Statistics and Its Interface Volume 13 (2020) 261–270

A spatial autoregression model with time-varying coefficients

Ke Xu*, Luping Sun†, Jin Liu*,
Xuening Zhu‡, and Hansheng Wang*

This article proposes a spatial autoregression (SAR) model with time-varying coefficients. The model incorporates both spatial dependence and the impacts of explanatory variables, and all the coefficients are allowed to flexibly vary according to time. This article further develops a kernel-smoothed estimator (KSE) to estimate the time-varying coefficients. Compared with the maximum likelihood estimator (MLE) obtained at discrete time points, the KSE utilizes the potentially useful information from time neighborhoods. We have theoretically proved the consistency of the proposed KSE. A number of simulation studies show that the KSE is more accurate and performs substantially better than the MLE. Moreover, a real data analysis for a ride-hailing platform in China also shows that the KSE is more stable and interpretable. The proposed model as well as the KSE can be widely applied to data with a large number of cross-sectional units and regularly spaced time points.

Keywords and phrases: Time-varying coefficients, Spatial autoregression model, Kernel-smoothed estimator, Maximum likelihood estimator.

1. INTRODUCTION

Online ride-hailing has become a multi-billion business in the U.S. as well as China. Well-known ride-hailing service providers include Lyft (www.lyft.com) and Uber (www.uber.com) in the U.S., and Didi Chuxing (www.didichuxing.com) and Yidao (www.yongche.com) in China. The ride-hailing service provider serves as a matchmaker between passengers and drivers. It profits mainly by charging a commission on the transactions made on its application-based ride-hailing platform. As a result, it is highly important for ride-hailing platforms to optimize their matchmaking service and raise the number of rides. To this end, ride-hailing platforms need to forecast and monitor the number of rides in different locations across the city in real time. Therefore, they have opportunities to efficiently allocate drivers (i.e., supply) to satisfy more rides (i.e., demand). This research intends to model the number of rides in different locations across time and provides guidance for ride-hailing platforms to improve their business.

However, modeling the number of rides in a particular location at a particular time point can be challenging because the influential factors could be multifold. The number of rides at particular time may be affected by both the observed cross-sectional variables and the unobserved spatial dependence. First, at a given time point, the number of rides might depend on the nearby ride-hailing supply and demand intensities, the economic incentives provided, and the geographic distance of the ride-hailing requests (i.e., observed cross-sectional variables). For example, a relatively low supply may lead to a small number of rides. Locations that have more requests with higher economic incentives (e.g., a 20 RMB bonus) may be associated with a larger number of rides. With regard to geographical distance, the number of rides is typically larger for requests with a larger distance (e.g., drivers may prefer a 10 km ride rather than a 2 km ride). Second, there may exist spatial dependence between adjacent regions. If the number of rides is relatively small in one region, the nearby regions are less likely to have a very large number of rides. Moreover, the number of rides changes dynamically over time as the drivers move around to look for potential passengers. This adds another layer of complexity.

To quantify the factors that influence the number of rides in a particular location at particular time, several classical models could be considered. First, a linear regression model can be readily employed to examine the impacts of the observed explanatory variables on the number of rides [8, 17]. But it cannot deal with the potentially important spatial dependence. Second, to capture the spatial dependence from

*Ke Xu is supported by "the Fundamental Research Funds for the Central Universities" in UIBE (No. 19QD22). Ke Xu, Jin Liu, and Hansheng Wang are supported by National Natural Science Foundation of China (Grant No. 11525101, No. 71332006, and No. 71532001), and China's National Key Research Special Program (No. 2016YFC0207704).
†Luping Sun is supported by National Natural Science Foundation of China (Grant No. 71972195 and No. 71502182) and Program for Innovation Research in Central University of Finance and Economics.
‡Xuening Zhu (Corresponding author) is supported by the National Natural Science Foundation of China (nos. 11901105, 71991472, 11971504, U1811461), the Shanghai Sailing Program for Youth Science and Technology Excellence (19YF1402700), and the Fudan-Xinzailing Joint Research Centre for Big Data, School of Data Science, Fudan University.

adjacent regions, a standard spatial autoregression (pure SAR) model can be adopted [1, 3]. However, this model only deals with spatial dependence and does not take the explanatory variables into consideration. In order to capture spatial correlation and the impacts of explanatory variables (e.g., the characteristics of rides) simultaneously, a mixed regressive SAR model could be used [2, 16]. One drawback of this model, however, is that it does not allow the coefficients to vary according to time. Consider the empirical context of the ride-hailing service. The impact of some variables (e.g., the economic incentive) on the number of rides may change over time. In addition, the extent of spatial dependence in the number of rides might also vary at different time points. Therefore, allowing the coefficients to be time-varying would make the SAR model more flexible. Unfortunately, the estimation of time-varying coefficients in the SAR model is technically difficult [14].

In this research, we introduce a SAR model with time-varying coefficients. The model allows the coefficients of both spatial dependence and explanatory variables to vary according to time. Moreover, we propose an estimation method to estimate the time-varying coefficients. The proposed method is applicable for data with a particular structure. First, the data should be collected at a large number of sequential time points on a regular basis. Second, the observations collected at each particular time point can be spatially allocated into different regions. Third, the spatial correlation between adjacent regions is important and has to be taken into account. Finally, the dependent variable of interest, as well as the explanatory variables, can be recorded at regional level. To estimate the coefficients, we first obtain a maximum likelihood estimator (MLE) for each time point using the cross-sectional data. Next, for each coefficient we further smooth its MLE estimators over time. This leads to a kernel-smoothed estimator (KSE) with fairly good finite sample performance.

The rest of the article is organized as follows. Section 2 introduces the SAR model with time-varying coefficients. Then, the KSE is proposed and its asymptotic properties are studied. Section 3 reports the numerical studies, including a number of simulation studies and a real data analysis. All technical details are provided in the appendices.

2. METHODOLOGY

2.1 Model and notations

We consider an area with N regions indexed by $i = 1, \cdots, N$. The spatial structure of the regions is captured by an adjacency matrix $A = (a_{i_1 i_2}) \in \mathbb{R}^{N \times N}$, where $a_{i_1 i_2} = 1$ if regions i_1 and i_2 are spatially adjacent to each other, and $a_{i_1 i_2} = 0$ otherwise. Note that the spatial structure is symmetric by nature. As a result, $a_{i_1 i_2} = a_{i_2 i_1}$. Furthermore, we follow the convention to define $a_{ii} = 0$ for any $1 \le i \le N$. In this work, we assume that A is non-stochastic and time-invariant. For region i, there exists a random process of the response variable $\{Y_i(t) \in \mathbb{R}^1, t \in [0,1]\}$ and a random process of explanatory variables $\{X_i(t) = (X_{i1}(t), \cdots, X_{ip}(t))^\top \in \mathbb{R}^p, t \in [0,1]\}$. Note that the explanatory variables $X_i(t)$ are not allowed to be endogenous.

