3322
				β_{SC}	0.6231	11.93	<0.0001
Experiment 5 $a_A=1, a_{SC}=0$ (Var=0)	Survival analysis	1072.66	0.19	β_A	12.9838	434.55	<0.0001
				β_{SC}	-0.6129	1.91	0.1671
	Logistic regression	2263.24	0.13	β_A	16.4833	252.02	<0.0001
				β_{SC}	-1.9216	16.19	<0.0001
	OLS regression		0.16	β_A	-1.8046	-19.09	<0.0001
				β_{SC}	-0.4744	-6.40	<0.0001

Note: the R^2 for survival analysis and logistic regression is calculated by $R^2 = 1-\exp((2Lp-2Lo)/N)$, where Lp and Lo are the log likelihood of the full model (with covariates) and the null model (without covariates), and N is the number of observations used in the model. This is called "generalized R^2", which cannot be explained as the proportion of variations explained by the covariates (Allison 1995).

3.1 Comparison between methods

When the two preference weights (a_A and a_{SC}) take non-zero values with zero variances, all the three statistical models gave significant coefficients (Figure 5(c)). Under the

extreme assumption that preference over distance to service center was the only factor affecting the home choice decision (Experiment 1: $\alpha_A = 0$, $\alpha_{SC} = 1$; var = 0), all the three models captured this assumption by showing insignificant coefficients of aesthetic quality ($p = 0.676$, 0.9279, and 0.3322), and significant coefficients of distance to service center ($p < 0.0001$ for all; Table 2). When the preference weights were opposite (Experiment 5: $\alpha_A = 1$, $\alpha_{SC} = 0$), only the survival model had a correct insignificant coefficient for distance to service center ($p = 0.1671$), and both logistic and OLS regressions masked this trend by giving significant coefficients ($p < 0.0001$ for both; Table 2). In addition, the survival analysis gave the highest R^2 (for cautions of using R^2 to compare models, see An and Brown in review).

3.2 Change of regression coefficients

The coefficients revealed by the statistical models in response to changes in variances of the agent preferences depended on the weights of preferences over the two variables (Figure 4). Since the coefficient signs only reflect the directions of the effects, we plotted $|\beta_{SC}|$ to display how the magnitudes of β_{SC} change in response to changes in the controlled variables. Later, we focus our discussion on the coefficients for aesthetic quality (β_A) and the absolute values of the coefficients for distance to service center ($|\beta_{SC}|$), and bold solid lines are used in Figure 4.

