THE DYNAMIC SPATIAL ORDERED PROBIT MODEL: METHODS FOR CAPTURING PATTERNS OF SPATIAL AND TEMPORAL AUTOCORRELATION IN ORDERED RESPONSE DATA, USING BAYESIAN ESTIMATION

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

Submitted for publication in the *Journal of Regional Science*, 7 February 2008

Abstract

Many databases involve ordered discrete responses in a temporal and spatial context, including, for example, land development intensity levels, vehicle ownership, and pavement conditions. An appreciation of such behaviors requires rigorous statistical methods, recognizing spatial effects and dynamic processes. This study develops a dynamic spatial ordered probit (DSOP) model in order to capture patterns of spatial and temporal autocorrelation in ordered categorical response data. This model is estimated in a Bayesian framework using Gibbs sampling and data augmentation, in order to generate all autocorrelated latent variables. It incorporates spatial effects in an ordered probit model by allowing for inter-regional spatial interactions and heteroskedasticity, along with random effects across regions or any clusters of observational units. The model assumes an autoregressive, AR(1), process across latent response values, thereby recognizing time-series dynamics in panel data sets. The model code and estimation approach is tested on simulated data sets, in order to reproduce known parameter values and provide insights into estimation performance, yielding much more accurate estimates than standard, non-spatial techniques. The proposed and tested DSOP model is felt to be a significant contribution to the field of spatial econometrics, where binary applications (for discrete response data) have been seen as the cutting edge. The Bayesian framework and Gibbs sampling techniques used here permit such complexity, in world of two-dimensional autocorrelation.

Key words: Ordered response data, Bayesian approach, MCMC sampling, spatial autocorrelation, dynamics

1. Introduction

In the fields of regional science and transportation, variables of interest often are discrete in nature and involve temporal and spatial relationships. For example, land use intensity, vehicle ownership, and roadway service levels often are measured (and/or coded) as ordered discrete responses, dependent on various influential factors. These discrete responses share a common feature: they all exhibit some degree of temporal and spatial dependence or autocorrelation. For example, in two slices of a panel survey of households, the count of vehicles owned by the same household will be highly correlated. This phenomenon is normally defined as temporal dependency or autocorrelation. Meanwhile, even after controlling for household attributes, auto ownership levels are expected to exhibit positive correlations in the spatial context. To some extent, such correlation patterns can be explained by uncertainty or proximity because, in reality, there are always influential factors that cannot be controlled (e.g. pedestrian friendliness of all neighborhoods). The sign and magnitude of such uncertainties tend to vary rather gradually over space. Of course in a spatial context, in contrast to time-series data, such dependencies are *two* dimensional – which adds complexity. Like temporal relationships, correlation tends to diminish with increases in distance between any two households/observed units.

Essentially, then, many phenomena involving ordered categorical data also have temporal and spatial relationships; yet the development of rigorous methods for analyzing these phenomena is still in its infantry. This paper presents a model that is appropriate for describing the temporal and spatial relationships that exist in ordered categorical data. Related issues also are explored, indicating model estimation techniques, model validation and model comparisons (with simplified, less behaviorally reasonable models). Such model specifications and estimation techniques may be viewed as breakthroughs in the area of spatial econometrics, and the results extendable to a wide range of topics, where dependent variables are ordered discrete values and may involve temporal and spatial dependencies across observations.

The following sections review existing studies, explain the intuition behind model specification and illustrate how to estimate the unknown parameters using Bayesian methods. Model performance is quantified using simulated datasets.

2. Literature Review

As in standard spatial econometrics, methods for dealing with spatial effects in discrete choice models can be categorized into three basic types. Geographically weighted regression (GWR) is most applicable when spatial variation in behavioral parameters is of strong interest. In an analysis on suburban subcenters and employment density, McMillen and McDonald (1998) propose the idea of applying standard logit or probit methods to distance weighted sub-samples of the data in place of least squares, essentially using GWR to deal with discrete responses. LeSage (1999) provided code for producing binary logit and probit GWR estimates, using crime data. Atkinson et al. (2003) also used a GWR binary logit model to explore relationships between the presence (or absence) of riverbank erosion and geo-morphological controls. Vanasse et al. (2005) incorporated GWR in a binary logit model to study spatial variation in the management and outcomes of acute coronary syndrome.

The second method, spatial filtering, has been applied more broadly. It saves much specification and estimation effort. In addition to several land use/land cover models (e.g., Nelson and Hellerstein, 1997, Wear and Bolstad, 1998, and Munroe et al., 2001), many other works rely on this method. For example, an early study by Boots and Kanaroglou (1988) introduced a measure of spatial structure and used it as an explanatory variable when considering spatial effects in Toronto's intra-metropolitan migration. Dugundji and Walker (2005) controlled for spatial network independencies in their mixed logit model when studying mode choice behavior. Coughlin et al. (2003) incorporated global and regional spatial effects into an analysis of state lotteries.

The third method incorporates spatial effects directly in a discrete choice model setting and is the focus of this study. This method can be distinguished by two approaches. The first considers spatial autocorrelation across choices or alternatives, as often discussed for location choice models. This approach extends the commonly used GEV model by allowing correlated alternative-specific error terms in a mixed logit framework. For example, Miyamoto et al. (2004) assumed that location choice follows an SAR process, and used the weight matrix as a multiplier on dependent variables. Bhat and Guo (2004) used a contiguity matrix on their latent dependent variables to represent alternative-zone correlation patterns.

The second approach considers spatial autocorrelation across observational units (or individuals), and is the focus of this work. Currently, studies recognizing such spatial autocorrelation are limited to binary choice settings. To some extent, Wang and Kockelman (2006)'s work on estimating urban land cover evolution seems an exception, because multiple choices are studied in a mixed logit framework. However, rather than permitting a more flexible SAR process, Wang and Kockelman used a direct representation method and assumed a specific distance-decay function for inter-observational correlations, making the spatial correlation pattern across observations rather arbitrary. All other existing spatial probit and logit work is binary in nature. Anselin (2001) reviewed such spatial probit models and notes that McMillen (1995) first used the EM algorithm to estimate a probit model with a SAR process. Beron and Vijverberg (2004) specified probit models with both spatial errors and spatial lags, and then estimated these models by using recursive importance sampling (RIS) to approximate the n-dimensional log-likelihood. LeSage (2000) specified a model with a spatially correlated error term and used Gibbs sampling for estimation. Smith and LeSage (2004) extended this study by incorporating a regional effect and used Bayesian techniques to analyze the 1996 presidential election results. Similar studies include Kakamu and Wago's (2007) Bayesian estimation of a spatial probit model for panel data to analyze the business cycle in Japan.

Another estimation approach is the generalized method of moments, or GMM. Pinkse and Slade (1998) first used GMM to estimate a probit model with spatial error components. Pinkse et al. (2005) refined that study by incorporating a dynamic structure for dependent variables and applying a one-step GMM. And Klier and McMillen (2007) used GMM to estimate a spatial logit model for analyzing the clustering of auto supplier plants in the U.S. However, the use of GMM is limited because it requires orthogonality conditions (as discussed in works like Klier and McMillen, 2007, Pinkse and Slade, 1998, and Pinkse et al., 2005), and standard errors must be derived. For this reason, it presently is applied only to binary response models; it has not yet been extended to multiple-response models. All the other estimation methods can be called simulation estimators. As Anselin (2001) concludes, all current simulation estimators are slow, but Gibbs sampler is relatively less slow. In other words, among all three general methods discussed above, the most promising one for a model of multiple discrete response with spatial effects (both autocorrelation and heteroskedasticity) is Gibbs sampling within a Bayesian framework.

In contrast to frequentist methods (i.e., classical statistical analysis), the Bayesian approach is rather straightforward in both model estimation and results interpretation. A primary motivation is rather direct interpretation of parameter estimates and probabilities. A Bayesian approach yields estimates of parameter *distributions* (rather than relying on asymptotics for normality). These distributions effectively define intervals that can be "regarded as having a high probability of containing the unknown quantit(ies) of interest" (Gelman et al., 2004). In contrast, frequentist methods focus on producing point estimates and rather standard confidence intervals, and resulting probabilities that are strictly interpreted as "long run (asymptotic) relative frequenc(ies)" (Koop et al., 2007).

In practice, an important advantage of a Bayesian framework is its flexibility, allowing it to deal with complex estimation problems more easily. In fact, this is the main reason for this study's choice of Bayesian framework – in addition to wanting to develop new methods of model estimation for regional and transportation sciences (where frequentist methods are the norm).

In general, Bayesian estimation via Markov chain Monte Carlo (MCMC) simulation relies on a set of conditional distributions to deduce each parameter's marginal distribution. In this way, models with many parameters and complicated multiple-layered probability specifications can be decomposed into a set of simpler sub-problems. By contrast, with frequentist methods, the models have to deal directly with any complicated model specification and any statistical problems arising from it. Of course, another well-understood advantage of using a Bayesian approach is that by having priors, one can make use of established intuition and experience to balance new information found in sample data. Thanks to its many advantages, a Bayesian approach has been used in various areas. For example, Wallerman et al. (2006) relied on Bayesian estimation for remote sensing data in forested areas, and Hamilton et al. (2005) used it to estimate expansion times and migration rates for Swiss populations.

Albert and Chib (1993) introduced the Bayesian approach for (stationary, non-spatial) discrete response data models. LeSage (2000) first extended Albert and Chib's approach to models involving spatial dependencies. Later work by Smith and LeSage (2004) further extended the model, by incorporating an error specification that allows both spatial dependencies and general spatial heteroscedasticity. All such studies, however, deal only with binary data. As previously discussed, many data sets offer multiple categories. No existing studies tackle such patterns in a spatial context. While Albert and Chib (1993) briefly mentioned possible extensions from binary data to ordered categorical data, they did not offer any methodological details. Several years later, Johnson and Albert (1999) suggested a detailed Bayesian framework for modeling ordinal data, and Cowles (1996) presented a method for accelerating MCMC convergence for models like the ordered probit. Girard and Parent (2001) even extended Albert and Chib's study (1993) to temporally autocorrelated ordered categorical data, but there is nothing spatial in these studies. This study is inspired by such studies but adds sophistication while combining space and time for ordered categorical data. It goes beyond a simple extension or combination of these works. The contribution of these prior studies will discussed in more detail in the next section, through illustrations of model specification and estimation techniques.

3 MODEL SPECIFICATION

3.1 Standard Ordered Probit (OP) Model

A standard ordered probit model has been used widely for estimating discrete responses of an ordinal nature (Greene, 2000). The model is built upon a latent regression that is expressed as follows:

$$U_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \xi_i \tag{1}$$

where i indexes observations, ($i = 1,\ldots,N$,) and U_i is a latent (unobserved) response variable for individual i . X_i is a $Q \times 1$ vector of explanatory variables, and β is the set of corresponding parameters. ξ_i stands for unobservable factors for observation i and (for a standard ordered probit model) is assumed to follow an iid standard normal distribution.

The observed response variable, y , for the i^{th} observation is as follows:

$y_i = s$ if $\gamma_{s-1} < U_i < \gamma_s$, $s = 1,\ldots,S$

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_i is allowed to vary between unknown boundaries $\gamma_0 < \gamma_1 < \cdots < \gamma_{S-1} < \gamma_S$, where γ_0 is $-\infty$ and γ_S is $+\infty$. If constants are to be included in the explanatory variables, γ_1 also is normalized to equal zero. The probabilities for these S outcomes are as follows:

$$\begin{aligned}
&\Pr\left(y_i = 1 \middle| \boldsymbol{X}_i\right) = \Phi\left(\gamma_1 - \boldsymbol{X}_i'\boldsymbol{\beta}\right) - \Phi\left(\gamma_0 - \boldsymbol{X}_i'\boldsymbol{\beta}\right)\\
&\Pr\left(y_i = 2 \middle| \boldsymbol{X}_i\right) = \Phi\left(\gamma_2 - \boldsymbol{X}_i'\boldsymbol{\beta}\right) - \Phi\left(\gamma_1 - \boldsymbol{X}_i'\boldsymbol{\beta}\right)\\
&\qquad\vdots\\
&\Pr\left(y_i = S \middle| \boldsymbol{X}_i\right) = \Phi\left(\gamma_S - \boldsymbol{X}_i'\boldsymbol{\beta}\right) - \Phi\left(\gamma_{S-1} - \boldsymbol{X}_i'\boldsymbol{\beta}\right)
\end{aligned} \tag{2}$$

where $\Phi(\bullet)$ is the cumulative distribution function (CDF) for a standard normal distribution.

3.2 Spatial Ordered Probit (SOP) Model

In many studies, individuals are surveyed from a region containing several sub-regions or neighborhoods. A certain number of observations is collected from each of these sub-regions. In such cases, the effects of different regions need to be considered. Smith and LeSage (2004) proposed the following form for the latent variable:

$$U_{ik} = \boldsymbol{X}_{ik}'\boldsymbol{\beta} + \xi_{ik}\text{, with } \xi_{ik} = \theta_i + \varepsilon_{ik} \tag{3}$$

where i now indexes regions (instead of individuals) ($i = 1,...,M$) and k indexes individuals inside each region (i.e., $k = 1,...,n_i$). In other words, there are M regions, each containing n_i observations, so that the total number of observations is $\sum_{i=1}^{M} n_i = N$.