To model the regression relationship between $Y_i(t)$ and $X_i(t)$ with spatial dependence, we propose the following SAR model with time-varying coefficients

$$Y_{i_1}(t) = \rho(t) n_{i_1}^{-1} \sum_{i_2=1}^{N} a_{i_1 i_2} Y_{i_2}(t) + X_{i_1}^\top(t)\beta(t) + \varepsilon_{i_1}(t), \tag{1}$$

where $1 \le i_1 \le N$ and $\varepsilon_{i_1}(t) \in \mathbb{R}^1$ is the error term. We assume that $E(\varepsilon_{i_1}(t)) = 0$ and $\mathrm{var}(\varepsilon_{i_1}(t)) = \sigma^2(t)$. Here, $\beta(t) = (\beta_1(t), \cdots, \beta_p(t))^\top \in \mathbb{R}^p$ includes the parameters associated with the explanatory variables. In the meanwhile, $\rho(t) \in \mathbb{R}^1$ is the time-varying spatial effect [21]. Assume that there exists a positive constant $\lambda \in (0,1)$ such that $\rho(t) \in [-\lambda, \lambda]$ for any $t \in [0,1]$. Note that $n_{i_1} = \sum_{i_2=1}^{N} a_{i_1 i_2}$ is region i_1's total number of spatial neighbors, and $n_{i_1}^{-1} \sum_{i_2=1}^{N} a_{i_1 i_2} Y_{i_2}(t)$ is the average response of region i_1's spatial neighbors. For convenience, we define $\theta(t) = (\rho(t), \beta^\top(t), \sigma^2(t))^\top \in \mathbb{R}^{p+2}$. Assume that $\theta(t)$ is bounded and twice continuously differentiable over $t \in [0,1]$. Further define $\mathbb{Y}(t) = (Y_1(t), Y_2(t), \cdots, Y_N(t))^\top \in \mathbb{R}^N$, $\mathbb{X}(t) = (X_1(t), X_2(t), \cdots, X_N(t))^\top \in \mathbb{R}^{N \times p}$, and $\mathcal{E}(t) = (\varepsilon_1(t), \varepsilon_2(t), \cdots, \varepsilon_N(t))^\top \in \mathbb{R}^N$. Model (1) can be rewritten in a vector form as

$$\mathbb{Y}(t) = \rho(t) W \mathbb{Y}(t) + \mathbb{X}(t)\beta(t) + \mathcal{E}(t), \tag{2}$$

where $W = (w_{i_1 i_2}) = \mathrm{diag}\{n_1^{-1}, \cdots, n_N^{-1}\} A \in \mathbb{R}^{N \times N}$ with $w_{i_1 i_2} = n_{i_1}^{-1} a_{i_1 i_2}$. Hence, W is the row-normalized spatial weight matrix [15]. In real practice, there might be different choices of W. In this research, we choose W in a similar fashion with prior literature [4, 5, 11]. Given model (2), we then discuss the estimation methods in the following section.

2.2 The estimation methods

Inspired by the empirical scenario, we assume the data are collected at a large number of time points between 0 and 1. These discrete time points are uniformly distributed over $[0,1]$ and indexed by $\mathcal{T} = \{t_k \in [0,1] : 1 \le k \le T\}$. At any time point t_k, we observe both $\mathbb{Y}(t_k)$ and $\mathbb{X}(t_k)$. This leads to

$$\mathbb{Y}(t_k) = \rho(t_k) W \mathbb{Y}(t_k) + \mathbb{X}(t_k)\beta(t_k) + \mathcal{E}(t_k). \tag{3}$$

As one can see, model (3) is a standard SAR model with explanatory variables [11]. By temporarily assuming that the error term $\mathcal{E}(t_k)$ follows a multivariate normal distribution, $\theta(t_k)$ can be estimated by the standard method of maximum likelihood estimation. This leads to the MLE at time point

t_k. Specifically, we write the log-likelihood function of model (3) at time point t_k as

$$\ell\{\theta(t_k)\} = \log|\mathbb{S}_k| - \frac{N}{2}\log\sigma^2(t_k) - \frac{1}{2\sigma^2(t_k)} \left\|\mathbb{S}_k\mathbb{Y}(t_k) - \mathbb{X}(t_k)\beta(t_k)\right\|^2, \tag{4}$$

where $\mathbb{S}_k = I - \rho(t_k)W$, $|\mathbb{S}_k|$ denotes the determinant of matrix $\mathbb{S}_k$, and $\|v\| = \sqrt{v^\top v}$ for an arbitrary vector v. Note that we omit the constants in (4) and those in the following log-likelihood functions for convenience. To optimize (4), we first fix $\rho(t_k)$ and $\beta(t_k)$, and then maximize (4) with respect to $\sigma^2(t_k)$. This leads to the following estimator

$$\widetilde{\sigma}^2(t_k) = \frac{1}{N}\left\|\mathbb{S}_k\mathbb{Y}(t_k) - \mathbb{X}(t_k)\beta(t_k)\right\|^2. \tag{5}$$

Next, apply (5) back to (4), and we obtain the profiled log-likelihood function as

$$\ell\{\rho(t_k), \beta(t_k)\} = \log|\mathbb{S}_k| - \frac{N}{2}\log\left\|\mathbb{S}_k\mathbb{Y}(t_k) - \mathbb{X}(t_k)\beta(t_k)\right\|^2. \tag{6}$$

Given $\rho(t_k)$, we further maximize (6) with respect to $\beta(t_k)$. This leads to

$$\widetilde{\beta}(t_k) = \{\mathbb{X}^\top(t_k)\mathbb{X}(t_k)\}^{-1}\{\mathbb{X}^\top(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}. \tag{7}$$

Lastly, apply (7) back to (6), and we get the final profiled log-likelihood function

$$\ell\{\rho(t_k)\} = \log|\mathbb{S}_k| - \frac{N}{2}\log\left\|\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\right\|^2, \tag{8}$$

where $\mathbb{Q}(t_k) = I - \mathbb{X}(t_k)\{\mathbb{X}^\top(t_k)\mathbb{X}(t_k)\}^{-1}\mathbb{X}^\top(t_k)$. We then maximize (8) by the Newton-Raphson method. This leads to $\widehat{\rho}_{\text{MLE}}(t_k) = \text{argmax}_{\rho(t_k)}\ell\{\rho(t_k)\}$. The computational details are provided in Appendix A. Lastly, replacing $\rho(t_k)$ with $\widehat{\rho}_{\text{MLE}}(t_k)$ in (7) allows us to get $\widehat{\beta}_{\text{MLE}}(t_k)$. Further apply $\widehat{\rho}_{\text{MLE}}(t_k)$ and $\widehat{\beta}_{\text{MLE}}(t_k)$ to (5), and we obtain $\widehat{\sigma}^2_{\text{MLE}}(t_k)$. Thus, we have $\widehat{\theta}_{\text{MLE}}(t_k) = (\widehat{\rho}_{\text{MLE}}(t_k), \widehat{\beta}^\top_{\text{MLE}}(t_k), \widehat{\sigma}^2_{\text{MLE}}(t_k))^\top$.

