

**APPLICATION OF THE DYNAMIC SPATIAL ORDERED PROBIT MODEL:  
PATTERNS OF LAND DEVELOPMENT CHANGE IN AUSTIN, TEXAS**

**Xiaokun Wang**

(corresponding author)

Assistant Professor

Department of Civil and Environmental Engineering

Bucknell University

Lewisburg, PA 17837, USA

(570) 577-1112

**Kara M. Kockelman**

Associate Professor & William J. Murray Jr. Fellow Department of  
Civil, Architectural and Environmental Engineering The University  
of Texas at Austin

6.9 ECJ, Austin, TX 78712-1076

kkockelm@mail.utexas.edu

The following paper is a pre-print and the final publication can be found in  
*Papers in Regional Science*, 88 (2):345-366, 2009.

# **APPLICATION OF THE DYNAMIC SPATIAL ORDERED PROBIT MODEL: PATTERNS OF LAND DEVELOPMENT CHANGE IN AUSTIN, TEXAS**

## **ABSTRACT**

The evolution of land development in urban area has been of great interest to policy makers and planners. Due to the complexity of the land development process, no existing studies is considered sophisticated enough. This research uses Dynamic Spatial Ordered Probit (DSOP) model to analyze Austin's land use intensity patterns over a 4-point panel. The observational units are 300m×300m grid cells derived from satellite images. The sample contains 2,771 such grid cells, spread among 57 zip code regions. The estimation suggests that increases in travel times to CBD substantially reduce land development intensity. More important, temporal and spatial autocorrelation effects are significantly positive, showing the superiority of the DSOP model.

**KEY WORDS:** spatial autocorrelation, temporal dependency, ordered discrete response data, land development

## **1. BACKGROUND**

In studies of social behaviors and human activities, many choices or attributes (e.g., religious beliefs, presidential election outcomes, and levels of crime) involve discrete responses in a temporal and spatial context. It is especially true for analysis of dynamics in land development intensity levels under the influence of geology, demographics, transportation conditions and other, socio-economic factors: land owners make development decisions based on their knowledge and prediction of neighboring land development (See, e.g., Waddell, 2002, and Candau et al., 2000). As a result, land development is often clustered. For example, one can expect that a parcel of land is more likely to be intensely developed if its neighborhood offers intensely developed land. Wang and Kockelman (2008) have developed a dynamic spatial ordered probit (DSOP) model aiming to analyzing the temporal and spatial relationships in ordered categorical data. This paper demonstrates how this model can be applied to the analysis of land development change.

The analysis relies on Austin, Texas data sets. Thanks to rapid population growth and economic expansion, the area has experienced some dramatic changes during the last two decades. As will be shown in more detail in the following sections, during this time period, the region's land development has both sprawled over space and escalated in intensity. One direct result of this development is congestion. The Texas Transportation Institute's urban mobility report (Schrank and Lomax, 2005) indicates that Austin ranks number 1 among all 30 medium-sized U.S. cities, in travel delay and wasted fuel per capita.

In this study, land development intensity is defined based on how much land is covered by manmade materials, which are characterized by higher reflectance levels and other visual clues provided via satellite images. These "intensity levels" are indexed as integers, and their order is key. This application, in addition to disclosing the spatial and temporal patterns of urban land

development change, illustrates the potential broad application of the dynamic spatial ordered probit model.

For urban areas, the evolution of land development intensity is a topic of interest to traffic demand modelers, policy makers, and land developers. Such changes influence regional economies and environmental conditions. For non-urban areas, analyzing the dynamics of land development intensity is also important: For example, undeveloped land around the world, including some precious lands like the Amazon rainforest, are being converted for agriculture and other human uses. Such changes can significantly contribute to climate change, desertification, resource depletion and loss of habitats and species.

Many studies have been conducted on the land change patterns. However, none of these models recognizes spatial or temporal autocorrelation in a statistically rigorous manner. In fact, many studies that recognize spatial effects have tried to either construct and control for a variety of neighborhood attributes or remove all spatial correlation through strategic sampling (to provide a dispersed sample, with minimal interactions). Some also attempt to recognize temporal dependencies by controlling for variables from previous periods. For example, Nelson and Hellerstein (1997) sampled selectively and created exogenous variables based on neighboring units' land cover data in order to study the deforestation effects of roadways via a multinomial logit model. Wear and Bolstad (1998) controlled for prior land uses in the neighborhood of each data cell in their study of southern Appalachian landscapes, which involved binary response data. Munroe et al. (2001) attempted to filter out spatial correlations through sampling and then removed the residual spatial dependence through a "trend surface" approach (Cliff and Ord, 1981). As with all other existing models dealing with discrete response data in a temporal and/or spatial context, the applicability of these methods is still limited because of the neglect of spatial effects (even intentionally) and data dynamics.

The following sections first introduce the specification and estimation of the DSOP model, then describes the datasets used in this study. The effects of different factors on land development intensity are discussed based on the estimation results. The estimates also are applied to predict land development intensity levels in the study area.

## 2. MODEL SPECIFICATION AND ESTIMATION

Wang (2007) and Wang and Kockelman (2008) discuss the DSOP's model specification and estimation in detail. This following discussion summarizes DSOP methodology and highlights key findings. In short, a dynamic ordered probit model with spatial and temporal autocorrelation can be described by extending existing specifications of spatial probit models, static and dynamic, ordered and categorical. The most closely related works are those by Wang (2007), Wang and Kockelman (2008), Smith and LeSage (2004) and Girard and Parent (2001).

The model specification is as follows:

$$U_{ikt} = \lambda U_{ikt-1} + \mathbf{X}_{ikt}' \boldsymbol{\beta} + \theta_{it} + \varepsilon_{ikt}, \quad t = 1, \dots, T \quad (1)$$

where  $i$  indexes regions ( $i = 1, \dots, M$ ),  $k$  indexes individuals inside each region, or neighborhood (i.e.,  $k = 1, \dots, n_i$ ), and  $t$  indexes time periods. In other words, there are  $M$  regions/neighborhoods, each containing  $n_i$  observations; so that the total number of observations is  $\sum_{i=1}^M n_i = N$ . In addition,  $\lambda$  is the temporal autocorrelation coefficient to be estimated. Each individual is observed  $T$  times, so the total number of observations is  $NT$ .  $U_{ikt}$  is a latent (unobserved) response variable for individual  $k$  from region  $i$  at period  $t$ .  $\mathbf{X}_{ikt}$  is a  $Q \times 1$  vector of explanatory variables, and  $\boldsymbol{\beta}$  is the set of corresponding parameters.  $\theta_{it}$  captures all common yet random components for observations within region  $i$  in period  $t$ , while remaining random information is captured by individual effect  $\varepsilon_{ikt}$  which is heteroscedastic with variance  $v_i$ .

This specification allows the model to reflect spatial autocorrelation across regions while recognizing intra-regional clustering. A spatial autoregressive process can be formulated here, as follows:

$$\theta_i = \rho \sum_{j=1}^M w_{ij} \theta_j + u_i, \quad i = 1, \dots, M \quad (2)$$

where weight  $w_{ij}$  reflects proximity, and can be derived based on contiguity and/or distance between regions. The magnitude of overall neighborhood influence is reflected by  $\rho$ , also called the spatial coefficient.  $u_i$  aims to capture any regional effects that are not spatially distributed, and is assumed to be iid normally distributed, with zero mean and common variance  $\sigma^2$ . The vector of regional effects will be a function of the weight matrix  $\mathbf{W}$ , which is composed of purely exogenous elements  $w_{ij}$ .

The use of such regional effects to capture certain spatial dependencies also enhances computational efficiency: normally, the number of regions is much lower than the total number of observations, allowing use of a weight matrix  $\mathbf{W}$  of relatively low rank. Thanks to a lower dimension, the inversion of  $\mathbf{W}$  and calculation of its eigenvalues are much less memory-intensive. Of course, both of these computations are necessary for parameter estimation. Furthermore, the specifications shown here allow the special case of every individual serving as a separate region, where  $n_i = 1, \forall i \in M$  (and  $M = N$ ). In this context, all individuals can be spatially autocorrelated without imposing regional boundaries. While computational burdens will increase, this approach certainly is feasible, assuming a reasonable sample size.

Equation (1)'s recursive time-space form implies that current response values depend on previous period values, along with various contemporaneous factors (Anselin, 1999). Furthermore, after controlling for all such temporally lagged and contemporaneous variables, the residuals remain spatially autocorrelated. The context of land development intensity levels fits this specification: land development depends on past and present conditions, including owner/developer experiences of local and regional congestion and population, as well as nearby development and variables like school access.

