Spatial Econometric Models for Panel Data: Incorporating Spatial and Temporal Data

By

Christopher Frazier, Graduate Student Researcher, The University of Texas at Austin.
6.9 E. Cockrell Jr. Hall, Austin, TX 78712-1076, simplexr@yahoo.com

Kara M. Kockelman, Clare Boothe Luce Assistant Professor of Civil Engineering
The University of Texas at Austin, 6.9 E. Cockrell Jr. Hall, Austin, TX 78712-1076
kkockelm@mail.utexas.edu, Phone: 512-471-0210, FAX: 512-475-8744
(Corresponding Author)

ABSTRACT

Cities are constantly evolving, complex systems; and modeling them, both theoretically and empirically, is a complicated task. However, understanding the manner in which developed regions change over time and space can be of great importance for transportation researchers and planners. In this paper, methodologies for modeling developed areas are developed while incorporating spatial and temporal effects of the data. The work emphasizes spatial relationships between various geographic, land-use, and demographic variables characterizing fine zones across regions. It derives and combines land cover data for the Austin, Texas region from a panel of satellite images and U.S. Census of Population data. Models for population, vehicle ownership, and developed, residential, and agricultural land cover are estimated; and the effects of space and time on the models are shown to be statistically significant. Simulations of population and land cover for the year 2020 help to illustrate the strengths and limitations of the models.

INTRODUCTION

Urban systems are intricate, multifaceted and constantly evolving. Their evolution is dictated by a large number of influences, including public policy, individual preferences and actions, the physical landscape, technology and history. All of these factors (and more) interact in myriad ways. Discerning how and why urban systems evolve is, from the start, an extremely difficult task.

There is great benefit to uncovering the dynamics underlying urban systems. Understanding the ways in which geographic, economic, demographic, political and other factors interact is of interest to transportation engineers and land use planners, economists as well as historians, policymakers and the public. Models that reliably track these interactions are of great interest to transportation planners, as they illuminate how, among other things, policy impacts land use and travel patterns, welfare and development, congestion and air quality.

Parker, et. al. (2003) discussed the wide range of many land-use/cover change (LUCC) models recently developed. They pointed out that, due to the complexity of the systems encompassing land-use/cover, no single existing model is of more use than others; thus, a wide range of models, from the theoretical to the empirical, are being investigated by a variety of researchers (see, e.g., Candau (2002); Clarke and Gaydos (1998); Parker, Berger and Manson (2001)). In this paper, a closer connection between the real world and the model, as opposed to largely theoretical work, is sought. This parallels some recent models, developed for use by planning organizations for regional forecasting and policymaking. The regional models most similar to the work undertaken here are UrbanSim, What If?, and CUF2.

UrbanSim (Waddell 2002) micro-simulates the effects of location, land use, and policy decisions by households, workers, developers and policymakers on the land use patterns and rents across a region. Land use and development is modeled at the level of single parcels. Others are modeled at the level of user-defined grid cells which have no lower bound. Klosterman's (1999) "What if?" model of land use assigns land uses to a set of homogeneous zones in a bottom-up fashion, derived from socioeconomic, geographic, transportation and zoning information. Landis and Zhang's (1998) California Urban Futures 2 (CUF2) model employs multinomial models of land-use change per hectare (or other unit of observation) to predict future land use patterns.

One of the major drawbacks to many of these models is that they fail to incorporate and integrate the spatial and temporal correlations that are present in urban systems. That is, on an intuitive level, it would be expected that plots of land which are "close," in either spatial or temporal dimensions, would have more similarities which would influence or be representative of their characteristics than those which are "far" away. Whereas panel data techniques that account for temporal correlations are in widespread use, the methods described in Anselin (1988) and Elhorst (2003) – used to account for correlations, or more correctly the autocorrelations, in the spatial dimension –are less well known. There have been a variety of studies accounting for spatial autocorrelation. For example Case (1992) examined the influence of neighbors on technological changes on Indonesian farms, Coughlin et al. (2003) looked at the effect of spatial dependence on state lotteries in the U.S., and Dubin (1991) studied the spatial autocorrelations of residential neighborhood qualities. However, most studies incorporating spatial autocorrelation do not incorporate temporal correlations, and their focus is not aimed at transportation-based applications.

Researcher and planners would like to obtain as much information as possible from the spatial and temporal characteristics of the urban landscape. A primary goal of this work is to develop methodologies to analyze urban growth that account for such characteristics and are of interest to transportation researchers and planners. These models are tested empirically using land-cover data derived from satellite images coupled with U.S. Census data. The following sections detail the data sets and their development, the applied methodologies, and results for an Austin, Texas application.[1]

DATA DESCRIPTION

The data used in this work is drawn primarily from satellite and U.S. Census data, which, in their original form, are spatially and temporally incongruous. This section discusses these data sources, as well as the methods used to integrate them into a single data set. It should be

---

[1] For additional depth on statistical specifications, additional model formulations and results (including sample selection methods and differential equation model approximations), readers may consult Frazier (2004).

noted that the term "land cover" is throughout the text as opposed to the more common term "land use." This is essentially because the data derived from the visual/spectral qualities of the land, rather than information on the manner in which humans actually use it.

Land Cover Data Derived from Satellite Imagery

Satellite data offer excellent opportunities and considerable challenges. A serious and recurring problem for modeling land use has been the lack of spatially detailed data. Remote sensing, imaging technology, and geographical information systems (GIS) are making accurate land cover maps far more accessible to the researcher, and to the public. In particular, global satellite imaging, initiated in the early 1970s, provides highly detailed images regularly. And image analysis software can classify these by various general categories. GIS software combines data maps of various types, dramatically facilitating spatial analysis.

The United States launched LandSat 1 in 1972. Passing over Austin every 18 days, this early satellite provides images with 79 m × 79 m pixel resolution. LandSat 4 was launched in 1982, and resulted in 185 km × 185 km images with 30 m × 30 m resolution with a repeat orbit cycle of 16 days. 1984's LandSat 5 and 1999's LandSat 7 have essentially identical orbit and image characteristics to LandSat 4. These imaging systems work by scanning multiple passes (each representing one pixel) over an area and recording the reflectance of seven distinct spectral bands (Richards and Jia 2000); six of these bands record with 30 m × 30 m resolution, while the seventh, a thermal band, records with 120 m × 120 m resolution (60 m × 60 m for LandSat 7).