Theoretically, $\widehat{\theta}_{\text{MLE}}(t_k)$ can be proved to be consistent and asymptotically normal under certain conditions [11, 19]. However, $\widehat{\theta}_{\text{MLE}}(t_k)$ only utilizes the information at the given time point t_k. The potentially useful information from the time neighborhoods of t_k has been ignored. In addition, $\widehat{\theta}_{\text{MLE}}(t_k)$ can only be obtained for discrete time points $\mathcal{T} = \{t_k \in [0,1] : 1 \le k \le T\}$. Thus, we further smooth the MLE over time by the method of kernel smoothing. This leads to the KSE for any continuous time point $t \in [0,1]$, which is denoted as $\widehat{\theta}_{\text{KSE}}(t) = (\widehat{\rho}_{\text{KSE}}(t), \widehat{\beta}^\top_{\text{KSE}}(t), \widehat{\sigma}^2_{\text{KSE}}(t))^\top$. Specifically,

$$\widehat{\theta}_{\text{KSE}}(t) = \Big\{\sum_{k=1}^{T} K_h(t_k - t)\Big\}^{-1}\Big\{\sum_{k=1}^{T}\widehat{\theta}_{\text{MLE}}(t_k)K_h(t_k - t)\Big\},$$

where $K_h(\cdot) = K(\cdot/h)/h$, h is the bandwidth, and $K(\cdot)$ is the kernel function, i.e., a symmetric and bounded probability density function. Throughout this article, we assume that $K(\cdot)$ satisfies the Lipschitz condition and has a compact support. We next investigate the asymptotic theory of the proposed estimator (i.e., KSE).

2.3 Asymptotic theory

To study the asymptotic properties of the proposed estimator, we call for a number of technical conditions. The details are given below.

(C1) Assume that $\{\varepsilon_i(t_k), i = 1, \cdots, N, k = 1, \cdots, T\}$ are independently and identically distributed with mean 0 and variance $\sigma^2(t)$. In addition, there exists a positive constant α such that $\sup_{t\in[0,1]}\max_{1\le i\le N}E(|\varepsilon_i(t)|^{4+\alpha}) < \infty$ and $\sup_{t\in[0,1]}\max_{1\le i\le N}E(\|X_i(t)\|^{4+\alpha}) < \infty$. Meanwhile, for $1 \le k \le T$, assume that $\lim_{N\to\infty} N^{-1}\mathbb{X}(t_k)^\top\mathbb{X}(t_k)$ exists and is nonsingular.

(C2) For the spatial weight matrix $W = (w_{i_1 i_2})$, there exists a constant C such that, for any $N > 0$, $\max_{i_2}\sum_{i_1=1}^{N} w_{i_1 i_2} + \max_{i_1}\sum_{i_2=1}^{N} w_{i_1 i_2} < C < \infty$.

(C3) There exists a positive constant $\lambda \in (0,1)$ such that $\rho(t) \in [-\lambda, \lambda]$ for any $t \in [0,1]$. In addition, $\beta(t)$ and $\sigma^2(t)$ are bounded and twice continuously differentiable over $t \in [0,1]$.

(C4) Define $\mathbb{M}(t) = (\mathbb{X}(t), W\mathbb{S}^{-1}\mathbb{X}(t)\beta(t))$. Assume that $\lim_{N\to\infty} N^{-1}\mathbb{M}(t)^\top\mathbb{M}(t)$ exists and is positive definite for any $t \in [0,1]$.

These conditions have been widely used in prior literature [6, 7, 10, 11, 13, 19, 22, 23, 24]. Condition (C1) contains the standard assumptions regarding the error term and explanatory variables. Condition (C2) is the technical assumption about the spatial weight matrix W. Condition (C3) ensures that the parameters are bounded and Condition (C4) is a sufficient condition for identifying $\theta(t)$.

Assume conditions (C1)-(C4) hold. By Theorem 3.2 in [11] and Theorem 4.3 in [19], we have

$$\sqrt{N}\{\widehat{\theta}_{\text{MLE}}(t_k) - \theta(t_k)\} \longrightarrow_d N\Big(0, \Xi(t_k)\Big), \tag{9}$$

as $N \to \infty$, where $\Xi(t_k) = \Sigma^{-1}_{\theta(t_k)}\Omega_{\theta(t_k)}\Sigma^{-1}_{\theta(t_k)}$, $\Sigma_{\theta(t_k)} = -\lim_{N\to\infty}E\{N^{-1}\ddot{\ell}(\theta(t_k))\}$, and $\Omega_{\theta(t_k)} = \lim_{N\to\infty}E\{N^{-1}\dot{\ell}(\theta(t_k))\dot{\ell}(\theta(t_k))^\top\}$. Here $\dot{\ell}(\theta(t_k)) \in \mathbb{R}^{p+2}$ and $\ddot{\ell}(\theta(t_k)) \in \mathbb{R}^{(p+2)\times(p+2)}$ are the first and second order derivatives of $\ell(\theta(t_k))$ with respect to $\theta(t_k)$, respectively. The detailed analytical derivations of $\dot{\ell}(\theta(t_k))$ and $\ddot{\ell}(\theta(t_k))$ are given in Appendix B. We next investigate the asymptotic properties of $\widehat{\theta}_{\text{KSE}}(t)$ in the following theorem.

Theorem 1. *Assume conditions (C1)-(C4) hold. Then, for any $t \in (0,1)$, we have*

$$\widehat{\theta}_{\text{KSE}}(t) - \theta(t) \longrightarrow_p 0,$$

as $\min\{N, T\} \to \infty$, $h \to 0$, and $NTh \to \infty$.

The proof of Theorem 1 is given in Appendix C. Theorem 1 suggests that the resulting estimator is consistent. As both N and T go to infinity, their relative divergence rate is unknown. This makes it very difficult to bound the order of the residual term in the Taylor expansion argument while examining the asymptotic distribution. Therefore, in this paper we focus only on the consistency of the proposed estimator.

3. NUMERICAL STUDIES

3.1 Simulation models

In this subsection, we demonstrate the finite sample performance of the proposed estimator with simulations. First, we generate a spatial adjacency matrix A as follows. We divide a squared area $[0,1]\times[0,1]$ into $N = m_1 \times m_2$ equally sized rectangle regions. Then we index these regions columnwise from 1 to N. The left panel of Figure 1 presents an illustrating example with $N = 3\times 3$. We define $a_{i_1 i_2} = 1$ if regions i_1 and i_2 share one common edge, and $a_{i_1 i_2} = 0$ otherwise. For example, we have $a_{12} = 1$ and $a_{13} = a_{15} = 0$ for the example in Figure 1.