The main difference between Equations (1) and (3) is that the unobserved factor ξ_{ik} is now composed of two parts: a "regional effect" θ_i and an individual effect ε_{ik}. The θ_i captures all unobserved, common features for observations within region i. To some extent, this specification is very close to a random effect in panel data, only here the "common factor" is cross-sectional rather than temporal. Of course, these regional effects are likely to exhibit spatial autocorrelation: individuals in region i are likely to be more similar to those in neighboring regions than those in more distant locations. Therefore, a spatial autoregressive process can be formulated here, where

$$\theta_i = \rho \sum_{j=1}^{M} w_{ij}\theta_j + u_i \,,\ i = 1,...,M \qquad (4)$$

and weight w_{ij} can be derived based on contiguity and/or distance. In addition, the weight matrix is row-standardized[1] so that $w_{ii} = 0$ and $\sum_{j=1}^{M} w_{ij} = 1$. The magnitude of overall neighborhood influence is thus reflected by ρ, 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 σ^2. Stacking all regions, then, the vector of regional effects can be formulated as

$$\boldsymbol{\theta} = \rho \boldsymbol{W}\boldsymbol{\theta} + \boldsymbol{u} \,,\ \boldsymbol{u} \square N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_M\right) \qquad (5)$$

Here, $\boldsymbol{W}$ is the exogenous weight matrix with elements w_{ij} and $\boldsymbol{I}_M$ is an identity matrix with rank M. Let $\boldsymbol{B}_\rho = \boldsymbol{I}_M - \rho \boldsymbol{W}$, where the subscript ρ means that $\boldsymbol{B}_\rho$ depends only on the unknown parameter ρ. Now, the vector of regional effects can be expressed as

$$\boldsymbol{\theta} = \boldsymbol{B}_\rho^{-1} \boldsymbol{u} \qquad (6)$$

In other words, the distribution of $\boldsymbol{\theta}$ depends on two unknown parameters: ρ and σ^2. It has a multivariate normal distribution:

$$\boldsymbol{\theta} \big| \left(\rho, \sigma^2\right) \square N\left[0, \sigma^2 \left(\boldsymbol{B}'_\rho \boldsymbol{B}_\rho\right)^{-1}\right] \qquad (7)$$

[1] The row-standardized approach is chosen because in this way the "*Wy* term becomes essentially a weighted average of observations at neighboring locations" (Anselin and Hudak, 1992). This leads to a more meaningful interpretation of ρ.

The intuition behind this “regional effect” can be explained as follows: In many cases, individuals in a region[2] share common features, yet these features differ from region to region. One source of such differences is policy variation by regions. For example, parcels subject to the same zoning constraints may share common features, but differ across zone boundaries. Animals enjoying the same habitat share experiences, thanks to vegetation and micro climates. Their settings shift across wide rivers, mountain ranges, or high-capacity freeways. Multiple regions may exist based on these physical boundaries. In short, there are reasons to believe that observations across space are influenced by “local effects”, which may exhibit spatial autoregressive patterns as a function of proximity. 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 $\boldsymbol{B}_\rho$ of relatively low rank. Thanks to a lower dimension, the inversion of $\boldsymbol{B}_\rho$ and calculation of its eigenvalues, are much less computer-memory-intensive. Of course, both of these computations are necessary for parameter estimation.

This “regional effect” offers an opportunity to make each individual a region, i.e., $n_i = 1 \;\; \forall i \in M$, (so $M = N$). This allows all individuals to be spatially auto-correlated without imposing regional boundaries. While increasing computational burdens, such a specification is definitely feasible with a reasonable sample size.

The final item requiring specification is the individual effect, ε_{ik} . It is computationally simplest to assume an iid distribution for ε_{ik} . And, within each region, it is behaviorally reasonable to make such assumptions (i.e., all ε_{ik} follow a normal distribution with zero mean and variance υ_i). Across regions, it seems reasonable to expect heteroscedasticity. Stacking all observations and denoting $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0},\boldsymbol{V}\right)$, one has

$$\boldsymbol{V} = \begin{pmatrix} \upsilon_1 \boldsymbol{I}_{n1} & & \\ & \ddots & \\ & & \upsilon_M \boldsymbol{I}_{nM} \end{pmatrix} \qquad (8)$$

which is an $N \times N$ matrix with non-zero elements only along its diagonal.

3.3 Dynamics to the Spatial Ordered Probit Model

In this study, it is assumed that a time-space recursive formulation (Anselin, 1999) is proper for specifying the dynamics and spatial autocorrelation in dataset, which means that the current value depends on the previous period’s value (at the same location, and thus affected by neighboring locations), along with various contemporaneous factors. Furthermore, after controlling for all these temporally lagged and contemporaneous variables, the residuals remain spatially autocorrelated:

[2] As used here, “region” means a cluster of observational units, within the same neighborhood or socially defined group (such as members of the same household or employees in the same firm).

$$U_{ikt} = \lambda U_{ikt-1} + \boldsymbol{X}_{ikt}{}'\boldsymbol{\beta} + \theta_{it} + \varepsilon_{ikt}\,,\ t = 1,...,T \tag{9}$$

where t indexes time periods and λ is the temporal autocorrelation coefficient to be estimated. The absolute value of this λ must be less than one in order to guarantee temporal stationarity. Each individual is now observed T times (the dataset is a balanced panel), and the total number of observations is NT. θ_{it} is assumed to iid distributed over t and so is ε_{ikt}. In other words, after controlling for lagged dependent variables (U_{ikt-1}), the error terms are sequentially uncorrelated and identically distributed. Though a more flexible framework is, of course, to allow θ_{it} and ε_{ikt} to exhibit sequentially dependencies or at least heteroscedasticity, it is reasonable enough to believe that after one controls for lagged latent dependencies (both spatial and temporal), the remaining error terms may be temporally constant, i.e.,

$$\theta_{it} \equiv \theta_i \text{ or } \boldsymbol{\theta}_t = \boldsymbol{\theta}\text{, for all } t = 1,...,T \tag{10}$$

and

$$\varepsilon_{ikt} \equiv \varepsilon_{ik} \text{ or } \boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}\text{, for all } t = 1,...,T \tag{11}$$

Equations (9) through (11) specify a dynamic spatial ordered probit (DSOP) model. Many examples in practice fit this specification. For example, land development decisions strongly depend on pre-existing and existing conditions, as well as owner/developer expectations of future conditions (such as local and regional congestion, population, and school access). These expectations can be approximated using contemporaneous measures of access and land use intensity, after which some spatial correlation in unobserved factors is likely to remain.

Another example is of air quality, and ozone concentration levels: changes are temporally continuous so inclusion of lagged values is wise. The impact of some factors, such as temperature, may be instantaneous, so their contemporaneous values should be used. The process of atmospheric transport and other unobserved factors may cause spatial dependence, so spatially autocorrelated effects (regional/clustered or observational in nature) should be considered. Certainly, recognition of such temporal dependencies and spatial autocorrelation (of nuisance terms) is behaviorally more convincing and statistically more rigorous than simply controlling for contemporaneous factors and ignoring other, underlying spatial dependencies.

The model specification can be expressed in vector form as follows: for each $t \in T$, observations can be stacked by region, then by individuals. The resulting vector of latent responses is expressed as:

$$\boldsymbol{U}_t = \lambda \boldsymbol{U}_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{L}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{12}$$

where $\boldsymbol{U}_t = \begin{bmatrix} \boldsymbol{U}_{1t} \\ \vdots \\ \boldsymbol{U}_{it} \\ \vdots \\ \boldsymbol{U}_{Mt} \end{bmatrix}$, with each $\boldsymbol{U}_{it} = \begin{bmatrix} U_{i1t} \\ \vdots \\ U_{ikt} \\ \vdots \\ U_{in_it} \end{bmatrix}$. It is similar with $\boldsymbol{X}_t$, only $\boldsymbol{X}_t$ is an $N \times Q$ matrix (instead of an $N \times 1$ vector). Here, $\boldsymbol{L} = \begin{bmatrix} \boldsymbol{l}_{n_1} & & \\ & \ddots & \\ & & \boldsymbol{l}_{n_M} \end{bmatrix}$, with each $\boldsymbol{l}_{n_i} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n_i}$ being a $n_i \times 1$ vector of 1's.

If observations over all time periods are stacked, the model can be written as

$$\boldsymbol{U}^{\lambda} = \boldsymbol{X\beta} + \Delta\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{13}$$

where $\boldsymbol{U}^{\lambda}$ is the vector of differences between adjacent time periods: $\boldsymbol{U}^{\lambda} = \left(\boldsymbol{U}_1^{\lambda}, \boldsymbol{U}_2^{\lambda}, \ldots \boldsymbol{U}_T^{\lambda}\right)'$, with each $\boldsymbol{U}_t^{\lambda} = \boldsymbol{U}_t - \lambda\boldsymbol{U}_{t-1}$, and

$$\Delta = \boldsymbol{l}_T \otimes \boldsymbol{L} \tag{14}$$

where $\boldsymbol{l}_T$ is a $T \times 1$ vector of 1's.

Here, $\boldsymbol{X}$ is an $NT \times Q$ matrix, and $\boldsymbol{\varepsilon}$ is an $NT \times 1$ vector with variance matrix

$$\boldsymbol{\Omega} = \boldsymbol{I}_T \otimes \boldsymbol{V} \tag{15}$$

The likelihood function is thus

$$\Pr\left(\boldsymbol{y} \middle| \boldsymbol{U}, \boldsymbol{\gamma}\right) = \prod_{t=1}^{T} \prod_{i=1}^{M} \prod_{k=1}^{n_i} \sum_{s=1}^{S} \delta\left(y_{ikt} = s\right) \cdot \Pr\left(y_{ikt} = s \middle| X_{ikt}\right) \tag{16}$$

where $\delta(A)$ is an indicator function equaling 1 when event A is true (and 0 otherwise). Now it is clear that the parameters of interest are $\left(\boldsymbol{\beta}, \lambda, \rho, \boldsymbol{V}, \sigma^2, \boldsymbol{\gamma}\right)$, together with unobserved ("nuisance") variables $\boldsymbol{\theta}$ and $\boldsymbol{U}$. One way to estimate these is via MCMC sampling under a Bayesian framework, as discussed below.

4. PARAMETER ESTIMATION VIA MCMC SIMULATION

As discussed above, MCMC simulation can be used in model estimation by sampling sequentially from the parameters' complete set of conditional distributions. Gelfand and Smith (1990) showed that MCMC sampling leads to consistent estimates of the true joint posterior distribution of all parameters (including "nuisance parameters", such as $\boldsymbol{V}, \sigma^2$ and θ). Using Bayes' basic rule, the following formulation always holds true:

$$\begin{aligned} & p\left(\boldsymbol{\beta}, \lambda, \rho, \boldsymbol{V}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{U}, \boldsymbol{U}_0 \middle| \boldsymbol{y}\right) \square p\left(\boldsymbol{y}\right) \\ & \quad = p\left(\boldsymbol{y} \middle| \boldsymbol{\beta}, \lambda, \rho, \boldsymbol{V}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{U}, \boldsymbol{U}_0\right) \square \pi\left(\boldsymbol{\beta}, \lambda, \rho, \boldsymbol{V}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{U}, \boldsymbol{U}_0\right) \end{aligned} \tag{17}$$

Here, $\boldsymbol{U}_0$ is a vector for all individuals' latent response levels in the initial period, $p(\bullet)$ indicates posterior densities, and $\pi(\bullet)$ stands for prior distribution assumptions. Assuming certain forms of independent priors, as discussed later, the posterior joint density $p\left(\boldsymbol{\beta},\lambda,\rho,\boldsymbol{V},\sigma^2,\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{U},\boldsymbol{U}_0\middle|\boldsymbol{y}\right)$ will exhibit the following proportionality:

$$\begin{aligned} p\left(\boldsymbol{\beta},\lambda,\rho,\boldsymbol{V},\sigma^2,\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{U},\boldsymbol{U}_0\middle|\boldsymbol{y}\right) &\propto p\left(\boldsymbol{y}\middle|\boldsymbol{U},\boldsymbol{\gamma}\right)\cdot\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi\left(\boldsymbol{\theta}\middle|\rho,\sigma^2\right) \\ &\cdot\pi(\boldsymbol{\gamma})\cdot\pi(\boldsymbol{U}_0)\cdot\pi(\boldsymbol{\beta})\cdot\pi(\rho)\cdot\pi\left(\sigma^2\right)\cdot\pi(\lambda)\cdot\pi(\boldsymbol{V}) \end{aligned} \tag{18}$$

From Equation (18), the conditional distributions can be derived as follows, for each parameter and variable of interest. The $\boldsymbol{\Theta}_{\beta}$ (or $\boldsymbol{\Theta}_{\lambda}$, $\boldsymbol{\Theta}_{\theta}$, etc.) in these formulations represents the set of conditional arguments for the conditional distribution of $\boldsymbol{\beta}$ (or λ, $\boldsymbol{\theta}$, etc.). It includes all arguments except $\boldsymbol{\beta}$ (or λ, $\boldsymbol{\theta}$, etc.). (For example, $\boldsymbol{\Theta}_{\beta}$ stands for the set $\left(\lambda,\rho,\boldsymbol{V},\sigma^2,\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{U},\boldsymbol{U}_0,\boldsymbol{y}\right)$.)