The land-cover data used in this work was derived from images taken by the LandSat 4, 5, and 7 satellite systems. Four images of Austin, Texas, taken in the years 2000, 1997, 1991, and 1983, were used. The image sections used are all 48.5 km × 55.8 km and have 30 m × 30 m resolution; each section thus contains just over three million pixels of data.

The derivation of land cover from the satellite images was achieved by a method called supervised image classification and was performed by University of Texas - Austin professor Dr. Barbara Parmenter and students in a graduate geography course. Supervised image classification basically uses the satellite image data from areas of known land cover to create a set of decision rules by which the rest of the image can be classified (Richards and Jia 1999). In the data used here, each satellite pixel was classified into one of nine land-cover types: water, barren, forest/woodland, shrubland, herbaceous natural/semi-natural, herbaceous planted/cultivated, fallow, residential, or commercial/industrial/transportation. In the context of this work, the second through fifth classifications are considered uninhabited land, the sixth and seventh are considered agricultural land, and the final two are developed land. Qualitative comparisons of the land cover classifications with aerial photography showed the results to be accurate, though no quantitative analysis of the quality of the classification was carried out. For more details concerning both the classification process and possible issues with the data, the reader is referred to Frazier (2004).

Derived Land Cover Data and Other Data Sources

Two spatial statistics were computed based on the land-cover data described above. These are land-cover mix and land-cover entropy. Land-cover mix (from here on called mix) characterizes the dissimilarity of the land-cover in a particular area: For a given pixel, mix is an index of adjacent pixels' dissimilarity; it measures the level of homogeneity between a central pixel's use type (x_0) and those of its neighbors (x_i) (Kockelman 1997, Cervero and Kockelman

1997). For this work, the neighborhood around a pixel was considered to be the eight pixels immediately surrounding it (see Figure 1). Mathematically, mix is defined by

$$\text{mix}(x_0) = \sum_{i=1}^{8} \frac{1 - \delta_{x_0, x_i}}{8}$$

(1)

where

$$\delta_{x_0, x_i} = \begin{cases} 1 & \text{if } x_i = x_0 \\ 0 & \text{otherwise} \end{cases}$$

(2)

As an average measure of dissimilarity, the mix index ranges from 0 to 1, with a higher numerical value corresponding to less similarity between a given pixel and its neighbors.

x_1	x_2	x_3
x_4	x_0	x_5
x_6	x_7	x_8

Figure 1 Reference diagram for pixel neighborhood used in calculation of land-cover mix statistic.

As a complement to mix, land-cover entropy (from here on called entropy) measures the level of land-cover variety of a particular neighborhood. Entropy is also called land-cover balance, and it essentially provides a measure for the level of heterogeneity of land-cover in the neighborhood (Kockelman 1997). Rather than comparing all the pixels in a neighborhood with the central one, as is done in mix, it instead compares all of the pixels with each other. If there are J possible land-cover types which a neighborhood may be made up of, then entropy is defined by:

$$\text{entropy}(x_i) = \frac{-1}{\ln(J)} \sum_{j=1}^{J} P_j \ln(P_j)$$

(3)

where P_j is the fraction of the neighborhood that is land-cover type j. Entropy also ranges from 0 to 1, with a higher value corresponding to a greater level of neighborhood land-cover heterogeneity. It equals 1 when all land cover types exist in a zone and all their proportions are equal (i.e., perfect "balance" in cover types). Because of this non-centralized nature of the statistic, it was calculated for 300 m × 300 m "neighborhoods" (which correspond to the combination grid cells as described in the next section) as opposed to the nine cell ones used for mix.

In addition to the land cover data and its derived statistics, Census of Population data was used. Statistics from both the 100%-sample Census (SF1) and the 17% sample (SF3) were used. These include population and household-level variables (such as household size and the number of vehicles per household). Data for Travis, Williamson, Bastrop and Hays Counties was collected so as to completely encompass the land-cover data region. Of course, the smallest areal unit for Census data is the block or block group, which typically encompass dozens of 30 m x 30 m pixel-based cells. So data had to be cleverly combined and then allocated to grid cells, as described in the following section. Finally, two Euclidean distance measures are used for analysis: distance to the central business district (CBD) and distance to the nearest highway.

Data Combination Methods
The fact that the years of the Census data do not align with the years of the satellite pictures, as well as the fact that the Census block groups do not line up with any grid system, necessitates the use of various methods to reasonably combine the data sets. That is, to use the various data sources collected for this work all together, the data must all be registered to the same temporal and spatial coordinate system.

To spatially combine the data, a grid that combines 100 of the pixel cells is used. This coarser grid is superimposed over the Census block groups, and the Census data allocated to each grid cell based on how much area each block group represents within the cell. For actual count variables, such as population, the fraction of the variable that corresponded to the fraction of the block-group in the cell was transferred; for the variables representing averages over the block-group, such as average household income, the transfer was done by (spatially) weighted summation of the Census values.

The new grid system has another benefit in that it reduces the large land cover data set. As noted earlier, each land cover data set has over 3 million pixels, which is an excessive amount of data, especially when compared to the thousand or so Census block groups. By using a combination grid whose cells are exactly ten pixels square (300 m × 300 m), the land cover data set was reduced by a couple orders of magnitude, while still retaining significant resolution of the region's land cover patterns. This coarsening of the grid system transformed the land cover data from a set of distinct, single-valued land cover types to a proportions data set (wherein each combination grid cell has a percentage of each land cover type associated with it).

In order to align the data sets temporally, an approximation method was applied to the Census data. Under the assumption that all Census variables roughly follow an exponential growth pattern with time, an approximation of the form

$$\bar{z}(t) = \alpha e^{\lambda t} \tag{4}$$

is used for each variable at an aggregate level, with parameters α and λ estimated using the 2000 and 1990 Census figures in a least squares framework and $\bar{z}(t)$ representing the average variable value in a grid cell at time t (the simple exponential form is motivated in Smith and Sincich 1992). Averages for off-Census years are then calculated and the values for the combination grid cells determined from (4) by using the deviations of each grid cell from the 2000 and 1990 means. That is, for each grid cell i, the value of the variable at time t is given by:

$$z_i(t) = \left(\frac{x_i(2000) + x_i(1990)}{\bar{x}(2000) + \bar{x}(1990)} \right) \bar{z}(t) \tag{5}$$

where $x_i(t)$ is the true Census level of the variable in grid cell i, and $\bar{x}(t)$ is the average across all grid cells.