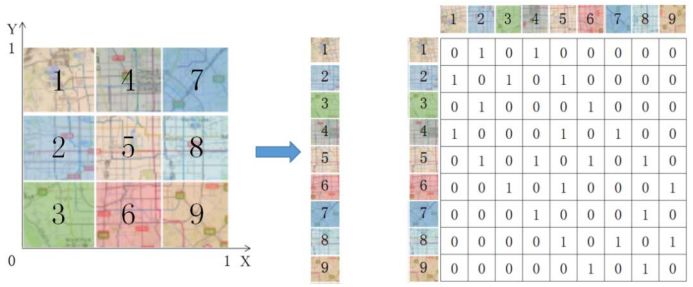


Figure 1. An illustration of how we get the adjacency matrix. In the left panel, the area $[0,1]\times[0,1]$ is divided into $3\times 3 = 9$ equally sized rectangle regions. The right panel shows the corresponding adjacency matrix for these nine regions. We define two regions to be adjacent to each other if they share one common edge.

Next, we assume that the process is observed at a set of regularly spaced time points $\mathcal{T} = \{t_k = k/T : k = 1, 2, \cdots, T\}$. For any $t \in \mathcal{T}$ and region i, the explanatory variables $X_i(t) \in \mathbb{R}^4$ are independently generated according to a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{i_1 i_2})$ with $\sigma_{i_1 i_2} = 0.5^{|i_1 - i_2|}$. The random noise $\varepsilon_i(t) = \sigma^2(t)E$, where E is independently sampled from a standardized exponential distribution and $\sigma^2(t)$ is one of the parameters that will be specified below. Note that the random noise is temporarily assumed to follow a normal distribution in prior sections. Our simulation results, however, are not sensitive to the distribution of the random noise. In the following simulation study, we only show the results obtained with the random noise following the exponential distribution. Specifically, we consider the following three examples.

Example 1 (Time-Varying Coefficients)**.** We first consider a case where all the parameters are time-varying, as in our model (2). Specifically, we follow the simulation setting of [20] to set $\rho(t) = 0.3\sin(t\pi)$, $\sigma^2(t) = \sin\{\pi(2t+1)/4\} + 1$, $\beta_1(t) = \sin(t\pi)$, $\beta_2(t) = \cos(t\pi)$, $\beta_3(t) = e^t$, and $\beta_4(t) = e^t(1+e^t)^{-1}$.

Example 2 (Time-Invariant Coefficients)**.** Next, we consider a case that all the parameters are constants [14]. Specifically, we fix $\rho(t) = 0.3$, $\sigma^2(t) = 1$, and $\beta_1(t) = \beta_2(t) = \beta_3(t) = \beta_4(t) = 1$.

Example 3 (A Standard SAR model)**.** Lastly, we consider a standard SAR model with no explanatory variables [11]. In the meanwhile, all the other parameters are left to be time-varying as in Example 1. Specifically, the parameters are given by $\rho(t) = 0.3\sin(t\pi)$, $\sigma^2(t) = \sin\{\pi(2t+1)/4\} + 1$, and $\beta_1(t) = \beta_2(t) = \beta_3(t) = \beta_4(t) = 0$.

Given the adjacency matrix, explanatory variables, random noise, and parameter specification, we can generate the response variable according to model (2).

3.2 Performance measurements and simulation results

For each simulation example, we consider different (N, T) combinations, where $N \in \{50, 100, 400\}$ and $T \in \{100, 200, 400, 1000, 2000\}$. Moreover, for each (N, T) combination, we randomly replicate the experiment for a total of $R = 500$ times. Let $\widehat{\rho}_d^{(r)}(t_k)$, $\widehat{\beta}_d^{(r)}(t_k)$, and $\widehat{\sigma}_d^{2(r)}(t_k)$ be the estimators obtained in the r-th replication, where $d \in \{\mathrm{MLE}, \mathrm{KSE}\}$ and $1 \le r \le R$. To obtain the KSE, we use a Epanechnikov kernel with bandwidth $h = 2.34(NT)^{-1/5}$ [18]. To gauge the finite sample performance of the estimators, we define the root mean squared error (RMSE) for different parameters as

$$\mathrm{RMSE}_\rho = \Big[(RT)^{-1}\sum_{r=1}^{R}\sum_{k=1}^{T}\{\widehat{\rho}^{(r)}(t_k) - \rho(t_k)\}^2\Big]^{1/2},$$

$$\mathrm{RMSE}_{\beta_j} = \Big[(RT)^{-1}\sum_{r=1}^{R}\sum_{k=1}^{T}\{\widehat{\beta}_j^{(r)}(t_k) - \beta_j(t_k)\}^2\Big]^{1/2},$$

$$\mathrm{RMSE}_{\sigma^2} = \Big[(RT)^{-1}\sum_{r=1}^{R}\sum_{k=1}^{T}\{\widehat{\sigma}^{2(r)}(t_k) - \sigma^2(t_k)\}^2\Big]^{1/2},$$

where $j = 1, \cdots, 4$. It is worth noting that the performance of MLE is also evaluated. The MLE and KSE may not be directly comparable since KSE utilizes more information than MLE. Thus, it is not surprising that the KSE may perform better than the MLE. But what remains unclear is that how much advantage the KSE can obtain by utilizing additional information. The detailed simulation results are given in Tables 1–3.

Consider the performance of KSE and MLE for Example 1; see Table 1. Take ρ as an example. First, we study the case that T is fixed and N goes to infinity. Let $T = 400$ and

Table 1. Simulation results for the parameters specified in Example 1. For each parameter, we report the RMSE of both KSE and MLE for different (N,T) combinations. When T is fixed and N goes to infinity, the RMSE of both MLE and KSE declines. When N is fixed and T goes to infinity, however, only the RMSE of KSE declines and that of MLE remains unchanged. Moreover, the RMSE of KSE is much smaller than that of MLE for each (N,T) combination

N	T	ρ		β_1		β_2		β_3		β_4		σ^2	
		MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE
50	100	0.1129	0.0320	0.2367	0.0702	0.2671	0.0802	0.2665	0.0789	0.2370	0.0679	0.7133	0.2707
	200	0.1164	0.0261	0.2349	0.0541	0.2665	0.0569	0.2678	0.0588	0.2389	0.0521	0.7137	0.2387
	400	0.1154	0.0201	0.2339	0.0401	0.2640	0.0450	0.2647	0.0431	0.2363	0.0385	0.7173	0.2269
	1000	0.1164	0.0145	0.2371	0.0282	0.2659	0.0307	0.2677	0.0317	0.2381	0.0263	0.7356	0.2048
	2000	0.1168	0.0117	0.2379	0.0220	0.2654	0.0224	0.2663	0.0240	0.2374	0.0205	0.7333	0.2017
100	100	0.0798	0.0276	0.1619	0.0559	0.1815	0.0623	0.1767	0.0626	0.1595	0.0538	0.5169	0.2045
	200	0.0810	0.0210	0.1620	0.0432	0.1798	0.0470	0.1820	0.0464	0.1630	0.0428	0.5203	0.1593
	400	0.0807	0.0159	0.1626	0.0313	0.1829	0.0353	0.1824	0.0365	0.1628	0.0333	0.5246	0.1371
	1000	0.0806	0.0113	0.1633	0.0220	0.1826	0.0240	0.1833	0.0248	0.1628	0.0221	0.5172	0.1189
	2000	0.0805	0.0088	0.1634	0.0172	0.1821	0.0190	0.1832	0.0189	0.1631	0.0167	0.5176	0.1116
400	100	0.0413	0.0198	0.0794	0.0375	0.0894	0.0412	0.0900	0.0433	0.0793	0.0383	0.2640	0.1281
	200	0.0416	0.0151	0.0791	0.0292	0.0889	0.0320	0.0895	0.0339	0.0800	0.0296	0.2666	0.0989
	400	0.0416	0.0114	0.0798	0.0221	0.0891	0.0244	0.0890	0.0247	0.0796	0.0223	0.2646	0.0765
	1000	0.0414	0.0079	0.0800	0.0152	0.0891	0.0173	0.0891	0.0169	0.0802	0.0153	0.2663	0.0551
	2000	0.0411	0.0059	0.0800	0.0115	0.0889	0.0126	0.0888	0.0128	0.0795	0.0116	0.2652	0.0455