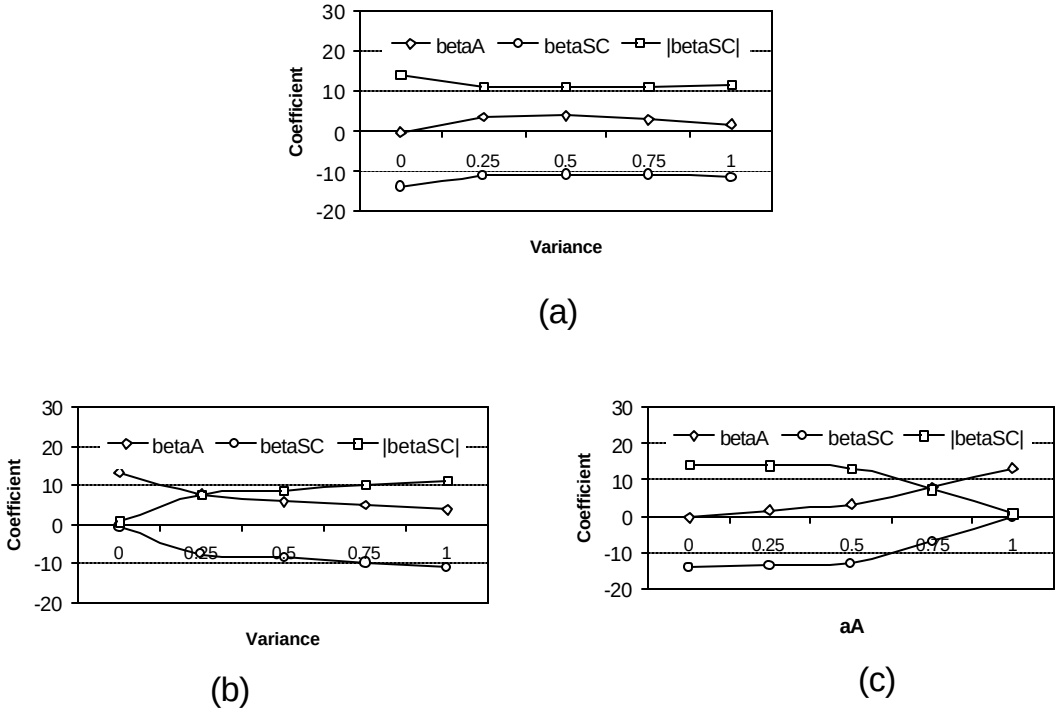

Figure 4. Regression coefficients (betas) curves under different conditions: (a) α_A =0 and α_{SC} =1; (b) α_A =1 and α_{SC} =0; and (c) variance =0.

When all the preference was placed on distance to service center (i.e., Experiments 1-4: $\alpha_A = 0$, $\alpha_{SC} = 1$), the rise in variance resulted in an increasing effect of aesthetic quality

and decreasing effects of distance to service center. At some point (approximately variance = 0.5), these trends leveled off and then went to the opposite direction (Figure 4 (a)). However, when all the preference was placed on aesthetic quality (i.e., Experiments 5-8: $a_A = 1$, $a_{SC} = 0$), the rise in variance led to monotonously decreasing effect of aesthetic quality and monotonously increasing effects of distance to service center. At some point (approximately variance = 0.25), these two curves intersected (Figure 4 (b)).

On the other hand, when variance was zero, the rise in the preference weight on aesthetic value a_a (corresponding to a reduction in the preference weight on distance to service center a_{SC}) gave rise to an increasing coefficient for aesthetic quality, and a decreasing coefficient for distance to service center, and these two curves intersect at $a_a = 0.7$ (Figure 4(c)).

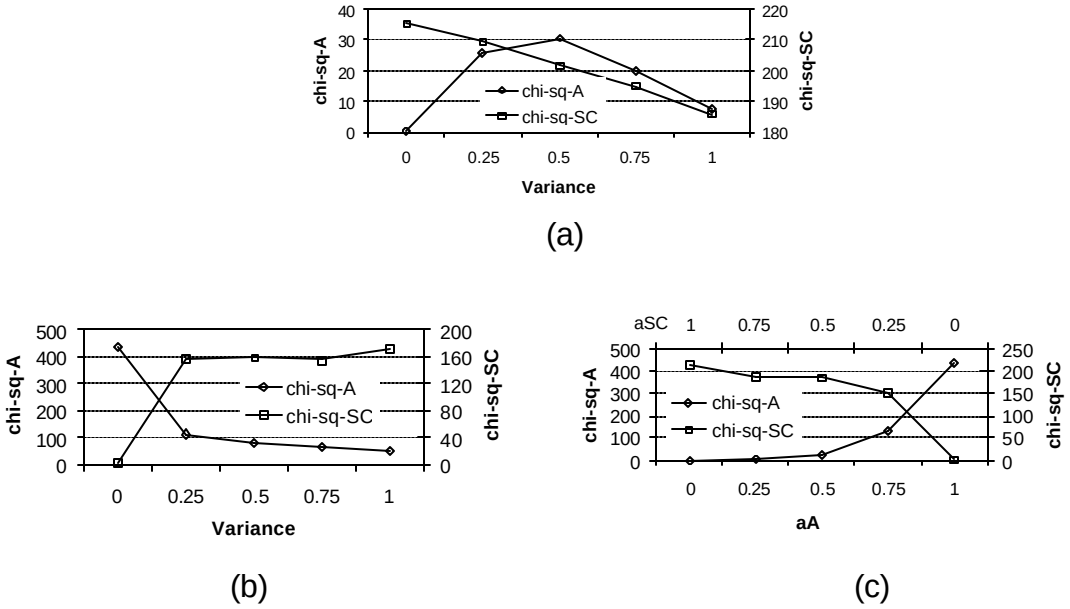

Figure 5. The significance levels (χ^2) of the regression coefficients for aesthetic quality and distance to service center under different conditions: (a) a_A =0 and a_{SC} =1; (b) a_A =1 and a_{SC} =0; and (c) variance =0.