METHODOLOGY
Spatial Linear Regression Model for Panel Data

The specification used for modeling continuous variables in this work's data sets is the panel-data spatial linear regression model. Examples of research using this model (though in different forms than that used in this work) include Dubin's (1991) study of residential home values and the study of national homicide rates in Messner and Anselin (2002).

In the context of this work, the general form of the model for an individual cell i (with N total cells and T total time periods) is:

$$y_{it} = \beta x_{it^*} + v_i + \theta_{it} \tag{6}$$

where y_{it} is the dependent variable at time t, v_i is an individual-specific effect assumed to be normally distributed with zero mean and variance σ_v^2, and x_{it^*} is a vector of exogenous explanatory variables, some of which may be time lagged (see Frazier (2004) for a discussion of exogeneity issues as they relate to this work). θ_{it} is an error term which, to capture spatial autocorrelation, is specified, in block matrix form, as follows (Anselin 1988):

$$\theta = \lambda W \theta + \xi \rightarrow \theta = (\mathbf{1} - \lambda W)^{-1} \xi \tag{7}$$

where ξ is a ($TN \times 1$) vector of which every element is distributed as Normal($0, \sigma^2$) and W is a ($TN \times TN$) block diagonal matrix with T copies of the ($N \times N$) spatial weight matrix $\widetilde{W}$ defined by:

$$\widetilde{W} = \begin{bmatrix} 0 & f(d_{12}) & \cdots & f(d_{1N}) \\ f(d_{21}) & 0 & \cdots & f(d_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ f(d_{N1}) & f(d_{N2}) & \cdots & 0 \end{bmatrix} \tag{8}$$

where $g(\cdot)$ is a function and d_{ij} is the distance between cells i and j. For this work, an inverse squared-distance measure was used in order to recognize greater autocorrelation present among cells close to each other, and a rapid reduction in such correlation with distance. Thus, the equation used is as follows (see Anselin 1988 for a discussion of other functional forms):

$$g(d_{ij}) = (d_{ij})^{-2} \tag{9}$$

To estimate the model parameters, a combination of feasible generalized least squares regression (FGLS) and maximum-likelihood estimation (MLE) can be used (Elhorst 2003). In the following derivation of the model, which closely follows Elhorst (2001 and 2003), it is first noted that the random effect can be realized as a variable-parameters model, with the constant variable, $X_{1it}=1$, having a variable coefficient $\beta_1 + v_i$. Furthermore, β is partitioned such that $\beta = [\beta_1, \beta_{-1}]$, the eigenvalues of $\tilde{W}$ are ω_i, the matrix of the eigenvectors of $\tilde{W}$ is Λ, and a parameter κ is defined such that

$$\kappa^2 = \frac{\sigma_v^2}{\sigma^2},$$ (10)

Moreover, R is defined as an $(N \times N)$ diagonal matrix whose ith diagonal element is given by $T\kappa^2 + (1 - \lambda\omega_i)^{-2}$. With these assumptions, the model's concentrated log-likelihood function is given by

$$\ln L = \frac{NT}{2}\left[\ln\left(NT/2\pi\right) - 1 - \ln\left(\sum_{t=1}^{T} c_t' c_t\right)\right] + \\ \sum_{i=1}^{N}\left[T\ln(1 - \lambda\omega_i) + \left(\tfrac{1}{2}\right)\ln\left(1 + T\kappa^2(1 - \lambda\omega_i)^2\right)\right]$$ (11)

where

$$c_t = (\mathbf{1} - \lambda\tilde{W})\left[Y_t - \overline{Y} - \left(\frac{1}{N}\sum_{i=1}^{N}\overline{y}_i\right)\iota - \left(X_t - \overline{X} - \left(\frac{1}{N}\sum_{i=1}^{N}\overline{x}_i\right)\iota\right)\beta_{-1}\right] + \\ R\Lambda\left[\overline{Y} - \left(\frac{1}{N}\sum_{i=1}^{N}\overline{y}_i\right)\iota - \left(\overline{X} - \left(\frac{1}{N}\sum_{i=1}^{N}\overline{x}_i\right)\iota\right)\beta_{-1}\right]$$ (12)

Here, Y_t is the $(N \times 1)$ vector of observed y values at time t, $\overline{Y}$ is the $(N \times 1)$ vector of time averages across Y_t, X_t is the $(N \times (K-1))$ matrix of exogenous variables minus the constant term, $\overline{X}$ is the $(N \times (K-1))$ matrix of time averages across X_t, and ι is an $(N \times 1)$ vector of ones. Equation 11 is called the concentrated log-likelihood function because β_1 and σ^2 have been factored out of the equation; they can be recovered by

$$\beta_1 = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i - \frac{1}{N}\sum_{i=1}^{N}\beta_{-1}'\overline{x}_i$$ (13)

and

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T} e_t' e_t$$ (14)

where

$$e_t = (1 - \lambda \widetilde{W})[Y_t - \bar{Y} - (X_t - \bar{X})\beta_{-1}] + R\Lambda[\bar{Y} - \beta_1 t - \bar{X}\beta_{-1}] \tag{15}$$

The e_t term in (14) and (15) is the vector of residuals or error estimates that correspond to (7)'s ξ term. To estimate the parameters λ, κ^2, and β_{-1}, a two-step iterative procedure is used (Elhorst 2003). First, values for λ and κ^2 are chosen, then β_{-1} is estimated using an ordinary least squares routine of X^* on Y^*, where both are stacked with elements given by

$$Y_t^* = (1 - \lambda \widetilde{W})\left[Y_t - \bar{Y} - \left(\frac{1}{N}\sum_{i=1}^{N}\bar{y}_i\right)\right] + R\Lambda\left[\bar{Y} - \left(\frac{1}{N}\sum_{i=1}^{N}\bar{y}_i\right)t\right] \tag{16}$$

and

$$X_t^* = (1 - \lambda \widetilde{W})\left[X_t - \bar{X} - \left(\frac{1}{N}\sum_{i=1}^{N}\bar{x}_i\right)\right] + R\Lambda\left[\bar{X} - \left(\frac{1}{N}\sum_{i=1}^{N}\bar{x}_i\right)t\right] \tag{17}$$

Next, given β_{-1}, λ and κ^2 are estimated using an MLE routine. The entire routine is iterated until suitable convergence is achieved.