$$p\left(\boldsymbol{\beta}\middle|\boldsymbol{\Theta}_{\beta}\right)\propto\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi(\boldsymbol{\beta}) \tag{19}$$

$$p\left(\boldsymbol{\theta}\middle|\boldsymbol{\Theta}_{\theta}\right)\propto\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi\left(\boldsymbol{\theta}\middle|\rho,\sigma^2\right) \tag{20}$$

$$p\left(\lambda\middle|\boldsymbol{\Theta}_{\lambda}\right)\propto\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi(\lambda) \tag{21}$$

$$p\left(\rho\middle|\boldsymbol{\Theta}_{\rho}\right)\propto\pi\left(\boldsymbol{\theta}\middle|\rho,\sigma^2\right)\cdot\pi(\rho) \tag{22}$$

$$p\left(\sigma^2\middle|\boldsymbol{\Theta}_{\sigma^2}\right)\propto\pi\left(\boldsymbol{\theta}\middle|\rho,\sigma^2\right)\cdot\pi\left(\sigma^2\right) \tag{23}$$

$$p\left(\boldsymbol{V}\middle|\boldsymbol{\Theta}_{V}\right)\propto\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi(\boldsymbol{V})=\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\prod_{i=1}^{M}\pi(\upsilon_i) \tag{24}$$

$$p\left(\boldsymbol{\gamma}\middle|\boldsymbol{\Theta}_{\gamma}\right)\propto p\left(\boldsymbol{y}\middle|\boldsymbol{U},\boldsymbol{\gamma}\right)\cdot\pi(\boldsymbol{\gamma}) \tag{25}$$

$$p\left(\boldsymbol{U}_0\middle|\boldsymbol{\Theta}_{U_0}\right)\propto\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi(\boldsymbol{U}_0) \tag{26}$$

$$p\left(\boldsymbol{U}\middle|\boldsymbol{\Theta}_{U}\right)\propto p\left(\boldsymbol{y}\middle|\boldsymbol{U},\boldsymbol{\gamma}\right)\cdot\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right) \tag{27}$$

The formulations found in Equations 19 through 27 involve three factors: $\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)$, $p\left(\boldsymbol{y}\middle|\boldsymbol{U},\boldsymbol{\gamma}\right)$ and $\pi\left(\boldsymbol{\theta}\middle|\rho,\sigma^2\right)$, and the following paragraphs discuss these factors in more detail.

From Equation (12), it can be observed that for all $t\neq 0, t=1...,T$, $\boldsymbol{U}_t\middle|\boldsymbol{U}_{\neq t},\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\sim N\left(\lambda\boldsymbol{U}_{t-1}+\boldsymbol{X}_t\boldsymbol{\beta}+\boldsymbol{L\theta},\boldsymbol{V}\right)$, so the conditional prior distribution can be expressed as follows:

$$\begin{aligned} \pi\left(\boldsymbol{U}_t\middle|\boldsymbol{U}_{\neq t},\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right) &= \left|\boldsymbol{V}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{U}_t-\lambda\boldsymbol{U}_{t-1}-\boldsymbol{L\theta}-\boldsymbol{X}_t\boldsymbol{\beta}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{U}_t-\lambda\boldsymbol{U}_{t-1}-\boldsymbol{L\theta}-\boldsymbol{X}_t\boldsymbol{\beta}\right)\right\} \\ &= \left|\boldsymbol{V}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{U}_t^{\lambda}-\boldsymbol{L\theta}-\boldsymbol{X}_t\boldsymbol{\beta}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{U}_t^{\lambda}-\boldsymbol{L\theta}-\boldsymbol{X}_t\boldsymbol{\beta}\right)\right\} \end{aligned} \tag{28}$$

Therefore, for $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots \boldsymbol{U}_T)'$, one has the following:

$$\pi\left(\boldsymbol{U} \middle| \boldsymbol{U}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \lambda, \boldsymbol{V}\right) = \prod_{t=1}^{T} |\boldsymbol{V}|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{U}_t^{\lambda} - \boldsymbol{L\theta} - \boldsymbol{X}_t\boldsymbol{\beta}\right)' \boldsymbol{V}^{-1}\left(\boldsymbol{U}_t^{\lambda} - \boldsymbol{L\theta} - \boldsymbol{X}_t\boldsymbol{\beta}\right)\right\}$$
$$= |\Omega|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{U}^{\lambda} - \Delta\boldsymbol{\theta} - \boldsymbol{X\beta}\right)' \Omega^{-1}\left(\boldsymbol{U}^{\lambda} - \Delta\boldsymbol{\theta} - \boldsymbol{X\beta}\right)\right\} \qquad (29)$$

Alternatively, this can be expressed as

$$\pi\left(\boldsymbol{U} \middle| \boldsymbol{U}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \lambda, \boldsymbol{V}\right) = \prod_{t=1}^{T}\prod_{i=1}^{M}\prod_{k=1}^{n_i}\left\{\upsilon_i^{-1/2} \exp\left[\frac{-1}{2\upsilon_i}\left(U_{ikt} - \lambda U_{ikt-1} - \boldsymbol{X}_{ikt}{}'\boldsymbol{\beta} - \theta_i\right)^2\right]\right\} \qquad (30)$$

Since $p\left(\boldsymbol{y} \middle| \boldsymbol{U}, \boldsymbol{\gamma}\right)$ is already given by Equation (16) and $\boldsymbol{\theta}\left|\left(\rho, \sigma^2\right)\right.$ is given by Equation (7), the following holds:

$$\pi\left(\boldsymbol{\theta} \middle| \rho, \sigma^2\right) = \sigma^{-M/2}\left|\boldsymbol{B}_\rho\right| \exp\left(\frac{-1}{2\sigma^2}\boldsymbol{\theta}'\boldsymbol{B}_\rho'\boldsymbol{B}_\rho\boldsymbol{\theta}\right) \qquad (31)$$

4.1 Prior Distributions for All Parameters

Diffuse priors are a valuable type of prior distribution commonly used in Bayesian statistics. These "non-informative" or "flat" priors reflect the notion of "letting the data speak for themselves." For studies with no established prior information (such as this study), diffuse priors are a necessary starting point. Gelman et al. (2004) pointed out that diffuse priors imply that the posterior distributions for all parameters are weighted averages of standard maximum likelihood estimators and prior mean values. This implies that when the study enjoys a large sample, the dataset will overcome all prior information, asymptotically. In this study, most priors take the forms assumed by Smith and LeSage (2004), while others are similar to those used in work by Girard and Parent (2001). All are diffuse.

Here, the parameter set $\boldsymbol{\beta}$ is assumed to have a multivariate normal conjugate prior:

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{c}, \boldsymbol{H}\right) \qquad (32)$$

where $\boldsymbol{H} = h\boldsymbol{I}_Q$. For small $\boldsymbol{c}$ and large h, this prior becomes diffuse. Of course, if one has valid reasons for specifying other values of $\boldsymbol{c}$ and h, it can be very helpful, particularly with small sample sizes (where priors carry more weight). Estimation may be improved through experience and intuition, which can impact selection of priors.

Prior assumptions are similar with the threshold parameters:

$$\boldsymbol{\gamma} \sim N\left(\boldsymbol{q}, \boldsymbol{G}\right) \cdot \delta\left(\gamma_1 < \gamma_2 < \ldots < \gamma_{S-1}\right) \qquad (33)$$

where $\boldsymbol{q}$ is a $S \times 1$ vector, with elements γ_{s0}, and $\boldsymbol{G}$ is a diagonal matrix, with elements g_s on its diagonal and zeros elsewhere. In this way, the threshold parameters also follow a normal conjugate prior, only now with one more constraint to ensure that all probabilities derived from these thresholds are positive. So as $\boldsymbol{q}$ approaches zero and g_s approaches infinity, this also becomes a diffuse prior.

The variances of regional effects σ^2 and individual effects υ_i are assumed to be conjugate inverse-gamma priors:

$$1/\sigma^2 \sim \Gamma(\alpha,\tau) \tag{34}$$

More specifically, σ^2 is given a diffuse prior by setting parameters $\alpha=\tau=0$. All υ_i are assumed to follow an inverse chi-square distribution with hyperparameter ϖ, which is a special case of the inverse gamma:

$$r/\upsilon_i \sim \chi^2(\varpi) \tag{35}$$

Here, the spatial autocorrelation coefficient ρ is given a uniform prior that is diffuse. As Sun et al. (1999) prove, the lower and upper bounds for ρ are determined by the inverse of eigenvalues from weight matrix $\boldsymbol{W}$. Let $\varsigma_{\min}$ and $\varsigma_{\max}$ denote the minimum and maximum eigenvalues; then,

$$\rho \sim U\left[\varsigma_{\min}^{-1},\varsigma_{\max}^{-1}\right] \tag{36}$$

In other words,

$$\pi(\rho)\propto 1 \tag{37}$$

Here, λ is specified to have a normal distribution but limited to the range $(-1,1)$ in order to ensure stationarity:

$$\lambda \sim N(\lambda_0,D)\cdot\delta(|\lambda|<1) \tag{38}$$

Selection of an initial value for the latent variable U is termed the “initial condition problem.” Many have discussed this complicated issue (e.g., Vishniac, 1993; Wooldrige, 2005; and Barlevy and Nagaraja, 2006). Here U_0 is assumed to be normally distributed, in order to be compatible with U’s distribution in other periods. It has the following prior:

$$\boldsymbol{U}_0 \sim N\left(a_0\boldsymbol{l}_N,d_0\boldsymbol{I}_N\right) \tag{39}$$

where $\boldsymbol{l}_N$ is a $N\times 1$ vector with all elements equal to 1 and $\boldsymbol{I}_N$ is an N-dimension identity matrix. Therefore, this distribution approximates a diffuse prior when a_0 is bounded and d_0 goes to infinity.

4.2 Full Conditional Posterior Distributions

Based on the conditional posterior distributions and the parameters’ prior distributions, this section explains how each parameter’s conditional posterior distribution can be mathematically derived.

4.2.1 Conditional Posterior Distribution of $\boldsymbol{\beta}$

From Equations (19) and (31), it can be derived that

$$\begin{aligned} p\left(\boldsymbol{\beta}\middle|\boldsymbol{\Theta}_{\beta}\right) &\propto \pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\Box\pi\left(\boldsymbol{\beta}\right) \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{c})'\boldsymbol{H}^{-1}(\boldsymbol{\beta}-\boldsymbol{c})\right\}\Box \\ &\quad \exp\left\{-\frac{1}{2}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}\right)'\Omega^{-1}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}\right)\right\} \end{aligned} \tag{40}$$

As many previous studies show (e.g., Gelman et al., 2004; and Smith and LeSage, 2004), this form can be simplified to

$$p\left(\boldsymbol{\beta}\middle|\boldsymbol{\Theta}_{\beta}\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{A}^{-1}\boldsymbol{b}\right)'\boldsymbol{A}\left(\boldsymbol{\beta}-\boldsymbol{A}^{-1}\boldsymbol{b}\right)\right] \tag{41}$$

where $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}+\boldsymbol{H}^{-1}$ (42)

and $\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}\right)+\boldsymbol{H}^{-1}\boldsymbol{c}$. (43)

These equations indicate that the posterior mean vector for $\boldsymbol{\beta}$ is $\boldsymbol{A}^{-1}\boldsymbol{b}$ and the variance-covariance matrix is $\boldsymbol{A}^{-1}$. In fact, as Gelman et al. (2004) show, such a posterior distribution is a weighted average of $\boldsymbol{\beta}$'s prior distribution and sample data information and the weights are the *inverse* of the variance-covariance matrices or associated "uncertainty" levels. Using maximum likelihood estimation methods, the estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{MLE} = \left(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\Omega^{-1}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}\right) \tag{44}$$

Here, the prior mean of $\boldsymbol{\beta}$ is assumed to be $\boldsymbol{c}$ and its prior variance is assumed to be ***H***. It is not difficult to show that the posterior mean can then be written as follows:

$$\begin{aligned} E\left(\boldsymbol{\beta}\middle|\boldsymbol{\Theta}_{\beta}\right) &= \boldsymbol{A}^{-1}\boldsymbol{b} \\ &= \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}+\boldsymbol{H}^{-1}\right)\Box\left[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}\right)+\boldsymbol{H}^{-1}\boldsymbol{c}\right] \\ &= \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}+\boldsymbol{H}^{-1}\right)\Box\left[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{H}^{-1}\boldsymbol{c}\right] \end{aligned} \tag{45}$$

In Equation (44), as sample size and information quality increase, the variance $\boldsymbol{\Omega}$ should decrease, which allows $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}$ to dominate, giving $\hat{\boldsymbol{\beta}}_{MLE}$ more weight.