3.3 Change of coefficient significance

When the distance to service center was the sole consideration in making residence decisions (Experiments 1-4: $a_a = 0$, $a_{SC} = 1$), the rise in variance resulted in a monotonous decrease in the significance level (measured by χ^2) of the service center coefficient (β_{SC}), and an inverted U-shaped significance curve for the aesthetic quality coefficient with a summit at approximately 0.5 (Figure 5(a)). When aesthetic quality played the only role (Experiments 5-8: $a_a = 1$, a_{SC} =0), the rise in variance resulted in a decline in the significance level of its coefficient, but an increasing significance level for

distance to service center (Figure 5(b)). When variance was zero, the significance level rose as the mean of the associated preference weights increased (Figure 5(c)).

4. Discussion

Data in land-change science are often characterized by data censoring and time-dependent variables. Fortunately, we have shown in this paper that survival analysis, among the three methods, can provide us with the capacity to analyze such data and detect the assumed mechanisms. This finding is consistent with what An and Brown (in review) found. This capacity still varied from situation to situation, however. For instance, the regression coefficients could largely reflect the mechanisms we assumed in the agent-based model in terms of the sign and significance when variance was zero. When variances were introduced, survival analysis gave coefficients for aesthetic quality with significance (χ^2 >7, see Figure 5 (a)), even if the average preference weight was zero, and coefficients for distance to service center with even higher significance (χ^2 >150, see Figure 5 (b)) when the preference weight was zero. This "malfunction" can be understood from the sampling of the associated weights: when the variance is positive, the resultant a is always positive due to the "resample-till-positive" process (see the section "The SOME Model"), thus the resultant average weight a is a positive number rather than zero. In this sense, the significant coefficients still reflect what is really in the SOME model.

The asymmetry between aesthetic quality and distance to service center deserves attention. When a_a = 0 and a_{SC} =1, the magnitude of the coefficient for distance to service center (>11.00) is much larger than that of the coefficient for aesthetic quality (<4; Figure 4 (a)); when a_a = 1 and a_{SC} =0, the magnitudes of these two coefficients are closer, ranging from 0.6 to 11.1 (Figure 4 (b)). This asymmetry manifests as well in terms of the χ^2 index: When aesthetic quality is ignored or less important (a_a = 0 and a_{SC} =1), the χ^2 values for aesthetic quality are below 31; when distance to service center is ignored or less important (a_a = 1 and a_{SC} = 0), the χ^2 values for distance to service center could still be as high as 172 (Figure 5 (b)). All these asymmetrical characteristics may have a single cause: the path-dependence in locating service centers in the SOME model. The service center, at the time of entering the landscape, is always placed near the residence, no matter how much emphasis the homebuyer places on this factor while choosing his/her residence location. Even when a_{SC} = 0, the service center and residence locations are still not independent, which can be seen in Figure 3: whenever there is a service center (black dot), there are residences around (white dots); but the reverse is not always true because the homebuyers under this situation did not take service centers into consideration when choosing residence locations.

We used the signs, the significance indices (χ^2 in particular—p-value is not reported by many software packages such as SAS when small enough), and the relative magnitudes of regression coefficients to evaluate the capacity of regression analysis to capture the "real" mechanisms, and these metrics turned out to be meaningful and useful. We expect, however, when the utility function in the ABM or the regression model is calibrated to conform to some protocol (e.g, both use the same function such as equation (3)), the

absolute magnitudes of the coefficients should make more sense and help us detect the information contained in the regression model.

The statistical models do not have a very good fit, indicated by the R^2 ranging from approximately 0.06 to 0.20 (though a generalized R^2 over 0.05 is considered acceptable in disciplines such as medical science and sociology when real data are used; see Allison 1995, An and Brown in review), which can be, at least partly, explained by the stochasticity existing in the SOME model. First, each time a homebuyer comes to the landscape, he/she can only evaluate ten randomly selected cells and choose the cell with the highest utility. On the one hand, the ten selected cells may have excluded some cells that are more suitable for residence in terms of our criteria, e.g., having a shorter distance to an existing service center or a higher aesthetic quality. On the other hand, the choice of the ten candidates, in real situations, may still depend on some variables (e.g., availability of informative realtors) that we do not know and thus have to be excluded in our SOME model. In this sense, our agent-based model is not a fully mechanistic model, but a combination of mechanistic and stochastic model, which is reasonable in the real situations because the homebuyers may not have complete information, and they may not be fully aware of their decision rules (often complex and difficult to formalize). So, we still cannot screen out all the uncertainties as we mentioned in the introduction. Second, the utility function (Equation (2)) is only one of several possible forms. For instance, the SOME model can also use an additive utility function:
\