For the case of an ordered probit specification, the observed response variable,  $y_{ikt}$ , is as follows:

$$y_{ikt} = s \text{ if } \gamma_{s-1} < U_{ikt} < \gamma_s \text{ for } s = 1, \dots, S \quad (3)$$

That is, the observed variable is a censored form of the latent variable, and its possible outcomes are integers ranging from 1 to  $S$ . The latent variable  $U_{ikt}$  is allowed to vary between unknown boundaries  $\gamma_0 < \gamma_1 < \dots < \gamma_{S-1} < \gamma_S$ , where  $\gamma_0$  is  $-\infty$  and  $\gamma_S$  is  $+\infty$ . If constants are to be included in the explanatory variables,  $\gamma_1$  also is normalized to equal zero. The probabilities for these  $S$  outcomes are as follows:

$$\begin{aligned} \Pr(y_i = 1 | \mathbf{X}_i) &= \Phi(\gamma_1 - \mathbf{X}_i' \boldsymbol{\beta}) - \Phi(\gamma_0 - \mathbf{X}_i' \boldsymbol{\beta}) \\ \Pr(y_i = 2 | \mathbf{X}_i) &= \Phi(\gamma_2 - \mathbf{X}_i' \boldsymbol{\beta}) - \Phi(\gamma_1 - \mathbf{X}_i' \boldsymbol{\beta}) \\ &\vdots \\ \Pr(y_i = S | \mathbf{X}_i) &= \Phi(\gamma_S - \mathbf{X}_i' \boldsymbol{\beta}) - \Phi(\gamma_{S-1} - \mathbf{X}_i' \boldsymbol{\beta}) \end{aligned} \quad (4)$$

where  $\Phi(\bullet)$  is the Cumulative Distribution Function (CDF) for standard normal distribution.

The resulting likelihood function is as follows:

$$\Pr(\mathbf{y} | \mathbf{U}, \boldsymbol{\gamma}) = \prod_{t=1}^T \prod_{i=1}^M \prod_{k=1}^{n_i} \sum_{s=1}^S \delta(y_{ikt} = s) \cdot \delta(\gamma_{s-1} < U_{ikt} < \gamma_s) \quad (5)$$

where  $\mathbf{y}$ ,  $\mathbf{U}$  and  $\boldsymbol{\gamma}$  are the vector forms of  $y_{ikt}$ ,  $U_{ikt}$  and  $\gamma_s$ .  $\delta(A)$  is an indicator function equaling 1 when event  $A$  is true (and 0 otherwise).

It should be noted that this model specification pivots off of existing functional forms, but is not a simple extension of current models. The simultaneous consideration of multiple discrete responses, temporal dependencies and spatial autocorrelation in the dataset present a significant innovation.

Allowance of all these features should make the model more statistically reliable in mimicking and forecasting the temporal and spatial evolution of ordered response variables, like land use intensity, tree cover, and home safety ratings. In addition, modifications of this model to allow for irregularly spaced panel data sets may be very helpful for practice, along with new methods for allowing non-sparse weight matrices and large sample sizes. Another important opportunity for exploration is prediction out of sample, which is rarely pursued in academic work but often critical in practice.

Estimation of the dynamic spatial ordered probit model is achieved in a Bayesian framework where each parameter has prior and posterior distributions. The posterior distributions are consistently derived using MCMC methods (Gelfand and Smith 1990), by sampling sequentially from the complete set of conditional distributions.

As Wang and Kockelman (2008) describe, most of the parameters follow standard distributions and can be conveniently generated using routines built in commercial mathematical packages

(such as Matlab and Gauss). The spatial coefficient  $\rho$ , however, follows a non-standard posterior distribution and has to be generated using numerical methods. The threshold parameter  $\gamma$  follows a multidimensional truncated normal distribution and the truncations co-vary. Therefore, the marginal distribution of each element in  $\gamma$  also is expected to be non-standard.

### 3. DATA DESCRIPTION

The data used for land development dynamics comes from multiple data sources, including satellite images, the Census of Population, City of Austin school district and employment data, as well as transportation and geographic data from the Capital Area Council of Governments (CAPCOG). The land cover information serves as the dependent variable, and all others serve as explanatory variables. These include total neighborhood population, number of workers living in the neighborhood, average household income and number of schools in the neighborhood, travel time to the nearest major highway (including U.S. Highway 290, U.S. Highway 79, U.S. Highway 183, State Highway 71, Interstate 35, Loop 1, and Loop 360, which did not change from 1983 to 2000), travel time to the region's CBD, travel time to major (Austin's 15 biggest) employers, travel time to the nearest airfield, average ground slope, and average elevation (of each 300m×300m grid cell). A set of rather standard routines were followed to integrate the various databases, as described in Wang (2007).

An interesting and important part of the data processing involves definition of “regions” and selection of cell samples. As discussed above, observations in the same region should share common latent features. In ecological and environmental studies, regional boundaries may derive from natural spatial partitions, such as rivers and mountain ranges, with observations in the same region sharing vegetation and micro climate. For human activities, boundaries are more likely to be administrative units, across which policies and practices can change, such as zoning and school administration.

In Austin's urban area, zoning is based on neighborhood planning areas (NPAs). Changes in zoning constraints often occur across these boundaries. However, information for many interesting variables is often organized based on zip codes. In order to be consistent with existing spatial units, study regions were based on 57 zip codes. These zip codes tend to align nicely with a single NPA, or the union of 2 to 4 NPAs. There are 57 of these, offering interesting regional variation while keeping computational burdens reasonable.

After defining these regions, the next step was to select observations (grid cells) in each region. Of course, one can use all 29,946 300m grid cells as observations. However, there are good reasons for selecting only a subset of these. First, the “boundary” of a region may be somewhat ambiguous and the differences between regions may be slight. If all grid cells are used, cells that are located in two different regions yet lie in close proximity may be more similar than grid cells that are far away from each other yet belong to the same region. The second reason is computational: 29,946 grid cells create a very large pool of observations with difficulty in parameter estimation due to large-matrix inversion for spatial covariance components. A 10% sampling rate ( $\sum n_i = 2,995$ ) is likely to return satisfactory estimation results with significantly reduced computation time and so was used here.

In order to ensure that observations in the same region are more alike than those in other regions, samples were selected around regional (zip code area) centroids. In this way, observations in the same region are spatially clustered (all contiguous), and thus expected to be more similar to each

other than to observations in other regions. Second, in order represent the entire study area, samples should be distributed as evenly as possible across space. If an equal number of 300m cell observations is selected in each region, smaller regions will get more weight (than they “deserve”) in the sample. In order to spatially balance the selection, the number of observations in each region was set proportional to region area. Finally, 224 sampled cells were removed, because, in the case of very narrow zip codes, they extended into neighboring regions or fell along edges of the study area (so neighborhood information could not be obtained), while others exhibited unrealistic elevation and slope values (Caused by missing value in the CAPCOG dataset). A total of 2,771 observations (per year) resulted from this processing. These observations are distributed across the 57 regions (zip code areas), with the number of grid-cell observations per region ranging from 2 to 333.

Table 1 summarizes definitions of all these variables, and Table 2 summarizes their statistics. Table 2’s statistics show trends that are expected: development intensity levels, population, number of workers, and average household income have all increased over time. Average (uncongested) travel times to major facilities and employers have fallen, thanks to road system expansions in peripheral zones.

**Table 1 Data Description for Land Development Intensity Level Analysis**

Variable	Description
INTLV	Development intensity level
ELEVTN	Average elevation of the 300m grid cell (km)
SLOPE	Average slope of the 300m grid cell (%)
NSCHOOL	Number of K-12 schools in the neighborhood
POP	Population (thousand) in the neighborhood
WORKER	Number of workers (thousand) living in the neighborhood
INC	Average household income (thousand dollars) in the neighborhood
EMPTT	Travel time to nearest major (top 15) employer (hours)
CBDTT	Travel time to CBD (hours)
AIRTT	Travel time to nearest airfield (hours)
RDTT	Travel time to nearest highway (hours)