4.2.2 Conditional Posterior Distribution of $\boldsymbol{\theta}$

Some manipulation of Equations (20) and (30) can show that

$$p\left(\boldsymbol{\theta}\middle|\boldsymbol{\Theta}_{\boldsymbol{\theta}}\right) \propto \pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\Box\pi\left(\boldsymbol{\theta}\middle|\rho,\sigma^2\right)$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}\right)'\,\Omega^{-1}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}\right)\right\}\cdot\exp\left(\frac{-1}{2\sigma^{2}}\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}\right)\\
&=\exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'\left(\sigma^{-2}\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\right)\boldsymbol{\theta}+\boldsymbol{\theta}'\Delta'\Omega^{-1}\Delta\boldsymbol{\theta}-2\left(\boldsymbol{U}^{\lambda}-\boldsymbol{X\beta}\right)\Omega^{-1}\Delta\boldsymbol{\theta}+\boldsymbol{C}\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'\left(\sigma^{-2}\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}+\Delta'\Omega^{-1}\Delta\right)\boldsymbol{\theta}-2\left(\boldsymbol{U}^{\lambda}-\boldsymbol{X\beta}\right)\Omega^{-1}\Delta\boldsymbol{\theta}\right]\right\} \qquad (43)
\end{aligned}
$$

where $\boldsymbol{C}$ stands for the constant term, which does not involve $\boldsymbol{\theta}$. Similar to the derivation of the conditional posterior distribution for $\boldsymbol{\beta}$, it can be shown that

$$
\boldsymbol{p}\left(\boldsymbol{\theta}\middle|\boldsymbol{\Theta_{\theta}}\right)\propto\exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{A_{\theta}}^{-1}\boldsymbol{b_{\theta}}\right)'\boldsymbol{A_{\theta}}\left(\boldsymbol{\theta}-\boldsymbol{A_{\theta}}^{-1}\boldsymbol{b_{\theta}}\right)\right] \qquad (47)
$$

where $\boldsymbol{A_{\theta}}=\sigma^{-2}\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}+\Delta'\Omega^{-1}\Delta$ (48)

and $\boldsymbol{b_{\theta}}=\Delta'\Omega^{-1}\left(\boldsymbol{U}^{\lambda}-\boldsymbol{X\beta}\right)$. (49)

These equations indicate that the mean vector for $\boldsymbol{\theta}$ is $\boldsymbol{A_{\theta}}^{-1}\boldsymbol{b_{\theta}}$ and the variance-covariance matrix is $\boldsymbol{A_{\theta}}^{-1}$. It should be noticed here, however, that $\boldsymbol{A_{\theta}}$ depends on $\boldsymbol{B}_{\rho}$, which depends on ρ. That is, each random draw involves a matrix inversion. This computation demands much memory, especially when the number of regions (M) is large. Therefore, an appropriate sampling approach is very important. There are two alternative ways to calculate this matrix inverse. One is to compute the inverse directly; the other way, as Smith and LeSage (2004) suggest (when M is larger), is to sample from univariate normal distributions for each θ_i conditional on all other elements of $\boldsymbol{\theta}$ (excluding the ith element), which is the approach used here.

4.2.3 Conditional Posterior Distribution of λ

From Equations (20), (29), and (30), one can obtain the full form of λ's conditional posterior distribution, written as follows:

$$
\begin{aligned}
p\left(\lambda\middle|\boldsymbol{\Theta}_{\lambda}\right)&\propto\pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_0,\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\pi\left(\lambda\right)\\
&\propto\exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{U}_t-\lambda\boldsymbol{U}_{t-1}-\boldsymbol{L\theta}-\boldsymbol{X}_t\boldsymbol{\beta}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{U}_t-\lambda\boldsymbol{U}_{t-1}-\boldsymbol{L\theta}-\boldsymbol{X}_t\boldsymbol{\beta}\right)\right\}\cdot\\
&\qquad\exp\left\{-\frac{1}{2}\left(\lambda-\lambda_0\right)'D^{-1}\left(\lambda-\lambda_0\right)\right\}\cdot\delta\left(\left|\lambda\right|<1\right)
\end{aligned} \qquad (50)
$$

This is another conjugate distribution; so, similar to $\boldsymbol{\beta}$'s conditional posterior distribution:

$$
p\left(\lambda\middle|\boldsymbol{\Theta}_{\lambda}\right)\propto\exp\left[-\frac{1}{2}\left(\lambda-A_{\lambda}^{-1}b_{\lambda}\right)'A_{\lambda}\left(\lambda-A_{\lambda}^{-1}b_{\lambda}\right)\right]\cdot\delta\left(\left|\lambda\right|<1\right) \qquad (51)
$$

where $A_{\lambda}=\sum_{t=1}^{T}\boldsymbol{U}_{t-1}{}'\boldsymbol{V}^{-1}\boldsymbol{U}_{t-1}+D^{-1}$ (52)

and $b_{\lambda}=\sum_{t=1}^{T}\boldsymbol{U}_{t-1}'\boldsymbol{V}^{-1}\left(\boldsymbol{U}_t-\boldsymbol{X}_t\boldsymbol{\beta}-\boldsymbol{L\theta}\right)+D^{-1}\lambda_0$. (53)

One evident difference between this distribution of λ and the distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is that this is a truncated normal. In each draw, the value of λ needs to be limited to $\left(-1,1\right)$.

4.2.4 Conditional Posterior Distribution of ρ

Equations (21), (30) and (36) lead to the following formulation for ρ's conditional posterior distribution:

$$
\begin{aligned}
p\left(\rho \middle| \boldsymbol{\Theta}_{\rho}\right) &\propto \pi\left(\boldsymbol{\theta} \middle| \rho,\sigma^2\right)\cdot\pi\left(\rho\right) \\
&\propto \left|\boldsymbol{B}_{\rho}\right|\cdot\exp\left(\frac{-1}{2\sigma^2}\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}\right) \qquad (54)
\end{aligned}
$$

and $\rho \in \left[\varsigma_{\min}^{-1},\varsigma_{\max}^{-1}\right]$. As Smith and LeSage (2004) point out, this expression cannot be simplified into a standard distribution. They further suggest that one may use univariate numerical integration to obtain this posterior density, as described below.

First, a range of ρ values between $\left[\varsigma_{\min}^{-1},\varsigma_{\max}^{-1}\right]$ is generated from a uniform distribution. Before MCMC sampling, a vector of determinant values for $\boldsymbol{B}_{\rho}$ corresponding to this range of ρ values can be constructed. Thus, during the iterative sampling process, only the second item ($\exp\left(\frac{-1}{2\sigma^2}\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}\right)$) needs to be updated for each draw. Equation (53) is then numerically integrated (via a sum of point-area estimates) over the range of ρ values. The normalizing constant is obtained, given the condition that ρ is limited to the interval $\left[\varsigma_{\min}^{-1},\varsigma_{\max}^{-1}\right]$, and this renders Equation (53)'s proportionality an equality. After this approximation for ρ's CDF is acquired, one can randomly draw the ρ value from its inversion. As Smith and LeSage (2004) have suggested, the advantage of this approach (over a standard Metropolis-Hastings approach) is that it is more efficient: each pass through the sampler produces a draw for ρ.

4.2.5 Conditional Posterior Distribution of σ^2

From Equations (22), (30) and (33), the following distribution for σ^2 can be obtained:

$$
\begin{aligned}
p\left(\sigma^2 \middle| \boldsymbol{\Theta}_{\sigma^2}\right) &\propto \pi\left(\boldsymbol{\theta} \middle| \rho,\sigma^2\right)\cdot\pi\left(\sigma^2\right) \\
&\propto \left(\sigma^2\right)^{-M/2}\cdot\exp\left(\frac{-1}{2\sigma^2}\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}\right)\cdot\left(\sigma^2\right)^{-(\alpha+1)}\cdot\exp\left(\frac{\tau}{\sigma^2}\right) \qquad (55) \\
&= \left(\sigma^2\right)^{-(M/2+\alpha+1)}\cdot\exp\left(-\frac{\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau}{2\sigma^2}\right)
\end{aligned}
$$

This is an inverse gamma distribution with shape parameter $-M/2+\alpha$ and scale parameter $\left(\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau\right)/2$.

Letting $\kappa = \left(\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau\right)/\sigma^2$, so that $\sigma^2 = \left(\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau\right)/\kappa$, and following the work of Geweke (1993), Equation (54) can be expressed as follows:

$$
\begin{aligned}
p\left(\sigma^{2}\middle|\boldsymbol{\Theta}_{\sigma^{2}}\right) &\propto \left(\left(\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau\right)/\kappa\right)^{-(M/2+\alpha+1)}\cdot\exp\left(-\frac{\kappa}{2}\right)\cdot\left|\frac{d\sigma^{2}}{d\kappa}\right| \\
&=\left(\left(\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau\right)/\kappa\right)^{-(M/2+\alpha+1)}\cdot\exp\left(-\frac{\kappa}{2}\right)\cdot\left|\frac{\left(\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau\right)}{\kappa^{2}}\right| \\
&\propto \kappa^{(M/2+\alpha-1)}\exp\left(-\frac{\kappa}{2}\right)
\end{aligned}
\tag{56}
$$

This density is proportional to a chi-square density with $M+2\alpha$ degrees of freedom (DOF). Alternatively, the conditional posterior of σ^2 can be expressed as

$$
\left.\frac{\boldsymbol{\theta}'\boldsymbol{B}_{\rho}'\boldsymbol{B}_{\rho}\boldsymbol{\theta}+2\tau}{\sigma^{2}}\right|\boldsymbol{\Theta}_{\sigma^{2}} \sim \chi^{2}\left(M+2\alpha\right)
\tag{57}
$$

4.2.6 Conditional Posterior Distribution of *V*

From Equations (23) and (34), it can be shown that

$$
\begin{aligned}
p\left(\boldsymbol{V}\middle|\boldsymbol{\Theta}_{V}\right) &\propto \pi\left(\boldsymbol{U}\middle|\boldsymbol{U}_{0},\boldsymbol{\beta},\boldsymbol{\theta},\lambda,\boldsymbol{V}\right)\cdot\prod_{i=1}^{M}\pi\left(\upsilon_{i}\right) \\
&\propto\left|\boldsymbol{\Omega}\right|^{-1/2}\cdot\exp\left\{-\frac{1}{2}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}\right)\right\}\cdot\prod_{i=1}^{M}\pi\left(\upsilon_{i}\right)
\end{aligned}
\tag{58}
$$

By letting $\boldsymbol{e}=\boldsymbol{U}^{\lambda}-\Delta\boldsymbol{\theta}-\boldsymbol{X\beta}$, the distribution of $\boldsymbol{V}$ can also be derived term by term for each i:

$$
\begin{aligned}
p\left(\upsilon_{i}\middle|\boldsymbol{\Theta}_{\upsilon_{i}}\right) &\propto\left|\boldsymbol{\Omega}\right|^{-1/2}\cdot\exp\left(-\frac{1}{2}\boldsymbol{e}'\boldsymbol{\Omega}^{-1}\boldsymbol{e}\right)\cdot\pi\left(\upsilon_{i}\right) \\
&=\prod_{j=1}^{M}\left[\upsilon_{j}^{-n_{j}T/2}\cdot\exp\left(-\sum_{t=1}^{T}\sum_{j=1}^{M}\frac{\boldsymbol{e}_{jt}'\boldsymbol{e}_{jt}}{2\upsilon_{j}}\right)\cdot\upsilon_{j}^{-\left(\frac{\varpi}{2}+1\right)}\cdot\exp\left(-\frac{\varpi}{2\upsilon_{j}}\right)\right] \\
&\propto\upsilon_{i}^{-n_{i}T/2}\cdot\exp\left(-\sum_{t=1}^{T}\frac{\boldsymbol{e}_{it}'\boldsymbol{e}_{it}}{2\upsilon_{i}}\right)\cdot\upsilon_{i}^{-\left(\frac{\varpi}{2}+1\right)}\cdot\exp\left(-\frac{\varpi}{2\upsilon_{i}}\right) \\
&=\upsilon_{i}^{-\left(\frac{\varpi+n_{i}T}{2}+1\right)}\cdot\exp\left(-\frac{\sum_{t=1}^{T}\boldsymbol{e}_{it}'\boldsymbol{e}_{it}+\varpi}{2\upsilon_{i}}\right)
\end{aligned}
\tag{59}
$$

Similar to the derivation of σ^2's posterior distribution, letting $\kappa_i=\dfrac{\sum_{t=1}^{T}\boldsymbol{e}_{it}'\boldsymbol{e}_{it}+\varpi}{\upsilon_i}$, then this can be shown as the following chi-square distribution:

$$p\left(\upsilon_i \middle| \boldsymbol{\Theta}_{\upsilon_i}\right) = \left(\frac{\sum_{t=1}^{T} \boldsymbol{e}_{it}{}' \boldsymbol{e}_{it} + \varpi}{\kappa_i} \right)^{-\left(\frac{\varpi + n_i T}{2} + 1\right)} \cdot \exp\left(-\frac{\kappa_i}{2}\right) \cdot \frac{\sum_{t=1}^{T} \boldsymbol{e}_{it}{}' \boldsymbol{e}_{it} + \varpi}{\kappa_i^{2}}$$

$$= \kappa_i^{\left(\frac{\varpi + n_i T}{2} - 1\right)} \cdot \exp\left(-\frac{\kappa_i}{2}\right) \tag{60}$$

Therefore, it follows a chi-square density with $r + n_i T$ degrees of freedom, i.e.,

$$\left. \frac{\sum_{t=1}^{T} \boldsymbol{e}_{it}{}' \boldsymbol{e}_{it} + \varpi}{\upsilon_i} \right| \boldsymbol{\Theta}_{\upsilon_i} \sim \chi^2\left(\varpi + n_i T\right) \tag{61}$$

Similar to $\boldsymbol{\beta}$, Smith and LeSage (2004) show that the posterior mean of υ_i is a weighted average of the maximum likelihood estimator $\hat{\upsilon}_i$ and the prior mean, μ_i, which equals $\varpi/(\varpi - 2)$. In the dynamic model, the weights can be calculated using a method very similar to that suggested by Smith and LeSage (2004). These weights are $n_i T$ and $\varpi - 2$, respectively.