$$u_{x,y}(t) = a_A(1 - A_{x,y})/2 + a_{SC}(1 - 1/SC_{x,y}(t))/2 \qquad (6)$$

To make the regression coefficients able to accurately reflect parameters that we assumed in the ABM, it would be ideal to choose the same equations (at least very close) in both the utility function and regression function. For instance, we can modify Equation (6) to be:

$$u_{x,y}(t) = \exp[a_A(1 - A_{x,y})/2 + a_{SC}(1 - 1/SC_{x,y}(t))/2] \qquad (7)$$

At the same time, while conducting the survival analysis, rather then regressing against $A_{(x,y)}$ and $SC_{(x,y)}(t)$ directly, we can regress against $1\text{-}A_{(x,y)}$ and $1\text{-}1/SC_{(x,y)}(t)$, then the hazard model for regression will change from Equation (3) to:

$$h_{(x,y)}(t) = h_0(t) \exp[b_A(1\text{-}A_{(x,y)}) + b_{SC}(1\text{-}1/SC_{(x,y)}(t))] \qquad (8)$$

thus the α's in Equation (7) and β's estimated from Equation (8) will be totally comparable because they should be proportional (if $h_0(t)$ is a constant, which is commonly accepted in many survival analysis studies). The only assumption is that higher utility lead to higher hazards of development, which is quite reasonable. Doing so we can interpret the regression coefficients (β's) as the preference weights or a multiplication of the weights with some constants.

In the future, efforts should be directed towards the following issues based on the above discussion. First, we may vary the number of cells that each homebuyer can evaluate so that we can reduce or increase the uncertainty, and test how different models can capture the mechanisms assumed in the ABM model. Second, we may try different utility forms in the ABM, especially those discussed above. Third, in light of the fact that the location rules for service centers cause path-dependence, we can place more than one initial service center (say, each quadrant has one) or place the service center in different locations on the landscape, and explore if, and how, the service center(s) would affect development patterns over time and the associated regression results based on these patterns. We could also allow the service centers to use a different placement mechanism. Last, it is promising to test whether statistical models (especially survival analysis) can detect dynamic preference weights (i.e., $\alpha = f(t)$, or $\alpha = f$ (some covariates)).

The significance of this research goes beyond the above-mentioned perspectives, contributing to the methodology of land-change science. As mentioned in the introduction, ABMs can take a set of model rules on a virtual landscape, and thus can not only screen out some variables and relationships of no interests, but also relieve the researcher of the burden of collecting reliable data. For instance, to study the effects of censored data on regression results, we need both censored data (as we showed above: we were supposed to know only the status of each cell at a fixed interval) and non-censored data (i.e., we knew the exact time that the specific cell was developed), and then to compare what bias will come out due to different levels of data uncertainty. The SOME model can easily generate such information and facilitate further research.

Furthermore, the integration of agent-based modeling and regression analysis (survival analysis, in particular) can be used for a variety of other purposes: (1) Verify an ABM model. If the regression coefficients based on emergent outcomes are inconsistent with some rules in the ABM, for instance, there may be bugs in the ABM. (2) Use regression results to confirm plausible hypotheses. The researcher can fit his/her hypothetical relationships into an ABM, analyze the emergent outcomes using regressions, and decide whether the regression results agree with our theory, experience, or experts' opinion. If not, these hypothetical relationships may be questionable. (3) Conduct relationship-, heterogeneity- or scale-sensitivity analysis. If one relationship is removed from an existing ABM, one type of agents has assumed homogenous characteristics for some variable(s), one type of agents is scaled up to a higher level of agents, but the regression results based on the emergent data remain largely unchanged, then we may conclude that this relationship, the heterogeneity in this variable, or the low scale of agents in the ABM are insignificant and can be neglected.

Acknowledgements

We are indebted to Rick Riolo, Derek Robinson, and Moira Zellner for their very helpful assistance. We benefited from the financial support from the National Science Foundation Biocomplexity in the Environment program (BCS-0119804). Special thanks

go to the Center for the Study of Complex Systems at the University of Michigan for its computational resources and technical assistance.

References

Allison, P.D., 1995, *Survival Analysis Using SAS®: A Practical Guide* (Cary, NC: SAS Institute Inc).

