Modeling Land Use Change Using an Eigenvector Spatial Filtering Model Specification for Discrete Response

Parmanand Sinha

The University of Tennessee, Knoxville
304 Burchfiel Geography Building
1000 Phillip Fulmer Way
Knoxville TN 37996-0925
Email: psinha5@utk.edu

1. Introduction

Analysis of land-use changes requires considering spatial effects that becomes challenging in a discrete choice framework. Ignoring the spatial effects will lead to inconsistent estimators. Recently many studies have been done in land use change mostly involving binary choice. The binary models are more widely used due to their computational simplicity relative to multinomial models. This comes at the expense of assuming the independence of irrelevant alternatives in every case. The recent advances in computational algorithms, it is now possible to implement multinomial models accounting for more than one unordered alternatives. This has been done using an Eigenvector spatial filtering (ESF) based multinomial logit model. ESF (Griffith 2000; Griffith 2003; Getis and Griffith, 2002). This paper is focused on finding a way to reduce the computation time for large multinomial data.

2. Multinomial Autologistic Regression for Land suitability

There is a finite set of land pixels, $|\mathrm{N}| = n$, and each pixel can be converted into one of the land uses in the set M $|\mathrm{M}| = m$. The logit equation can be written as

Ln (odds of land use change from vacant to j for location i at time t_2) = f (Factors of the Built Environment, and the natural environment, and Socioeconomic Characteristics)

where Ln denotes the natural logarithm and the *f* denotes a function. Included covariates measure, to an extent, some of the potential spatial dependence. How well they account for it can be seen from the statistical significance of the estimate of the spatial lag parameter in the MNL model:

$$U_{ij} = \rho \sum_{k=1}^{n} w_{ik} U_{ik} + X_i \beta_j ,$$

$$P(Y_{ij} = 1) = \frac{\exp\left(U_{ij}\right)}{\sum_{j=1}^{m} \exp\left(U_{ij}\right)}$$

$$P(Y_{ij} = 1) = \frac{\exp\left(\rho \sum_{k=1}^{n} w_{ik} Y_{ik} + X_i \beta_j\right)}{\sum_{j=1}^{m} \exp\left(\rho \sum_{k=1}^{n} w_{ik} Y_{ik} + X_i \beta_j\right)}$$

where U_{ij} is a latent dependent variable representing the underlying utility from choosing a given alternative j, $P(Y_{ij} = 1)$ is the probability of land parcel i having land use j, X_i is a vector of explanatory variables, w_{ik} are elements of the spatial weight matrix W, and parameter ρ measures the nature and degree of spatial dependence. The spatial weight matrix is typically row standardized such that $\sum_{k=1}^{n} w_{ik} = 1$ for $k \neq i$, and $w_{ii} = 0$.

3. Method of Estimation

To calculate the estimators of discrete choice models, the conventional estimators become infeasible in large data sets due to their requirement of inversion of large matrices. Griffith (2004) estimates parameters of a spatial lagged autologistic model using Eigenvector Spatial Filtering (ESF), which is a comparatively newer technique that does not involve inversion of matrices, and is relatively simpler to apply.

$$\mathbf{U}_{ij} = \mathbf{X}_i \boldsymbol{\beta}_j + \mathbf{E}_{Ki} \boldsymbol{\beta}_{Ej}$$

$$\mathrm{P}(\mathbf{Y}_{\mathrm{ij}} = 1) = \frac{\exp(\mathbf{U}_{\mathrm{ij}})}{\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathbf{U}_{\mathrm{ij}}}$$

where U_{ij} is a latent dependent variable representing the underlying utility from choosing a given alternative j, $P(\boldsymbol{Y}_{ij} = 1)$ is the probability of land parcel i having land use j, $\boldsymbol{X}_i$ is a vector of explanatory variables, $\boldsymbol{E}_K$ is an n-by-K matrix containing K eigenvectors, β_E is the corresponding vector of regression parameters, and ε_{ij} is a vector of identically distributed errors.