Table 2. Simulation results for the parameters specified in Example 2. When T is fixed and N goes to infinity, the RMSE of both MLE and KSE decreases. When N is fixed and T goes to infinity, only the RMSE of KSE declines while that of MLE remains almost the same. Moreover, the RMSE of KSE is much smaller than that of MLE for all (N,T) combinations

N	T	ρ		β_1		β_2		β_3		β_4		σ^2	
		MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE
50	100	0.2021	0.0616	0.1888	0.0544	0.2117	0.0606	0.2123	0.0618	0.1883	0.0536	2.4260	0.8428
	200	0.2114	0.0488	0.1911	0.0425	0.2139	0.0472	0.2128	0.0469	0.1907	0.0417	2.5679	0.6713
	400	0.2125	0.0387	0.1921	0.0323	0.2109	0.0351	0.2143	0.0357	0.1921	0.0323	2.7144	0.5536
	1000	0.2074	0.0270	0.1907	0.0217	0.2119	0.0243	0.2124	0.0243	0.1908	0.0224	2.5147	0.3632
	2000	0.2122	0.0221	0.1914	0.0166	0.2132	0.0186	0.2124	0.0185	0.1912	0.0165	2.6159	0.2851
100	100	0.0630	0.0227	0.1188	0.0402	0.1333	0.0449	0.1337	0.0453	0.1201	0.0409	0.3819	0.1736
	200	0.0669	0.0186	0.1200	0.0309	0.1349	0.0348	0.1338	0.0344	0.1201	0.0311	0.4495	0.1728
	400	0.0633	0.0136	0.1197	0.0234	0.1336	0.0259	0.1330	0.0263	0.1194	0.0232	0.3924	0.1220
	1000	0.0648	0.0096	0.1196	0.0162	0.1338	0.0180	0.1337	0.0180	0.1195	0.0163	0.4116	0.0953
	2000	0.0661	0.0078	0.1197	0.0124	0.1339	0.0137	0.1338	0.0137	0.1198	0.0122	0.4415	0.0862
400	100	0.0284	0.0134	0.0581	0.0275	0.0649	0.0305	0.0652	0.0310	0.0582	0.0274	0.1406	0.0678
	200	0.0283	0.0102	0.0583	0.0210	0.0651	0.0235	0.0648	0.0234	0.0582	0.0209	0.1403	0.0522
	400	0.0284	0.0078	0.0581	0.0160	0.0648	0.0178	0.0649	0.0178	0.0580	0.0159	0.1404	0.0405
	1000	0.0284	0.0054	0.0582	0.0110	0.0649	0.0123	0.0651	0.0124	0.0583	0.0111	0.1404	0.0294
	2000	0.0283	0.0041	0.0582	0.0083	0.0650	0.0093	0.0650	0.0093	0.0581	0.0083	0.1405	0.0238

N increase from 50 to 400. We find that the RMSE of the MLE decreases from 0.1154 to 0.0416. In the meanwhile, the RMSE of the KSE drops from 0.0201 to 0.0114. This suggests that both estimators are consistent as N goes to infinity. Next, we consider the case that N is fixed and T goes to infinity. Let $N = 400$ and T increase from 100 to 2000. We find that the RMSE of the MLE fluctuates between 0.0411 and 0.0416. This is expected because the consistency of the MLE is only driven by N, and thus the RMSE of the MLE cannot converge to 0 with a fixed N. The RMSE of the KSE, however, declines from 0.0198 to 0.0059, suggesting that the KSE is still consistent as T goes to infinity.

Moreover, for each (N,T) combination in Example 1, the RMSE of KSE is much smaller than that of MLE. For example, the RMSE of KSE for $(N,T) = (50, 2000)$ is 0.0117, which is much smaller than that of MLE (i.e., 0.1168). We observe similar patterns for the estimation results of the other parameters. As a result, the overall performance of the KSE is substantially better than that of the MLE for Example 1. Qualitatively similar findings are observed for Examples 2–Examples 3 (see Tables 2–3), which will not be discussed in detail.

To get a more intuitive understanding, we fix $(N,T) = (400, 1000)$ for Example 1. Next, we plot the MLE and KSE

Table 3. Simulation results for the parameters specified in Example 3. When T is fixed and N goes to infinity, the RMSE of both MLE and KSE decreases. When N is fixed and T goes to infinity, only the RMSE of KSE declines while that of MLE remains almost at the same level. Moreover, the RMSE of KSE is much smaller than that of MLE for all (N, T) combinations

N	T	ρ		β_1		β_2		β_3		β_4		σ^2	
		MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE	MLE	KSE
50	100	0.1746	0.0525	0.2354	0.0685	0.2646	0.0785	0.2630	0.0764	0.2355	0.0677	0.7181	0.2758
	200	0.1766	0.0398	0.2326	0.0514	0.2638	0.0561	0.2633	0.0567	0.2365	0.0511	0.7159	0.2430
	400	0.1737	0.0282	0.2315	0.0383	0.2615	0.0443	0.2615	0.0422	0.2344	0.0384	0.7083	0.2298
	1000	0.1769	0.0209	0.2346	0.0272	0.2632	0.0299	0.2636	0.0298	0.2355	0.0260	0.7162	0.2081
	2000	0.1767	0.0165	0.2354	0.0209	0.2629	0.0218	0.2624	0.0226	0.2350	0.0200	0.7160	0.2081
100	100	0.1250	0.0423	0.1614	0.0551	0.1813	0.0622	0.1762	0.0624	0.1588	0.0536	0.5188	0.2059
	200	0.1252	0.0330	0.1614	0.0428	0.1788	0.0467	0.1809	0.0459	0.1625	0.0425	0.5225	0.1617
	400	0.1252	0.0242	0.1619	0.0305	0.1821	0.0350	0.1817	0.0360	0.1621	0.0331	0.5263	0.1391
	1000	0.1254	0.0179	0.1625	0.0215	0.1819	0.0239	0.1819	0.0240	0.1622	0.0220	0.5193	0.1216
	2000	0.1252	0.0138	0.1626	0.0166	0.1814	0.0188	0.1820	0.0184	0.1623	0.0166	0.5196	0.1142
400	100	0.0642	0.0308	0.0793	0.0372	0.0893	0.0411	0.0896	0.0431	0.0792	0.0383	0.2642	0.1281
	200	0.0654	0.0237	0.0790	0.0290	0.0887	0.0320	0.0891	0.0337	0.0800	0.0296	0.2671	0.0994
	400	0.0654	0.0180	0.0797	0.0220	0.0889	0.0243	0.0887	0.0246	0.0795	0.0223	0.2650	0.0770
	1000	0.0654	0.0126	0.0800	0.0151	0.0890	0.0172	0.0888	0.0169	0.0801	0.0153	0.2666	0.0555
	2000	0.0652	0.0096	0.0798	0.0114	0.0888	0.0126	0.0886	0.0127	0.0794	0.0116	0.2655	0.0462