Panel Data Spatial Logistic Regression Model
        Because it must lie within the [0,1] interval, fractional land-cover data should not be modeled using the spatial linear regression model described above. However, a modification of that model can be applied that allows for fractional response in a fairly straightforward manner. The technique used to model the proportion land-cover data is a new technique, representing an extension of the logistic regression method (see, e.g., Greene (2000)). This method models binary data, so it is applied here when modeling one land use type versus another (for example "developed" vs. "undeveloped"). Because of space considerations, many of the methodological details are not included here; those interested are referred to Frazier (2004).
        The technique begins by using the inverse of the logistic cumulative distribution function (CDF):

$$F^{-1}(P_{it}) = \ln\left(\frac{P_{it}}{1 - P_{it}}\right) \tag{18}$$

to transform the proportions data, P_{it}, to the $(-\infty, \infty)$ interval. This variable, with certain assumptions concerning the random effects term and appropriate corrections for heteroscedasticity (see Frazier (2004) for details), can be modeled using the panel data spatial regression technique described previously. Using earlier definitions for X, v, and ξ, the final model form is:

$$QF^{-1}(P) = QX\beta + v + (1 - \lambda W)^{-1}Q\xi \tag{19}$$

where Q is a variance-normalizing diagonal matrix defined as:

$$Q_{it,it} = \sqrt{F(x'_{it}\beta)(1 - F(x'_{it}\beta))} \tag{20}$$

with $F(\cdot)$ being the logistic CDF:

$$F(x'_{it}\beta) = \frac{e^{x'_{it}\beta}}{1 + e^{x'_{it}\beta}} \tag{21}$$

This technique works only for *binary* proportions data. That is, binary distinctions such as developed versus undeveloped can be modeled, but residential versus commercial versus undeveloped cannot. In order to distinguish more than two categories of land-cover, this technique may be performed iteratively. For example, if the residential proportion of cell i in time t is P_{it}^{Res}, and the developed proportion is P_{it}^{Dev}, then the quantity

$$P_{it}^{\text{Res}|\text{Dev}} = \frac{P_{it}^{\text{Res}}}{P_{it}^{\text{Dev}}} \tag{22}$$

can be modeled using the methods described above. However, because $P_{it}^{\text{Res}|\text{Dev}} \propto \left(P_{it}^{\text{Dev}}\right)^{-1}$, estimates of $\left(P_{it}^{\text{Dev}}\right)^{-1}$ from the developed versus undeveloped model results should be used to instrument the sub-model (i.e., this inverse probability should act as an explanatory variable), since leaving these out potentially would deprive the model of important information. Doing so requires further assumptions concerning the random effects term and a more complicated version of the variance-normalizing matrix (equation 20), but the method is essentially the same as described above (see Frazier (2004) for further details).

Time Adjustment

Because time differences between successive satellite images is not constant (one gap is three years, one is six years, and one is eight years), simply using time-lagged variables without accounting for this difference may lead to inaccuracies and/or misleading results. In order to account for this, a "time adjustment" factor is introduced for the coefficients of all time-lagged variables. If τ_t is the time difference between panel t and the previous panel, then an estimated parameter from the models representing explanatory variable k and time period t is transformed according to:

$$\beta_{k,t} \rightarrow \beta_{k,t}(a_k)^{\tau_t} \tag{35}$$

where a_k is the time adjustment factor. For variables that are not time lagged, a_k is equal to one; for time-lagged variables, a_k is estimated. To simplify estimation, a_k is assumed to be the constant across all time-lagged variables in each model.

MODEL RESULTS

In this section the results are presented for applications of the spatial panel data regression model as applied to population and vehicles per household variables; as well as for land cover (developed, residential developed, and agricultural undeveloped) as modeled by

spatial logistic models for panel data. Because of data set size, sampling had to be employed before model calibration; this technique is discussed as well.

Linear Regression Model for Spatial Panel Data

Two dependent variables are modeled using the spatial panel data linear regression model; they are population and the number of vehicles per household. Though not reported here due to space restrictions, models without lagged variables or time adjustment also have been estimated, and the results suggest that the models perform similarly, generally with only small changes for the effect of time lags and adjustment (Frazier 2004).

Due to computational demands in finding eigenvectors and eigenvalues of a spatial weight matrix involving all observations available (30,000 grid cells translates to size 30,000 × 30,000 matrices), cell sampling is used to reduce the burden. All results reported are the means from 25 models run on 25 random samples of 1,000 observations each. With the exception of the parameters relating to random effects and spatial autocorrelation, the means are consistent estimators of the population parameters (Greene 2000). The means of the standard errors and t-statistics are not consistent estimates of these secondary parameters, but they do provide an idea of statistical significance. At a 95% confidence level, some of the parameter estimates for some of the samples do not differ (in a statistical sense) from zero; however, these are still included in the final models (see Table 1) because in some of the samples they *were* statistically significant (i.e., t-statistic > 1.64) and because the only risk of leaving these variables in the model is possible model over-specification or "over-interpretation." Also reported are elasticities for the variables for the three years modeled (the final year, 1983, was dropped to permit use of time-lagged variables).

The random-effects and spatial autocorrelation parameters are specific to each random sample of 1000 observations, and this must be taken into account before using the results reported below for predictions or simulations., The reason for this is that the effects only account for the error terms from a random sample of observations, and not from the entire data set.

The population model uses the natural log of population as the response variable, in order to ensure non-negativity of predictions and to recognize the fact that population may have an exponential relationship with some or all of the independent variables (as with time, for example). The results of the spatial regression model with lagged independent variables and time adjustment are presented in Table 1. The distance measures are not time lagged because, at least in the scale of this work, they are time invariant. A square-root of the distance measure is used as an explanatory variable, since it is expected that there should be some added dampening of its effect. (For example, the effect on cell population of moving one kilometer away from the CBD is expected to be much more pronounced the closer that cell is to the CBD; intuitively, this is because the effect of a change in distance (to the CBD or nearest highway) matters at a relative, as opposed to an absolute, level (Frazier and Kockelman 2003).)