For binary dataset, ESF requires shorter and more straightforward computation compared to other estimation techniques (Wang, Kockelman, and Wang 2013). In the case of a multinomial dataset, the computation time significantly increases. The K eigenvectors are selected from the candidate set using stepwise multinomial logistic regression maximizing model fit at each step, which requires lots of computation time as n increases. One possible way to reduce this computation time is to reduce the number of eigenvectors in the candidate set. This can be done by making the spatial resolution coarser for the eigenvectors while keeping the same spatial resolution for the response variable as shown in figure 1 (a) and nine cells of eigenvectors have been aggregated to one for the spatial effect as shown in figure 1 (b).

Figure 1. (a) showing finer spatial resolution for response variableFigure (b) showing coarser spatial resolution for eigenvectors

4. Study area and Data

Collin County, TX is selected as a study area for this research which is divided into approximately 101,874 square cells (pixels) each with dimension 150-by-150 meters. Table 1 and figure 2 shows the land use change between 2000-2010.

Table 1. Land use distribution in 2000 and 2010

Land Use	Land use in 2000	Land use in 2010	Change in land use 2000-2010	% change in 10 years
Single Family	14745	32660	17915	121.50%
Multi Family	1141	3407	2266	198.60%
Commercial	2080	4828	2748	132.12%
Industrial	604	1415	811	134.27%
Parks & open spaces	1094	3113	2019	184.55%
Vacant	70129	44370	25759	

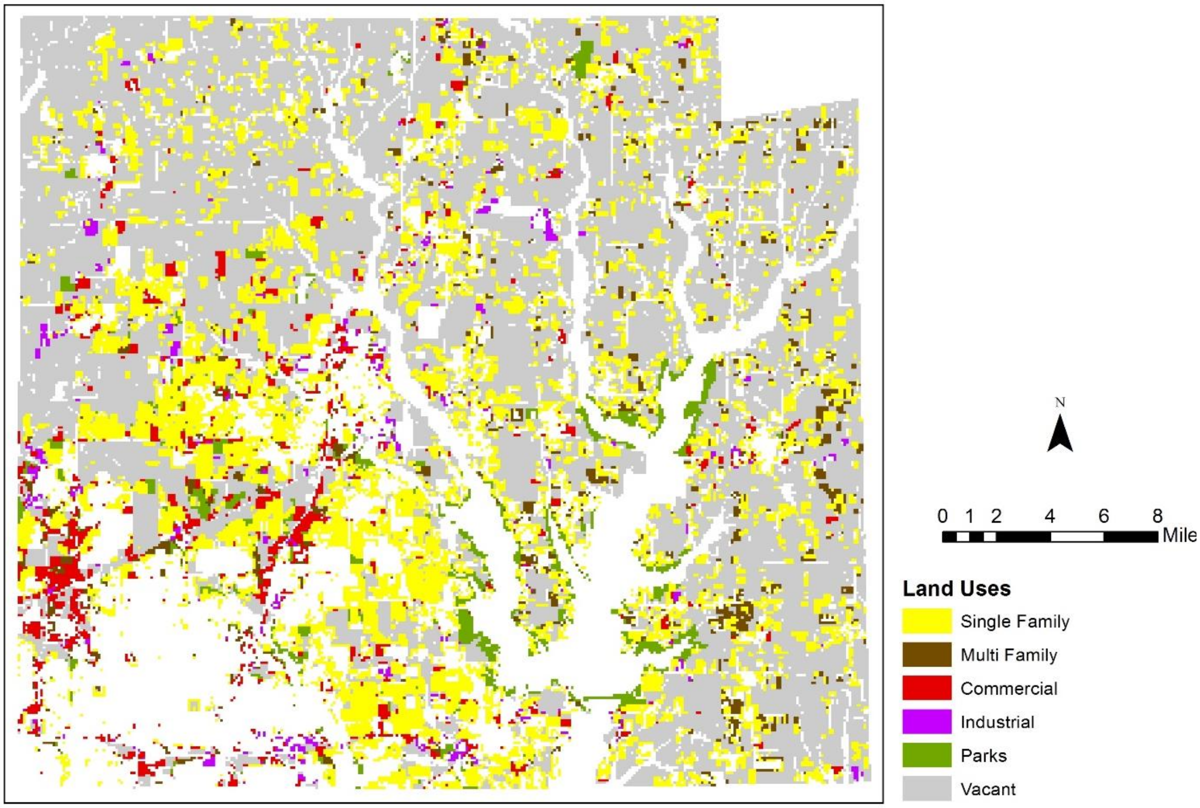

Figure 2. land use change between 2000-2010

Table 2 displays the descriptive statistics of explanatory variables used in this study for the year 2000.

