Transportation Research Part B 94 (2016) 440–455

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Discrete choice with spatial correlation: A spatial autoregressive binary probit model with endogenous weight matrix (SARBP-EWM)

Yiwei Zhou[a], Xiaokun Wang[b,*], José Holguín-Veras[c]

[a] Senior Traffic Engineer, Precision Systems, Inc., 80 M Street, SE, Washington, DC 20003, USA
[b] Department of Civil and Environmental Engineering, 4032 JEC Building, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590, USA
[c] Director of the Center for Infrastructure, Transportation, and the Environment, Department of Civil and Environmental Engineering, 4032 JEC Building, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590, USA

ARTICLE INFO

Article history:
Received 24 February 2016
Revised 17 October 2016
Accepted 18 October 2016
Available online 28 October 2016

Keywords:
Bayesian MCMC
Probit model
Endogenous weight matrix
Choice behavior

ABSTRACT

Discrete choice modeling is widely applied in transportation studies. However, the need to consider correlation between observations creates a challenge. In spatial econometrics, a spatial lag term with a pre-defined weight matrix is often used to capture such a correlation. In most previous studies, the weight matrix is assumed to be exogenous. However, this assumption is invalid in many cases, leading to biased and inconsistent parameter estimates. Although some attempts have been made to address the endogenous weight matrix issue, none has focused on discrete choice modeling. This paper fills an existing gap by developing a Spatial Autoregressive Binary Probit Model with Endogenous Weight Matrix (SARBP-EWM). The SARBP-EWM model explicitly considers the endogeneity by using two equations whose error terms are correlated. Markov Chain Monte Carlo (MCMC) method is used to estimate the model. Model validation with simulated data shows that the model parameters can converge to their true values and the endogenous weight matrix can be reliably recovered. The model is then applied to a simplified firm relocation choice problem, assuming that similar size firms influence one another. The model quantifies the peer effect, and takes into consideration other independent variables including industry type and population density. The estimation results suggest that peer influence among firms indeed affect their relocation choices. The application results offer important insights into business location choice and can inform future policy making. The sample size for applying the model is currently limited to hundreds of observations. This paper contributes to the existing literature on discrete choice modeling and spatial econometrics. It provides a new tool to discover spatial correlations that are hidden in a wide range of transportation issues, such as land development, location choice, and various travel behavior. Those hidden spatial correlations are otherwise difficult to identify and estimation results may be biased. Establishing a new model that explicitly considers endogenous weight matrix and applying the model to a real life transportation issue represent a significant contribution to the body of literature.

* Corresponding author.
E-mail addresses: zywadam@gmail.com (Y. Zhou), wangx18@rpi.edu (X. Wang), jhv@rpi.edu (J. Holguín-Veras).

http://dx.doi.org/10.1016/j.trb.2016.10.009
0191-2615/

Introduction

A classic assumption in travel behavior modeling is that observations are independent, which has proven to be unrealistic and restrictive in many transportation study contexts. Decision makers' behavior, ranging from long term, relatively stable activities (e.g., residential location choice Guo and Bhat, 2004), to daily activity patterns (e.g., making shopping trips with friends and family Zhou and Wang, 2014a, 2014b; Wang and Zhou, 2015), are correlated. Such correlations are attributed to explicit socioeconomic interaction with each other (Zhou and Wang, 2014a) or the shared unobserved effects in the data (Guo and Bhat, 2007). Unfortunately, previous studies mainly analyze travel behavior without considering the influence of such correlations in decision making. This prevented the research and practice communities from fully understanding people's activities, leading to biased estimation results, inaccurate interpretation and misleading policy measures (Cao, 2015).

In response to the research needs, some studies sought to address the correlations in the decision making process. Such form of dyadic dependency between agents in close social or spatial proximity is also referred to as the endogenous interaction effect (Elhorst, 2010 and Manski, 1993). There are many valid approaches to capture such endogenous interactions. Several studies used linear-in-means model and assumed peer social interaction effects within exclusive groups (Blume et al., 2011). Lee et al. (2010) used a spatial network autoregressive weight structure for correlations within network, as well as group-specific unobserved effects. At the same time, some researchers start considering spatial correlations in discrete choice models (Brock and Durlauf, 2001, 2006, 2007; Soetevent and Kooreman, 2007; Krauth, 2006; Zhang and Wang, 2016a, 2016b; Zou et al., 2015). Recently, Bhat (2015) accommodated spatial correlation effects, while allowing a global spatial structure on the individual-specific unobserved response sensitivity to exogenous variables. The latter two effects are referred to as spatial drift effects. In most of these studies, the weight matrix represent the spatial correlations. The weight matrix indicates relative weight of social or spatial interactions between agents. Because of the high dimension of correlation caused by spatial interactions, most of these spatial models have difficulty handling large sample sizes.

In general, there has been limited empirical work on the observation correlation, particularly in discrete form (Bhat et al., 2015). Two major challenges exist in addressing observation correlations: First, in order to consider extensive observation correlations, the model structure requires a new theoretical framework, and the model estimation and interpretation could become extremely difficult (Bhat et al., 2014a). Second, it is difficult to measure the level of connection between two observations. Spatial econometrics often relies on Tobler's first law of geography and assumes that close objects are more related than distant objects. Most spatial econometric models then assume an exogenous weight matrix. However, in many transportation problems, the weight matrix entry should not be exogenous, especially when the dependency is caused by social interaction or if it is subject to the influence of many unobserved effects. The misspecification of weight matrix may lead to erroneous estimation results, and possibly misleading policy assessments. Many studies have acknowledged these problems and highlighted the importance of tackling them (Anselin, 2010; Corrado and Fingleton, 2011; Pinkse and Slade, 2010), but due to the complication resulting from the unit dependency, model nonlinearity, and weight matrix endogeneity, few studies have successfully addressed these challenges.

In this paper, we propose to capture the observation interaction effect through the definition of an endogenous weight matrix in a spatial lag structure. The spatial lag structure has been used in many previous studies to analyze observation interdependency (Chakir and Parent, 2009; Bhat, 2011; Sidharthan and Bhat, 2012; Zhang and Wang, 2016a, 2016b; Ni et al., 2016). A typical example is the spatial autoregressive model (SAR), where the dependent variable is function of a linear combination of neighboring observations, as well as exogenous variables and an error term. The weight matrix entries are used to measure the relative level of social and spatial interactions between agents. There are many approaches to define the weight matrix (LeSage and Pace, 2009). It can be defined based on geographic approximation or social connection depending on need (Dugundji, 2013; Leenders, 2002), which makes it applicable to many social and economic issues. For example, in social networking, the weight matrix can be defined as binary peer matrix, which measures peer effect (LeSage and Pace, 2009). In this paper, we go further to allow the weight matrix entries to be endogenous. The weight elements are specified as functions of exogenous variables and a stochastic term. By allowing correlation between the weight elements' stochastic term and the error terms in the decision making models, it relaxes the constraint of an exogenous weight matrix.

Consequently, this paper develops a Spatial Autoregressive Binary Probit Model with Endogenous Weight Matrix (SARBP-EWM). Bayesian Markov Chain Monte Carlo (MCMC) method is used for estimation. The model is developed based on SAR model while allowing the weight matrix to be endogenous. It can be applied to many transportation phenomena with individuals choosing among alternatives in discrete form, including impact of transportation infrastructure on urbanization, social network, travel behavior, residential location choice, etc. This paper also applies the model to a firm relocation issue to investigate the observation interaction and other influential factors such as industry type and population density in firm relocation decision. The application of the SARBP-EWM model can quantify intensity of interaction among agents, and identify influential factors in the decision-making process. It will add to the existing literature by addressing observation correlation in discrete choice models.

Literature review

In a spatial econometric model, weight matrix represents the relative strength of connection between each pair of spatial units. Traditionally, the weight matrix in spatial econometric models is treated as exogenous. This assumption is true in certain circumstances. For example, when spatial weight is defined as the geographic distance between different spatial units,

the weight matrix is exogenous. As Anselin and Bera (1998) stated, “in the standard estimation and testing approaches, the weight matrix is taken to be exogenous". In many studies, researchers just simply treat weight matrix as exogenous. It is under the exogenous weight matrix assumption that many estimation methods and estimators are developed. Kelejian and Prucha (1998) developed a generalized spatial two-stage least squares (2SLS) procedure to estimate SAR model with autoregressive disturbances, where the weight matrix is set as known constants. Similarly, Lee (2004) investigated asymptotic properties of both maximum likelihood (MLE) estimator and the quasi-maximum likelihood estimator for the SAR model, where weight matrix is specified as constants. Lee (2007) proposed a GMM procedure for the estimation of the mixed regressive, spatial autoregressive model that combines both advantages of computational simplicity and efficiency over the conventional maximum likelihood (MLE) method. The proposed GMM estimators were shown to be consistent and asymptotically normal.

However, the exogenous weight matrix assumption is invalid in many cases, leading to bias and inconsistent parameter estimates. For example, in a study of traveler behaviors where the weight matrix captures the intensity of social connections, the weight matrix is very likely to be endogenous: people behaving similarly tend to form stronger social connections. Similarly, low travel impedance between two locations allows strong dependency of land development between them. In return, the co-development tends to attract more investment in transportation infrastructure that further reduces the travel impedance.

Addressing the endogenous weight matrix issue in spatial model has gained increasing attention in recent years, and researchers have studied a variety of approaches. Kelejian and Piras (2012) used instrumental variables to deal with the endogenous weighting matrix in their spatial panel data. The results are shown to be consistent and asymptotically normal. Chandrasekhar and Lewis (2011) attempted to correct the estimation biases caused by endogeneity that arises from missing network data. Two strategies were proposed: a two-step estimation procedure using graphical reconstruction and a set of analytical corrections for commonly used network statistics. Masten (2012) developed a linear simultaneous equations model to study social interactions between firms. Random coefficients were used with the endogenous variables. The model was estimated using a nonparametric sieve maximum likelihood estimator. Bhat and Guo (2007) used a joint mixed multinomial logit-ordered model to study the impact of built environment on household residential choice and auto ownership levels. They explicitly considered unobserved heterogeneity by proposing a method controlling the self-selection of individuals into neighborhoods. LeSage and Pace (2008) used spatial lags of dependent variable to quantify the endogenous interaction of commuter flow between origin and destination. Recently, Bhat et al. (2014b) discussed why identification of spatial interaction effects and exogenous interaction effects are possible in discrete choice and nonlinear models using a different specification than the ones discussed above. Bhat (2015) continued to formulate a model that extends the traditional panel discrete choice model to include social/spatial dependencies in the form of dyadic interactions between each pair of decision-makers. He studied the spatial correlation effects as well as global spatial structure, which together were referred to as “spatial drift effects”. Bhat’s (2011) maximum approximate composite marginal likelihood (MACML) method is used for model estimation. Results indicated MACML approach recovers the model parameters very well.