As expected, this means that more weight is given to the sample information as sample size n_i or the panel length, T, increases. ϖ needs to be larger than 2, but also needs to be kept small if one wants to use a diffuse prior. Here the hyperparameter ϖ is assumed to be 4.

4.2.7 Conditional Posterior Distribution of $\boldsymbol{\gamma}$

Equations (16), (24), and (32) lead to the following formulation for the conditional posterior distribution of $\boldsymbol{\gamma}$:

$$\begin{aligned} p\left(\boldsymbol{\gamma} \middle| \boldsymbol{\Theta}_{\gamma}\right) &\propto p\left(\boldsymbol{y} \middle| \boldsymbol{U}, \boldsymbol{\gamma}\right) \cdot \pi\left(\boldsymbol{\gamma}\right) \\ &\propto \left[\prod_{t=1}^{T} \prod_{i=1}^{M} \prod_{k=1}^{n_i} \sum_{s=1}^{S} \delta\left(y_{ikt} = s\right) \cdot \delta\left(\gamma_{s-1} < U_{ikt} < \gamma_s\right) \right] \cdot \\ &\quad N\left(\boldsymbol{q}, \boldsymbol{G}\right) \cdot \delta\left(\gamma_1 < \gamma_2 < \ldots < \gamma_{S-1}\right) \end{aligned} \tag{62}$$

This equation can be considered term by term for $s = 1, \ldots S - 1$, by only extracting terms that involve γ_s:

$$\begin{aligned} p\left(\gamma_s \middle| \boldsymbol{\Theta}_{r_s}\right) &\propto \prod_{t=1}^{T} \prod_{i=1}^{M} \prod_{k=1}^{n_i} \delta\left(U_{ikt} < \gamma_s \middle| y_{ikt} = s\right) \cdot \prod_{t=1}^{T} \prod_{i=1}^{M} \prod_{k=1}^{n_i} \delta\left(U_{ikt} > \gamma_s \middle| y_{ikt} = s + 1\right) \cdot \\ &\quad N\left(\gamma_{s0}, g_s\right) \cdot \delta\left(\gamma_{s-1} < \gamma_s < \gamma_{s+1}\right) \\ &= \delta\left(\gamma_s^{\inf} < \gamma_s < \gamma_s^{\sup}\right) \cdot \exp\left\{ -\frac{1}{2 g_s} \left(\gamma_s - \gamma_{s0}\right)^2 \right\} \end{aligned} \tag{63}$$

With $\gamma_s^{\inf} = \max\left\{\max\left\{U_{ikt} : y_{ikt} = s\right\}; \gamma_{s-1}\right\}$ (64)

and $\gamma_s^{\sup} = \min\left\{\min\left\{U_{ikt} : y_{ikt} = s+1\right\}; \gamma_{s+1}\right\}$, for $\forall i \in M, k \in n_i, t \in T$. (65)

Similar to the derivation for λ, this is a truncated normal distribution. The normalizing constant can be found using a univariate normal distribution, with the given lower and upper bounds. The major difference is, however, that these lower and upper bounds are interdependent, which may make the final posterior distribution multimodal.

4.2.8 Conditional Posterior Distribution of $\boldsymbol{U}_0$

Substituting Equations (29) and (38) into Equation (25), one can get the following formulation:

$$\begin{aligned} p\left(\boldsymbol{U}_0 \middle| \boldsymbol{\Theta}_{U_0}\right) &\propto \pi\left(\boldsymbol{U} \middle| \boldsymbol{U}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \lambda, \boldsymbol{V}\right) \cdot \pi\left(\boldsymbol{U}_0\right) \\ &\propto \prod_{t=1}^{T}\prod_{i=1}^{M}\prod_{k=1}^{n_i}\left\{\exp\left[-\frac{1}{2\upsilon_i}\left(U_{ikt} - \lambda U_{ikt-1} - \theta_i - \boldsymbol{X}_{ikt}\boldsymbol{\beta}\right)^2\right] \cdot \exp\left[-\frac{1}{2d_0}\left(U_{ik0} - a_0\right)^2\right]\right\} \\ &\propto \prod_{i=1}^{M}\prod_{k=1}^{n_i}\exp\left[-\frac{\left(U_{ik1} - \lambda U_{ik0} - \theta_i - \boldsymbol{X}_{ik1}\boldsymbol{\beta}\right)^2}{2\upsilon_i} - \frac{\left(U_{ik0} - a_0\right)^2}{2d_0}\right] \end{aligned} \tag{66}$$

Deriving U_0 term by term for each *i* and *k*, by extracting only items involving U_{ik0}, Equation (66) reduces to:

$$p\left(U_{ik0} \middle| \boldsymbol{\Theta}_{U_{ik0}}\right) \propto \exp\left[-\frac{\left(U_{ik1} - \lambda U_{ik0} - \theta_i - \boldsymbol{X}_{ik1}\boldsymbol{\beta}\right)^2}{2\upsilon_i} - \frac{\left(U_{ik0} - a_0\right)^2}{2d_0}\right] \tag{67}$$

This is a univariate normal distribution. Thus, similar to the posterior distribution calculations for $\boldsymbol{\beta}$, the distribution is as follows:

$$U_{ik0} \middle| \boldsymbol{\Theta}_{U_{ik0}} \sim N\left(A_{U0}^{-1} b_{U0}, A_{U0}^{-1}\right) \tag{68}$$

where $A_{U0} = \lambda^2 \upsilon_i^{-1} + d_0^{-1}$ (69)

and $b_{U0} = \lambda \upsilon_i^{-1}\left(U_{ik1} - \theta_i - \boldsymbol{X}_{ik1}\boldsymbol{\beta}\right) + d_0^{-1} a_0$. (70)

4.2.9 Conditional Posterior Distribution of $\boldsymbol{U}$

For latent variables other than the initial status, Equations (15), (26), and (29) lead to the following formulation:

$$\begin{aligned} p\left(\boldsymbol{U} \middle| \boldsymbol{\Theta}_U\right) &\propto p\left(\boldsymbol{y} \middle| \boldsymbol{U}, \boldsymbol{\gamma}\right) \cdot \pi\left(\boldsymbol{U} \middle| \boldsymbol{U}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \lambda, \boldsymbol{V}\right) \\ &\propto \prod_{t=1}^{T}\prod_{i=1}^{M}\prod_{k=1}^{n_i}\left\{\begin{array}{l} \left[\sum_{s=1}^{S}\delta\left(y_{ikt} = s\right) \cdot \delta\left(\gamma_{s-1} < U_{ikt} < \gamma_s\right)\right] \cdot \\ \upsilon_i^{-1/2} \cdot \exp\left[-\frac{1}{2\upsilon_i}\left(U_{ikt} - \lambda U_{ikt-1} - \theta_i - \boldsymbol{X}_{ikt}\boldsymbol{\beta}\right)^2\right] \end{array}\right\} \end{aligned} \tag{71}$$

U_{ikt} appears in the formulation for both periods t and $t+1$. Therefore, for any i,k,t observation, by extracting only items involving U_{ikt}, the posterior distribution for U_{ikt} can be expressed as follows:

$$p\left(U_{ikt}\middle|\boldsymbol{\Theta}_{U_{ikt}}\right) \propto \left\{\sum_{s=1}^{S}\left[\delta\left(y_{ikt}=s\right)\square\delta\left(\gamma_{s-1}<U_{ikt}<\gamma_{s}\right)\right]\right\}\square\upsilon_i^{-1}\square \qquad (72)$$

$$\exp\left\{-\frac{1}{2\upsilon_i}\left[\left(U_{ikt}-\lambda U_{ikt-1}-\theta_i-\boldsymbol{X}_{ikt}\boldsymbol{\beta}\right)^2+\left(U_{ikt+1}-\lambda U_{ikt}-\theta_i-\boldsymbol{X}_{ikt+1}\boldsymbol{\beta}\right)^2\right]\right\}$$

This is a truncated normal distribution. The first expression in Equation (3.74),

$$\sum_{s=1}^{S}\left[\delta\left(y_{ikt}=s\right)\square\delta\left(\gamma_{s-1}<U_{ikt}<\gamma_{s}\right)\right]$$

indicates that if $y_{ikt}=s$, the distribution is truncated on the left by γ_{s-1} and on the right by γ_s. The last item in Equation (71), $\exp\left\{-\frac{1}{2\upsilon_i}\left[\left(U_{ikt}-\lambda U_{ikt-1}-\theta_i-\boldsymbol{X}_{ikt}\boldsymbol{\beta}\right)^2+\left(U_{ikt+1}-\lambda U_{ikt}-\theta_i-\boldsymbol{X}_{ikt+1}\boldsymbol{\beta}\right)^2\right]\right\}$ suggests that the un-truncated part is a normal distribution. This part has mean a_{ikt} and variance b_{ikt}. (Readers may wish to see Appendix B for more details on this.)

Here, $a_{ikt}=\left[\lambda U_{ikt+1}+\lambda U_{ikt-1}+\left(1-\lambda\right)\theta_i+\left(\boldsymbol{X}_{ikt}-\lambda\boldsymbol{X}_{ikt+1}\right)\boldsymbol{\beta}\right]\Big/\left(1+\lambda^2\right)$ (73)

and $b_{ikt}=\upsilon_i\Big/\left(1+\lambda^2\right)$. (74)

Therefore, for each i,k and each $t=1,\ldots T-1$,

$$U_{ikt}\middle|\boldsymbol{\Theta}_{U_{ikt}} \square N\left(a_{ikt},b_{ikt}\right)\square\sum_{s=1}^{S}\left[\delta\left(y_{ikt}=s\right)\square\delta\left(\gamma_{s-1}<U_{ikt}<\gamma_{s}\right)\right] \qquad (75)$$

A special case is that, when $t=T$, U_{ikT} only appears in the exponential term with U_{ikT-1}. That is,

$$p\left(U_{ikT}\middle|\boldsymbol{\Theta}_{U_{ikT}}\right) \propto \left\{\sum_{s=1}^{S}\left[\delta\left(y_{ikT}=s\right)\square\delta\left(\gamma_{s-1}<U_{ikT}<\gamma_{s}\right)\right]\right\}\square\upsilon_i^{-1/2}\square$$

$$\exp\left[-\frac{1}{2\upsilon_i}\left(U_{ikT}-\lambda U_{ikT-1}-\theta_i-\boldsymbol{X}_{ikT}\boldsymbol{\beta}\right)^2\right] \qquad (76)$$

This also is a truncated normal distribution. The (un-truncated) normal distribution has a mean $a_{ikT}=\lambda U_{ikT-1}+\theta_i+\boldsymbol{X}_{ikT}\boldsymbol{\beta}$ and variance υ_i; and, if $y_{ikT}=s$, the distribution is truncated on the left by γ_{s-1} and on the right by γ_s.

4.3 MCMC SAMPLER

The MCMC sampling process begins with an initial parameter set $\left(\boldsymbol{\beta}^0,\lambda^0,\rho^0,\sigma^0,\boldsymbol{V}^0,\boldsymbol{\gamma}^0,\boldsymbol{U}^0\right)$, where superscripts indicate the current number of draws, or iteration step (for value updating.) All parameters or variables of interest are sampled sequentially and updated parameter values are then used to replace the initial values. The whole process is carried out iteratively by always using the most recent values of the parameters and variables, until the desired number of draws is achieved. The flowchart for sampling the parameters of interest is shown as Figure 1.