An, L., Linderman, M. A., Shortridge A., Qi, J., Liu, J., 2005, Exploring Complexity in a Human-Environment System: an Agent-based Spatial Model for Multidisciplinary and Multi-scale Integration. *Annals of Association of American Geographers*, **95 (1)**, 54-79.

An, L., and Brown, DG (in review). Exploring temporal complexity in land-use and land-cover transitions: integrating survival analysis with GIS and remote sensing. Submitted to *Landscape Ecology*.

Arthur, W.B., 1999, Complexity and the economy. *Science*, **284 (2)**, 107-109.

Axelrod, R., and Cohen, M.D., 2000, *Harnessing Complexity: Organizational Implications of a Scientific Frontier* (New York: Basic Books).

Baltzer, H., 2000, Markov chain models for vegetation dynamics. *Ecological modeling*, **126(2-3)**, 139-154.

Brown, D.G., Pijanowski, B.C., and Duh, J.D., 2000, Modeling the relationships between land use and land cover on private lands in the Upper Midwest, USA. *Journal of Environmental Management*, **59**, 247-263.

Brown, D.G., Page, S.E., Riolo, R.L., and Rand, W., 2004, Agent Based and Analytical Modeling to Evaluate the Effectiveness of Greenbelts. *Environmental Modelling and Software,* **19(12)**, 1097-1109.

Brown, D.G., Page, S.E., Riolo, R.L., Zellner, M., and Rand, W., 2005, Path dependence and the validation of agent-based spatial models of land-use. *International Journal of Geographical Information Science,* **192 (2)**, 153-174.

Cantor, A.B., 2003, SAS® *Survival Analysis Techniques for Medical Research*, Second Edition (Cary, NC: SAS Institute Inc).

Cox, D.R., 1972, Regression models and life tables (with discussion). *Journal of the Royal Statistical Society,* **B34**, 187-220.

Epstein, J.M., 1999, Agent-based models and generative social science. *Complexity*, **4(5)**, 41-60.

Epstein, J.M., 2002, Modeling civil violence: an agent-based computational approach. *Proceedings of the National Academy Of Sciences of the United States of America*, **99**, 7243-7250 (Suppl. 3, May 14 2002)

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2000. *Quantitative Geography: Perspectives on Spatial Data Analysis* (London, UK: SAGE Publications).

Klein, J.P., and Moeschberger, M.L., 1997, *Survival Analysis: Techniques for Censored and Truncated Data* (New York. Springer-Verlag).

Kollman, K., Miller, J.H., and Page, S.E., 2003, *Computational models in political economy* (Cambridge, Massachusetts: MIT Press, 2003).

Lim K., Deadman, P.J., Moran, E., Brondizio, E., and McCracken, S., 2002, Agent-based simulations of household decision-making and land use change near Altamira, Brazil. In *Integrating geographic information systems and agent-based techniques*

for simulating social and ecological processes, ed. H. R. Gimblett, (New York: Oxford University Press), pp. 277-308.

Mertens, B., Sunderlin, W.D., Ndoye, O., and Lambin, E.F., 2000. Impact of macroeconomic change on deforestation in south Cameroon: integration of household survey and remotely-sensed data. *World Development,* **28(6)**, 983-999.

Mertens, B., and Lambin, E.F., 2000. Land-cover-change trajectories in southern Cameron. Annals of the Association of American Geographers **90(3)**: 467-494.

Müller, D., and Zeller, M., 2002. Land-use dynamics in the central highlands of Vietnam: a spatial model combining village survey data with satellite imagery interpretation. *Agricultural Economics*, **27 (2002)**, 333-354.

Parker, D.C., and Meretsky, V., 2004. Measuring pattern outcomes in an agent-based model of edge-effect externalities using spatial metrics. *Agriculture, Ecosystems and Environment,* **101 (2004)**, 233-250.

Rand, W., Zellner, M., Page, S.E., Riolo, R., Brown, and D.G., Fernandez, L.E, 2002. The complex interaction of agents and environments: an example in urban sprawl. *Proceedings of Agent 2002, Social Agents: Ecology, Exchange, and Evolution* (Chicago, 2002).

Rand, W., Page, S.E., Brown, D.G., Riolo, R., Fernandez, L.E., and Zellner, M., 2003. Statistical validation of spatial patterns in agent-based models. Proceedings, *Agent-based Simulations* (Montpellier, France, April 2003).