These studies conclude that there are two major ways of addressing the endogeneity problem: use of instrumental variables and development of a model structure with explicit consideration of endogeneity. Naturally, the instrumental variable approach requires that the instrumental variable exist. However, it is often difficult to find appropriate instrumental variables. Besides, whether information for instrumental variables can be collected or not remains a problem, even when they exist. The second approach directly addresses the endogeneity problem in the model structure. For example, Han (2014) explicitly addressed the endogeneity issue in a spatial autoregressive (SAR) model. In his study, A SAR model and an entry equation, which defines the weight matrix W are presented. The endogeneity occurs when the error terms in two equations are correlated. A Bayesian MCMC method was used to estimate the model. The model was applied on Medicaid spending across states. Results indicated both geographical distance and economic distance have significant effects on the interaction strength of state Medicaid related spending. Han’s paper sets a good example for the study of weight matrix endogeneity problem. However, his model only focuses on the continuous variable, while many transportation issues, such as land development, location choice, and mode choice are typically in discrete form.

Inspired by Han’s work, this paper aims to develop a Spatial Autoregressive Binary Probit Model with Endogenous Weight Matrix (SARBP-EWM). As discussed above, there have been some studies on spatial model with endogenous weight matrix. However, to the authors’ best knowledge, none has focused on discrete form. This paper fills the gap in existing literature by extending the spatial model with endogenous weight matrix in discrete form.

Model development

Spatial probit model

The usual approach of dealing with binary variable is to use a latent variable (usually utility) to indicate choice:

$$\Pr(y_{it} = 1) = \Pr(U_{it1} > U_{it0}) = \Pr(y_{it}^{*} > 0) \tag{1}$$

where y_{it} is the observed choice of observation i at time t, U_{it1} is the utility of alternative 1 for observation i at time t, and y_{it}^{*} is the latent variable that measures the utility difference between two alternatives for observation i at time t. When decision makers are allowed to be interdependent subject to a spatial autoregressive (SAR) process, a spatial probit model

can be expressed as (Smith and LeSage, 2004),

$$Y_t^* = \rho W_t Y_t^* + X\beta + M + ln + \mathrm{E}_t,\ t = 1, 2, ..., T \tag{2}$$

where $Y_t^* = (Y_{1t}^*, y_{2t}^*, ..., y_{Nt}^*)'$ is a $N \times 1$ vector for utility differences; the connection between Y_{Nt}^* and the observed choice variable Y_{Nt}remains the same as Eq. (1). W_t is an $N \times N$ weight matrix that captures the relative weight among agents, which will be detailed in the next session; X is $N \times k_1$ matrix of exogenous and time-varying independent variables. β is $k_1 \times 1$ the vector of parameters associated with the independent variables; $M=(m_{11}, m_{21},..., m_{N1})'$ is a $N \times 1$ vector for geographic fixed effect where m_{i1} represents the state fixed effect for observation i; l is a $n \times 1$ vector of ones and n_{1t} is scalar representing time fixed effect at time t; $E_t = (\varepsilon_{1t}, \varepsilon_{2t}, ..., \varepsilon_{Nt})'$ is an $N \times 1$error term vector.

Definition of weight matrix

The element on the ith row and jth column of weight matrix W_{Nt} is denoted as $w_{ij,\,t}$, indicating the level of connection between unit i and j at time t. Unlike traditional spatial econometric models where $w_{ij,\,t}$ is exogenously defined as a function of the geographic distance between i and j, the study defines $w_{ij,\,t}$ based on the two units' socioeconomic characteristics so that

$$w_{ij,t} = F\left(z_{it}, z_{jt}\right) \qquad i, j \quad \in N \tag{3}$$

where $F(\bullet)$ is either a proximity or an interaction function with estimable parameters. The function can take various forms such as a generalized Euclidean distance, a normalized Euclidean distance, or a gravity model. $z_{it}=(z_{i1,\,t}, z_{i2,\,t}, ..., z_{ip,\,t})$ is a $1 \times p$ vector of observation i's demographic or economic characteristics at time t. For example, Conley and Topa (2002) defined a "race and ethnicity distance" by calculating the difference between two Census tracts' race and ethnicity characteristics. They used the resulting weight matrix to investigate the spatial pattern of unemployment. Lee and Yu (2012) further extended the definition of $w_{ij,\,t}$ by combining the demographic and economic distance with geographical distance, and incorporating estimable parameters. Following Han's definition, this paper defines weight element $w_{ij,\,t}$ by incorporating geographic distance d_{ij} with exponential coefficient $-\gamma_0$ and economic distance $|z_{it}-z_{jt}|$ between pairs with exponential coefficients $-\gamma_1$, $-\gamma_2$,..., $-\gamma_p$

$$w_{ij,t} = d_{ij}^{-\gamma_0} \cdot \left|z_{i1,t} - z_{j1,t}\right|^{-\gamma_1} \cdot \left|z_{i2,t} - z_{j2,t}\right|^{-\gamma_2} ... \left|z_{ip,t} - z_{jp,t}\right|^{-\gamma_p}, \qquad i \neq j, t = 1, 2, ..., T \tag{4}$$

where d_{ij} is the geographical distance between observation i and j, which does not vary over time; $|\bullet|$ is the Euclidean distance function so that $|z_{is,\,t}-z_{js,\,t}|$ measures the difference of feature s (e.g., population density) between observation i and j at time t; and $\gamma=(\gamma_0, \gamma_1, ..., \gamma_p)$is the $(p+1) \times 1$vector of estimable parameters. The matrix is then row-normalized, leading to the W_{Nt} used in Eq. (2). This weight matrix definition proposed by Lee and Yu (2012) is both comprehensive and flexible, and is thus adopted by this study.

Spatial autoregressive binary probit model with endogenous weight matrix (SARBP-EWM)

In many circumstances, the socioeconomic factors z_{it} are rather endogenous, and may be correlated with the decision maker's choice y_{it}. For example, a company's employment size may be correlated with its location choice; and a resident's vehicle ownership may be correlated with his/her mode choice. Therefore, the problem of endogenous weight matrix will arise when using the employment size to measure two companies' similarity and further create weight matrix W_{Nt}, or when using vehicle ownership to define two traveler's socioeconomic connection. In order to address the endogeneity issue, the correlation between z_{it} and y_{it}^* must be explicitly addressed. The resulting model structure is similar to the one proposed by Han (2014), who addressed the endogeneity weight matrix issue for continuous response variables using an additional entry equation:

$$Z_t = \tilde{X}\tilde{\beta} + \tilde{M} + l \otimes \tilde{n}' + \Delta_t, t = 1, 2, ..., T \tag{5}$$

where $Z_t=(z'_{1t}, z'_{2t},..., z'_{Nt})'$ is a $N \times p$ matrix representing the $p$dimensional economic characteristics of all observations at time t. Eq. (5) is called entry equation because it provides entry to the W in Eq. (2). In other words, W is defined by Eq. (5). X is an $N \times k_2$ matrix of exogenous variables allowed to be time variant. $\tilde{\beta}$ are the $k_2 \times p$ coefficients matrix associated with exogenous variables The definition of $\tilde{M}$ and $\tilde{n}$ are similar to those for Eq. (2); and $\Delta_t=(\delta'_{1t}, \delta'_{2t},..., \delta'_{Nt})'$ is a $N \times p$ matrix of error terms. The endogeneity occurs when the E_{nt} in Eq. (2) and Δ_{nt} are allowed to be correlated. Assuming an i.i.d. jointly normal distribution for ε_{it} and δ_{it} across all i's and j's with mean 0 and variance-covariance matrix V, the multivariate normal distribution can be written as

$$(\varepsilon_{it}, \delta_{it}) \sim i.i.d\ \ N_{p+1}\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma'_{\varepsilon\delta} \\ \sigma_{\varepsilon\delta} & V_\delta \end{pmatrix}\right) = N_{p+1}(0, V) \tag{6}$$

or

$$\varepsilon_{it}|\delta_{it} \sim N(\sigma'_{\varepsilon\delta}V_\delta^{-1}\delta_{it}, \sigma_\varepsilon^2 - \sigma'_{\varepsilon\delta}V_\delta^{-1}\sigma_{\varepsilon\delta}) \tag{7}$$

Substituting Eq. (7) into Eq. (2), the equation can be rewritten as

$$(I_N - \rho W_t)Y_t^* = X\beta + M + ln + \left(Z_t - \tilde{X}\tilde{\beta} - \tilde{M} - l \otimes \tilde{n}'\right)\eta + \xi_t, t = 1, 2, ldots, T \tag{8}$$

where $\eta = V_\delta^{-1}\sigma_{\varepsilon\delta}$ is a $p \times 1$ vector. Error term $\xi_t \sim N(0, \sigma_\xi^2 I_N)$ follows the normal distribution with mean 0 and variance $\sigma_\xi^2 = \sigma_\varepsilon^2 - \sigma'_{\varepsilon\delta}V_\delta^{-1}\sigma_{\varepsilon\delta}$. In particular, ξ_t is independent of Y_t^*. The conditional likelihood function can be generally written as

$$f(Y_t^*|Z_t, \rho, \gamma, \beta, V) \propto \left(\left(\sigma_\xi^2\right)^{-\frac{N}{2}} \times |S_t(\rho, \gamma)| \times \exp\left(-\frac{H'_t H_t}{2\sigma_\xi^2}\right)\right) \tag{9}$$

where $S_t(\rho, \gamma) = I_N - \rho W_t(\gamma)$, and $H_t = (I_N - \rho W_t)Y_t^* - X\beta - M - ln - (Z_t - \tilde{X}\tilde{\beta} - \tilde{M} - l \otimes n')\eta$

Model estimation

Given the multi-layer form of the likelihood function, a Bayesian Markov Chain Monte Carlo (MCMC) method is adopted to estimate the model. Essentially, MCMC decomposes the estimation of a complex model into a sequence of sub layers, each addressing one parameter (LeSage and Pace, 2009). Therefore, a great advantage of MCMC method is that the posterior distributions of most parameters may remain the same even when additional considerations lead to a more complex model form. For example, when extending a linear regression model to a binary choice model, most parameters' posterior distributions will stay the same by simply treating the latent variable (utility) as the original continuous dependent variable. The key change is an additional layer of posterior distribution that links the latent variable Y_t^* and the binary response Y_t. In fact, the posterior distributions for ρ, γ, β, V take the same forms as those explained in previous study (Han, 2014). Therefore, the paper skips the derivation process and simply summarizes the posterior distributions below:

The key, additional, layer that enables the new SARBP-EWM model is the generation of Y_t^* from observed binary response Y_t, which is also a major challenge in the SARBP-EWM model estimation. Since there is no way that the utility difference could be obtained, the latent dependent variable Y_t^* has no prior information. It must be derived from the discrete dependent variable Y_t, input Z_t and other parameters.