Table 2. A list of explanatory variables for the year 2000

Variable	Mean	Std. Dev
Built environmental Factors		
Proximity to shopping center weighted by GLA	109476	442908
Proximity to employment center weighted by number of employees	171.779	840.298
Distance to School (Categorical)	0.13166	0.33812
Distance to highways	23962.6	18031.2
Distance to arterial roads (more than 1/4th miles from highways)	3348.68	2960.7
Distance to city center	15020.3	7129.86
Proximity to city center weighted by population of city	0.65431	2.40597
Distance from single family land use	1648.95	1426.82
Distance from multifamily land use	5189.4	3681.6
Distance from commercial land use	7839.61	5898.4
Distance from industrial land use	10803.8	7202.47
Distance from parks and open spaces	15879.7	13268.7
Distance to DART/public transit hubs	92099.6	36530.5
Distance to flood plain/Distance to highway	0.26037	0.84364
Socio-economic Factors		
EASI Total Crime Index 2008 (Block Group)	33.2778	23.9865
Median Rent (Block Group)	514.134	203.766

Median Value Owner Households (Block Group)	131203	68595.6
Median Year Built (Block Group)	1985.63	20.7755
Household Inc., Median, 2000 (Block Group)	63408.8	18439.3
% Employment, Travel Time Less than 15 Min, 2000 (Block Group)	14.8175	6.46538
% Employment, Retail Trade, 2000 (Block Group)	11.9721	2.81368
Population (block)	67.7337	137.337
% Housing, Occ (block)	0.78815	0.33636
Natural environmental Factors		
DEM	192.163	21.6404
Distance from floodplain	1470.34	1290.97
Distance to waterbodies	11544.2	8442.61

5. Results

The Non-spatial MNL model

Figure 3 shows the predicted value of 2000-2010 land use change using the non-spatial MNL model.