As expected, population is predicted to fall with distance to the CBD and rise with entropy and mix statistics, and with residential and commercial land coverage. Interestingly, it also is predicted to rise slightly with agricultural land coverage and with distance to the nearest highway (perhaps due to highway externalities, particular after having controlled for a distance-to-CBD variable, which may account for many network intensity effects). From the reported elasticities, it is evident that the distance to the CBD is the variable with the greatest impact on

the model, followed by the distance to the nearest highway. This indicates that the location of a cells, as opposed to its land cover levels, is the most important factor determining its population.

More importantly, it is seen that the parameters measuring the spatial autocorrelation (λ), random effects (κ), and the time adjustment of lagged variables are all highly statistically significant, as is the time adjustment factor (estimated to be 0.943). As expected, the effect of spatial autocorrelation is positive, which indicates that neighboring cells tend to have similar populations.

Table 2 presents the results from the vehicle ownership model (for average vehicles per household per zone). Ownership is estimated to increase with distance to the CBD, distance to the nearest highway, and land cover mix. It is estimated to fall rather quickly as the fraction of land in commercial use increases, as one might expect (since households may be smaller in more commercially developed locations and rely less on vehicles for access to commercial services and employment). It also falls slightly with residential and agricultural land coverages. Again, the parameters representing the effects of spatial autocorrelation and random effects are highly statistically significant. And, as with the population model, the time adjustment factor is estimated to be less than one, implying that the magnitude of the effects of past land cover on the present level of vehicle ownership decrease with time (see Frazier (2004) for a more detailed discussion of issues concerning and interpretations of the time adjustment factor).

Logistic Regression Model for Spatial Panel Data

Three models of land cover proportions, based on two binary-split levels (one conditioned on the other, for a total of four land cover classifications), were run using a logistic model for spatial panel data. The first split is for the fraction of land that is developed; the second conditions on that information for the proportions that are residential/non-residential (given the proportion that is developed) and agricultural/non-agricultural (given the proportion that is undeveloped). As in the estimation of the previously discussed models, separate models were run for each of 25 randomly selected sets of 1000 cell observations, and their estimates averaged. The same caveats discussed previously, concerning the primary parameter estimates, standard deviations, and t-statistics, hold here as well. Tables 3 through 5 present the average results from the three models. It should be noted that for the proportion of land that is residential given the proportion that is developed (hereafter called the residential model), three of the samples were thrown out because the maximum likelihood procedure's Hessian calculation failed (and extra-long estimation , of 5 to 12 hours per sample) prevented re-estimation of the models for these samples).

All of the models' random effects, spatial autocorrelation, and time adjustment parameters are statistically significant. Interestingly, the average level of spatial correlation is nearly identical for all three models, indicating that there may be similar unobserved spatial information across all models that is not being accounted for by the explanatory variables.

As expected, the fraction of land that is developed is predicted to fall with distance to the CBD, along with that that is residential in nature. Developed land is predicted to rise with distance to the nearest highway, however, probably to counter the effects of the increasing distance-to-CBD term. Residential land falls with this distance.

For the agricultural model component of this two-tiered model system, agricultural land will tend to lie farther from the CBD, but closer to highways, than non-agricultural, undeveloped land. It should be noted that the reported levels for the land cover mix and entropy variables for this model are heavily skewed by one of the 25 sample model results. Removing that sample's

results from the averages causes the mix and entropy parameters to not only be smaller in magnitude, but also to change sign, indicating that the sample may have introduced significant estimator bias. However, there is no mathematical or statistical reason to drop the sample from the averages, so it was left in (see Frazier (2004) for more information).

SIMULATIONS FOR PREDICTION

To test the practical performance of the estimated models, simulations were run to develop predictions for population and developed land cover for Austin's downtown in the year 2020. A 15 km × 15 km (or 50 cells × 50 cells) section of the CBD was selected for application, rather than the entire region, in order to economize on calculation times. (The results presented here took about 1 day each, for population and land cover predictions, due to the necessity of inverting a large number of 1,000 × 1,000 cell matrices.)

Though random effects were used to estimate the time-constant parts of the models, a method more akin to "fixed effects" is used to generate predictions for the population and developed land cover. The random effect method estimates the parameters for a normal distribution which best fits the time-constant parts of the dependent variables in the model. One natural method for using this information in a prediction would be to take a random draw from the estimated distribution for each cell to estimate the time-constant effect. However, this disregards important information contained in the data used to estimate the models. That is, it is possible to extract the exact time-constant effect for every cell from which the random-effect distribution is estimated. The estimate for the time-constant effect, denoted as v^*, is determined by using the estimates for the β coefficients vector, denoted as $\beta^*(t)$ (which, for simplicity, is assumed to include the time-adjustment factor from equation 35), in the following equation:

$$v_i^* = \bar{y}_i - \beta_1^* - \overline{x_i' \beta_{-1}^*(t)} \qquad (36)$$

Here the overbar indicates an average over time. To generate predictions, the above approximation of the random effect, along with the 2000 data, the coefficient estimates using the correct time-adjustment factor (to predict ahead the desired number of years), and an approximation for the random error, θ^* (discussed below) are used together:

$$y_i^{\text{Pred}}(t) = \beta_1^* + x_{i,2000}' \beta_{-1}^*(t) + v_i^* + \theta_i^* \qquad (37)$$

The process for simulating the spatial autocorrelation required maintaining the sampling strategy used in the model estimations. To begin with, a single random term for each cell in the entire data region was drawn from the distribution estimated in the models to generate the error terms, ξ, from equation 7. Then, 999 cells were randomly sampled from the entire data region for every one of the 2,500 downtown cells, and their sampled error terms, ξ, used to account for spatial autocorrelation. Because the calculation of the spatial autocorrelation for every cell required the inversion of a 1000 × 1000 matrix, the process of generating the predictions took a large amount of computing time, which is why the prediction region was limited to a 50 × 50 cell region.