from one arbitrarily selected random replication in Figure 2. Take $\rho(t)$ as an example; see Figure 2(a). We find that the KSE of $\rho(t)$ matches the true parameter very well over time. However, the MLE is highly unstable. This further confirms that the KSE is much more accurate than the MLE. We find qualitatively similar results for other parameters; see Figures 2(b)–2(f).

3.3 Real data analysis

In this subsection, we apply our model to a real dataset provided by a ride-hailing service provider in Beijing. For illustration purpose, we divide the Beijing urban area into $8 \times 8 = 64$ equally sized regions in the same way as Figure 1. The regions are indexed by $i = 1, \cdots, 64$. Then, the adjacency matrix $A \in \mathbb{R}^{64\times 64}$ and the row-normalized spatial weight matrix $W = (w_{i_1 i_2}) = \text{diag}\{n_1^{-1}, \cdots, n_N^{-1}\}A$ can be readily obtained.

We collected region-level data every five minutes from 6:00 to 24:00 of the day. So each time point corresponds to a five-minute time interval. This leads to a total of $T = 216$ discrete time points. Here we choose T subjectively and how to select the optimal T by data-driven methods would be an interesting question for future study.

For each region i at time point t_k, we recorded the number of rides achieved. Note that the number of rides can only be nonnegative integers and its distribution can be heavily skewed; see Figure 3(a). Therefore, we transform the number of rides with $\log(1 + x)$ and standardization, which leads to our response variable $Y_i(t_k)$. The distribution of $Y_i(t_k)$ is shown in Figure 3(b), which is approximately symmetric.

For each ride, the platform automatically records its departure and destination locations. Therefore, we can calculate the geographical distance of each ride. The distance can be averaged over all rides in region i at time point t_k. This leads to our first explanatory variable – Average Distance, denoted as $X_{i1}(t_k)$. For each ride, the platform may offer a subsidy. It can be considered as an economic incentive to solicit drivers to respond to the ride request. Again, we calculate the average economic incentive over all the rides in region i at time t_k. This leads to our second explanatory variable – Economic Incentive, denoted as $X_{i2}(t_k)$. The number of rides in a region can also be influenced by the local demand. It is measured by the overall number of available passengers in region i at time t_k. This leads to our third explanatory variable – Local Demand, denoted as $X_{i3}(t_k)$. Lastly, the number of rides in a region is also influenced by the local supply. It is measured by the overall number of available drivers in region i at t_k. We use $X_{i4}(t_k)$ to represent Local Supply. Together, these four explanatory variables are denoted as $X_i(t_k) = (X_{i1}(t_k), X_{i2}(t_k), X_{i3}(t_k), X_{i4}(t_k))^\top$.

To obtain the KSE, we use a Epanechnikov kernel with bandwidth $h = 2.34(NT)^{-1/5}$. The estimation results are shown in Figure 4. For comparison purpose, both the KSE (i.e., the solid line) and the MLE (i.e., the dotted line) are presented. Similar to our simulation results, we find that the MLE is highly unstable and has many estimation results that can hardly be correct. For example, the MLE of $\beta_2(t)$ (i.e., the effect of Economic Incentive) is negative at some time points; see Figure 4(c). A negative estimate of $\beta_2(t)$ implies a negative correlation between the Economic Incentive and the number of rides in a particular region. This is rather counterintuitive and can hardly be correct. The KSE of $\beta_2(t)$, however, is fairly stable and is positive over all the time periods, which is quite consistent with our empirical experience. The findings of the other parameters are similar; see Figures 4(a), 4(b), 4(d), 4(e), and 4(f). This

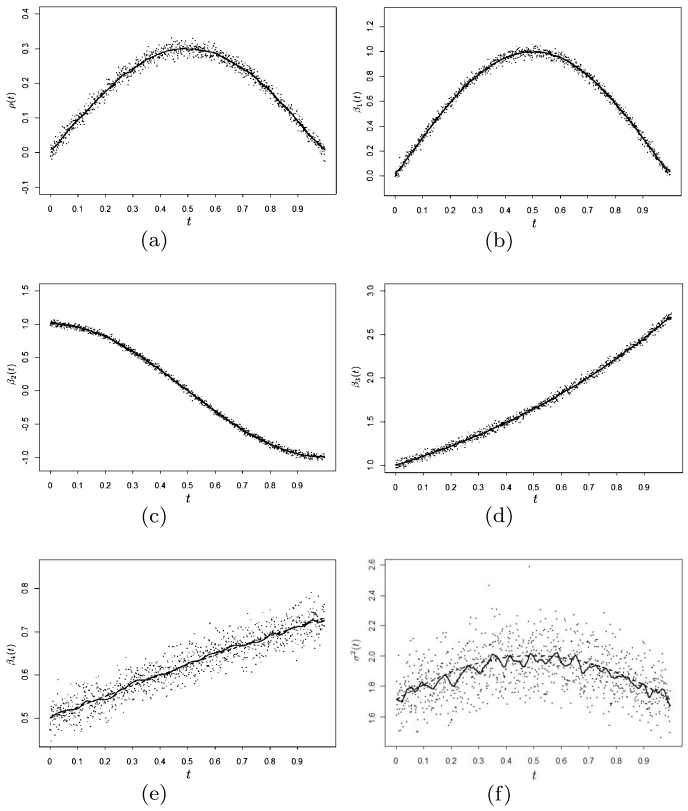


Figure 2. The KSE and MLE of each parameter specified in Example 1. Note that this figure shows the estimators from an arbitrarily selected replication for $(N, T) = (400, 1000)$. The dotted line represents the true parameter over time. The solid line is the KSE, which is relatively stable and coincides with the true parameter very well. The scattered points are the MLEs at discrete time points. Compared with the KSE, the MLE is less stable and may generate highly inaccurate estimates for some parameters (e.g., $\beta_4(t)$).