Denote $\theta = (\rho, \gamma, \beta, V)$as the parameters of the model. Using Bayesian's theory, the joint posterior density can be expressed below:

$$\begin{aligned} &p(\theta, Y_t^*|Y_t, Z_t) \\ &\propto f(Y_t|Y_t^*) \times f(Y_t^*|Z_t, \theta) \times f(Z_t|\theta) \times \pi(\theta) \end{aligned} \tag{10}$$

where $p(\theta, Y_t^*|Y_t, Z_t)$ indicates posterior density on condition of Y_t, Z_t. For simplicity, exogenous variables $X, \tilde{X}$ are not included in the above equations. The posterior density is proportional to parameters' prior density $\pi(\theta)$ and likelihood function.

Inspired by previous study on Bayesian probit model (Smith and LeSage, 2004), the posterior distribution $p(Y_t|Y_t^*)$ of input variable Y_t conditional on latent dependent variable Y_t^* can be written in terms of $\prod_{t=1}^{T}\prod_{i=1}^{N}\{\delta(y_{it} = 1)\delta(y_{it}^* > 0) + \delta(y_{it} = 0)\delta(y_{it}^* \le 0)\}$. Using Bayes' theorem and Eq. (10), the posterior distribution of the dependent variable Y_{Nt}^* can be written as

$$\begin{aligned} &p(Y_t^*|Z_t, Y_t, \rho, \gamma, \beta, V) \propto p(Y_t|Y_t^*) \times f(Y_t^*|Z_t, \theta) \\ &\propto \prod_{t=1}^{T}\prod_{i=1}^{N}\left\{\delta(y_{it} = 1)\delta\left(y_{it}^* > 0\right) + \delta(y_{it} = 0)\delta\left(y_{it}^* \le 0\right)\right\} \\ &\times \prod_{t=1}^{T}\left\{\left(\sigma_\xi^2\right)^{-\frac{N}{2}} \times |S_t(\rho, \gamma)| \times \exp(-\frac{H'_t H_t}{2\sigma_\xi^2})\right\} \end{aligned} \tag{11}$$

where the first element of the right part is simply a censoring function and the second element is essentially a normal distribution $N(S_t^{-1}H_t, \sigma_\xi^2(S'_tS_t)^{-1})$. By letting $y_{-it}^* = (y_{11}^*, \ldots, y_{i,t-1}^*, y_{i,t+1}^*, \ldots, y_{NT}^*)$, the conditional posterior of y_{it}^* thus follows a truncated normal distribution below

$$p\left(y_{it}^*|Z_t, Y_t, y_{-it}^*, \rho, \gamma, \beta, V\right) \sim \begin{cases} N_i(S_t^{-1}H_t, \sigma_\xi^2(S'_tS_t)^{-1}), \text{left truncated at 0, if } y_{it} = 1 \\ N_i(S_t^{-1}H_t, \sigma_\xi^2(S'_tS_t)^{-1}), \text{right truncated at 0, if } y_{it} = 0 \end{cases} \tag{12}$$

where $N_i(S_t^{-1}H_t, \sigma_\xi^2(S'_tS_t)^{-1})$ refers to the ith element of the multivariate normal distribution $N(S_t^{-1}H_t, \sigma_\xi^2(S'_tS_t)^{-1})$.

Model validation

To validate the SARBP-EWM model, simulated data is used to analyze the model performance. The simulated dataset contains 20 different units ($N = 20$) across 20 time periods ($T = 20$). To simplify the problem, we assume both the SAR equation and the entry equation have only one independent variable ($k_1 = 1$, $k_2 = 1$) and the entry equation dependent variable has only two dimensions ($p = 2$). The independent variable X is randomly generated from uniform distribution between -1

Table 1
Parameters' posterior distributions.

Variable	Prior distribution	Posterior distribution	Sampling method
ρ, γ	$\rho \sim U(-1,1)$ $\gamma \sim N_{p+1}(\gamma_o, R_0)$ $\rho_o = 0.1$ $\gamma_o : (p+1) \times 1$vectorof0 $R_0 : I_{p+1} \times 10^{12}$	$f(\{Y_t^*\}\|\{Z_t\}, \rho, \gamma, \beta, V) \propto \prod_{t=1}^{T} \|S_t(\rho,\gamma)\| \times \exp(-\frac{H'_t H_t}{2\sigma_\xi^2})$	Metro-polis Hasting (M-H)
β	$(\beta, \tilde{\beta}) \sim N_{k_1+pk_2}(\beta_o, B_0)$ $\beta_o : (k_1+pk_2) \times 1$vectorof0 $B_0 : I_{k_1+pk_2} \times 10^{12}$	$N_{k_1+pk_2}(T_\beta, A_\beta^{-1})$ where $T_\beta = A_\beta^{-1}(B_0^{-1}\beta_0 + \sum_{t=1}^{T}\sum_{i=1}^{n} x'_{it\|\beta} V^{-1} y^*_{it\|\beta})$ $A_\beta^{-1} = (B_0^{-1} + \sum_{t=1}^{T}\sum_{i=1}^{n} x'_{it\|\beta} V^{-1} x_{it\|\beta})^{-1}$	Multi-variate normal distri-bution
V	$V \sim W_{p+1}^{-1}(\Psi, \nu)$ $\Psi : I_{p+1}$ $\nu : 10$	$W_{p+1}^{-1}(\Psi + \sum_{t=1}^{T}\sum_{i=1}^{n} h_{it} h'_{it}, \nu + nT)$ where h_{it} is the i^{th} element of H_t at time t	Inverse-Wishart distri-bution

Note: More detailed derivation process of posterior distribution functions can be found in Han and Zhou's work (Han, 2014 and Zhou, 2015) Posterior distributions of Y_t^* $(Y_t^*|Z_t, Y_t, \rho, \gamma, \beta, V)$.

and 1. M,$\tilde{M}$,n and $\tilde{n}$ are held at zero in both creation and estimation process to expedite the estimation process. The true value of $\gamma = (\gamma_0, \gamma_1, \gamma_2)$ is set to be [0.9; 1.2; 1.3]. The geographic distance d_{ij} corresponding to γ_0 is pre-defined randomly and fixed in the simulation process. The true value of β is [0.5; 0.5; 0.5]. The true values for the diagonal elements in the variance-covariance matrix $V$are set to be [1.6; 0.8; 0.8], and for identification purpose, the first diagonal element of matrix V(i.e., σ_ε^2) is fixed at 1.6. Furthermore, to investigate the model performance with respect to different spatial autocorrelation and covariance levels, different ρ and covariance $\sigma_{\varepsilon\delta}$ values are also tested: Validation is performed using 3 different spatial correlation factor values (ρ = 0.3, 0.5, 0.8) and 4 different covariance values ($\sigma_{\varepsilon\delta}$ = 0, 0.3, 0.5, 0.8), leading to a total of 12 different scenarios. In addition, to compare the performance of SARBP-EWM model with simple spatial models ignoring endogeneity, simulations are also run for the case (ρ = 0.3, $\sigma_{\varepsilon\delta}$ = 0.8) with $\sigma_{\varepsilon\delta}$ fixed at 0.

In addition to calculating mean and standard deviation of parameters, two other criteria, mean absolute percentage error (MAPE) and percentage mean square error (PMSE) are introduced. MAPE measures the percentage deviation from its true value in the linear form. PMSE measures the percentage of square deviation from its true value where larger forecast errors are penalized in quadratic form.

$$MAPE = 100 \times \sum_{i=1}^{N} \left[\frac{|y_{it} - y_{it}^0|}{y_{it}^0} \right] \Big/ N \tag{13}$$

$$PMSE = 100 \times \sum_{i=1}^{N} \left[\frac{y_{it} - y_{it}^0}{y_{it}^0} \right]^2 \Big/ N \tag{14}$$

Where

y_{it}^0: True value for observation i at time t

y_{it}: Estimated value for observation i at time t

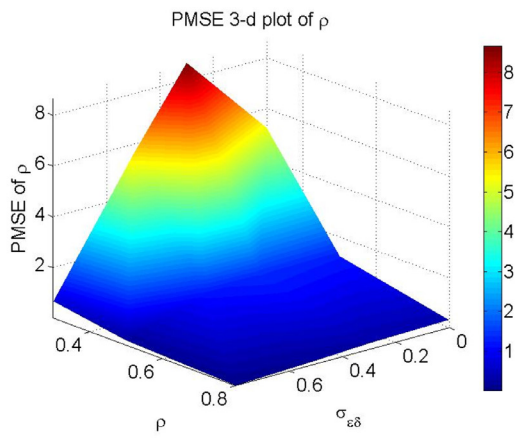
N: Number of observations

As randomness of one sample may obscure the properties of parameter estimates, 100 random samples are generated for each scenario following the same specification. Parameter estimates are obtained by taking the average of parameter values over 100 samples. For each sample, the model is run for 6000 times and the first 4000 iterations are set as the "burn-in". The "burn-in" iterations allow parameter estimates to gradually converge from their starting values to their true values. The estimated parameter values are obtained by taking the average of last 2000 iterations after the "burn-in" part. Parameter estimation results are presented in Table 2 through Table 4. The PMSE under all scenarios are also presented in Fig. 1.

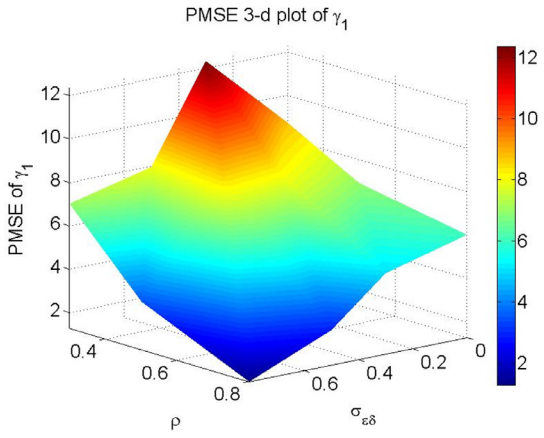
Table 5 shows estimation results from SARBP-EWM and the ones from simple spatial models ignoring endogeneity for the case ρ=0.3 and $\sigma_{\varepsilon\delta}$ = 0.8.