**Table 2 Summary Statistics for Land Development Intensity Analysis**

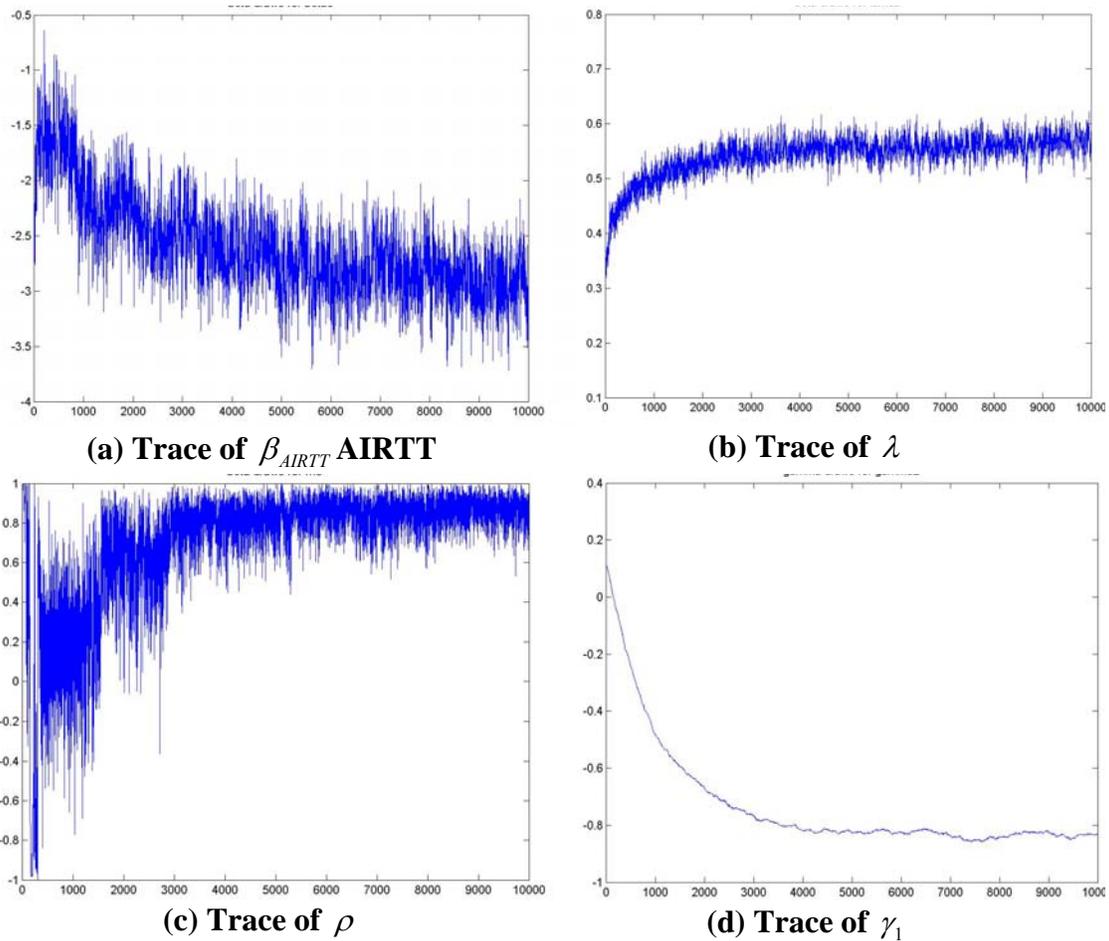
	Variable	Minimum	Maximum	Mean	Std. Deviation
Constant through Years	ELEVTN	0.136	0.390	0.251	0.061
	SLOPE	0.034	17.328	2.699	2.196
	NSCHOOL	0.000	7.000	1.208	1.377
1983	INTLV	0.000	3.000	0.826	0.774
	POP	0.225	37.531	4.632	7.298
	WORKER	0.121	19.997	2.408	3.918
	INC	17.330	88.941	45.368	15.109
	EMPTT	0.004	1.115	0.453	0.223
	CBDDT	0.000	0.358	0.154	0.070
	AIRTT	0.005	0.784	0.345	0.157
	RDDT	0.002	0.498	0.111	0.093
1991	INTLV	0.000	3.000	0.948	0.874
	POP	0.203	51.310	6.860	10.424
	WORKER	0.121	27.633	3.624	5.652
	INC	20.540	105.412	53.844	17.766
	EMPTT	0.004	0.733	0.298	0.149
	CBDDT	0.000	0.339	0.148	0.068
	AIRTT	0.004	0.630	0.259	0.120
	RDDT	0.002	0.430	0.092	0.082
1997	INTLV	0.000	3.000	1.300	0.827
	POP	0.389	64.873	8.007	12.615
	WORKER	0.211	35.220	4.240	6.900
	INC	23.332	119.738	61.077	20.341
	EMPTT	0.001	0.313	0.112	0.060
	CBDDT	0.000	0.308	0.142	0.065
	AIRTT	0.004	0.628	0.227	0.116
	RDDT	0.002	0.385	0.086	0.074
2000	INTLV	0.000	3.000	1.359	0.929
	POP	0.478	64.629	9.131	13.153
	WORKER	0.238	36.238	4.836	7.278
	INC	15.869	125.094	65.024	22.635
	EMPTT	0.001	0.182	0.070	0.037
	CBDDT	0.000	0.266	0.126	0.057
	AIRTT	0.005	0.437	0.154	0.070
	RDDT	0.002	0.251	0.054	0.044

#### 4. MODEL ESTIMATION

This section applies the DSOP model (see Wang and Kockelman [2008] for more discussion on the model specification and estimation techniques) to the land development intensity levels analysis. As noted in the section of “Model Specification and Estimation”, explanatory variables for both analyses include temporally lagged latent variables and various contemporaneous variables. The following sections discuss the model estimation and results. First, the number of burn-in samples is determined. Estimate means, standard deviations, posterior distributions, and their marginal effects are then calculated and discussed. The performance of DSOP model with this dataset is also compared to those with simpler models. Finally, model estimates are used to predict response variables’ values under hypothetical scenarios. The predictions can be visualized via a “most likely” result and an “uncertainty index.”

Figure 1 shows several typical estimation traces (convergence patterns) for parameters in the development intensity model. These patterns are representative, and the traces of other parameter estimations are all similar to them. Rigorous proof of convergence is a complicated topic, so here “convergence” is based on the trace of variable estimates. If, after a certain number of iterations, parameter estimates stabilize, the estimation is assumed to have converged. Results of iterations before this turning point are omitted and all inferences are drawn based on the converged iterations.

The model begins with diffuse priors and iterates 10,000 times. As observed in Figure 1, different parameters start “converging” after different numbers of runs. However, after 6000 runs, all traces appear stable, indicating an overall model convergence. Hence, the first 6000 runs were omitted (as a “burn-in” sample), and the model uses the next 4000 draws to estimate parameter means and standard deviations, as shown in Table 3.



**Figure 1 Convergence Patterns of Development Intensity Level Estimation**

According to the results, neighborhood population and worker counts do not have statistically significant impacts on land development intensity levels. Average household income, by contrast, appears to generally boost such levels. Distances to major employers, Austin’s CBD, and the nearest airfield all have statistically and practically significant effects on land development: the farther the cells lie from these attractions, the less likely they are to develop intensely. Interestingly, Euclidean distance to nearest highway is estimated to have a negative marginal effect on intensity, implying that (in the study area) development is more likely to occur at locations far from major roads. Considering that distances to the CBD and major employers already have been controlled for, this result can be interpreted as such: after access to work and the region’s core are determined, developers tend to choose locations some distance away from the highway (and its noise, pollutants and safety issues). The result also suggests that locations with more neighborhood schools are more likely to be intensely developed while elevation is not a statistically influential factor, locations with steeper slopes are less attractive to land development.

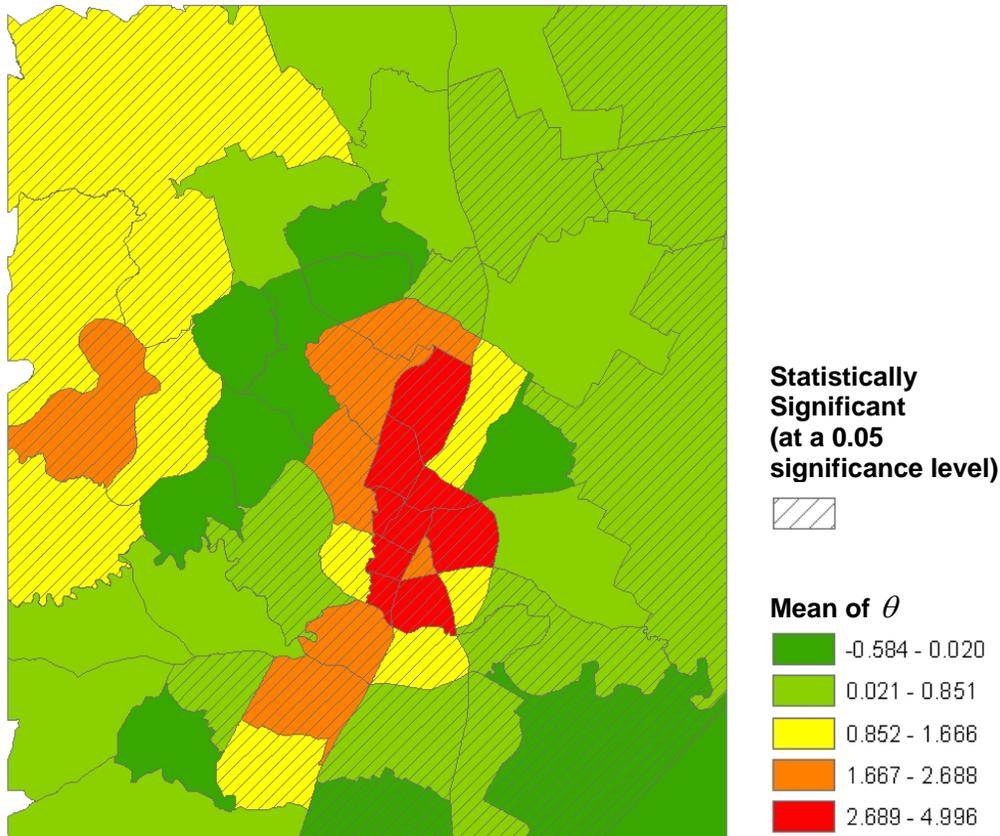
**Table 3 Estimation Results for Model of Land Development Intensity Levels**