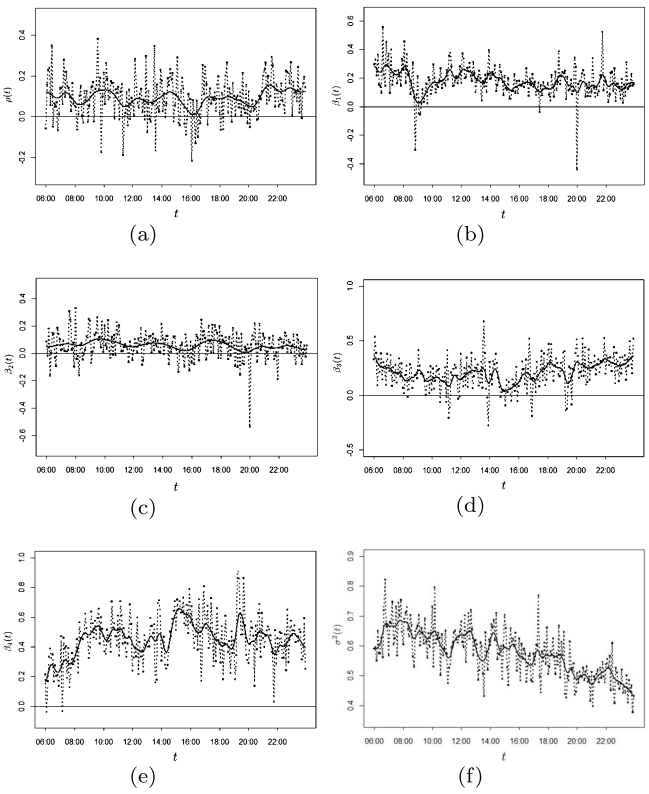
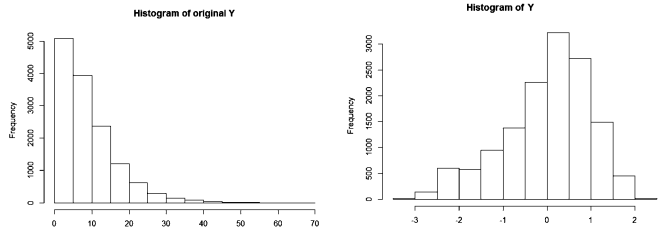


Figure 4. The KSE and MLE of the parameters in real data analysis. The dotted line represents the MLE at discrete time points, while the solid line represents the KSE. Similar with our simulation results, the KSE is more stable than the MLE. Moreover, the KSE is much easier to interpret in the empirical setting. By contrast, the MLE generates some abnormal estimates for almost every parameter (e.g., the negative estimate for $\beta_2(t)$), which can hardly be explained from a practical point of view.



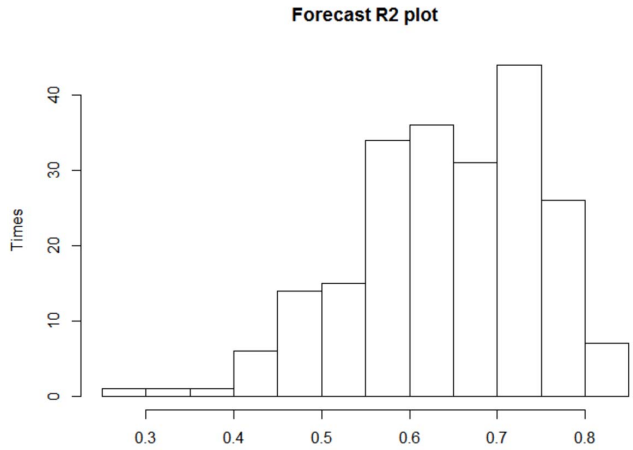
(a) Histogram of the original number of rides.

(b) Histogram of the response variable $Y_i(t_k)$.

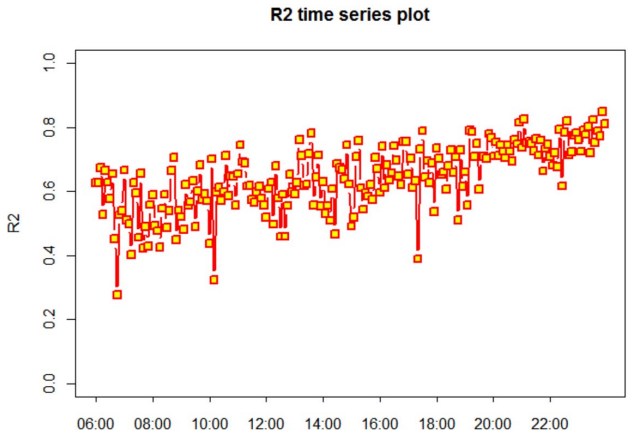
Figure 3. The histograms of the original number of rides and the response variable $Y_i(t_k)$. The left-hand panel shows the histogram of the number of rides in different regions, which is heavily right-skewed. The right-hand panel shows the histogram of the response variable $Y_i(t_k)$, which is approximately symmetric.

implies that the KSE exhibits more consistent performance and has much better interpretations than the MLE in the empirical study.

With the estimated parameters, we can use the information of neighbouring regions to forecast the number of rides in any target region i_1 at any particulate time point t (i.e., $\widehat{Y}_{i_1}(t)$). Define $Y_{(-i_1)}(t) = (Y_1(t), \cdots, Y_{i_1-1}(t), Y_{i_1+1}(t), \cdots, Y_N(t))$. Following similar techniques used by [9], one can verify that the conditional expectation of $Y_{i_1}(t)$ is $E\{Y_{i_1}(t)|Y_{(-i_1)}(t), \mathbb{X}(t)\} = \mu_{i_1}(t) + \sum_{i_3 \neq i_1} \alpha_{i_1 i_3}(t)\{Y_{i_3}(t) - \mu_{(-i_1)}(t)\}$. The definitions of the notations in the conditional expectation are given in Appendix D. Replace $\rho(t)$ and $\beta(t)$ in $E\{Y_{i_1}(t)|Y_{(-i_1)}(t), \mathbb{X}(t)\}$ with $\widehat{\rho}_{\mathrm{KSE}}(t)$ and $\widehat{\beta}_{\mathrm{KSE}}(t)$, respectively, and then we can obtain $\widehat{E}\{Y_{i_1}(t)|Y_{(-i_1)}(t), \mathbb{X}(t)\}$. For convenience, we denote $\widehat{Y}_{i_1}(t) = \widehat{E}\{Y_{i_1}(t)|Y_{(-i_1)}(t), \mathbb{X}(t)\}$. Then, the predicted values at t_k are $\widehat{\mathbb{Y}}(t_k) = (\widehat{Y}_1(t_k), \widehat{Y}_2(t_k), \ldots, \widehat{Y}_N(t_k))^\top \in \mathbb{R}^N$.



(a) The histogram of $R^2_{out}(t_k)$



(b) The time series plot of $R^2_{out}(t_k)$

Figure 5. The histogram and time series plot of $R^2_{out}(t_k)$. The left panel shows the histogram of $R^2_{out}(t_k)$, which suggests that $R^2_{out}(t_k)$ is mostly between 60% and 75%. The right panel shows the time series plot of $R^2_{out}(t_k)$, which indicates that the lowest $R^2_{out}(t_k)$ occurred at around 06:50 am, 10:15 am, and 5:25 pm.