In general, the parameters' estimation traces (not shown here due to space limitation) suggest convergence trend towards their true values. Although convergence diagnosis methods are not used, validation results already show that parameter estimates are close to their true values with relatively low variation. However, estimation results vary across different scenarios. Some key findings are summarized below.

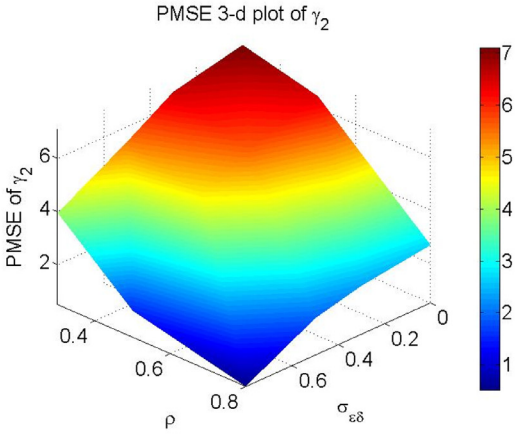
First, scenarios with high covariance values yield generally better estimation results compared to low covariance. Parameter estimates are closer to their true values and have less variation when covariance $\sigma_{\varepsilon\delta}$ is high, which are indicated by the smaller MAPE and PMSE values. For example, for the 4 scenarios with spatial autocorrelation factor ρ = 0.8, all elements in γ have smaller MAPE and PMSE values when covariance $\sigma_{\varepsilon\delta}$ values are higher. When the spatial autocorrelation factor is lower, such trend becomes less obvious. However, the general trend that MAPE and PMSE values are smaller when covariance is high can still be observed. Similar trends are also found in other parameters such as ρ, β and V. Intuitively,



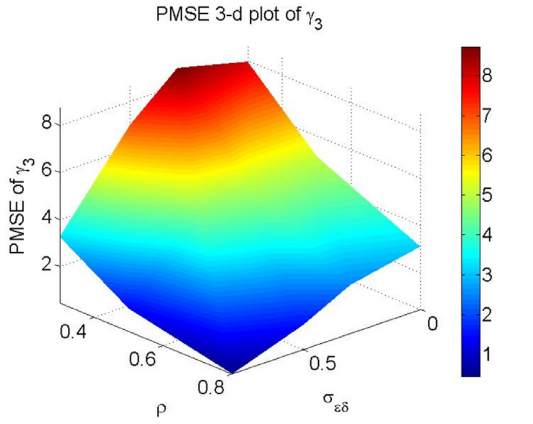
(a) PMSE of ρ for all scenarios



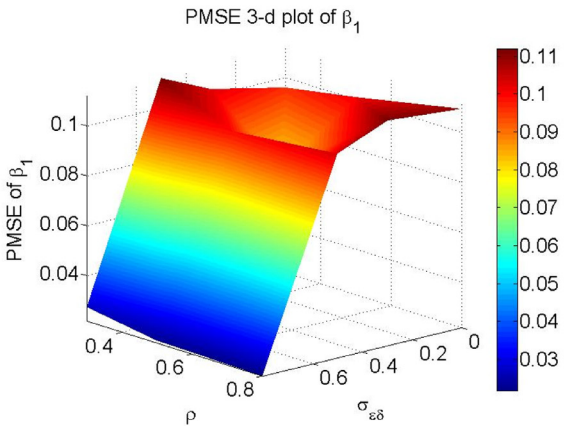
(b) PMSE of γ_1 for all scenarios



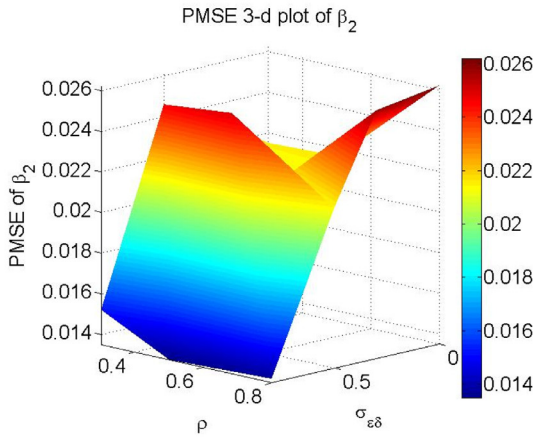
(c) PMSE of γ_2 for all scenarios



(d) PMSE of γ_3 for all scenarios

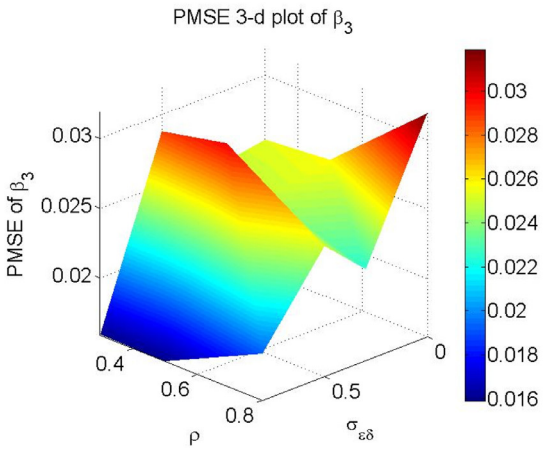


(e) PMSE of β_1 for all scenarios

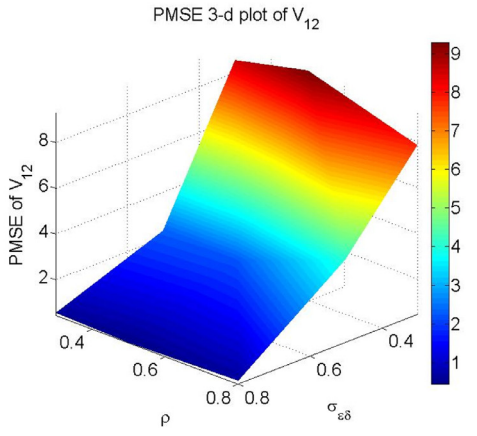


(f) PMSE of $\tilde{\beta}_1$ for all scenarios

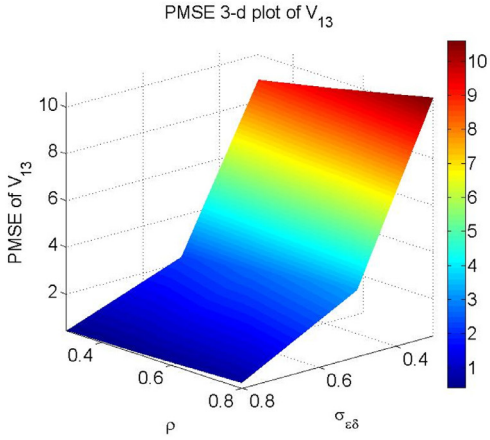
Fig. 1. PMSE for all scenarios with 100 random samples.



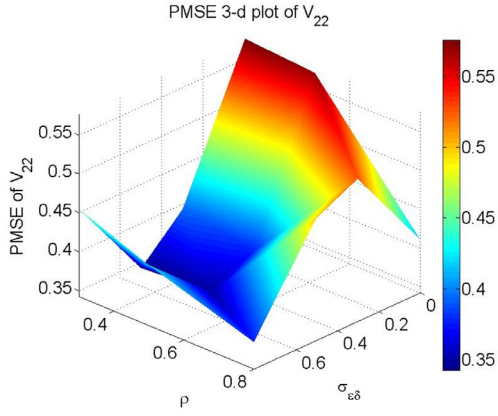
(g) PMSE of $\tilde{\beta}_2$ for all scenarios



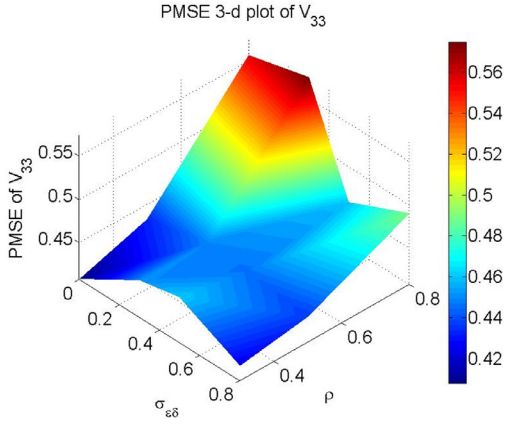
(h) PMSE of V_{12} for all scenarios



(i) PMSE of V_{13} for all scenarios



(j) PMSE of V_{22} for all scenarios



(k) PMSE of V_{33} for all scenarios

Fig. 1. Continued

Table 2
Validation results of 100 samples for $\rho = 0.3$.