Variable	Mean	Std. Dev.	t-stat.
POP	-0.024	0.036	-0.668
WORKER	0.089	0.067	1.327
INC	0.019	0.002	9.143
EMPTT	-0.232	0.130	-1.778
CBDTT	-4.365	0.851	-5.126
AIRTT	-2.867	0.248	-11.550
RDTT	2.309	0.385	6.001
NSCHOOL	0.039	0.017	2.305
ELEV	-0.239	0.696	-0.343
SLOPE	-0.034	0.010	-3.394
$\lambda$	0.561	0.019	30.005
$\rho$	0.857	0.074	11.612
$\sigma^2$	0.871	0.222	3.931
$\gamma_1$	-0.834	0.011	-77.231
$\gamma_2$	2.235	0.031	71.393
$\gamma_3$	4.361	0.034	130.167

Unlike slope coefficients in a standard linear model, beta values in a model involving latent response cannot be interpreted so directly. Moreover, as Greene (2005) explains, parameter signs in a model of ordered categorical response only indicate changes in likelihood of the two extreme outcomes ( $y = 1$  and  $4$ ), rather than shifts in all outcomes. Section 5.1 of this paper quantifies the marginal effects of all control variables.

Another important estimation result is the practical and statistical significance of both the temporal autocorrelation coefficient ( $\lambda$ ) and the spatial autocorrelation coefficient ( $\rho$ ). These suggest that prior-period information has a very important influence on the (current) latent variable's value (mean  $\lambda=0.561$ ) and that, even after controlling for various neighborhood characteristics, residuals remain strongly and positively correlated across space (mean  $\rho=0.857$ ). These results support the notion that land development decisions depend heavily on neighboring conditions, and that spatial relationships should be reflected in model specification.

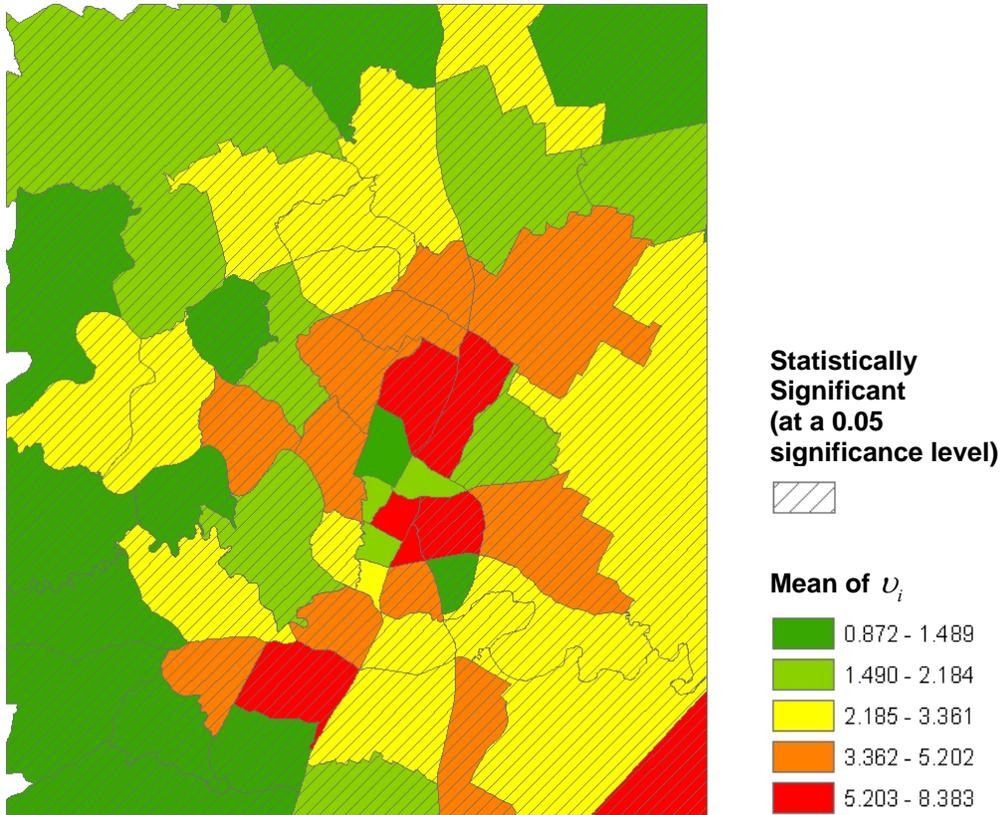
As a further confirmation, the mean values of regional specific error ( $\theta_i$ ) estimates (and their statistical significance) are shown in Figure 2. A clustering pattern (where similar values tend to co-locate, rather than lie randomly distributed across space) is clearly visible in this figure, so the spatial autocorrelation of these regional-specific error terms was tested using Moran's I (Moran, 1950), in ArcMap. It should be noted that the weight matrix used in ArcMap is based on the inverse of distance, not the contiguity approach used in this study. And, or course, methodologically, Moran's I is quite different from a Bayesian approach. Therefore, any similarity with  $\rho$  may be limited to signs and general statistical significance.



**Figure 2 Distribution of Region-Specific Error Term Estimates ( $\theta$ ) for Land Development Intensity Levels**

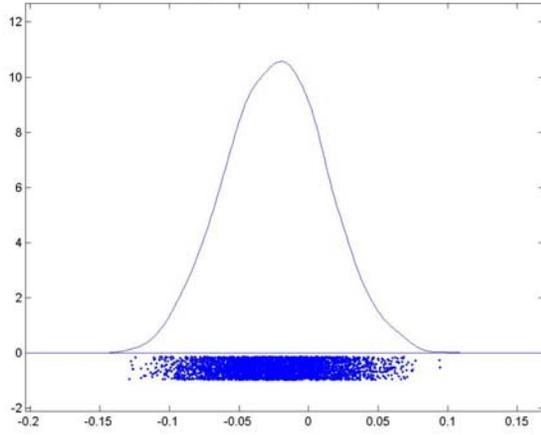
As expected, the Moran's I test results in ArcMap indicate clustering (i.e., positive spatial autocorrelation) of the  $\theta$  values, across space. (Moran's I value is very high: 0.56 with a Z score of 6.7.)

Figure 3 shows the estimation results for variances of these individual specific errors ( $\nu_i$ ). Except for downtown regions, where only a few grid cell observations exist per region, all variance estimates are statistically significant. "City edges" (i.e., areas between Austin's central, highly developed area and the outer, less developed areas) tend to have larger variances because these are where new developments are most likely to occur.

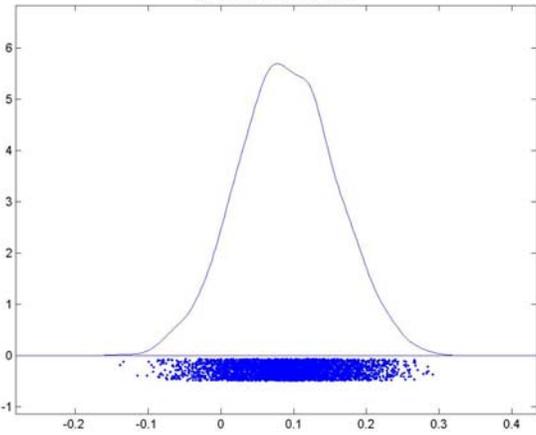


**Figure 3 Distribution of the Variances of Individual Specific Error Term Estimates ( $v$ ) for Land Development Intensity Levels**

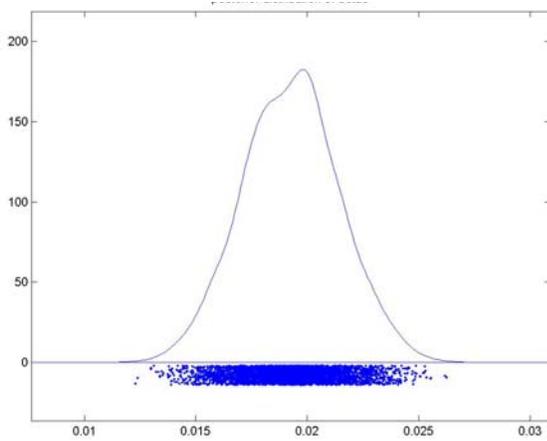
Figure 4 shows the posterior distributions of all parameters, based on the final 4000 draws. As discussed previously, all exogenous control variables are specified to follow normal posterior distributions.  $\lambda$  has a truncated normal distribution,  $\rho$  has a non-standard distribution, and  $\sigma^2$  follows a Chi-square distribution. The posterior distributions of threshold parameters  $\gamma$  are very interesting. As described in Wang and Kockelman (2008), they are shown to follow a normal distribution mixed with a multivariate uniform distribution. According to Figure 4's graphs (n) through (p), the resulting distributions are multimodal. Additionally, the shapes of  $\gamma$ s present some similarity, suggesting that their values are co-dependent. This is to be expected: based on the  $\gamma$  posterior distribution, it is clear that  $\gamma_s$ 's left threshold depends on  $\gamma_{s-1}$  and its right threshold depends on  $\gamma_{s+1}$ . This dependency can also be explained intuitively: the gap between  $\gamma_s$  and  $\gamma_{s+1}$  determines the probability of  $y = s$ . In order to maintain a generally constant gap, the values of  $\gamma_s$  and  $\gamma_{s+1}$  must move together.