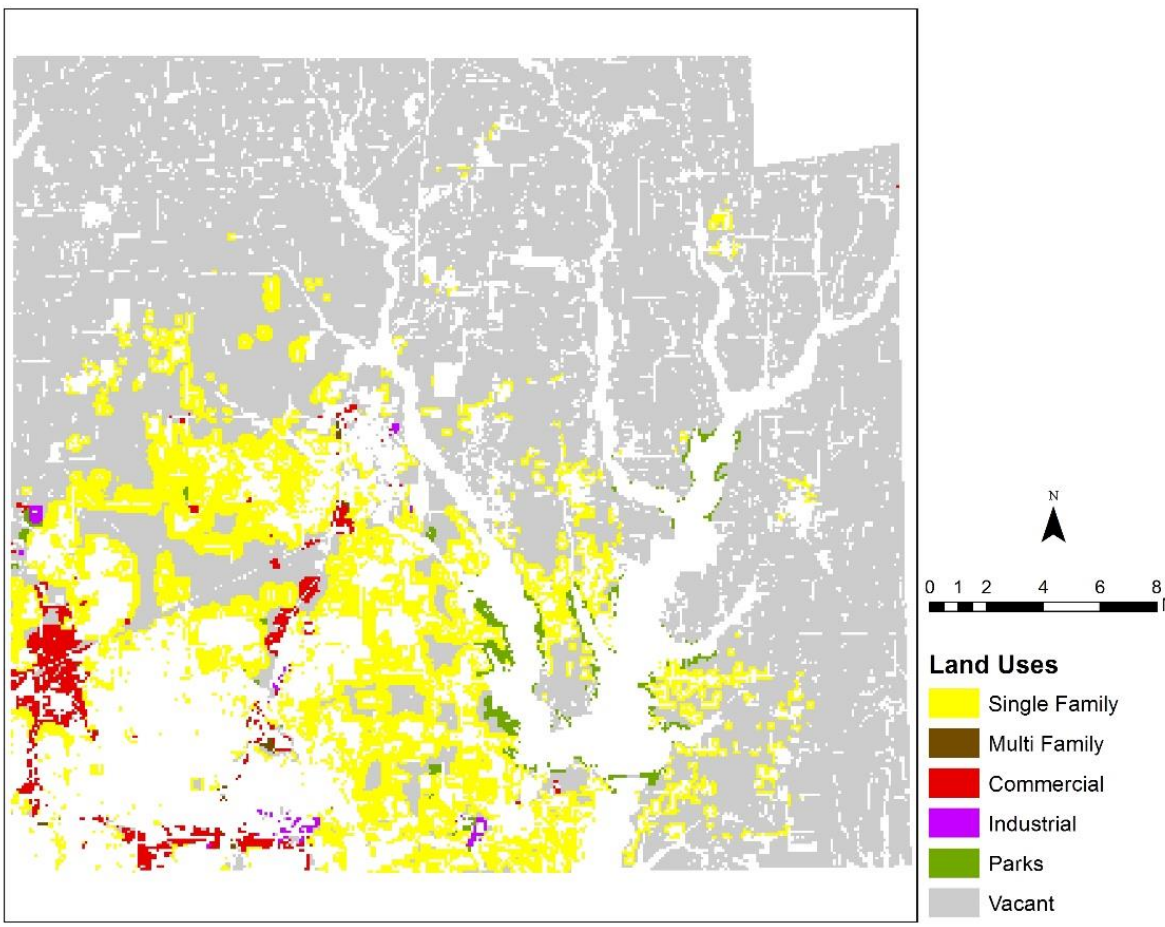

Figure 3. Prediction without spatial effects

Comparing predicted to actual change in land use, the non-spatial MNL model is not able to predict multi-family, commercial, industrial, and open space use correctly. It tends to correctly predicts single family, commercial, and vacant land use. Also, the predictions are clustered more towards the southeast part of the county.

Multi-family and industrial land uses are predicted poorly compared with commercial and open space land use. The single family land use category is slightly under-predicted, while the number of vacant pixels is over-predicted. Overall, with a kappa of 0.3565 and pseudo-R^2 of 0.39, the accuracy of prediction is poor.

The Spatial MNL model

The ESF model is used to account for spatial autocorrelation in data as part of the multinomial logistic regression equation specification. The next coarser pixel resolution of eigenvectors, 450m-by-450m, has been used and the result is shown in figure 4 and table 3.