Parameter	$\sigma_{\varepsilon\delta} = 0.0$					$\sigma_{\varepsilon\delta} = 0.3$				
	True value	Mean	S.D.	MAPE	PMSE	True value	Mean	S.D.	MAPE	PMSE
ρ	0.3000	0.2997	0.0071	18.77	5.236	0.3000	0.2787	0.0060	21.95	8.661
γ_1	0.9000	0.9512	0.0094	23.46	8.748	0.9000	0.9533	0.0046	26.45	12.35
γ_2	1.2000	1.0377	0.0065	21.13	7.096	1.2000	1.0207	0.0056	20.67	6.524
γ_3	1.3000	1.0442	0.0045	22.72	7.965	1.3000	1.0598	0.0072	23.86	8.725
β	0.5000	0.5027	0.0016	2.353	0.0961	0.5000	0.4977	0.0015	2.555	0.1027
$\tilde{\beta}_1$	0.5000	0.5008	0.0008	1.165	0.0215	0.5000	0.4994	0.0008	1.163	0.0235
$\tilde{\beta}_2$	0.5000	0.5001	0.0008	1.353	0.0254	0.5000	0.5003	0.0007	1.321	0.0246
$V_{12}(\sigma_{\varepsilon\delta})$	0.0000	-0.0061	0.0086	–	–	0.3000	0.3033	0.0083	23.58	8.541
$V_{13}(\sigma_{\varepsilon\delta})$	0.0000	0.0030	0.0087	–	–	0.3000	0.3117	0.0084	24.96	8.909
$V_{22}(\sigma_\delta^2)$	0.8000	0.7889	0.0056	5.982	0.5762	0.8000	0.7904	0.0055	5.100	0.3938
$V_{33}(\sigma_\delta^2)$	0.8000	0.7885	0.0055	4.898	0.4078	0.8000	0.7881	0.0054	5.423	0.4501
	$\sigma_{\varepsilon\delta} = 0.5$					$\sigma_{\varepsilon\delta} = 0.8$				
ρ	0.3000	0.2849	0.0058	18.12	5.629	0.3000	0.2995	0.0037	6.657	0.6806
γ_1	0.9000	0.9080	0.0084	21.31	7.937	0.9000	0.9009	0.0034	20.75	7.011
γ_2	1.2000	1.0745	0.0065	19.36	5.450	1.2000	1.0855	0.0122	15.62	3.965
γ_3	1.3000	1.0733	0.0037	21.86	7.012	1.3000	1.1705	0.0212	14.93	3.277
β	0.5000	0.4991	0.0014	2.690	0.1119	0.5000	0.5000	0.0012	1.314	0.0275
$\tilde{\beta}_1$	0.5000	0.4996	0.0007	1.246	0.0246	0.5000	0.4999	0.0007	0.9594	0.0152
$\tilde{\beta}_2$	0.5000	0.5007	0.0008	1.330	0.0288	0.5000	0.4998	0.0007	1.0228	0.0160
$V_{12}(\sigma_{\varepsilon\delta})$	0.5000	0.5117	0.0076	12.56	2.311	0.8000	0.7593	0.0057	6.139	0.5682
$V_{13}(\sigma_{\varepsilon\delta})$	0.5000	0.5162	0.0073	11.76	2.241	0.8000	0.7768	0.0053	5.268	0.416
$V_{22}(\sigma_\delta^2)$	0.8000	0.7918	0.0057	4.717	0.3425	0.8000	0.7828	0.0055	5.404	0.455
$V_{33}(\sigma_\delta^2)$	0.8000	0.7871	0.0056	5.314	0.4602	0.8000	0.7904	0.0054	5.280	0.425

Table 3
Validation results of 100 samples for $\rho = 0.5$.

Parameter	$\sigma_{\varepsilon\delta} = 0.0$					$\sigma_{\varepsilon\delta} = 0.3$				
	True value	Mean	S.D.	MAPE	PMSE	True value	Mean	S.D.	MAPE	PMSE
ρ	0.5000	0.4936	0.0054	8.567	1.251	0.5000	0.4788	0.0050	9.673	1.387
γ_1	0.9000	0.8843	0.0086	21.40	6.921	0.9000	0.8984	0.0161	23.08	8.221
γ_2	1.2000	1.0168	0.0070	20.52	6.423	1.2000	1.0731	0.0077	19.38	5.644
γ_3	1.3000	1.1163	0.0079	19.11	5.115	1.3000	1.1397	0.0138	18.10	5.072
β	0.5000	0.5026	0.0016	2.367	0.1010	0.5000	0.4979	0.0015	2.296	0.0800
$\tilde{\beta}_1$	0.5000	0.5007	0.0008	1.156	0.0216	0.5000	0.4990	0.0008	1.113	0.0209
$\tilde{\beta}_2$	0.5000	0.5000	0.0008	1.368	0.0258	0.5000	0.5002	0.0008	1.325	0.0248
$V_{12}(\sigma_{\varepsilon\delta})$	0.0000	-0.0151	0.0088	–	–	0.3000	0.3153	0.0082	25.38	9.291
$V_{13}(\sigma_{\varepsilon\delta})$	0.0000	0.0075	0.0090	–	–	0.3000	0.3097	0.0084	24.53	9.549
$V_{22}(\sigma_\delta^2)$	0.8000	0.7899	0.0055	5.934	0.5688	0.8000	0.7925	0.0056	4.954	0.3760
$V_{33}(\sigma_\delta^2)$	0.8000	0.7915	0.0055	5.029	0.4328	0.8000	0.7889	0.0054	5.438	0.4494
	$\sigma_{\varepsilon\delta} = 0.5$					$\sigma_{\varepsilon\delta} = 0.8$				
ρ	0.5000	0.4885	0.0048	8.562	1.146	0.5000	0.5012	0.0026	3.257	0.1779
γ_1	0.9000	0.8640	0.0113	23.46	8.627	0.9000	0.8483	0.0041	15.33	3.483
γ_2	1.2000	1.0928	0.0094	16.93	4.650	1.2000	1.1434	0.0119	9.901	1.514
γ_3	1.3000	1.1419	0.0158	17.82	4.594	1.3000	1.2492	0.0153	9.834	1.394
β	0.5000	0.4981	0.0014	2.610	0.0994	0.5000	0.4999	0.0009	1.184	0.0226
$\tilde{\beta}_1$	0.5000	0.4996	0.0008	1.216	0.0249	0.5000	0.5000	0.0007	0.8926	0.0135
$\tilde{\beta}_2$	0.5000	0.5008	0.0007	1.379	0.0298	0.5000	0.4999	0.0007	1.0269	0.0159
$V_{12}(\sigma_{\varepsilon\delta})$	0.5000	0.5020	0.0076	12.08	2.205	0.8000	0.7677	0.0054	5.512	0.4487
$V_{13}(\sigma_{\varepsilon\delta})$	0.5000	0.5032	0.0079	13.20	2.646	0.8000	0.7703	0.0054	5.613	0.4811
$V_{22}(\sigma_\delta^2)$	0.8000	0.7900	0.0056	4.749	0.3481	0.8000	0.7836	0.0052	5.171	0.4203
$V_{33}(\sigma_\delta^2)$	0.8000	0.7843	0.0056	5.299	0.4461	0.8000	0.7916	0.0055	5.314	0.4367

when the covariance level is higher and the correlation is explicitly recognized, the information in the entry equation adds explanatory power to the SAR model and vice versa, resulting in a better model performance.

Second, ρ and γ have lower MAPE and PMSE values as spatial autocorrelation factor ρ increases. For example, in all the 3 scenarios with $\sigma_{\varepsilon\delta} = 0.8$, MAPE and PMSE values decrease significantly for all elements in γ as ρ increases. Similarly, ρ itself also has significantly lower MAPE and PMSE values as ρ increases. Such trend can still be observed, although less evident for β and V. Intuitively, the high ρ value amplifies the effects of γ through the multiplication term $\rho W_t(\gamma)$, leading to better estimation results. Such effect becomes indirect for other parameters β and V. Another possible reason that ρ and γ show a clear trend with the change of ρ could be that they are estimated using the M-H method. Random samples with larger variance have higher probability of not being selected for next step.

Table 4
Validation results of 100 samples for $\rho = 0.8$.

Parameter	$\sigma_{\varepsilon\delta} = 0.0$					$\sigma_{\varepsilon\delta} = 0.3$				
	True value	Mean	S.D.	MAPE	PMSE	True value	Mean	S.D.	MAPE	PMSE
ρ	0.8000	0.7954	0.0001	4.399	0.327	0.8000	0.7696	0.0027	4.322	0.2591
γ_1	0.9000	0.9061	0.0049	18.82	6.017	0.9000	0.8820	0.0054	17.30	5.003
γ_2	1.2000	1.2107	0.0073	13.25	2.681	1.2000	1.1495	0.0090	12.57	2.325
γ_3	1.3000	1.3051	0.0085	13.37	3.104	1.3000	1.2399	0.0095	12.31	2.485
β	0.5000	0.4998	0.0010	2.610	0.1093	0.5000	0.4988	0.0015	2.664	0.1120
$\tilde{\beta}_1$	0.5000	0.4995	0.0007	1.274	0.0262	0.5000	0.5000	0.0008	1.275	0.0257
$\tilde{\beta}_2$	0.5000	0.4993	0.0007	1.461	0.0319	0.5000	0.4997	0.0008	1.239	0.0224
$V_{12}(\sigma_{\varepsilon\delta})$	0.0000	0.7583	0.0061	–	–	0.0000	0.2881	0.0097	22.13	7.835
$V_{13}(\sigma_{\varepsilon\delta})$	0.0000	0.7706	0.0059	–	–	0.0000	0.3006	0.0095	25.28	10.61
$V_{22}(\sigma_\delta^2)$	0.8000	0.7813	0.0053	5.231	0.4113	0.8000	0.7938	0.0056	5.807	0.5256
$V_{33}(\sigma_\delta^2)$	0.8000	0.7959	0.0054	6.036	0.5557	0.8000	0.7929	0.0054	6.079	0.5744
	$\sigma_{\varepsilon\delta} = 0.5$					$\sigma_{\varepsilon\delta} = 0.8$				
ρ	0.8000	0.7841	0.0024	2.770	0.1259	0.8000	0.7954	0.0001	1.1280	0.0223
γ_1	0.9000	0.8757	0.0059	13.50	2.873	0.9000	0.9061	0.0049	8.970	1.274
γ_2	1.2000	1.1638	0.0062	11.33	1.901	1.2000	1.2107	0.0073	5.687	0.5148
γ_3	1.3000	1.2693	0.0061	9.941	1.488	1.3000	1.3051	0.0085	5.224	0.4445
β	0.5000	0.5003	0.0014	2.633	0.1037	0.5000	0.4998	0.0010	1.200	0.0218
$\tilde{\beta}_1$	0.5000	0.4995	0.0008	1.177	0.0214	0.5000	0.4995	0.0007	0.9173	0.0137
$\tilde{\beta}_2$	0.5000	0.5008	0.0007	1.292	0.0252	0.5000	0.4993	0.0007	1.153	0.0192
$V_{12}(\sigma_{\varepsilon\delta})$	0.5000	0.5094	0.0087	16.44	4.071	0.8000	0.7583	0.0061	6.610	0.6175
$V_{13}(\sigma_{\varepsilon\delta})$	0.5000	0.5194	0.0085	14.37	3.254	0.8000	0.7706	0.0059	6.185	0.6201
$V_{22}(\sigma_\delta^2)$	0.8000	0.7941	0.0054	5.753	0.4993	0.8000	0.7813	0.0053	5.117	0.3771
$V_{33}(\sigma_\delta^2)$	0.8000	0.7954	0.0058	5.625	0.4591	0.8000	0.7959	0.0054	5.673	0.4896

Table 5
Comparison with base case for $\rho = 0.3$ and $\sigma_{\varepsilon\delta} = 0.8$.