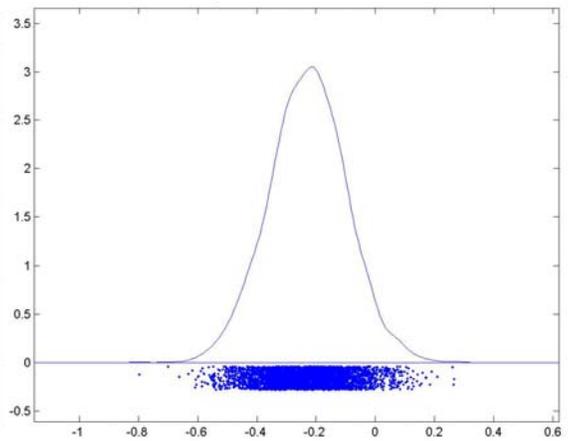
(a) Posterior Distribution of  $\beta_{POP}$



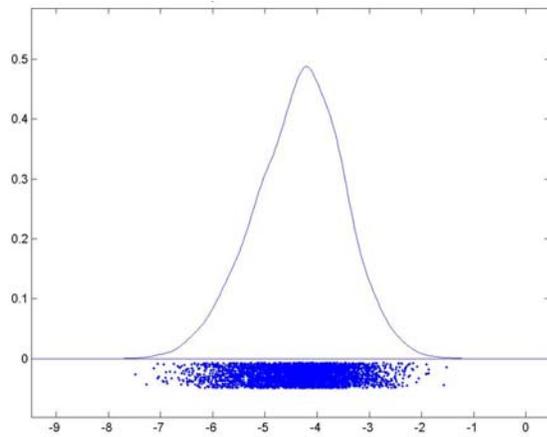
(b) Posterior Distribution of  $\beta_{WORKER}$



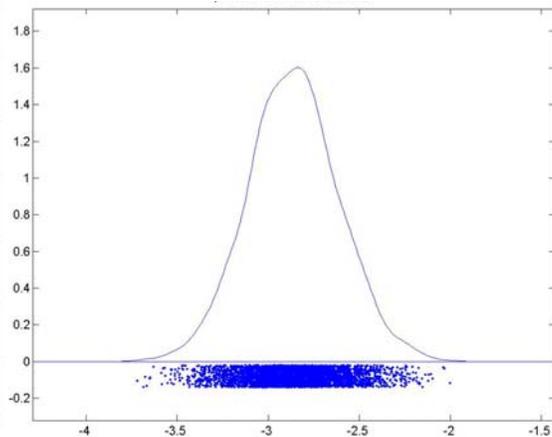
(c) Posterior Distribution of  $\beta_{INC}$



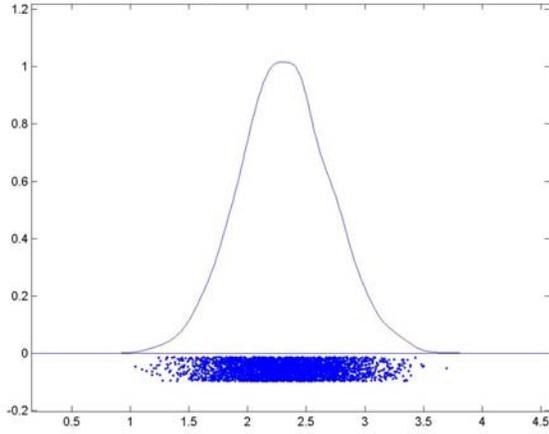
(d) Posterior Distribution of  $\beta_{EMPT}$



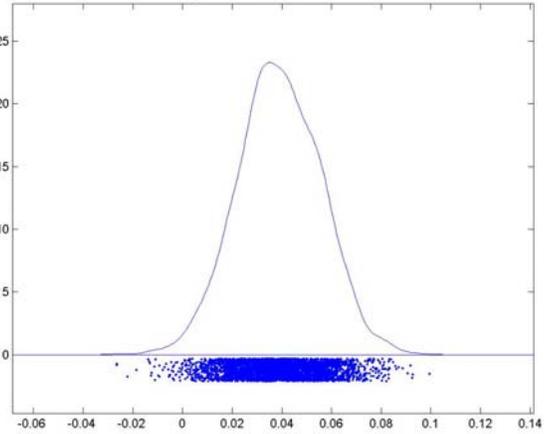
(e) Posterior Distribution of  $\beta_{CBDT}$



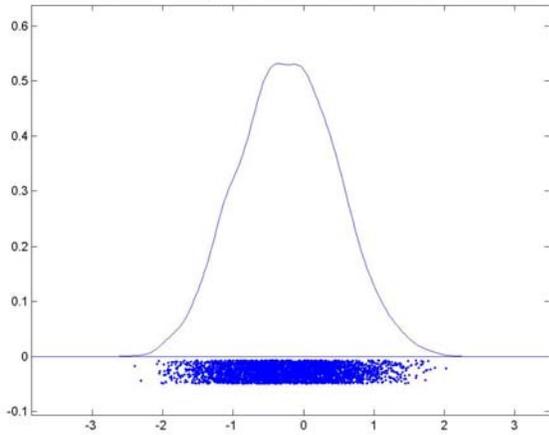
(f) Posterior Distribution of  $\beta_{AIRT}$



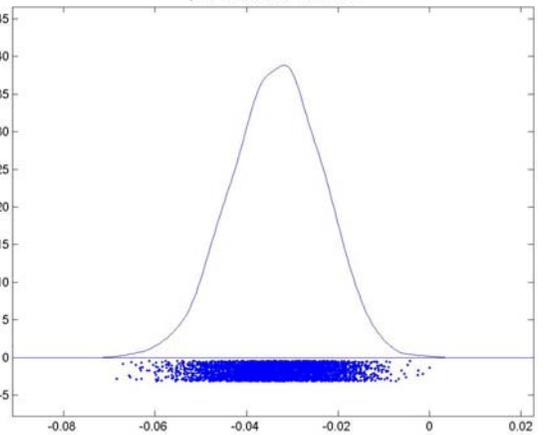
**(g) Posterior Distribution of  $\beta_{RDTT}$**



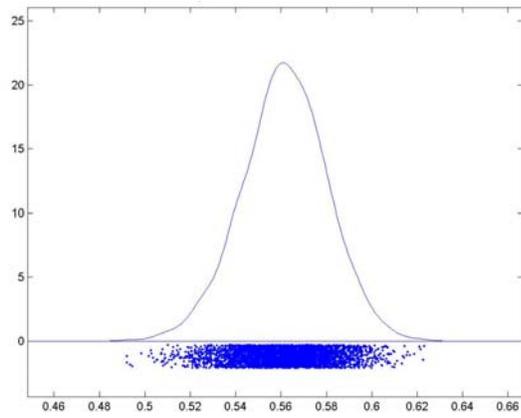
**(h) Posterior Distribution of  $\beta_{NSCHOOL}$**



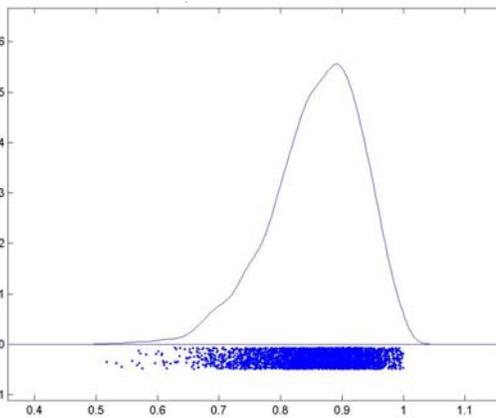
**(i) Posterior Distribution of  $\beta_{ELEVXN}$**