The results of the population prediction, along with 2000 Census data for reference, are presented in Figure 2. Because of the logarithmic form of the model in which small errors can grow exponentially, some of the predictions are unrealistically large (over 3000 persons for a
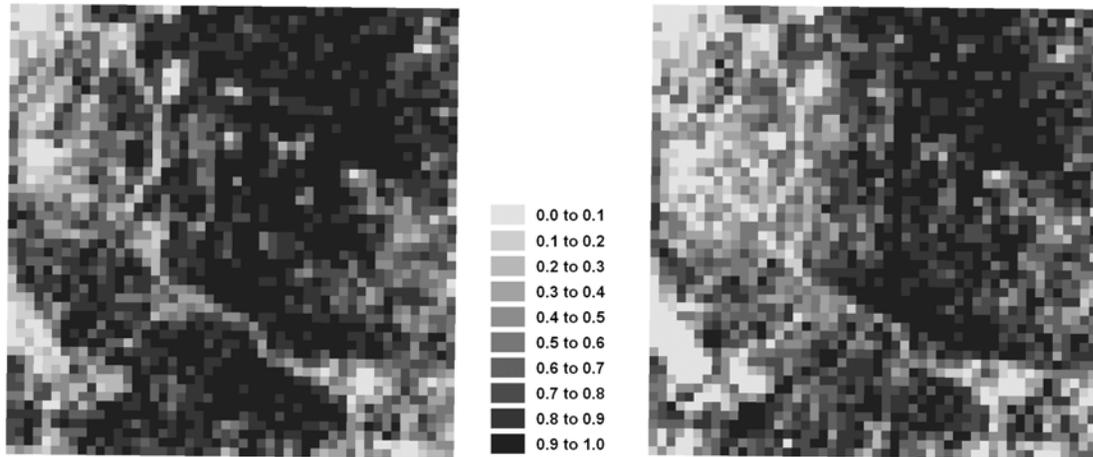
300 m x 300 m grid cell). Such predictions (there were 5 of these) were removed and replaced with averages of neighboring cell predictions for purposes of plotting the population simulation results. Quite clearly, the present distribution of downtown population in this 1.5 km x 1.5 km neighborhood is preserved in the 20-year predictions. However, the population is, in general, expected to *decrease* over the region (total population for the region dropped from 309,361 in 2000 to 239,892 in 2020). This indicates that though the model itself is able to account well for the population distribution in the region, the time dynamics of population change are not being correctly accounted for, at least with respect to predictions. Further analysis is required to determine why this non-intuitive result occurred, but most likely it is due to a misunderstanding of how the time adjustment factor actually affects the model. In future research, different ways to account for the time differences in the lagged variables will be examined.
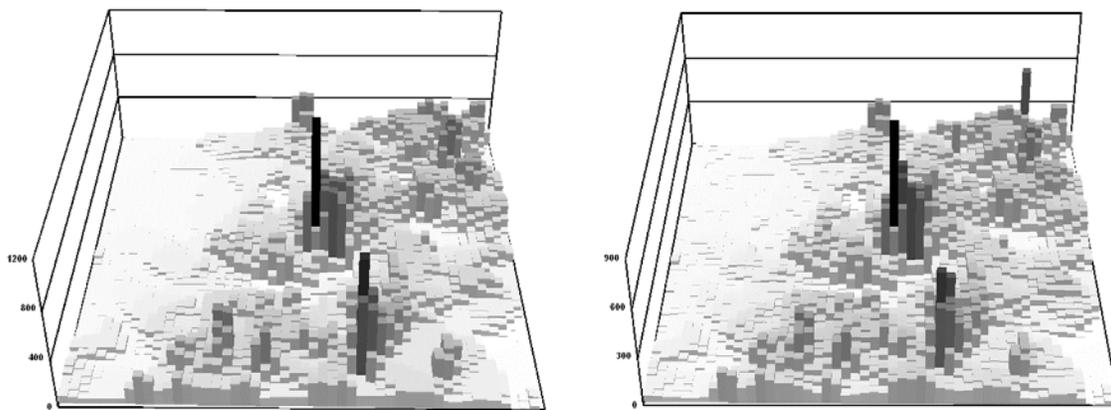
Figure 2. Population plot for downtown Austin: actual data from 2000 (left) versus average prediction for 2020 (right).
Note: Area of plot is 15 km x 15 km. Each plotted point covers 300 m x 300 m. Darker areas represent higher population levels.

The results of the developed land cover predictions, along with 2000 reference data, are presented in Figure 3. As with the population predictions, the distribution of the proportions of developed land cover are well maintained, but the effects of time seem to be incorrectly accounted for. That is, the proportion of developed land cover is expected to *decrease* from 2000 to 2020 (average developed proportion across the region dropping from 0.709 in 2000 to 0.630 in 2020). Again, this indicates that the way that time is accounted for in the model is flawed in some respect, and further analysis is required to find out exactly why this is happening.

	0.0 to 0.1
	0.1 to 0.2
	0.2 to 0.3
	0.3 to 0.4
	0.4 to 0.5
	0.5 to 0.6
	0.6 to 0.7
	0.7 to 0.8
	0.8 to 0.9
	0.9 to 1.0

Figure 3 Proportion of developed land cover in the downtown Austin area: actual data from 2000 (left) versus average prediction for 2020 (right).
Note: Area of plot is 15 km x 15 km. Each plotted point covers 300 m x 300 m; darker areas represent higher proportions of developed land cover.

From the simulation results it is obvious that using the models presented in this work do not perform well in a predictive capability. Though they capture the spatial distribution of the variables well, they do not account for the expected growth over time of population and urban development. As mentioned before, it is unclear why these results emerge as they do, but it is most likely due to inadequacies associated with the time adjustment factor. One possible cause is the fact that a single time adjustment factor was used, as opposed to having a separate one for each lagged variable.

Other issues also may have affected the predictions. An extension of this work might further investigate the methodology used to incorporate spatial autocorrelation in the predictions. Another issue is the fact certain, potentially important information was left out of the model; for example, population levels in the proportion of developed land cover model. Another would be accounting for the possibility that different model forms might exist for areas with different characteristics; e.g., distinct population models might exist for areas of high and low levels of development (see Frazier (2004) for examples of such models using sample selection methods).

Despite all of these issues, what the simulations do show is that the models, despite their flaws in the temporal dimension, perform very well in capturing the spatial diversity and distribution of variables across the region. Obviously future work on the models is required before they can actually be applied in a practical setting, but these results provide a promising start towards that end.

CONCLUSIONS

This paper presents a variety of innovative models for land cover and other data important for transportation engineers, geographers and planners. The work rigorously recognizes both space and time effects by incorporating spatial autocorrelation, temporal random effects, and adjustments for differences in time lags into linear regression and logistic regression model forms. Using both Census data and land cover data derived from satellite imagery, models for population, average vehicles per household, and developed, residential, and agricultural land cover are developed. Because of computational difficulties, a series of samples

were used for estimation. Not only were the results of the models informative, but the spatial and temporal effects were shown to be highly statistically significant, suggesting that their recognition and formal inclusion in the models is likely to be of great value. Positive spatial autocorrelation shows that, for example, areas of similar population or land cover proportions have a tendency to cluster. Also, the adjustment factor for the differences in time lags, though statistically significant, indicates that the effects of these differences are not that large (at least not in the time scale of the data).