Parameter		SARBP-EWM				MO($\sigma_{\varepsilon\delta}$ fixed at 0)			
	True value	Mean	S.D.	MAPE	PMSE	Mean	S.D.	MAPE	PMSE
ρ	0.3000	0.2995	0.0037	6.657	0.6806	0.3304	0.0210	24.9514	9.4284
γ_1	0.9000	0.9009	0.0034	20.75	7.011	0.8970	0.0164	16.3406	4.3763
γ_2	1.2000	1.0855	0.0122	15.62	3.965	0.9631	0.0218	26.4767	9.0886
γ_3	1.3000	1.1705	0.0212	14.93	3.277	0.9908	0.0225	24.1121	8.5858
β	0.5000	0.5000	0.0012	1.314	0.0275	0.4936	0.0051	3.4746	0.1864
$\tilde{\beta}_1$	0.5000	0.4999	0.0007	0.9594	0.0152	0.4964	0.0025	1.6977	0.0444
$\tilde{\beta}_2$	0.5000	0.4998	0.0007	1.0228	0.0160	0.5016	0.0025	1.4833	0.0347
$V_{12}(\sigma_{\varepsilon\delta})$	0.8000	0.7593	0.0057	6.139	0.5682	0(fixed)	N/A	N/A	N/A
$V_{13}(\sigma_{\varepsilon\delta})$	0.8000	0.7768	0.0053	5.268	0.416	0(fixed)	N/A	N/A	N/A
$V_{22}(\sigma_\delta^2)$	0.8000	0.7828	0.0055	5.404	0.455	0.7920	0.0177	8.7026	1.1605
$V_{33}(\sigma_\delta^2)$	0.8000	0.7904	0.0054	5.280	0.425	0.8149	0.0185	8.3371	1.0839

Third, The SARBP-EWM indeed achieves better estimation results compared to simple spatial models, ignoring endogeneity. As shown in Table 5, estimation results without recognizing the endogeneity effect (i.e., $\sigma_{\varepsilon\delta} = 0$ when the true value is 0.8) show significantly higher MAPE and PMSE values than the results estimated with SARBP-EWM. This suggests that estimation results that ignore endogenous weight matrix will have significant bias. The comparison gives a compelling reason for considering spatial correlations using endogenous weight matrix.

Overall, the parameter values, including the endogenous weight matrix structure, are reliably recovered with the SARBP-EWM model. The model successfully extends the spatial model with endogenous weight matrix from continuous form (Han, 2014) to binary form, while quantifying the functional structure of the weight matrix. The successful validation of the model paves way for further application of the model on transportation problems involving interdependencies, as shown below.

Application on firm relocation choice

We apply the SARBP-EWM model to a simplified firm relocation problem to further illustrate the model's behavior background and its capability to disclose the intertwining relationship between decision makers' connection and their choice decisions.

Background

Location choice is an important issue in transportation research. Firms make location choices to maximize their utilities, which are influenced by transportation conditions, markets, labor, materials, capital, government policy, etc. (Hayter, 1997).

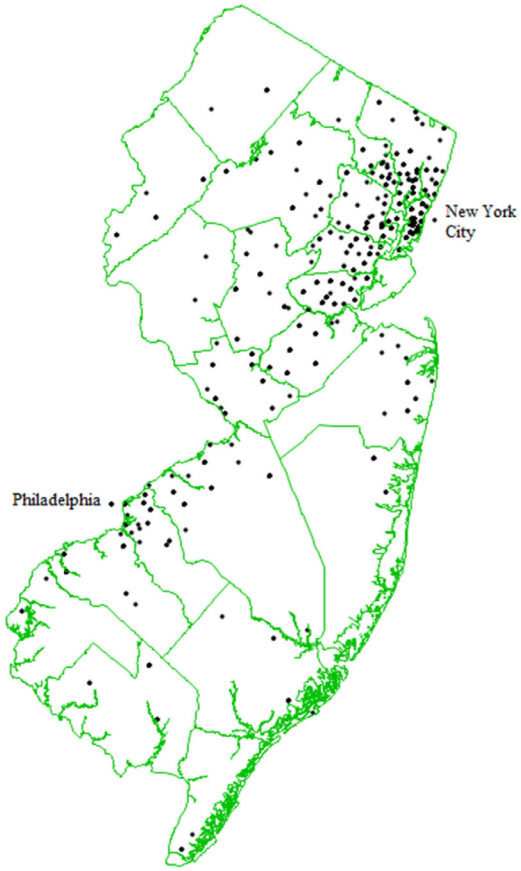


Fig. 2. Firm relocation choice in New Jersey.

For example, Mejia-Dorantes, Paez, and Vassallo (2012) used a multinomial logit model to evaluate the impacts of Madrid's metro line 12 expansion on business location patterns using a micro-level database. Results indicate that urban accessibility and urban agglomeration have great impacts on business location pattern. Holguin-Veras et al. (2005) conducted an in-depth investigation on firm relocation pattern using data for firms relocated to New Jersey between 1990 and 1999. Results showed that different industrial sectors had different elasticities with respect to accessibility variables. This paper further investigates firm location choice, considering the fact that firms often make decision in correlation with their peers. Firm location, employment size, type of industry, land use, transportation infrastructure and economic condition etc. are all influential factors. The mutual influence between firms needs to be considered in the location choice analysis to avoid biased estimation results. Meanwhile, the type and the level of firm interdependency are influenced by factors such as the firms' industrial sectors, employment size, floor space, and targeted customers, some of which may be correlated with the firm location, as decision makers tend to make joint decisions on all these factors. Such intertwining relationship between firm location choice is explored using the SARBP-EWM model.

Data description

This application uses a subsample of the dataset used by Holguin-Veras, et al. (2005). The full dataset contains information of 1017 firms relocated into New Jersey from outside the state from 1990 to 1999 (intrastate moves are not included). The 1017 firms cover 74 industrial sectors, which are represented by North American Industrial Classification System (NAICS) codes. 897 of these firms (88%) originated in the U.S., with 548 (54%) from New York State and 99 (10%) from Pennsylvania. Other states contributed the rest. Regarding the firms' destinations in New Jersey, Hudson County received 220 firms (22%) and Bergen County received 152 firms (15%). Both counties are directly connected to New York City. Hudson County is connected to midtown and lower Manhattan through Port Authority Trans-Hudson (PATH), Lincoln and Hudson Tunnels. Bergen County is connected to northern Manhattan via George Washington Bridge. Fig. 2 further illustrates the firm relocation choices. There are two clear clusters: New York City in the northeast, and Philadelphia in the southwest. Another clear trend is that many firms choose to relocate along transportation corridors, such as the New Jersey Turnpike (mostly I-95), which connects northeast and southwest of New Jersey. A few other firms choose to relocate following the Garden State

Table 6
Summary of variables in the firm relocation dataset.

Variable names	Description	Mean	Std. Dev	Min	Max
location	1 if firm chooses to relocate closer to New York City; 0 if firm chooses to relocate closer to Philadelphia	0.794	0.405	0	1
employment	Number of employees	111.85	127.17	1	1041
popden	Population density (1000 per square mile)	4.372	4.267	0.248	11.80
hhinc	Household income ($1000)	40.69	8.142	29.99	56.27
Naics2	North American Industry Classification System (NAICS) code in 2 digits that firm belongs to (Base case: construction sector (naics2_23))				
construction	1 if construction sector; 0 otherwise (naics2_23)	0.009	0.093	0	1
manufacturing	1 if manufacturing sector; 0 otherwise (naics2_3133)	0.232	0.423	0	1
wholesale	1 if wholesale trade sector; 0 otherwise (naics2_42)	0.066	0.248	0	1
retail	1 if retail trade sector; 0 otherwise (naics2_4445)	0.123	0.329	0	1
transware	1 if transportation and warehousing sector; 0 otherwise (naics2_4849)	0.250	0.434	0	1
information	1 if information sector; 0 otherwise (naics2_51)	0.057	0.232	0	1
finins	1 if finance and insurance sector; 0 otherwise (naics2_51)	0.105	0.308	0	1
realestate	1 if real estate and rental and leasing sector; 0 otherwise (naics2_53)	0.004	0.066	0	1
scitech	1 if professional, scientific, and technical services sector; 0 otherwise (naics2_54)	0.066	0.248	0	1
management	1 if management of companies and enterprises sector; 0 otherwise (naics2_55)	0.009	0.093	0	1
admin	1 if administrative and support and waste management and remediation services sector; 0 otherwise (naics2_56)	0.035	0.184	0	1
health	1 if health care and social assistance sector; 0 otherwise (naics2_62)	0.018	0.132	0	1
entertain	1 if arts, entertainment, and recreation sector; 0 otherwise (naics2_71)	0.004	0.066	0	1
food	1 if accommodation and food services sector; 0 otherwise (naics2_72)	0.009	0.093	0	1
other	1 if other services (except public administration) sector; 0 otherwise (naics2_81)	0.057	0.232	0	1
pubadm	1 if public administration sector; 0 otherwise (naics2_92)	0.004	0.066	0	1

Parkway along the east coast. Besides the firm relocation data, Holguin-Veras, et al. (2005) also compiled rich built environment information such as transportation accessibility and zonal level socioeconomic factors such as population density, area size and median income. The full dataset was integrated into the Geographic Information Systems (GIS) for spatial analyses.

This paper used a subsample of this original dataset by only focusing on firms that relocated to New Jersey during years 1998 and 1999. After data cleaning, the final dataset contains 183 valid records. One important issue is that the application only uses cross sectional data instead of panel data. Although the theoretical model indicates that the method can be applied to panel data, applicability cannot be validated due to data limitation. The model's applicability to panel data should be examined in the future when empirical panel data becomes available. Table 6 summarizes all variables in the dataset.

It should be noted that the original dataset contains more variables. This study only uses selected variables to demonstrate application of the model in capturing peer influence among firms using real data.

Model specification

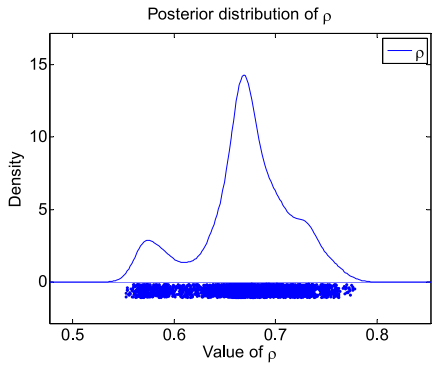
Variable "location" is used as the dependent variable Y, indicating whether the firm chose to relocate closer to New York City or Philadelphia. "Employment" is selected as the variable Z in the SARBP-EWM entry equation (Eq. 5). The number of employees is a typical indicator of firm size. Intuitively, firms with similar sizes tend to have stronger connections than firms with different sizes. In short, as the SARBP-EWM model implies, the variable Z influences the spatial correlation among dependent variable Y. The independent (and exogenous) variables for the location choice equation (i.e., X_1) include a set of industrial sector indicator variables and population density (popden). The independent variables for the employment size equation(X_2) include mainly the industrial sector indicator variables.