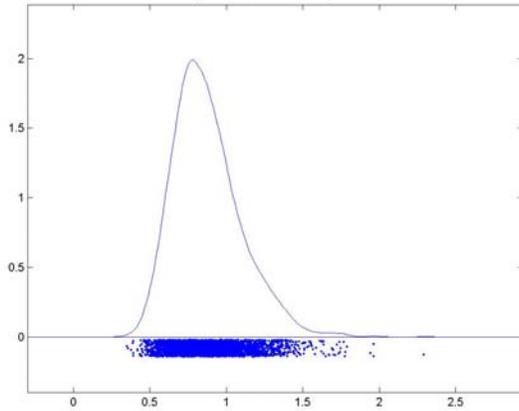
**(j) Posterior Distribution of  $\beta_{SLOPE}$**



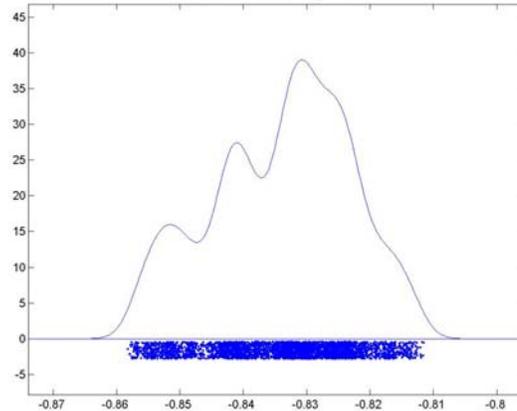
**(k) Posterior Distribution of  $\lambda$**



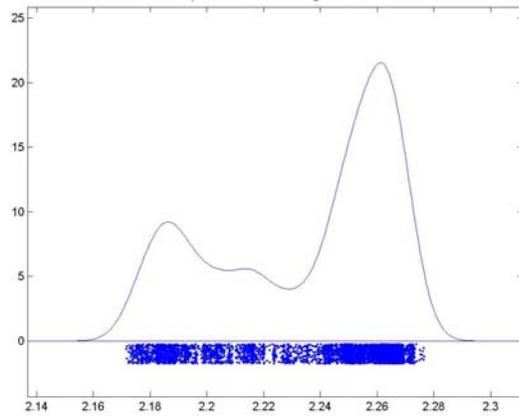
**(l) Posterior Distribution of  $\rho$**



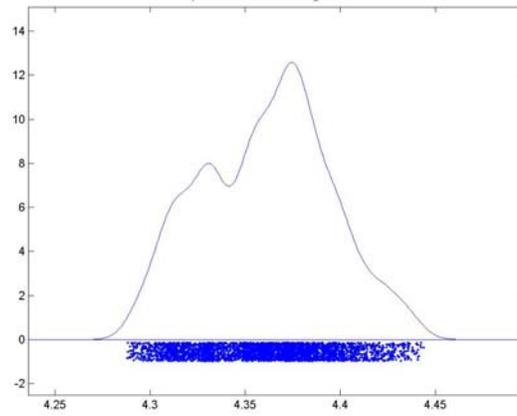
**(m) Posterior Distribution of  $\sigma^2$**



**(n) Posterior Distribution of  $\gamma_1$**



**(o) Posterior Distribution of  $\gamma_2$**



**(p) Posterior Distribution of  $\gamma_3$**

**Figure 4 Posterior Distributions of Land Development Intensity Level Model Parameters**

Table 4 shows the correlation between parameters, underscoring the high dependency among  $\gamma$  values. One also finds that the population and worker variables are highly correlated, indicating a potential multi-collinearity problem.

**Table 4 Correlation between Parameters**

	POP	WOR-KER	INC	EMPTT	CBDTT	AIRTT	RDTT	NSCH-OOL	ELEV	SLOPE	$\lambda$	$\rho$	$\sigma^2$	$\gamma_1$	$\gamma_2$	$\gamma_3$
POP	1.000	-0.993	0.193	0.028	0.070	0.112	-0.130	-0.096	-0.061	-0.015	-0.084	-0.055	0.028	-0.008	-0.029	-0.020
WORKER	-0.993	1.000	-0.174	-0.009	-0.057	-0.114	0.126	0.091	0.043	0.017	0.043	0.039	-0.030	0.024	0.035	0.034
INC	0.193	-0.174	1.000	0.146	-0.077	0.325	-0.078	0.034	-0.330	-0.114	-0.320	-0.040	0.115	0.014	-0.075	0.031
EMPTT	0.028	-0.009	0.146	1.000	-0.068	-0.365	-0.137	-0.002	0.042	0.002	-0.050	-0.011	0.019	0.205	0.116	0.174
CBDTT	0.070	-0.057	-0.077	-0.068	1.000	-0.155	-0.434	-0.040	-0.256	-0.026	0.191	-0.017	0.004	0.108	0.111	0.144
AIRTT	0.112	-0.114	0.325	-0.365	-0.155	1.000	-0.288	0.032	-0.054	-0.117	0.070	-0.059	0.017	0.069	0.030	0.044
RDTT	-0.130	0.126	-0.078	-0.137	-0.434	-0.288	1.000	0.019	-0.019	0.066	-0.095	0.052	0.011	-0.173	-0.132	-0.168
NSCHOOL	-0.096	0.091	0.034	-0.002	-0.040	0.032	0.019	1.000	-0.059	-0.036	-0.059	-0.006	-0.016	-0.016	0.007	-0.030
ELEV	-0.061	0.043	-0.330	0.042	-0.256	-0.054	-0.019	-0.059	1.000	0.022	0.046	-0.107	-0.034	0.141	0.133	0.068
SLOPE	-0.015	0.017	-0.114	0.002	-0.026	-0.117	0.066	-0.036	0.022	1.000	0.092	-0.007	-0.008	-0.051	-0.028	-0.058
$\lambda$	-0.084	0.043	-0.320	-0.050	0.191	0.070	-0.095	-0.059	0.046	0.092	1.000	0.000	-0.066	-0.031	0.026	-0.016
$\rho$	-0.055	0.039	-0.040	-0.011	-0.017	-0.059	0.052	-0.006	-0.107	-0.007	0.000	1.000	-0.261	-0.037	-0.044	0.010
$\sigma^2$	0.028	-0.030	0.115	0.019	0.004	0.017	0.011	-0.016	-0.034	-0.008	-0.066	-0.261	1.000	0.025	-0.029	0.056
$\gamma_1$	-0.008	0.024	0.014	0.205	0.108	0.069	-0.173	-0.016	0.141	-0.051	-0.031	-0.037	0.025	1.000	0.789	0.823
$\gamma_2$	-0.029	0.035	-0.075	0.116	0.111	0.030	-0.132	0.007	0.133	-0.028	0.026	-0.044	-0.029	0.789	1.000	0.544
$\gamma_3$	-0.020	0.034	0.031	0.174	0.144	0.044	-0.168	-0.030	0.068	-0.058	-0.016	0.010	0.056	0.823	0.544	1.000

Note: Values above 0.5 have been shaded.

## 5. Model Comparisons

The performance of the DSOP model can be compared to a standard ordered probit (OP) model, a dynamic ordered probit (DOP) model, and a spatial ordered probit (SOP) model. 10,000 draws were used in all these models, with the first 6,000 draws omitted (as the burn-in sample).

**Table 5 Goodness of Fit and Prediction Rates using Different OP Model Specifications**

Models	DIC	Predicted $y$ Value	Actual $y$ Value				% Cases Correctly Predicted (%)
			1	2	3	4	
DSOP	22587.9	1	1106	1417	179	47	48.8
		2	1354	3281	767	238	
		3	165	780	435	237	
		4	41	188	258	591	
DOP	23080.3	1	1120	1379	171	40	48.7
		2	1310	3237	750	236	
		3	208	841	479	279	
		4	28	209	239	558	
SOP	23091.3	1	1080	1379	187	49	47.8
		2	1294	3261	778	252	
		3	235	783	417	268	
		4	57	243	257	544	
OP	22800.0	1	992	1606	258	57	42.5
		2	1307	2913	770	324	
		3	273	822	371	299	
		4	94	325	240	433	

Table 5 provides the DIC<sup>1</sup> values and predictive accuracy with the four methods. While no model is clearly superior, the DSOP model outperforms the other models, even after being penalized for using more parameters. Interestingly, the DIC values suggest that the OP model may be preferred to the SOP and DOP models, due in part to its simpler model specification. Yet the predictive accuracy values tell a different story: the standard OP model only correctly predicts dependent values for 42.5% of the observations. The SOP model increases this percentage to 47.8%. The DOP and DSOP models have quite close prediction rates: 48.7 and 48.8%. These results imply that, recognition of dynamic effects may well be worth the added estimation complexity for these types of land use data sets.