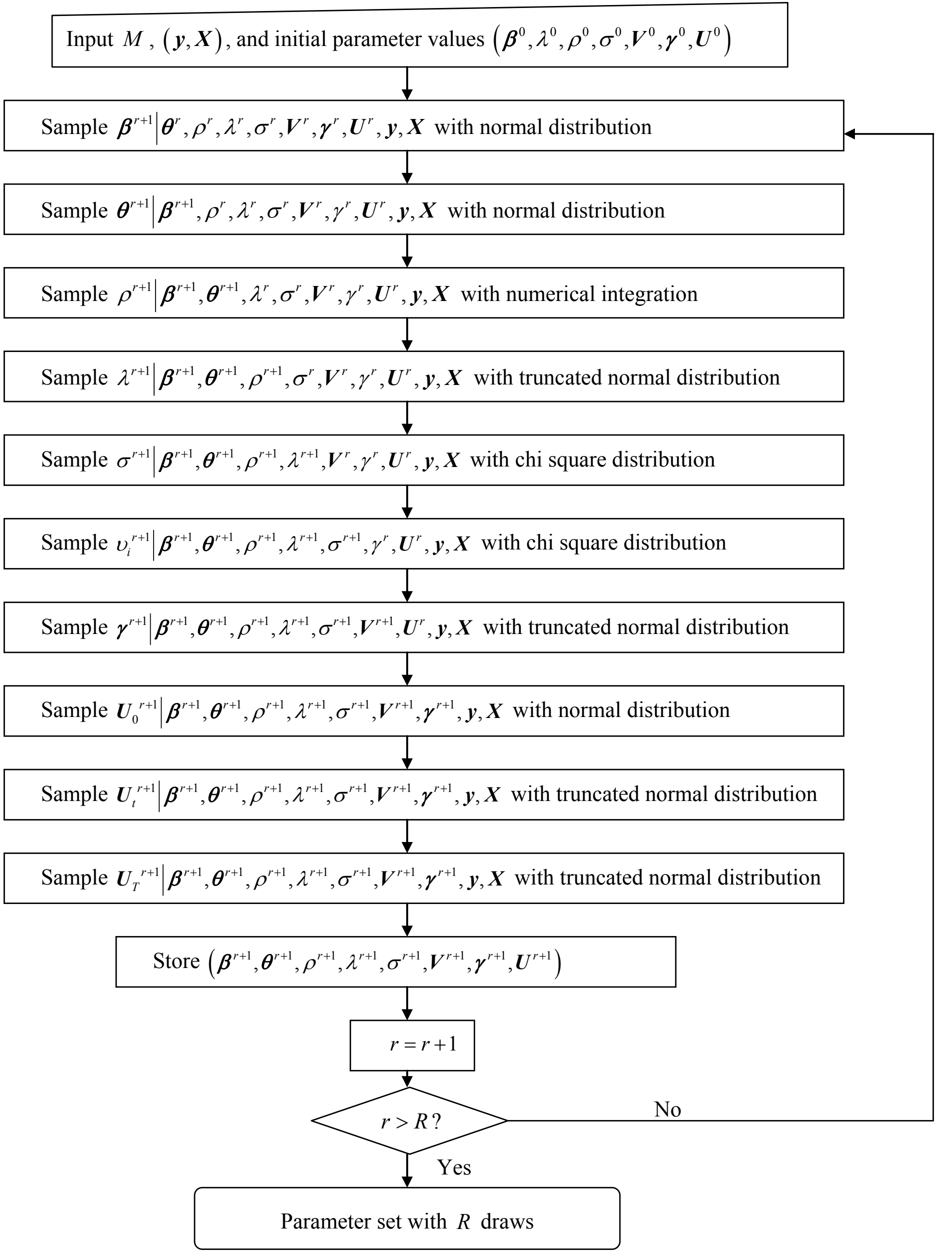
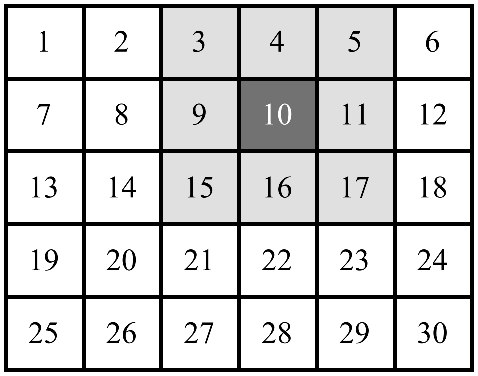

Input M, $(\boldsymbol{y},\boldsymbol{X})$, and initial parameter values $(\boldsymbol{\beta}^0,\lambda^0,\rho^0,\sigma^0,\boldsymbol{V}^0,\boldsymbol{\gamma}^0,\boldsymbol{U}^0)$
Sample $\boldsymbol{\beta}^{r+1}|\boldsymbol{\theta}^r,\rho^r,\lambda^r,\sigma^r,\boldsymbol{V}^r,\boldsymbol{\gamma}^r,\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with normal distribution
Sample $\boldsymbol{\theta}^{r+1}|\boldsymbol{\beta}^{r+1},\rho^r,\lambda^r,\sigma^r,\boldsymbol{V}^r,\gamma^r,\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with normal distribution
Sample $\rho^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\lambda^r,\sigma^r,\boldsymbol{V}^r,\gamma^r,\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with numerical integration
Sample $\lambda^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\sigma^r,\boldsymbol{V}^r,\gamma^r,\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with truncated normal distribution
Sample $\sigma^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\boldsymbol{V}^r,\gamma^r,\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with chi square distribution
Sample $\upsilon_i^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\sigma^{r+1},\gamma^r,\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with chi square distribution
Sample $\boldsymbol{\gamma}^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\sigma^{r+1},\boldsymbol{V}^{r+1},\boldsymbol{U}^r,\boldsymbol{y},\boldsymbol{X}$ with truncated normal distribution
Sample $\boldsymbol{U}_0^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\sigma^{r+1},\boldsymbol{V}^{r+1},\boldsymbol{\gamma}^{r+1},\boldsymbol{y},\boldsymbol{X}$ with normal distribution
Sample $\boldsymbol{U}_t^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\sigma^{r+1},\boldsymbol{V}^{r+1},\boldsymbol{\gamma}^{r+1},\boldsymbol{y},\boldsymbol{X}$ with truncated normal distribution
Sample $\boldsymbol{U}_T^{r+1}|\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\sigma^{r+1},\boldsymbol{V}^{r+1},\boldsymbol{\gamma}^{r+1},\boldsymbol{y},\boldsymbol{X}$ with truncated normal distribution
Store $(\boldsymbol{\beta}^{r+1},\boldsymbol{\theta}^{r+1},\rho^{r+1},\lambda^{r+1},\sigma^{r+1},\boldsymbol{V}^{r+1},\boldsymbol{\gamma}^{r+1},\boldsymbol{U}^{r+1})$
$r=r+1$
$r>R$?
No
Yes
Parameter set with R draws

Figure 1 Flowchart for the MCMC Simulation

5. SIMULATED DATASET

Instead of using empirical data, the dynamic spatial ordered probit (DSOP) model is tested using a simulated dataset. Because such self-generated data have known parameter values and controlled interactions, they are more reliable for evaluating performance of the model specification and the proposed estimation techniques.

In the simulated dataset, there are 30 regions, each containing 10 individuals observed over 8 time periods. Each individual has a response level of 1, 2, or 3. That is, $M = 30$ and $n_i = 10, \forall i$, (so that $N = 300$), $T = 8$ and $S = 3$. There are $300 \times 8 = 2400$ observed responses in total and 3 possible levels. Figure 2 shows the location of these 30 regions. The weight matrix is generated based on (queen) contiguity. For example, region 10 is considered contiguous with regions 3, 4, 5, 9, 11, 15, 16 and 17. It is then row-standardized.



1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30

Figure 2 Location of Regions in Simulated Dataset

Note: Region 10 and its contiguous neighbors are shown in grey.

The region-specific effect is generated using the following formulations:

$$\boldsymbol{\theta} = \left(\boldsymbol{I}_M - \rho \boldsymbol{W}\right)\boldsymbol{u} \qquad (77)$$

$$\boldsymbol{u} \sim N\left(0, \sigma^2 \boldsymbol{I}_M\right) \qquad (78)$$

where the spatial autocorrelation coefficient ρ is set to be 0.1, 0.6, 0.7 and 0.9 in different experiments. The variance σ^2 is equal to 1 so that $\boldsymbol{u}$ for each region follows an iid standard normal distribution.

The individual-specific variables are normally distributed independently and heteroskedastic over the regions. Assumed values of variance v_1 through v_{30} are shown in Figure 3:

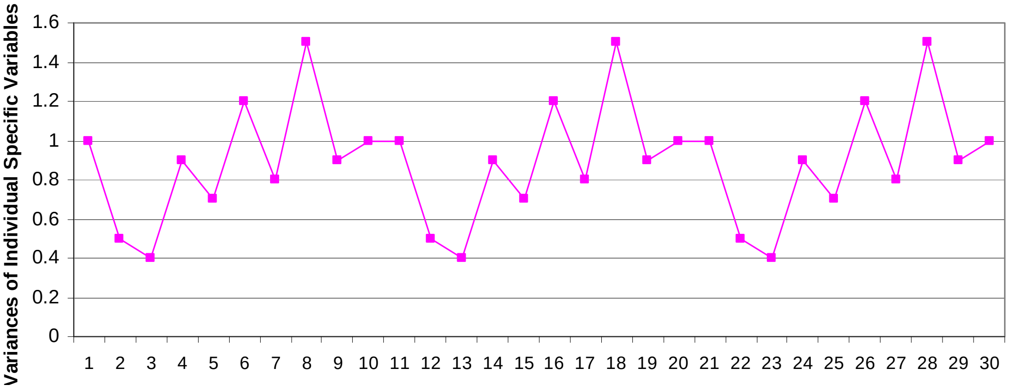


Figure 3 Assumed Variances of Individual Specific Effects across Regions

The specific value of each variance is set arbitrarily, between 0.4 and 1.5. This range of magnitudes helps ensure that the uncertainties caused by individual specific errors are important but do not overwhelm/dominate latent variables' effects, which is felt to be the most common case in reality. The variance for region 1 is fixed at its true value in estimation, which is necessary for identification.[3]

The explanatory variables include the lagged utility (unobserved dependent variable) U_{t-1} and four other observed values. The temporal autocorrelation coefficient λ (i.e., the parameter for the lagged dependent variable) is set to equal 0.1, 0.5 and 0.9 in different experiments. The four variables are generated using a standard uniform distribution (bounded between 0 and 1). Their corresponding parameters (slope coefficients) are arbitrarily set as -1.7, 2, 1 and 0.5, respectively. There are S=3 ordered categories, with thresholds γ_1=0 and γ_2=2.1. To summarize, the dataset is generated using the following model assumptions:

$$\boldsymbol{U} = 0.5\boldsymbol{U}_{t-1} - 1.7\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + \boldsymbol{x}_3 + 0.5\boldsymbol{x}_4 + \boldsymbol{\theta} + \boldsymbol{\varepsilon} \qquad (79)$$

where $\boldsymbol{y} = 1$ if $\boldsymbol{U} \le 0$, $\boldsymbol{y} = 2$ if $0 < \boldsymbol{U} \le 2.1$, and $\boldsymbol{y} = 3$ if $\boldsymbol{U} > 2.1$

and $\boldsymbol{\theta}$ is multivariate normal vector of region-specific effects with zero mean and variance matrix $\left(\boldsymbol{I}_M - \rho \boldsymbol{W}\right)$, where ρ is set to equal 0.1, 0.6, 0.7 and 0.9, across separate experiments. As noted above, $\boldsymbol{\varepsilon}$ is a normally distributed individual-specific error term with zero mean and variable variance (heteroskedastic) across regions. (The variances of these error terms range from 0.4 to 1.5. (Figure 3))

6 MODEL ESTIMATION AND VALIDATION

The simulated data samples were analyzed using the DOSP model. The resulting estimates are compared here to their true values, in order to appreciate the model's estimation ability. However, in addition to the identification problem mentioned above, the use of small simulated data samples involves other potential problems. These problems need to be carefully handled before a robust model evaluation can be achieved.

[3] This can be inferred from Figure 21.4 in Greene (2002): in an OP model, parameters and variances can be scaled simultaneously (so that the normal curve becomes flatter or sharper), with probabilities remaining constant. In other words, it is necessary to normalize at least one of the parameters or variances for the purpose of identification.

The first problem lies in the simulated sample data itself: in the process of random number generation, extreme values can appear. To address this, researchers often use a high number of draws (to try to avoid the influence of extreme values). Here, however, the simulated sample data also are randomly generated. (As noted in Section 5, $\boldsymbol{x}$ was generated from a standard uniform distribution, and $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ were generated using a standard normal distribution.) Unlike the number of draws used for estimation, the sample size here cannot be too large because a linear increase in sample size leads to an quadratic increase in computational burden. With 2,400 data points, the influence of extreme values is almost inevitable and "bad" samples are very likely to be generated. For example, the individual effect error term can become so large that it masks the contribution of explanatory variables and regional effects, leading to the conclusion that spatial autocorrelation or the influence of certain variables is insignificant.

Another example is that the values of explanatory variables and error terms may happen to be large for all data points, leading to a set of high latent dependent variable values, which means that few cells get labeled as Level 0. With such skewed data, the estimation may yield unreliable results. In order to neutralize this effect, for each parameter set, the data was re-generated 50 times, producing 600 samples (50 replicates × 4 ρ values × 3 λ values = 600). The averages of their estimated means and standard deviations are discussed below.

A second problem is estimation convergence. With a Bayesian approach, proof of convergence is a complicated issue. In this study, the estimation is assumed to converge when sampled parameter distributions appear to stabilize. Ideally, the number of draws (*R*) should be set as high as possible, but computational time and memory requirements also need to be taken into account. Especially when 600 samples (each containing 2,400 data points with complicated interactions) are to be analyzed, computational efficiency is an important consideration. Several *R* values were examined first here, for a small number of samples. Their estimation performances and computational intensities were evaluated, and the final selection was *R* = 2,000 since, beyond this number of draws, model the results no longer noticeably improved. Furthermore, after 1,000 runs, the distributions of all parameters appear stable. Therefore, the first 1,000 runs were omitted (burn-in) and the mean and standard deviation are both calculated based on the final 1,000 draws.