Results analysis

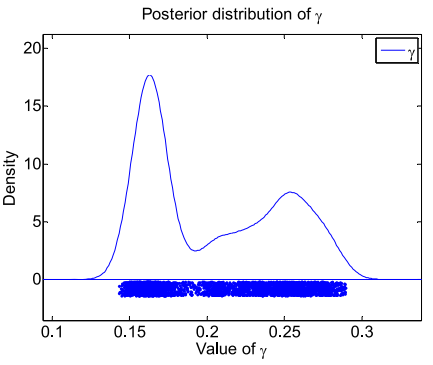
The model was run for 10,000 iterations, and in general all parameters show convergence trends. β and V converge quickly and remain stable after the first 1000 iterations ρ and γ require more iterations to converge with the M-H method, but after 6000 iterations their values also become stable. Therefore, the first 6000 iterations are treated as the "burnt-in" runs and the coefficient estimates are obtained by taking statistics of the remaining 4000 iterations.

Using a desktop with quad-core CPU, the estimation takes about 20 h to complete the 10,000 iterations. As the computational difficulty greatly increases with n, it would be very challenging to increase the value of n to be more than 500. If larger sample size analysis is needed, parallel computation with super computers may be considered.

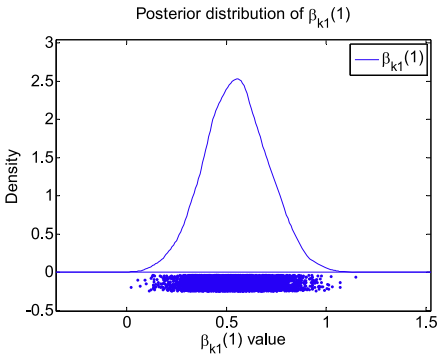
The estimated posterior distributions of all parameters are presented in Fig. 3, and their values are summarized in Table 7.



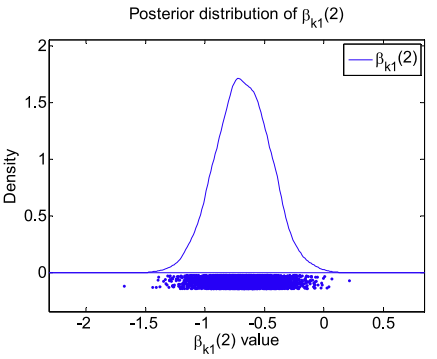
(a) Posterior distribution of ρ



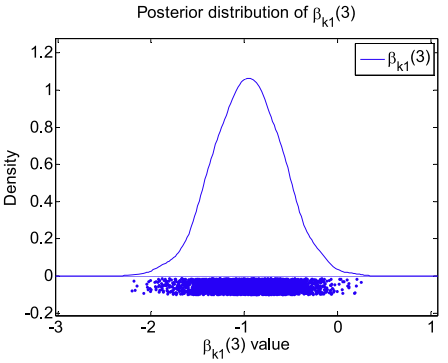
(b) Posterior distribution of γ



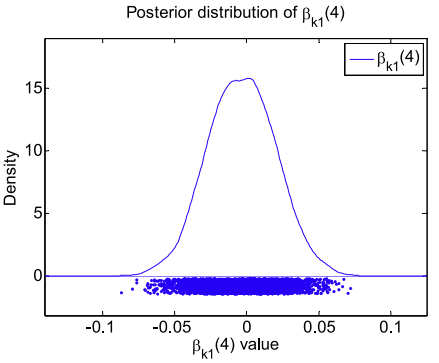
(c) Posterior distribution of β_1



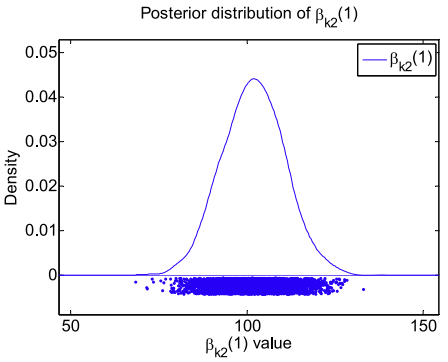
(d) Posterior distribution of β_2



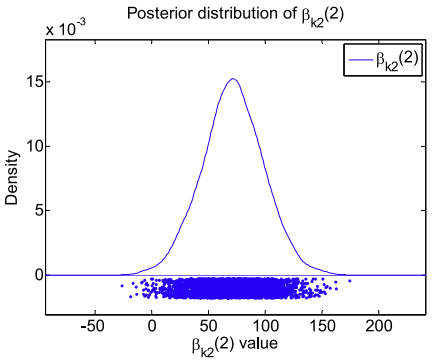
(e) Posterior distribution of β_3



(f) Posterior distribution of β_4

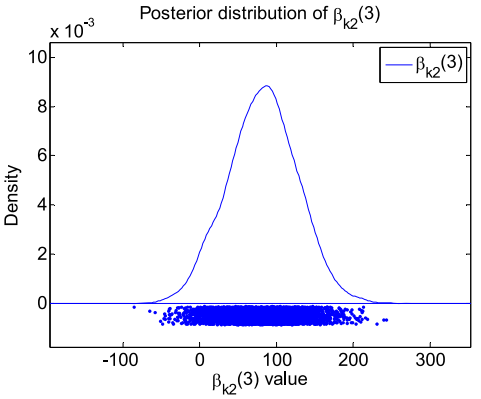


(g) Posterior distribution of $\tilde{\beta}_1$

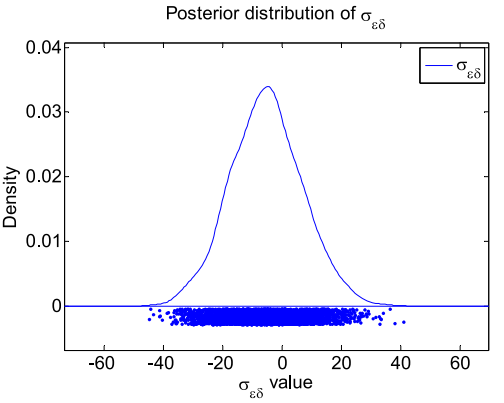


(h) Posterior distribution of $\tilde{\beta}_2$

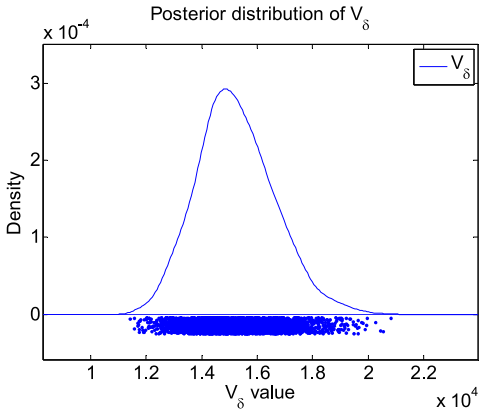
Fig. 3. Posterior distributions of parameters for firm relocation choice.



(i) Posterior distribution of $\tilde{\beta}_3$



(j) Posterior distribution of $\sigma_{\varepsilon\delta}$



(k) Posterior distribution of V_δ

Fig. 3. Continued

Table 7
Estimation results for firm location choice.

Parameters		Mean	Standard deviation	Pseudo t value	95% confidence interval
Spatial autocorrelation coef. (ρ)		0.6698	0.0445	15.052	(0.583, 0.757)
Coef. for weight matrix (γ)		0.2025	0.0442	4.581	(0.116, 0.289)
Coef. for location choice (β)	constant	0.5453	0.1568	3.478	(0.238, 0.853)
	transware	−0.6880	0.2254	−3.052	(−1.130, −0.246)
	scitech	−0.9670	0.3648	−2.651	(−1.682, −0.252)
	popden	−0.0040	0.0235	−0.170	(−0.050, 0.042)
Coef. for employment size ($\tilde{\beta}$)	constant	101.5984	8.816	11.524	(84.3, 118.9)
	finins	70.7318	27.03	2.617	(17.8, 123.7)
	admin	80.7986	45.06	1.793	(−7.52, 169.12)
Covariance matrix (V)	σ_ε^2	1(fixed)	0	–	–
	$\sigma_{\varepsilon\delta}$	−5.478	12.09	−0.453	(−29.174, 18.218)
	V_δ	15,207	1396	10.893	(12,471, 17,943)

ρ has a nonstandard distribution with mean value around 0.6698. It is estimated to be statistically significant with high pseudo t-value and 95% confidence interval between 0.664 and 0.676. This value indicates a strong correlation among firms and an ineligible impact on firms' relocation choices. γ value ranges between 0.15 and 0.3, with mean value 0.2025 and a high pseudo t-statistics. This suggests that the employment size indeed helps explain the weight matrix or the interdependency structure among firms.

Other independent variables also have strong effects on firms' relocation choices. The interpretation of these coefficients is similar to the interpretation for a standard spatial model: For β_{k_1}, the results suggest that compared to the base case (all other industry sectors), firms in the transportation and warehousing sector and the professional, scientific, and technical

services sector tend to relocate closer to Philadelphia, rather than New York City. This is consistent with expectation and previous findings (Holguin-Veras, et al, 2005): Transportation and warehouse companies require much land and are sensitive to transportation accessibility. It is likely that the travel time and land costs around New York City are much higher than other areas, which deters firms from relocating close to New York City. Firms in the scientific and technical sector also tend to relocate closer to Philadelphia than New York City. It is possible that Philadelphia has many top research institutions and universities, hence attracting these high-tech firms.

As for the factors influencing employment size, the estimates of $\tilde{\beta}$ indicate that firms in the finance and insurance sector (finins) and administrative and support sector (admin) tend to have higher numbers of employees. For example, compared to the base case (construction sector), a firm in the finance and insurance sector tends to have 70 more employees.

As for the covariance matrix V, the mean value of $\sigma_{\varepsilon\delta}$ is estimated to be -5.478 and the variance for the employment size equation is estimated to be 15,207, implying a correlation coefficient value -0.044. The value is both practically and statistically insignificant, suggesting that unobserved variables in firms relocation choice and employment size are at least linearly uncorrelated, and very likely, independent.