### 5.1 Marginal Effects

Based on the model specification, the marginal effects of explanatory variables ( $X$ ) on the probabilities of each outcome level can be defined as follows:

<sup>1</sup> The deviance information criterion (DIC) is a generalization of the Akaike information criterion (AIC) and Bayesian information criterion (BIC). In addition, In order to accommodate the bimodal posterior distributions in the DSOP model, this study uses the modified DIC calculation method for mixture models proposed by Celeux et al. (2006).

$$\frac{\partial P(y_{ikt} = s)}{\partial x_{iktq}} = \left( \phi \left( \frac{\mu_{s-1} \lambda U_{ikt-1} - X_{ikt} \beta - \theta_i}{\nu_i} \right) - \phi \left( \frac{\mu_s - \lambda U_{ikt-1} - X_{ikt} \beta - \theta_i}{\nu_i} \right) \right) \cdot \beta_q \quad (6)$$

This marginal effect indicates the effect that a one-unit change in explanatory variable  $x_{iktq}$  has on the probability of different discrete outcomes,  $s$ . As noted earlier, the marginal effects on intermediate probabilities are not obvious at first glance, since a shift in the distribution can cause the probability of intermediate response types to fall or rise, depending on the positioning of the average response (see, e.g., Wang and Kockelman, 2005).

As Equation (6) suggests, one variable's marginal effect is related not only to its own coefficient, but also to the values of all other coefficients. Moreover, each observation and each period carry a distinct set of marginal effect values. In practice, marginal effects are generally calculated using the parameters' final point estimates and average variable values. In this study, the marginal effects are calculated separately for every observation, in each period and every iteration. The results are then averaged in order to provide a single, average response estimate, for every variable, recognizing cumulative effects across the region. Results obtained in this way anticipate more global changes for the population of points and respect the multivariate distribution of parameter values. This latter benefit is an advantage of using a Bayesian approach: derived statistics can be calculated on the heels of estimation, within the iterative Gibbs sampling process.

Table 6 shows the estimates of these final marginal effects – and explains the magnitude of “one unit” changes in different  $X$  variables – relative to their standard deviations. As one example, when each neighborhood's average household income increases by \$1,000, the sample population's average probability of intense development is estimated to rise by just 0.26% and the estimated probability of remaining undeveloped falls by 0.523%. In other words, incomes (which are rising in Austin at about \$1,300 per household per year) are estimated to have a practically negligible effect on land use intensities.

Travel time to the nearest top employer also appears to have a practically negligible effect, *ceteris paribus*: a 10 minute (0.17 hour) increase in this variable across all zones causes the region's average probability of remaining undeveloped to rise by 1.1%, and Level 2 through Level 4 probabilities to fall by 0.1%, 0.4%, and 0.6%, respectively. In contrast, travel time to the region's CBD is estimated to have an impressive effect: A 10 minute increase is linked to a 10% decrease in the probability of Level 4 development across the sample. (The probabilities of Levels 2 and 3 also fall, while the probability of finding undeveloped land is estimated to rise by 20%.) As supported by other land use research (e.g., Kalmanje and Kockelman, 2004, and Zhou and Kockelman, 2007), this distance-to-CBD variable regularly offers more predictive power than any other measures of access.

Travel time to an airfield appears to have a moderate impact on land development: A 10 minute increase is associated with a decrease in development levels 2, 3, and 4 by 1.4%, 5.1%, and 6.4%, respectively. Travel time to highways is predicted to have the reverse impact, *ceteris paribus*: after controlling for travel times to major employers, the region's CBD and all airfields, a 10 minute increase in travel time to the closest highway is associated with 4.1% more Level 2 development (which is very likely to be residential, commercial, or industrial uses, dotted with vegetation) and 5.2% more Level 3 development (which tends to be densely developed residential, commercial, and/or industrial land).

The number of schools also is a practically insignificant covariate, while a 10-degree increase in ground slope is associated with 1.0%, 3.6%, and 4.5% decreases in the estimated probability of Level 2 through Level 4 development, respectively. (However, when one considers that the average slope in the grid-cell sample is only 2.7 degrees, the impact of slope appears insignificant in practice.)

**Table 6 Marginal Effects of Changes in Covariate Values, as Computed over All Observations**

Variable	Ratio to Std. Dev.	Marginal Effect (%)			
		Level 1	Level 2	Level 3	Level 4
POP	0.092	0.652	-0.070	-0.257	-0.324
WORKER	0.168	-2.417	0.261	0.955	1.201
INC*	0.053	-0.523	0.057	0.207	0.259
EMPTT*	8.529	6.309	-0.682	-2.492	-3.135
CBDTT*	15.39	118.6	-12.87	-46.92	-58.80
AIRTT*	8.634	77.86	-8.467	-30.84	-38.55
RDTT*	13.65	-62.77	6.804	24.83	31.13
NSCHOOL*	0.726	-1.048	0.114	0.415	0.519
ELEVTN	16.39	6.429	-0.713	-2.515	-3.202
SLOPE*	0.455	0.912	-0.099	-0.362	-0.451

Notes: \* indicates statistical significance at the 0.05 level. Change in each variable is one unit (e.g., 1 hour in case of travel times (TT)). “Ratio to Std. Dev.” is the ratio of one unit (e.g., 1 hour) to the standard deviation observed in the data set, for each variable.

In summary, most of the contemporaneous variables are practically insignificant. This suggests that when developers make decisions, past land conditions (represented by the lagged, latent dependent variables) are a more important consideration than current conditions. However, transportation conditions (especially travel time to the CBD) appear highly influential, consistent with expectations regarding the location preferences of households and businesses, and the profit-maximizing nature of developer objectives.

## 5.2 Model Prediction

One important model application is scenario-based prediction of development intensities across the region. For the 2,771 grid cells in the selected sample areas, one potential scenario is a (one-step) doubling in population from the year 2000 and a 30% increase in all travel times – to major employers, the CBD, nearest highway and nearest airfield (in order to reflect added congestion).

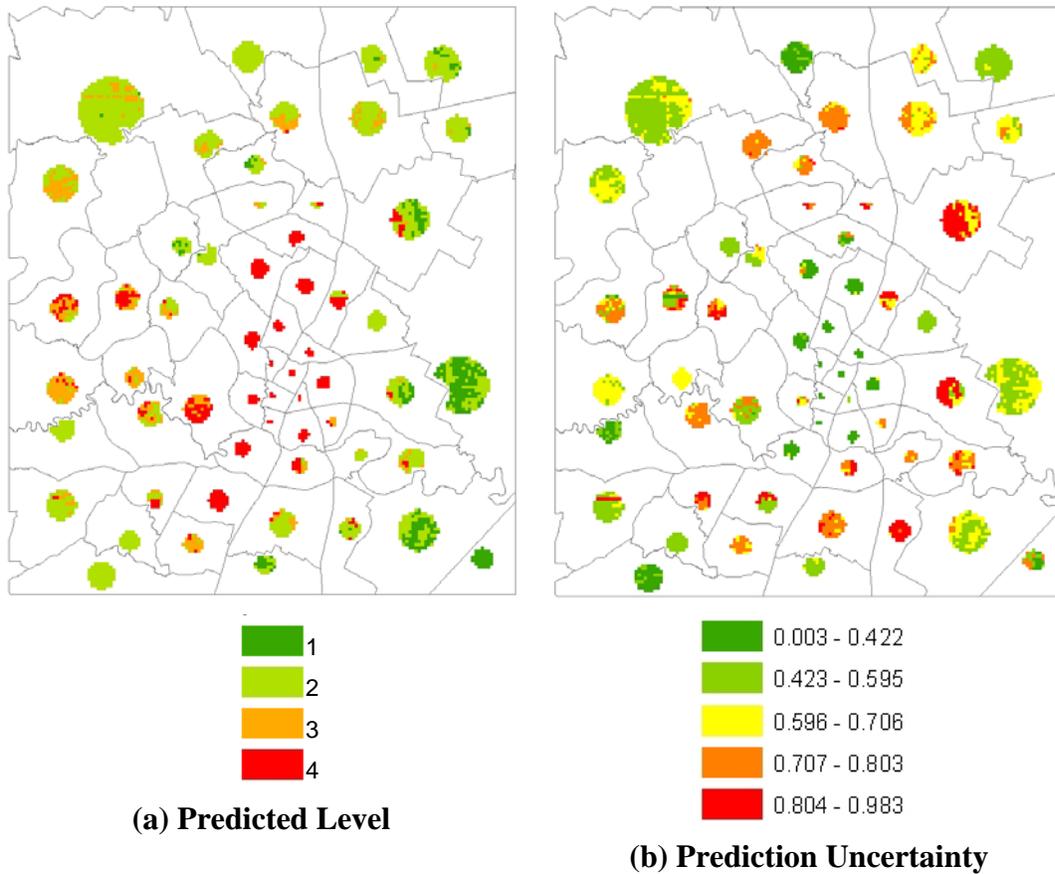
Similar to the calculation of marginal effects, predictions can be computed alongside during model estimation, recognizing the sampling distribution of parameter values. Over the final 4000 MCMC draws, estimates of latent dependent variable values for the next time step (approximately 7 years forward) are based on draws of latent dependent variable values for the year 2000, estimated parameter and error term values, and scenario control-variables. These latent variables then are compared to the threshold parameter values in each run, and development intensity levels for each location are calculated. Thus, for each of the 4000 draws and for each cell, there is a predicted development intensity level. The most common (frequently appearing) land development intensity levels in these 4000 runs for each sampled cell are shown

in Figure 5 (a). As expected, more intensely developed land is predicted to appear around Austin’s downtown.