As described above, 600 simulated data samples were generated and their parameters then estimated with diffuse priors. The averages of all parameter estimates' means are shown in Table 1. Table 1 also uses root mean squared errors (RMSE)[4] to describe estimation accuracy for each parameter set. As can be observed, all RMSEs lie below 1. Considering the magnitudes of the parameter values, the estimation results are quite close to true values.

Some interesting tendencies are apparent. As the temporal autocorrelation coefficient (λ) increases, the magnitudes of coefficients and variances for both individual and regional specific effects tend to exhibit higher bias (as shown in Table 1 and Figure 4). One reason for this phenomenon is that, as λ increases, the influence of temporally lagged, latent response values

[4] RMSE is the square root of mean squared error (MSE), which is (an estimate of) the expected value of the squared "error" (i.e., the difference between estimated and true values). This indicator is often used in assessing a forecasting model's predictive power (Greene, 2002). It also can be used to evaluate estimation accuracy when true parameter values are known. A larger RMSE value indicates an increase in variations that the model does not account for.

rises, adding uncertainty to the right-hand side of the model. In the estimation process, this uncertainty will be partially ascribed to the error terms, which leads to larger estimates of σ^2 and υ_i, $\forall i \in M$. As mentioned in Section 5, this process will in turn produce higher $\boldsymbol{\beta}$ estimates (to accommodate the increase in scale).

An increase in the spatial autocorrelation coefficient (ρ) also leads to greater bias. As can be expected, when positive spatial correlation exists but is not fully recognized, the coefficients tend to be more biased because areas with higher response magnitudes will have a greater impact on model estimates.

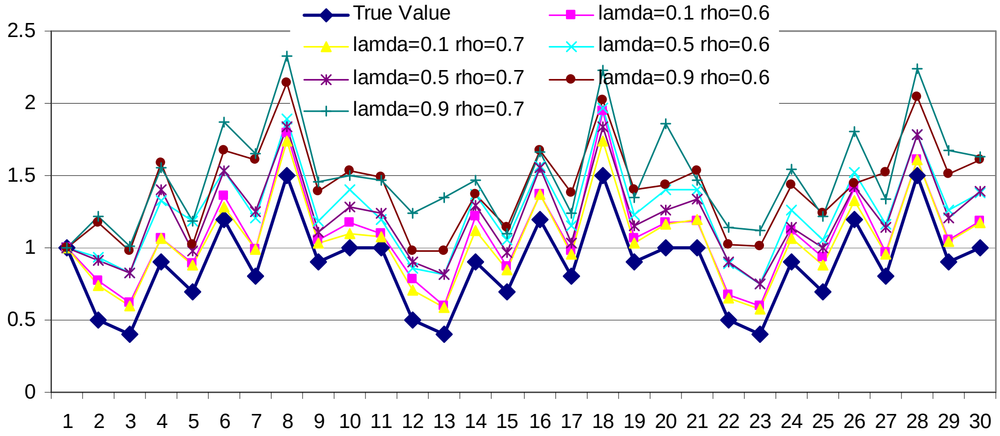


Figure 4 Variances of Individual Specific Errors with 70 Samples

In fact, this consistency problem is very common for nonlinear panel data models and dynamic models (see, e.g., Neyman and Scott, 1948) and has been studied for many years. A larger sample (larger *N*) and longer panel (larger *T*) may reduce this bias (Arellano and Hahn, 2005). Researchers also have proposed various approaches to reduce bias and achieve consistency with smaller *N* and *T* values (see, for example, Alvarez and Arellano, 2003, and Bester and Hansen, 2007). An efficient bias-reduction technique for the DSOP model makes an interesting topic for future study, but is not the focus here. In fact, such overestimation (due to increases in λ) appears to be slight here: all biases in slope parameters lie below 10%. Bias in estimates of the variances of individual specific errors (υ_i) are higher. However, as can be observed in Figure 4, with the exception of the extreme case (where both λ and ρ are 0.9), biases in all other cases lie well below 100% and their relative magnitudes appear close to the true pattern.

In summary, the DSOP model performs well with the simulated data. It satisfactorily detects the temporal and spatial interaction effects as well as the influence of different variables.

Table 1 Estimation Results using Simulated Data (Averages from 50 Samples)

Parameter	True Value		Average of Means from Estimated Parameter Distributions										
	λ	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5	0.9	0.9	0.9	0.9
	ρ	0.1	0.6	0.7	0.9	0.1	0.6	0.7	0.9	0.1	0.6	0.7	0.9
β_1	-1.7	-1.701	-1.715	-1.712	-1.888	-1.727	-1.776	-1.801	-1.887	-1.841	-1.833	-1.881	-1.822
β_2	2.0	1.965	1.984	2.019	2.125	2.046	2.049	2.097	2.278	2.191	2.140	2.105	2.154
β_3	1.0	0.965	0.972	1.004	1.012	1.012	1.030	1.033	1.129	1.046	1.090	1.072	1.086
β_4	0.5	0.519	0.519	0.543	0.551	0.542	0.518	0.554	0.646	0.561	0.545	0.539	0.647
λ	_	0.097	0.101	0.099	0.097	0.492	0.507	0.514	0.511	0.919	0.921	0.909	0.863
ρ	_	0.048	0.452	0.572	0.845	0.039	0.494	0.623	0.863	-0.001	0.498	0.616	0.855
σ^2	1.0	1.054	1.091	1.117	1.832	1.158	1.217	1.302	1.768	1.290	1.232	1.307	1.498
γ_1	0.0	-0.223	-0.133	-0.112	0.006	0.094	-0.333	-0.358	-0.351	-0.202	-0.330	-0.177	-0.138
γ_2	2.1	1.818	1.933	2.009	2.276	2.190	1.803	1.834	2.075	2.090	1.956	2.130	2.342
Average RMSE		0.371	0.278	0.231	0.883	0.225	0.517	0.566	0.930	0.445	0.492	0.429	0.630

7. MODEL COMPARISONS

To further validate the DSOP model, its performance is compared to those of simpler models (all estimated using a Bayesian approach), based on a data set that provides a balanced mix of the three levels of the 50 samples. These simpler models include a standard ordered probit (OP) model; a dynamic ordered probit (DOP) model, which still allows for spatial heterogeneity but not spatial autocorrelation; and a spatial ordered probit (SOP) model, which incorporates all spatial effects but does not consider the temporal dependency. Data statistics for this sample are shown in Table 2, and the histogram of y values (Figure 5) indicates that enough observations exist for each level.

Table 2 Summary Statistics for One Sample

Variable	Mean	Standard Deviation	Minimum	Maximum
x_1	0.4978	0.2873	6.579E-04	9.995E-01
x_2	0.4994	0.2893	5.842E-04	9.990E-01
x_3	0.4936	0.2901	4.174E-05	9.999E-01
x_4	0.4877	0.2895	3.049E-05	9.998E-01

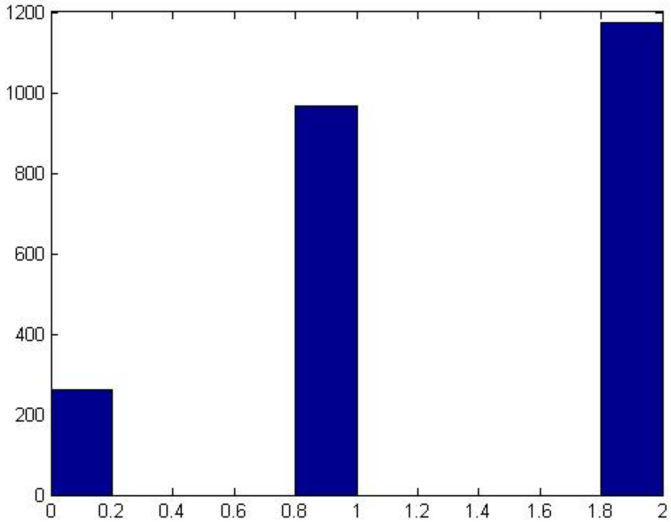


Figure 5 Histogram of Dependent Variable Values

As before, these models are run with 2,000 draws of which the first 1,000 draws are omitted (as a burn-in sample). As an example, Figure 6 shows the estimation convergence pattern for β_1. Estimates of other parameters follow a similar pattern. The figure suggests that after the first 1,000 draws, the estimation becomes stable and may be convergent.

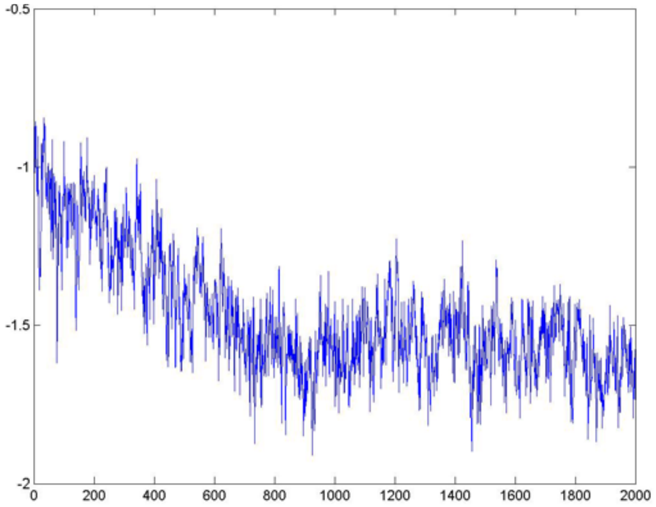


Figure 6 Estimation Convergence Pattern for β_1

Table 3 shows the estimation results for this sample. In addition, Figure 7 depicts estimates of υ_i ($\forall i \in M$), using the DSOP model, where lower and higher bounds are defined as 1st percentile and 99th percentile values. Mean estimates lie quite close to true values. Considering that only 80 observations are effectively used to estimate each υ_i, the standard deviations are understandably large.

Table 3 Estimation Results using One Sample and Different Specifications

Param.	True Value	OP		DOP		SOP		DSOP	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
β_1	-1.7	-0.807	0.079	-1.581	0.104	-1.621	0.117	-1.608	0.119
β_2	2.0	1.727	0.078	2.201	0.112	2.150	0.128	2.166	0.128
β_3	1.0	0.999	0.079	1.043	0.107	1.014	0.099	1.000	0.097
β_4	0.5	0.634	0.076	0.502	0.089	0.461	0.092	0.469	0.098
λ	0.1	---	---	0.131	0.023	---	---	0.110	0.021
ρ	0.7	---	---	---	---	0.769	0.098	0.751	0.098
σ^2	1.0	---	---	2.182	0.742	1.323	0.426	1.116	0.342
γ_2	0.0	-0.203	0.014	-0.372	0.038	-0.226	0.084	0.081	0.057
γ_3	2.1	1.264	0.009	1.811	0.029	1.980	0.032	2.261	0.013
RMSE		1.769		1.472		0.463		0.293	
DIC		4360.4		3098.0		3106.1		3070.7	

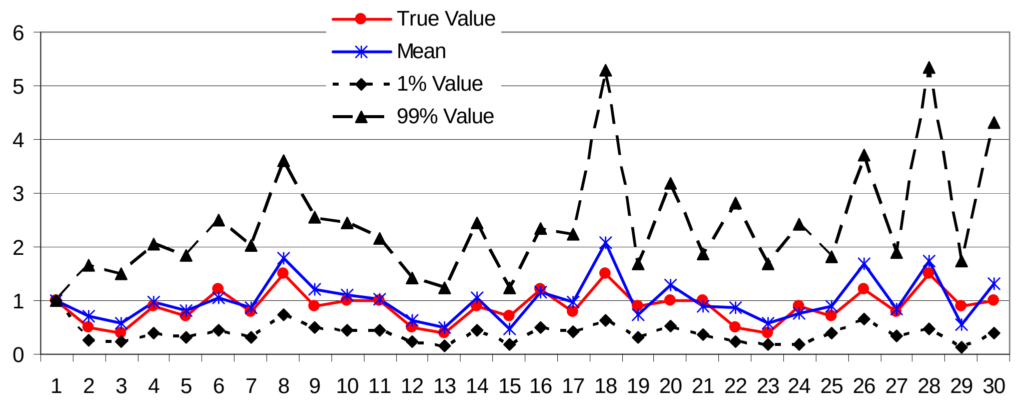


Figure 7 Variances of Individual Specific Errors with One Sample

Table 3 also shows RMSE and deviance information criteria (DIC)[5] values for each specification. As before, RMSE indicates estimator accuracy. The DIC is an indicator of model fit. Both suggest that the DSOP model more accurately estimates the underlying parameters, with high statistical significance and fit of the sample data. In contrast, because of the inability to detect λ and ρ, the OP model's estimates are highly unsatisfactory. As shown in Table 3, it returns the appropriate signs and relative magnitudes for β parameters, but estimates deviate from true values quite a bit. The performances of the DOP and SOP models lie in-between. Though inferior to the DSOP model results, they are rated better than the OP model. RMSE measures suggest that the SOP model yields much more accurate estimates than the DOP model, which is quite understandable given the fact that λ is only 0.1 and ρ is 0.7 in this particular sample. In other words, ignoring the temporal autocorrelation (i.e., restricting a 0.1 parameter to equal 0) should typically have less of an impact than a situation where one ignores a spatial autocorrelation term of 0.7. Interestingly, the DIC fit measure, suggests that the DOP model is very slightly preferred to the SOP model. The DOP model's smaller DIC value implies that, while the DOP model is not as able to produce accurate parameter estimates, it still fits sample data better than the SOP model, because it still accounts for spatial heterogeneity.