In the estimated models, Census data is not used as explanatory information. The motivation behind this was that the potential error introduced by the approximation for non-Census years could cloud evaluation of model performance. Furthermore, a structural equations framework integrating the models is also not explored. Both of these issues would serve as interesting investigations for future research.

Applying the model results in a practical application (simulating population and developed land cover levels in 2020) exposes both strengths of the models and some potential problems. Specifically, the local spatial diversity of the region is accounted for fairly well in the predictions, however the effects of time on the region's development are not intuitively captured.

Notwithstanding the issues raised by prediction results, the models' ability to explore interesting aspects of the data and rigorously accommodate panel data and spatial interactions is of substantial value. They provide important information about relationships among demographic and geographic variables at both general and regional levels. This information can be of great use for transportation researchers and planners; it leads to an improved understanding of the interrelations which affect the development of urban regions which, in turn, can lead to more informed and improved policy decisions. Moreover, the statistical methodologies used in this work for spatial panel data analysis are largely new; they can be viewed as stepping stones towards models that more fully account for spatial and temporal heterogeneities and effects in transportation data. Though they suggest a need for future research (to more fully explore the power and practicality of these methods), the results are very promising.

REFERENCES

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Press.

Candau, Jeanette Therese. 2002. *Temporal Calibration Sensitivity of the SLEUTH Urban Growth Model*. Masters Thesis. University of California, Santa Barbara.

Cervero, Robert, and Kara Kockelman. 1997. Travel Demand and the Three Ds: Density, Diversity, and Design. *Transportation Research D* 2(3): 199-219.

Clarke, Keith C., and Leonard Gaydos. 1998. Loose-Coupling a Cellular Automaton Model and GIS: Long-Term Urban Growth Prediction for San Francisco and Washington/Baltimore. *International Journal of Geographical and Information Science* 12(7): 699-714.

Dubin, Robin A. 1992. Spatial Auto Correlation and Neighborhood Quality. *Regional Science and Urban Economics* 22: 433-452.

Elhorst, J. Paul. 2001. Panel Data Models Extended to Spatial Error Autocorrelation or a Spatially Lagged Dependent Variable. University of Groningen, Research Institute SOM Research Paper 01C05. Accessed March 1, 2004: http://www.ub.rug.nl/eldoc/som/c/01C05/01C05.pdf

Elhorst, J. Paul. 2003. Specification and Estimation of Spatial Panel Data Models. *International Regional Science Review* 26: 244-268.

Frazier, Christopher. 2004. *Spatial Econometric Models for Land Use/Land Cover Data: Theory and Application using Satellite Images for the Austin, Texas Region*. Masters Thesis. Department of Civil Engineering. The University of Texas at Austin.

Frazier, Chris, and Kara Kockelman. 2003. Cities and Satellite Imagery: Models for Regional Change. Presented at the 2004 INFORMS conference in Atlanta Georgia. Accessed March 3, 2004: http://www.ce.utexas.edu/prof/kockelman/public_html/TRB04SatData.pdf

Greene, William. 2000. *Econometric Analysis*. Upper Saddle River: Prentice-Hall.

Klosterman, R. E. 1999. What if?: Collaborative Planning Support System. *Environment and Planning B* 26: 393-408.

Kockelman, Kara M. 1997. Travel Behavior as a Function of Accessibility, Land Use Mixing, and Land Balance: Evidence from the San Francisco Bay Area. *Transportation Research Record* 1607: 117 – 125.

Landis, J. and M. Zhang. 1998. The Second Generation of the California Urban Futures Model: Part 1: Model Logic and Theory. *Environment and Planning B* 30: 657 – 666.

Messner, Steve, and Luc Anselin. 2002. Spatial Analyses of Homicide with Areal Data. Working paper. Accessed July 12, 2004: http://agec221.agecon.uiuc.edu/users/anselin/papers/smla.pdf

Parker, Dawn C., Steven M. Manson, Marco A. Janssen, Matthew J. Hoffmann, and Peter Deadman. 2003. Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review. *Annals of the Association of American Geographers*, 93(2): 314-340.

Parker, Dawn C., Thomas Berger, and Steven M. Manson, eds. 2001. Agent Based Models of Land-Use and Land-Cover Change: Proceedings of an International Workshop, October 4-7, 2001. CIPEC Collaborative Report CCR-3. Accessed July 10, 2004: http://www.csiss.org/events/other/agent-based/additional/proceedings.pdf

Richards, John A., and Xiuping Jia. 1999. *Remote Sensing Digital Image Analysis*. Berlin: Springer-Verlag.

Smith, Stanley K., and Terry Sincich. 1992. Forecasting State and Household Populations: Evaluating the Forecast Accuracy and Bias of Alternative Population Projections for States. *International Journal of Forecasting* 8: 495-508.

Waddell, Paul. 2002 UrbanSim: Modeling Urban Development for Land use, Transportation, and Environmental Planning. *The Journal of the American Planning Association* 68(3): 297-314.

Variable	Beta	S.E.	T-statistic	Estimation Sample Properties			Elasticities		
				Max	Min	Standard Error	2000	1997	1991
Constant	4.985	0.174	28.660	5.347	4.485	0.187			
Square root of Distance to CBD	-0.717	0.0347	20.689	-0.629	-0.782	0.0325	-1.416	-1.536	-1.726
Square root of Distance to Nearest Highway	0.125	0.0463	2.696	0.194	0.0523	0.0360	0.115	0.125	0.140
Proportion of Commercial Land Cover*	0.224	0.0652	3.439	0.345	0.100	0.0585	0.00600	0.00499	0.00580
Proportion of Residential Land Cover*	0.575	0.0444	12.997	0.786	0.404	0.103	0.0407	0.0236	0.00852
ln(Proportion of Rural Land Cover)*	0.00096	0.00163	1.030	0.00404	-0.00388	0.00181	-0.00120	-0.00150	-0.00144
ln(Land Cover Mix)*	-0.00751	0.00533	1.429	-0.00165	-0.0141	0.00367	0.00336	0.00368	0.00361
Land Cover Entropy*	0.232	0.0453	5.079	0.738	0.122	0.119	0.010	0.00748	0.00662
κ	8.360	0.232	35.993	9.538	4.455	0.956			
λ	5.363	0.032	212.524	6.794	3.940	0.747			
Time Adjustment	0.943	0.0108	5.235	0.985	0.876	0.027			
Error Variance	0.032								
Random Effect Standard Deviation	1.439								
R-Squared	0.5533								
Number of Valid Samples	25								

Table 1. Results for spatial linear regression model of $Y = \ln(\text{population})$.