Of course, this single “most likely” pattern may not occur with a high likelihood. There is great flexibility and uncertainty in the future of these 2,771 grid cells. To help planners appreciate (and visualize) such uncertainty, an entropy statistic is used (see, e.g., Wang and Kockelman 2006, McKay 1995 and Kotz and Johnston 1982). The uncertainty associated with the set of 4 potential land covers in cell  $i$  is specified as follows:

$$uncertainty_i = \frac{-1}{\ln(4)} \sum_{s=1}^4 P_{is} \ln(P_{is}) \quad (7)$$

This formulation generates a value between 0 and 1 for each cell. The higher the value, the more uncertain the prediction for that cell. When all four future land development intensity levels have equal probabilities ( $P_{is} = 0.25 \forall s$ ), uncertainty entropy equals 1, indicating maximum uncertainty. When the same land intensity level emerges in all 4000 simulations, this uncertainty value is 0. As illustrated in Figure 5 (b), higher uncertainty appears around the intermediate areas of the study area, or the central-city’s edge. At these locations, the potential for variation in future development patterns is relatively large, resulting in a higher degree of uncertainty.



**Figure 5 Most Likely Development Intensity Levels Prediction and Uncertainty (following an assumed doubling of population)**

Table 7 compares these predictions to the year 2000 situation. While strong similarity is evident, one finds some “backward” changes – from higher intensity to lower intensity levels. This is to be expected, since some locations presently are more developed than trend behaviors would suggest. Moreover, some locations may lose their attraction due to increases in travel time.

**Table 7 Comparison of Base Year and Predicted Land Development Intensity Levels**

		Most Likely Intensity Levels in Future Scenario				Total
		1	2	3	4	
Base Year Intensity Levels (Year 2000)	1	374	103	0	0	477
	2	11	1280	22	5	1318
	3	0	166	344	27	537
	4	0	2	71	366	439
Total		385	1551	437	398	2771

## Conclusions

This study uses a dynamic spatial ordered probit model to analyze land development intensity levels in Austin, Texas. The estimation indicates that the temporal autocorrelation coefficient is highly practically and statistically significant. This implies that existing land conditions (represented by temporally lagged latent dependent values) offer high predictive power, as one might expect: land development is a costly and involved process, and existing development cannot be easily demolished or intensified.

Other control variables exhibit much smaller marginal effects, suggesting that an AR(1)-type approach with spatial lags can be key to land development prediction. Estimates of transportation condition effects, especially the influence of travel time to Austin’s CBD, highlight the important role of access.

Even after controlling for various neighborhood and access characteristics, along with lagged latent response levels, estimation residuals are high in this model, and positively correlated across space. This statistical result confirms the common intuition that land development tends to cluster rather than randomly distribute itself over space, and that a variety of unobserved variables (such as soil conditions and local aesthetics) play a role in development decisions.

One of the potential extensions of this study relates to variable time gaps in the panel data. The four data years are 1983, 1991, 1997, and 2000, with gaps of 8, 6, and 3 years, respectively. Intuitively, when the gap is longer, the temporal dependencies should be weaker. A more appropriate model specification would control for gaps variations in some way, by exploiting time series analysis tools for variable gap lengths. For example, one approach may be to express the temporal coefficient as an exponential function of the time gap. Another extension is to develop dynamic spatial models for multinomial (unordered) discrete response data, which will be more useful for land use type analysis. Improving data quality, including further screening problematic data and enlarging sample size are also important. However, in order to enlarge the sample size, the matrix inversion issue first needs to be solved. Finally, how to reasonably apply the calibrated model and spatially interpolate land development intensity level for out-of-sample area will be an interesting extension of this study.

While the data sets used here are imperfect and sample size issues remain a challenge, the application of such a statistically rigorous model in land development change analysis is new and useful for a variety of data contexts. Application of Wang and Kockelman's (2008) DSOP model to the Austin, Texas context also discloses some interesting patterns in the evolution of urban land development intensity. Moreover, this study demonstrates how one can capitalize on the existence of satellite data. As more frequent and accurate satellite images become available, this evolving data source will be used for more extensive topics, such as global climate change, loss of Amazon rainforest, Africa's desertification, human migration, and even real-time traffic condition forecasting. It is important that regional scientists, spatial econometricians and others unleash their potential, by recognizing the spatial relationships that exist and by exploiting their presence. The DSOP model is one such tool.

## References

- Anselin, L. (1999) *Spatial Econometrics*. Working paper. Accessed July 10, 2005: [http://www.csiss.org/learning\\_resources/content/papers/baltchap.pdf](http://www.csiss.org/learning_resources/content/papers/baltchap.pdf).
- Candau, J., Rasmussen, S. and Clarke, K. C. (2000) "A coupled cellular automaton model for land use/land cover dynamics". Accessed July 10, 2005: <http://www.geog.ucsb.edu/~kclarke/ucime/banff2000/533-jc-paper.htm>.
- Celeux, G., Forbes, F., Robert, C. P. and Titterton, D. M. (2006) "Deviance information criteria for missing data models." Technical report. Accessed May 10, 2007: <http://www.ceremade.dauphine.fr/~xian/cfirt05.pdf>.
- Gelfand, A. E. and Smith, A. F. M. (1990) "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association* 85: 398-409.
- Girard P. and Parent E. (2001) "Bayesian analysis of autocorrelated ordered categorical data for industrial quality monitoring." *Technometrics* 43(2): 180-191.
- Kotz, S. and Johnson, N. (1982) *Encyclopedia of Statistical Sciences*. New York: John Wiley & Sons
- McKay, M. D. (1995) "Evaluating prediction uncertainty". Technical Report NUREG/CR-6311, U.S. Nuclear Regulatory Commission and Los Alamos National Laboratory.
- Moran, P.A.P. (1950) "Notes on continuous stochastic phenomena." *Biometrika* 37(1-2): pp.17-23.
- Munroe, D., Southworth, J. and Tucker, C. M. (2001) "The dynamics of land-cover change in western Honduras: Spatial autocorrelation and temporal variation". *Conference Proceedings. American Agricultural Economics Association. AAEA-CAES 2001 Annual Meeting*. Accessed July 10, 2004: [http://agecon.lib.umn.edu/cgi-bin/pdf\\_view.pl?paperid=2611](http://agecon.lib.umn.edu/cgi-bin/pdf_view.pl?paperid=2611)
- Nelson, G. C., and Hellerstein, D. (1997). "Do roads cause deforestation: Using satellite images in econometric analysis of land use". *American Journal of Agricultural Economics* 79: 80-88.
- Schrank, D. and Lomax, T. (2005) *The 2005 Urban Mobility Report*. Accessed June 1, 2007: [http://tti.tamu.edu/documents/mobility\\_report\\_2005.pdf](http://tti.tamu.edu/documents/mobility_report_2005.pdf).

- Smith, T. E. and LeSage, J. P. (2004) "A Bayesian probit model with spatial dependencies." In Pace, R. K. and LeSage, J. P. (Eds.), *Advances in Econometrics Volume 18: Spatial and Spatiotemporal Econometric*. Oxford: Elsevier Ltd.
- 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.
- Wang, X. (2007) *Capturing Patterns of Spatial and Temporal Autocorrelation in Ordered Response Data: A Case Study of Land Use and Air Quality Changes in Austin, Texas*. Ph.D. Dissertation, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.
- Wang, X. and Kockelman, K. (2005) "Occupant injury severity using a heteroscedastic ordered logit model: distinguishing the effects of vehicle weight and type." *Transportation Research Record* 1908: 195-204.
- Wang, X. and Kockelman, K. (2006) "Tracking land cover change in a mixed logit model: recognizing temporal and spatial effects." *Transportation Research Record* 1977: 112-120.
- Wang, X. and Kockelman, K. (2008) "The Dynamic Spatial Ordered Probit Model: Methods for Capturing Patterns of Spatial and Temporal Autocorrelation in Ordered Response Data, using Bayesian Estimation." Under review for publication in the *Journal of Regional Science*.
- Wang, X. (2007) *Capturing Patterns of Spatial and Temporal Autocorrelation in Ordered Response Data: A Case Study of Land Use and Air Quality Changes in Austin, Texas*. Ph.D. Dissertation, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.
- Wear, D. N. and Bolstad, P. (1998) "Land-use changes in southern Appalachian landscapes: Spatial analysis and forecast evaluation." *Ecosystems* 1: 575-594.