The application results offer insights into firm's location choice problem. More importantly, it presents an example for the application of SARBP-EWM model in real life transportation issues. It is expected that the SARBP-EWM model can be applied to many transportation issues, and the results will provide important insights that facilitate policy-making.

Conclusions

This paper investigates the issue of interdependencies in discrete choice modeling. A Spatial Autoregressive Binary Probit Model with Endogenous Weight Matrix (SARBP-EWM) is developed and Bayesian MCMC method is used for model estimation. The model is validated with simulated data. Estimation results show that parameters converge to their true values and endogenous weight matrix can be reliably recovered. In fact, the model performs better when the level of spatial autocorrelation and the level of endogeneity are higher. The development of this model adds great value to the existing spatial econometrics literature.

The model was applied to a simplified firm relocation choice problem to demonstrate how the model structure and estimated coefficients can deepen understanding of the decision-making behavior. With two years' New Jersey firm location choice data, the estimation results suggest that there is strong interdependency across firms. Their relocation choices are indeed correlated with their peers, and firms with similar sizes tend to have stronger connections with each other. Besides such "peer effect," the industrial sector also has strong impacts on firm's choice decisions. Firms in the transportation and warehousing sector and the professional, scientific, and technical services sector tend to relocate closer to Philadelphia than New York City. Firms in the finance and insurance sector and the administrative service related sectors tend to have more employees. Such application results offer interesting insights into firm location choice problems and informs policy making.

The model can be applied to a wide range of transportation issues where discrete responses may be subject to the influence of spatial, social, and economic connections, such as land development, location choice, and travel behavior. The successful accommodation of endogeneity weight matrix and the comprehensive identification of influential factors contribute to informed decision making. This paper advances efforts to study discrete choice modeling with interdependencies. There are few limitations: First, as other Bayesian estimation methods, the criterion for convergence is subjective, at least as noted by classical statisticians. Second, the application uses cross-sectional data. Future applications could be explored using panel data. Third, the sample size is limited to hundreds of observations due to the complexity of the model. If the sample size becomes too large, there would be computational issues. Future studies can address model refinement or study new applications. For example, future research on model improvement could be done to allow ordered/multinomial dependent variable and experiment with different weight matrix definitions such as gravity model. Besides, model performance could be tested, by examining its sensitivity to other parameter values.

References

Anselin, L., 2010. Thirty years of spatial econometrics. Papers Reg. Sci. 89 (1), 3–25.
Anselin, L., Bera, A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A., A. Giles, D.E. (Eds.), Handbook of Applied Economic Statistics. Marcel Dekker, New York, NY, pp. 237–290.
Bhat, C.R., 2015. A new spatial (social) interaction discrete choice model accommodating for unobserved effects due to endogenous network formation. J. Reg. Sci. 54 (3), 462–502.
Bhat, C.R., Dubey, S.K., Alam, M.J.B., Khushefati, W.H., 2015. A new spatial multiple discrete-continuous modeling approach to land use change analysis. J. Reg. Sci. 55 (5), 801–841. doi:10.1111/jors.12201.
Bhat, C.R., Astroza, S., Sidharthan, R., Bhat, P.C., 2014a. A multivariate hurdle count data model with an endogenous multiple discrete–continuous selection system. Transp. Res. Part B 63, 77–97.
Bhat, C.R., Paleti, R, Singh, P., 2014b. A spatial multivariate count model for firm location decisions. J. Reg. Sci. 54 (3), 462–502.
Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transp. Res. Part B 45 (7), 923–939.
Bhat, C.R., Guo, J.Y., 2007. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. Transp. Res. Part B 41 (5), 506–526.
Blume LE, Brock WA, Durlauf SN, Ioannides YM (2011). Identification of social interactions. In: Behhabib J, Jackson, MO, Bisin A (eds) Handbook of Social Economics Volume 1B, Chapter 18: 853-964, North-Holland, San Diego, CA.
Brock, W., Durlauf, S., 2001. Discrete choice with social interactions. Rev. Econ. Stud. 68, 235–260.
Brock, W, Durlauf, S, 2006. Multinomial choice with social interactions. In: Blume, L, Durlauf, S (Eds.). The Economy as an Evolving Complex System, 3. Oxford University Press, New York.

Brock, W., Durlauf, S., 2007. Identification of binary choice models with social interactions. J. Econom. 140, 52–75.
Cao, X., 2015. Examining the impacts of neighborhood design and residential self-selection on active travel: a methodological assessment. Urban Geogr. 36 (2), 236–255.
Chandrasekhar, A.G., Lewis, R., 2011. Econometrics of Sampled Networks (Working paper). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary .
Chakir, R., Parent, O., 2009. Determinants of land use changes: a spatial multinomial probit approach. Pap. Reg. Sci. 88 (2), 327–344.
Conley, T.G., Topa, G., 2002. Socio-economic distance and spatial patterns in unemployment. J. Appl. Econom. 17 (4), 303–327. doi:10.1002/jae.670.
Corrado, L., Fingleton, B., 2011. Where is the economics in spatial econometrics. J. Reg. Sci. 51 (2), 1–30.
Dugundji, E., 2013. Socio-Dynamic Discrete Choice: Theory And Application. Doctoral Thesis from University of Amsterdam.
Elhorst, J.P., 2010. Applied spatial econometrics: raising the bar. Spat. Econ. Anal. 5 (1), 10–28.
Guo, J.Y., Bhat, C.R., 2007. Operationalizing the concept of neighborhood: application to residential location choice analysis. J. Transp. Geogr. 15 (1), 31–45.
Guo, J.Y., Bhat, C.R., 2004. Modifiable areal units: problem or perception in modeling of residential location choice. Transp. Res. Rec. 1898, 138–147.
Han, X., 2014. Three essays on spatial econometrics: Specification, estimation and model selection for spatial models. The Ohio State University Doctoral dissertationRetrieved from ProQuest Dissertations & Theses database. (UMI No. 3671367).
Hayter, R., 1997. The Dynamics of Industrial Location : the Factory, the Firm and the Production System. John Wiley & Sons Ltd, Chichester, UK.
Holguin-Veras, J., Xu, N., Levinson, H., Paaswell, R., McKnight, C., Weiner, R., Ozmen-Ertekin, D., 2005. An investigation on the aggregate behavior of firm relocations to New Jersey (1990–1999) and the underlying market elasticities. Netw. Spat. Econ. 5 (3), 293–331. doi:10.1007/s11067-005-3037-z.
Kelejian, H.H., Piras, G., 2012. Estimation of spatial models with endogenous weighting matrices and an application to a demand model for cigarettes. Paper presented at the 59th North American meetings of the RSAI.
Kelejian, H.H., Prucha, I.R., 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. J. Real Estate Finance Econ. 17 (1), 99–121. doi:10.1023/a:1007707430416.
Krauth, B., 2006. Social interactions in small groups. Can. J. Econ. 39, 414–433.
Lee, L.-F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica 72 (6), 1899–1925. doi:10.2307/3598772.
Lee, L.-f., 2007. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. J. Econom. 137 (2), 489–514.
Lee, L-F, Liu, X, Lin, X., 2010. Specification and estimation of social interaction models with network structure: contextual factors, correlation and fixed effects. Econom. J. 13, 145–176.
Lee, L.-f., Yu, J., 2012. QML estimation of spatial dynamic panel data models with time varying spatial weights matrices. Spat. Econ. Anal. 7 (1), 31–74. doi:10.1080/17421772.2011.647057.
LeSage, J.P., Pace, R.K., 2009. Introduction to Spatial Econometrics. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton.
LeSage, J.P., Pace, R.K., 2008. Spatial econometric modeling of origin-destination flows. J. Reg. Sci. 48 (5), 941–967.
Leenders, R.T.A., 2002. Modeling social influence through network autocorrelation: constructing the weight matrix. Soc. Netw. 24 (1), 21–47.
Manski, C.F., 1993. Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. 60 (3), 531–542.
Masten, M.A. (2012). Random coefficients on endogenous variables in simultaneous equations models. [Job market paper].
Mejia-Dorantes, L., Paez, A., Vassallo, J.M., 2012. Transportation infrastructure impacts on firm location: the effect of a new metro line in the suburbs of Madrid. J. Transp. Geogr. 22, 236–250.
Ni, L., Wang, X., Zhang, D., 2016. Impacts of information technology and urbanization on less-than-truckload freight flows in China: an analysis considering spatial effects. Transp. Res. Part A 92, 12–25.
Pinkse, J., Slade, M.E., 2010. The Future of Spatial Econometrics. J. Reg. Sci. 50 (1), 103–117.
Sidharthan, R., Bhat, C.R., 2012. Incorporating spatial dynamics and temporal dependency in land use change models. Geogr. Anal. 44 (4), 321–349.
Smith, T.E., LeSage, J.P., 2004. A Bayesian probit model with spatial dependencies. Adv. Econom. 18, 127–160.
Soetevent, A, Kooreman, P, 2007. A discrete-choice model with social interactions: with an application to high school teen behavior. J. Appl. Econom. 22, 599–624.
Wang, X., Zhou, Y., 2015. Deliveries to residential units: a rising form of freight transportation in the US. Transp. Res. Part C 58, 46–55.
Zhang, D., Wang, X., 2016a. Investigating the dynamic spillover effects of low-cost airlines on airport airfare through spatio-temporal regression models. Netw. Spat. Econ. 16 (3), 821–836.
Zhang, D., Wang, X., 2016b. Analyses considering partner selection and joint decision making: investigation of freight demand with spatial matching models. In: Transportation Research Board 95th Annual Meeting, pp. 16–6449.
Zhou, Y., Wang, X., 2014a. Decision-making process for developing urban freight consolidation centers: analysis with experimental economics. J. Transp. Eng. 140 (2) 04013003. doi:10.1061/(ASCE)TE.1943-5436.0000632.
Zhou, Y., Wang, X., 2014b. Explore the relationship between online shopping and shopping trips: an analysis with the 2009 NHTS data. Transp. Res. Part A 70, 1–9.
Zhou, Y., 2015. Discrete Choice Modeling with Interdependencies: A Spatial Binary Probit Model with Endogenous Weight Matrix. Rensselaer Polytechnic Institute Doctoral dissertation.
Zou, W., Wang, X., Zhang, D., 2015. Truck accident severity in new york city: an investigation of the spatiotemporal effects and vehicle weight. In: Transportation Research Board 94th Annual Meeting, pp. 15–4249.