Variable	Beta	S.E.	T-statistic	Estimation Sample Properties			Elasticities		
				Max	Min	Standard Error	2000	1997	1991
Constant	1.810	0.0358	50.640	1.846	1.740	0.0266			
Square root of Distance to CBD	0.0383	0.00700	5.484	0.0527	0.0290	0.00637	0.0887	0.0876	0.0855
Square root of Distance to Nearest Highway	0.00818	0.00917	1.137	0.0222	-0.0095	0.00848	0.00883	0.00873	0.00852
ln(Proportion of Commercial Land Cover)*	-0.170	0.0631	2.738	-0.0116	-0.359	0.0835	0.352	0.277	0.166
Proportion of Residential Land Cover*	-0.0101	0.00163	6.162	-0.00456	-0.0160	0.00313	-0.00409	-0.00487	-0.00523
Proportion of Rural Land Cover*	-0.00911	0.0281	0.720	0.0548	-0.0599	0.0288	-7.04E-04	-2.80E-04	-3.32E-04
Land Cover Mix*	0.145	0.0949	1.525	0.295	-0.0382	0.0894	0.0186	0.0103	0.00710
Land Cover Entropy*	-0.166	0.0829	1.946	2.98E-04	-0.432	0.107	-0.0622	-0.0768	-0.0837
κ	2.150	0.0640	33.583	2.588	1.954	0.152			
λ	5.283	0.0411	182.255	8.082	4.111	1.127			
Time Adjustment	0.871	0.0208	6.308	0.980	0.787	0.048			
Error Variance	0.0186								
Random Effect Standard Deviation	0.292								
R-Squared	0.9593								
Number of Valid Samples	25								

Table 2. Results for spatial linear regression model of Y = average number of vehicles per household.

Variable	Beta	S.E.	T-statistic	Estimation Sample Properties			Elasticities		
				Max	Min	Standard Error	2000	1997	1991
Constant	0.293	0.226	1.331	0.549	-0.0779	0.174			
Square root of Distance to CBD	-0.332	0.0403	8.254	-0.271	-0.404	0.0326	0.720	0.691	0.367
Square root of Distance to Nearest Highway	0.00736	0.0532	0.632	0.096	-0.103	0.0447	-0.00745	-0.00716	-0.00380
Land Cover Mix*	3.629	0.622	5.834	4.932	1.868	0.781	-0.697	-0.600	-0.308
Land Cover Entropy*	-1.732	0.555	3.007	-0.592	-9.207	1.638	0.105	0.086	0.0418
κ	0.0966	0.034	9.262	0.154	0.0558	0.0206			
λ	4.311	0.0812	59.814	4.808	3.892	0.286			
Time Adjustment	1.018	0.00938	1.789	1.083	0.966	0.0264			
Error Variance	1.555								
Random Effect Standard Deviation	0.387								
R-Squared	0.08999								
Number of Valid Samples	25								

Table 3. Results from panel data spatial logistic regression model run on land cover proportion variables: Proportion of developed land cover.

Variable	Beta	S.E.	T-statistic	Estimation Sample Properties			Elasticities		
				Max	Min	Standard Error	2000	1997	1991
Instrument Variable	-0.378	0.168	3.357	1.011	-1.895	0.618	-0.263	-0.252	-0.208
Constant	0.249	0.387	2.019	2.439	-1.718	0.953			
Square root of Distance to CBD	-0.0157	0.106	2.345	0.906	-0.845	0.363	-0.0158	-0.0140	-0.0113
Square root of Distance to Nearest Highway	-0.0522	0.0620	1.155	0.103	-0.229	0.0782	-0.0244	-0.0217	-0.0176
Land Cover Mix*	1.651	1.589	2.258	12.071	-6.460	4.247	0.135	0.0984	0.0729
Land Cover Entropy*	-0.928	0.991	1.483	2.951	-4.692	1.585	-0.0239	-0.0166	-0.0117
κ	0.0164	0.0705	2.108	0.0828	0.000	0.0205			
λ	4.512	0.0388	135.447	5.708	3.966	0.489			
Time Adjustment	0.990	0.0196	0.686	1.117	0.883	0.0675			
Error Variance	1.488								
Random Effect Standard Deviation	0.153								
R-Squared	0.1095								
Number of Valid Samples	22								

Table 4 Results from panel data spatial logistic regression model run on land cover proportion variables: Proportion of developed land cover that is residential.

Variable	Beta	S.E.	T-statistic	Estimation Sample Properties			Elasticities		
				Max	Min	Standard Error	2000	1997	1991
Instrument Variable	1.793	0.654	3.265	4.466	-3.340	1.592	0.622	0.525	0.429
Constant	-6.626	1.418	4.997	4.923	-12.638	3.491			
Square root of Distance to CBD	0.569	0.128	4.754	1.122	-0.439	0.303	-0.605	-1.436	-0.681
Square root of Distance to Nearest Highway	-0.00123	0.0745	0.878	0.208	-0.449	0.123	6.10E-04	0.00145	6.87E-04
Land Cover Mix*	-1.321	6.982	5.094	201.795	-20.174	42.563	0.105	0.190	0.0778
Land Cover Entropy*	-22.941	19.370	5.995	12.538	-729.427	147.206	0.577	0.989	0.385
κ	0.0385	0.0474	4.642	0.0881	0.000	0.0266			
λ	4.600	0.0399	125.784	5.898	3.971	0.551			
Time Adjustment	0.963	0.0130	2.768	1.076	0.393	0.124			
Error Variance	1.253								
Random Effect Standard Deviation	0.218								
R-Squared	0.204								
Number of Valid Samples	25								

Table 5. Results from panel data spatial logistic regression model run on land cover proportion variables: Proportion of undeveloped land cover that is rural.