Table 4 illustrates predictive accuracy using the four methods. The standard OP model only correctly predicts dependent values for 47.0% of the 2,400. observations. The DOP model increases this percentage to 60.8%. The SOP model's prediction rate is quite close to that of the DSOP model: 66.4%. Such a percentage is fairly satisfactory, given the presence of three response levels and considerable randomness in the sample dataset (σ^2 and υ_i in the simulated data have similar magnitudes as all slope parameters, causing regional-specific and individual-specific errors to have a similar level of influence on latent response values).

Table 4 Prediction Rates using Different OP Model Specifications

[5] The deviance information criterion (DIC) is a generalization of the Akaike information criterion (AIC) and Bayesian information criterion (BIC). It is particularly useful for Bayesian model comparison and selection (see Gelman et al., 2004, and Spiegelhalter et al., 2002). However, one limitation of the standard DIC is that it is only valid when posterior distributions are approximately multivariate normal. For models involving extremely asymmetric or bimodal posterior distributions (which happens for the DSOP model), some modified DIC need to be used instead. This study uses the DIC calculation method for mixture models proposed by Celeux et al. (2006).

			Actual			Total	% Cases Correctly Predicted
Response Value (y)			1	2	3		
Predicted	OP	1	55	145	100	300	47.0%
		2	106	372	372	850	
		3	100	448	702	1250	
	DOP	1	124	154	97	375	60.8
		2	116	511	253	880	
		3	21	300	824	1145	
	SOP	1	121	133	20	274	65.1
		2	124	536	249	909	
		3	16	296	905	1217	
	DSOP	1	121	112	17	250	66.4
		2	119	583	268	970	
		3	21	270	889	1180	
Total			261	965	1174	2400	

Such comparisons, of prediction rates, RMSE and DIC values, suggest that the DSOP model is superior to all the simpler models, as anticipated. It is followed by the SOP model, indicating the importance of recognizing the spatial autocorrelation in the dataset. Recognizing temporal dependency also significantly improves model performance, relative to a standard OP model. In this example study, this improvement is not as evident as recognizing the spatial autocorrelation, but this is partially due to the small true value of the temporal coefficient, λ. The OP model, though easy to specify and estimate, does not adequately utilize the observed information, thus returning inaccurate parameter estimates and response predictions.

8. CONCLUSIONS

Many data sets involve latent (unobserved) variables exhibiting underlying spatial interactions and temporal dependency patterns. Examples include land use change, voting outcomes, destination and location choices, crash counts (over a network), and traffic condition ratings. These examples all present two common features. First, the variables of interest are indicators or censored versions of unobserved variables. Second, they all exhibit certain degrees of temporal and spatial autocorrelation. Such phenomena also exist in other fields, like ecology, biology and anthropology. To capture these temporal and spatial patterns and accurately estimate the impacts of potentially influential factors, a rigorous statistical method for analyzing such data is needed. The dynamic spatial ordered probit (DSOP) model established in this study meets this need.
The DSOP model analyzes ordered response data based on latent variables exhibiting and spatial dependencies as well as individual heterogeneity. First, as in Smith and LeSage (2004), the model incorporates spatial effects by allowing for both regional spatial interactions and heteroskedasticity across observations from different regions. Second, the model allows for an AR(1) process via the latent, lagged dependent variable, thus recognizing dynamic features. Third, when compared to existing spatial discrete choice models, the DSOP model is the first to emerge from an ordered probit model, where multiple levels of ranked categorical data can be analyzed.

The models developed here were estimated in a Bayesian framework using MCMC sampling and data augmentation techniques (to generate the autocorrelated latent variables). The estimation process approximates the parameter set's joint probability using a set of conditional distributions. To achieve this, proper prior distributions for parameters and nuisance terms (latent dependent variables and variances) were assumed and their posterior distributions then derived. These posterior distributions include common distributions (like the truncated normal and chi square), mixture distributions (combining a normal and multivariate uniform), and nonstandard distributions (offering no closed-form expressions for hyperparameters). Matlab code was developed to draw from these distributions.

This study also renders some general insights into the pragmatic advantages of a Bayesian framework over a frequentist method[6]. For this type of work, the Bayesian approach appears more straightforward and much easier to apply than maximum (simulated) likelihood estimation (MSLE). Especially for models involving complicated statistical distributions and multi-layered specifications (as with the DSOP model), the advantage of a Bayesian framework is evident. By using "conditional" distributions, the Bayesian approach decomposes the joint estimation of many variables into much simpler, sequential simulations. In contrast, maximum (simulated) likelihood estimation (MSLE) must tackle an intractable likelihood function (and its gradients and possibly its Hessian matrix, with respect to the parameter set) (see, e.g., Wang and Kockelman [2008]). With a Bayesian framework, a slight change in model specification only requires modifying a part of the simulation procedure. With MSLE, on the other hand, the model estimation method may need to be completely overhauled. However, the Bayesian approach also has its limitations. For example, in this study, because the estimation involves simulating latent variables and one (multivariate) posterior distribution (for threshold terms) is multimodal, marginal effects and the model's goodness of fit need to be calculated simultaneously with the simulation. Otherwise, if an indicator (such as the deviance information criterion) needs to be obtained afterwards, the model must be completely re-run, which can be rather time consuming.

The DSOP model specification and estimation methods were validated using 50 simulated datasets, for each of the 12 parameter sets. The results produced estimates that are quite close to true values. The comparison highlighted the accuracy of the DSOP model, while recognizing report temporal and spatial autocorrelation patterns. As detailed spatial data sets become available to regional scientists and others, it behooves us to unleash their potential, by recognizing the spatial relationships that exist and exploiting their presence. The DSOP model and the estimation methods described here offer us the opportunity of a more appropriate approach.

[6] Of course, much has been written (e.g., Geweke, 1993; Gelman et al., 2004; and Koop et al. 2007) about the differences in classical and Bayesian statistical viewpoints. Much of the discussion is somewhat "philosophical" in nature, and "superiority" has never been conclusively determined (Gelman et al., 2004).

REFERENCES

Albert, J. H. and Chib, S. (1993) "Bayesian analysis binary and polychotomous response data." *Journal of the American Statistical Association* 88: 669-679.

Alvarez, J. and Arellano, M. (2003) "The time series and cross-section asymptotics of dynamic panel data estimators." *Econometrica* 71(4): 1121-1159.

Anselin, L. (1999) *Spatial Econometrics*. Working paper. Accessed July 10, 2005: http://www.csiss.org/learning_resources/content/papers/baltchap.pdf.

Anselin, L. (2001) "Issues in spatial probit models." Workshop on Qualitative Dependent Variable Estimation and Spatial Effects, College of ACES, University of Illinois, April 20, 2001.

Anselin, L. and Hudak, S. (1992) "Spatial econometrics in practice: A review of software options." *Regional Science and Urban Economics* 22(3): 509-536.

Anselin, L., Florax, R., and Rey, S. (2004) (Eds.), *Advances in Spatial Econometrics. Methodology, Tools and Applications.* Berlin: Springer-Verlag.

Arellano, M. and Hahn, J. (2005) "Understanding bias in nonlinear panel models: Some recent developments." Invited lecture. Econometric Society World Congress, London.

Atkinson, P. M., Clark, M. J., German, S. E. and Sear, D. A. (2003) "Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression." *Geographical Analysis* 35(1): 58-82.

Beron, K. and Vijverberg, W. (2004) "Probit in a spatial context: a Monte Carlo analysis". In Anselin, L. Florax, R. and Rey, S. (Eds.), *Advances in Spatial Econometrics*. Heidelberg: Springer-Verlag.

Bester, C. A. and Hansen, C. (2007) "A penalty function approach to bias reduction in non-linear panel models with fixed effects." Accessed June 1, 2007: http://faculty.chicagogsb.edu/christian.hansen/research/bh_penalizedfe_jul06.pdf

Bhat, C. and Guo, J. (2004) "A mixed spatially correlated logit model: Formulation and application to residential choice modeling." *Transportation Research Part B* 38: 147-168.

Boots, B.N. and Kanaroglou, P.S. (1988) "Incorporating the effects of spatial structure in discrete choice models of migration." *Journal of Regional Science* 28(4): 495-509.

Coughlin, C. C., Garrett, T. A. and Hernández-Murillo, R. (2003) "Spatial probit and the geographic patterns of state lotteries." Accessed March 1, 2006: http://research.stlouisfed.org/wp/2003/2003-042.pdf.

Cowles, M. K. (1996) "Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models." *Statistics and Computing* 6(2): 101-110.

Dugundji, E. R. and Walker, J. L. (2005) "Discrete choice with social and spatial network interdependencies." *Transportation Research Record* 1921: 70-78.

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.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., and Raton, B. (2004) *Bayesian Data Analysis (2nd Edition.).* Florida: Chapman and Hall/CRC Press.

Geweke, J. (1993) "Bayesian treatment of the independent Student-t linear model." *Journal of Applied Econometrics* 8(S): 19-40.

Girard P. and Parent E. (2001) "Bayesian analysis of autocorrelated ordered categorical data for industrial quality monitoring." *Technometrics* 43(2): 180-191.

Greene, W. (2000) *Econometric Analysis*. Upper Saddle River: Prentice-Hall.

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M. and Excoffier, L. (2005) "Bayesian estimation of recent migration rates after a spatial expansion." *Genetics* 170(1): 409-417.

Johnson, V. E. and Albert, J. H. (1999) *Ordinal Data Modeling*. New York: Springer.

Kakamu, K. and Wago, H. (2007) "Bayesian spatial panel probit model with an application to business cycle in Japan." Working paper. Accessed May 10, 2007: http://www.mssanz.org.au/modsim05/proceedings/papers/kakamu_2.pdf

Klier, T. and McMillen, D. P. (2007) "Clustering of auto supplier plants in the U.S.: GMM spatial logit for large samples." *Journal of Business and Economic Statistics* (forthcoming). Accessed May 10, 2007: http://tigger.uic.edu/~mcmillen/papers/Clustering%20of%20Auto%20Supplier%20Plants.pdf.

Koop, G. M., Poirier, D. J. and Tobias, J. L. (2007) *Bayesian Econometric Methods*. New York: Cambridge University Press.

LeSage, J. P. (1999) "The theory and practice of spatial econometrics." Manuscript. Accessed May 10, 2007: http://www.spatial-econometrics.com.

LeSage, J. P. (2000) "Bayesian estimation of limited dependent variable spatial autoregressive models." *Geographical Analysis* 32(1): 19-35.

McMillen, D. P. (1995) "Spatial effects in probit models: A Monte Carlo investigation." In Anselin, L. and Florax, R. (Eds.), *New Directions in Spatial Econometrics*. Heidelberg: Springer-Verlag.

McMillen, D. P., and McDonald, J. F. (1998) "Suburban subcenters and employment density in metropolitan Chicago." *Journal of Urban Economics* 43(2): 157-180.

Miyamoto, K., Vichiensan, V., Shimomura, N. and Paez, A. (2004). "Discrete choice model with structuralized spatial effects for location analysis" *Transportation Research Record* 1898: 183-190.

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

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.

Neyman, J. and Scott, E. L. (1948) “Consistent estimates based on partially consistent observations.” *Econometrica* 16(1): 1-32.

Pinkse, J. and Slade, M.E. (1998) “Contracting in space: An application of spatial statistics to discrete-choice models.” *Journal of Econometrics* 85(1): 125-54.

Pinkse, J., Slade, M.E. and Shen, L. (2005) “Dynamic spatial discrete choice using one step GMM: An application to mine operating decisions.” *Spatial Economic Analysis* 1(1): 53-99.

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.

Sun, D., Tsutakawa, R.K. and Speckman, P. L. (1999) “Posterior distribution of hierarchical models using CAR(1) distributions.” *Biometrika* 86(2): 341-350.

Vanasse, A., Niyonsenga, T., Courteau, J., Gregoire, J. P., Hemiari, A., Loslier, J. and Benie, G. (2005) “Spatial variation in the management and outcomes of acute coronary syndrome.” *BMC Cardiovascular Disorders* 5: 21.

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.

Wallerman1, J., Vencatasawmy, C. P., and Bondesson, L. (2006) “Spatial simulation of forest using Bayesian state-space models and remotely sensed data.” 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Lisbon, Portugal. July 5-7, 2006. is there a URL for this? we need to share that, b/c it’s very hard for people to get conf papers on their own.

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. (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. (2008) “Maximum Simulated Likelihood Estimation with Spatially Correlated Observations: A Comparison of Simulation Techniques.” Presented at the RSAI's 53rd Annual Meeting, Toronto (in 2006), and forthcoming in *Transportation Statistics* (J. Ross Publishing).

Wear, D. N. and Bolstad, P. (1998) “Land-use changes in southern Appalachian landscapes: Spatial analysis and forecast evaluation.” *Ecosystems* 1: 575-594.