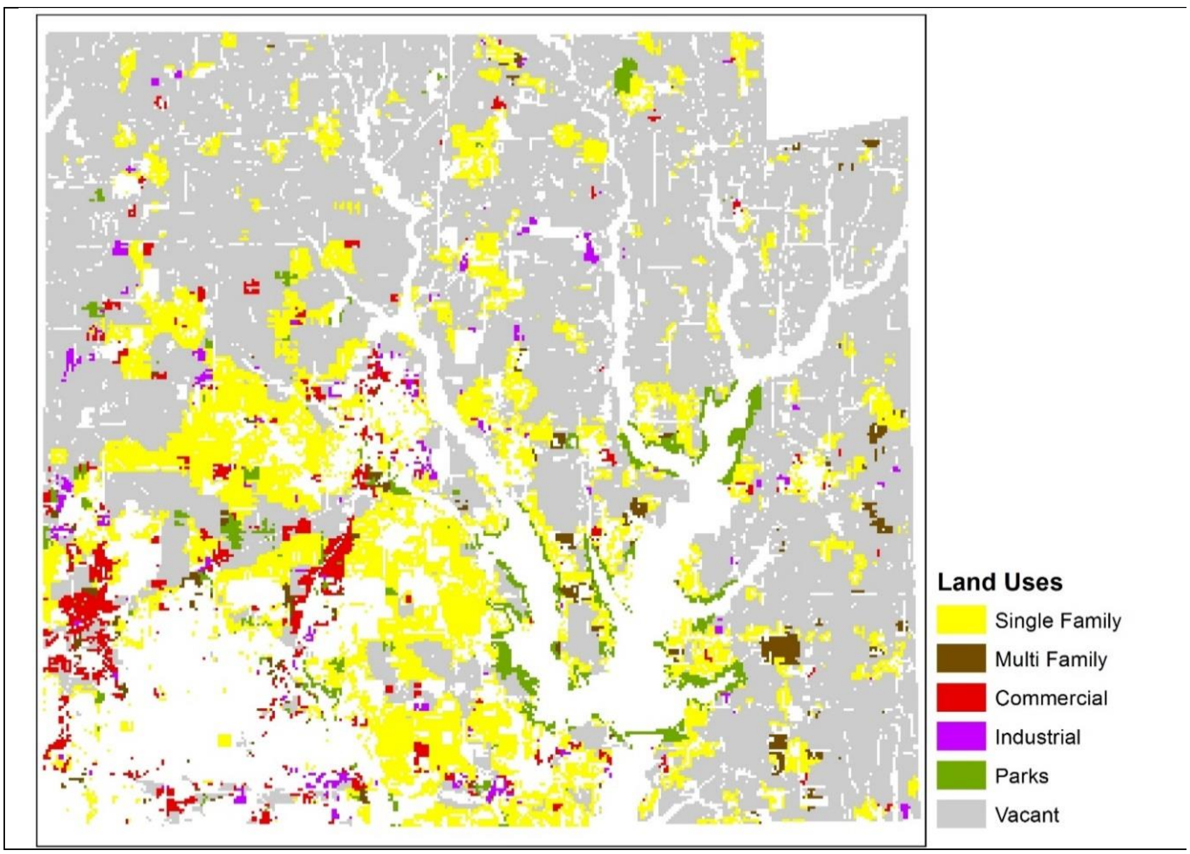

Figure 4. Predictions with spatial effects

Table 3. Cross-tabulation of actual versus predicted land use spatial effects

Table of _FROM_ by _INTO_								
Response		Predicted response						Total
		SF	MF	C	I	OS	V	
SF	Frequency	10375	203	221	60	236	6820	17915
MF	Frequency	413	569	93	10	58	1123	2266
C	Frequency	521	29	1378	56	37	727	2748
I	Frequency	64	1	33	511	9	193	811
OS	Frequency	273	23	25	8	1506	184	2019
V	Frequency	3098	169	324	110	136	40533	44370
Total	Frequency	14744	994	2074	755	1982	49580	70129
	Percent	21.02	1.42	2.96	1.08	2.83	70.7	100

The multi-family land use prediction is only 25.11% accurate and has been misclassified mostly into single family and vacant land uses. This finding could be attributable to the absence of an exploratory variable that explains multifamily land use change. Single family land use has been predicted with 58%, commercial land use with 50.14%, industrial land use with 60.3%, open spaces with 74.6% accuracy.

Table 8. Accuracy of predicted land use

	Without spatial effect	With spatial effect
Pseudo-R^2	0.3907	0.6289
Kappa	0.3565	0.5618
PCC	70.86	78.24
PCC (without vacant)	33.99	55.67

6. Conclusion

It is clear that spatial MNL gives a better prediction than a non-spatial. The coefficients for independent variables vary with the consideration of spatial effects, but show the same effect in both cases. In the case of a large spatial dataset, the candidate set of eigenvectors can be decreased by increasing the coarseness of the eigenvectors resolution.

After a certain number of eigenvectors entered into a model, increasing the number of eigenvectors in the model by reducing the criteria of selection may not improve the accuracy in the same proportion.

The ESF specification helps to understand the marginal effect of exploratory variables more precisely. Compared to other methods, such as those based on auto-models, ESF based specification is relatively easier to apply in a spatial discrete choice model for large datasets and involves comparatively straightforward computation.

7. References

Getis, Arthur, and Daniel A Griffith. 2002. "Comparative Spatial Filtering in Regression Analysis." *Geographical Analysis* 34: 130–40.

Griffith, Daniel A. 2004. "A Spatial Filtering Specification for the Autologistic Model." *Environment and Planning A* 36 (10): 1791–1811.

———. 2000. "Eigenfunction Properties and Approximations of Selected Incidence Matrices Employed in Spatial Analyses." *Linear Algebra and Its Applications* 321 (1-3): 95–112.

———. 2003. *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*. Springer.

Wang, Yiyi, Kara M. Kockelman, and Xiaokun (Cara) Wang. 2013. "Understanding Spatial Filtering for Analysis of Land Use-Transport Data." *Journal of Transport Geography* 31 (July): 123–31.