To examine the forecasting accuracy, we use the method of Leave-One-Out cross-validation. To this end, we define out-sample R-square at time point t_k as $R^2_{out}(t_k) = 1 - \sum_{i=1}^{N}(\widehat{Y}_i(t_k) - Y_i(t_k))^2 / \sum_{i=1}^{N}(Y_i(t_k) - \bar{Y}(t_k))^2$, where $\bar{Y}(t_k) = \sum_{i=1}^{N} Y_i(t_k)/N$. We summarize $R^2_{out}(t_k)$ with a histogram (see Figure 5(a)) and a time series plot (see Figure 5(b)). Overall, the forecasting performance is fairly good with most $R^2_{out}(t_k)$ in the range between 60% and 75%. However, there are three exceptions with $R^2_{out}(t_k)$ smaller than 40%. These exceptions occurred in business districts at around 06:50 am, 10:15 am, and 5:25 pm.

To gain more insights, we can readily identify the regions with poor forecasting accuracy at these three time points. Remember that we divided the Beijing urban area into 64 regions. We index these regions columnwise, which is shown in Figure 6. Then, for each region, we calculate the relative difference between the true value $Y_i(t_k)$ and the forecasted value $\widehat{Y}_i(t_k)$. Here, for region i, we define $df_i(t_k) = \|\{\widehat{Y}_i(t_k) - Y_i(t_k)\}/Y_i(t_k)\|$. The result shows that No.21, No.39, and No.24 regions have the lowest forecasting accuracy at 06:50 am, 10:15 am, and 5:25 pm, respectively. Though we don't know the specific reason behind the poor forecasting accuracy (due to the limitations of the data), we do believe that some unexpected events may have occurred in these regions, which can provide meaningful information for the online ride-hailing platform to adjust its strategies accordingly.

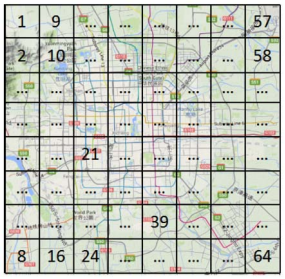


Figure 6. The 64 regions of Beijing urban area.

APPENDIX A. THE DERIVATIVES OF $\ell\{\rho(T_K)\}$ WITH RESPECT TO $\rho(T_K)$

In this appendix, we provide the first and second order derivatives of $\ell\{\rho(t_k)\}$ with respect to $\rho(t_k)$. They are the keys to the Newton-Raphson method.

The first order derivative of $\ell\{\rho(t_k)\}$ with respect to $\rho(t_k)$ is

$$\dot{\ell}\{\rho(t_k)\} = \frac{N\{\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}^\top\{\mathbb{Q}(t_k)W\mathbb{Y}(t_k)\}}{\{\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}^\top\{\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}} - \mathrm{tr}(\mathbb{S}_k{}^{-1}W).$$

The second order derivative of $\ell\{\rho(t_k)\}$ with respect to $\rho(t_k)$ can be expressed as

$$\begin{aligned}\ddot{\ell}\{\rho(t_k)\} =& \frac{2N[\{\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}^\top\mathbb{Q}(t_k)W\mathbb{Y}(t_k)]^2}{[\{\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}^\top\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)]^2} \\ &- \frac{N\{\mathbb{Q}(t_k)W\mathbb{Y}(t_k)\}^\top\mathbb{Q}(t_k)W\mathbb{Y}(t_k)}{\{\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)\}^\top\mathbb{Q}(t_k)\mathbb{S}_k\mathbb{Y}(t_k)} \\ &- \mathrm{tr}(\mathbb{S}_k{}^{-1}W\mathbb{S}_k{}^{-1}W).\end{aligned} \tag{A.1}$$

APPENDIX B. THE DERIVATIVES OF $\ell\{\theta(T_K)\}$ WITH RESPECT TO $\theta(T_K)$

In this appendix, we provide the first and second order derivatives of $\ell\{\theta(t_k)\}$ in (4) with respect to $\theta(t_k)$.

The first-order derivative of $\ell\{\theta(t_k)\}$ with respect to $\theta(t_k)$ is

$$\dot{l}\{\theta(t_k)\} = \frac{1}{\sigma^2(t_k)} \tag{B.1}$$

$$\begin{pmatrix} A \\ \mathbb{X}^\top(t_k)\mathcal{E}(t_k) \\ \{\mathcal{E}^\top(t_k)\mathcal{E}(t_k) - N\sigma^2(t_k)\}\{2\sigma^2(t_k)\}^{-1} \end{pmatrix},$$

where $A = \{\mathbb{Z}_k\mathbb{X}(t_k)\beta(t_k)\}^\top\mathcal{E}(t_k) + \mathcal{E}^\top(t_k)\mathbb{Z}_k\mathcal{E}(t_k) - \sigma^2(t_k)\mathrm{tr}(\mathbb{Z}_k)$, $\mathbb{Z}_k = W\mathbb{S}_k^{-1}$.

The second-order derivative of $\ell\{\theta(t_k)\}$ with respect to $\theta(t_k)$ is $\ddot{l}\{\theta(t_k)\}$

(B.2)
$$= -\frac{1}{\sigma^2(t_k)} \begin{pmatrix} B & \mathbb{X}^\top(t_k)\mathcal{M}_k & \frac{(t_k)\mathcal{M}_k^\top\mathcal{E}(t_k)}{\sigma^2} \\ \mathbb{X}^\top(t_k)\mathcal{M}_k & \mathbb{X}^\top(t_k)\mathbb{X}(t_k) & \sigma^{-2}(t_k)\mathbb{X}^\top(t_k)\mathcal{E}(t_k) \\ \frac{(t_k)\mathcal{M}_k^\top\mathcal{E}(t_k)}{\sigma^2} & \frac{(t_k)\mathbb{X}^\top(t_k)\mathcal{E}(t_k)}{\sigma^2} & C \end{pmatrix},$$

where $B = \sigma^2(t_k)\mathrm{tr}(\mathbb{Z}_k^2) + \mathcal{M}_k^\top\mathcal{M}_k$, $C = \frac{\mathcal{E}^\top(t_k)\mathcal{E}(t_k)}{\sigma^4(t_k)} - \frac{N}{2\sigma^2(t_k)}$, and $\mathcal{M}_k = W\mathbb{Y}(t_k)$.

APPENDIX C. PROOF OF THEOREM 1

Let $\theta(t) = (\rho(t), \beta^\top(t), \sigma^2(t))^\top$ be the vector of true parameters at time t. To examine the asymptotic properties of $\widehat{\theta}_{\mathrm{KSE}}(t)$, we rewrite $\widehat{\theta}_{\mathrm{KSE}}(t) - \theta(t)$ as follows.

$$\begin{aligned}\widehat{\theta}_{\mathrm{KSE}}(t) - \theta(t) =& \frac{1}{T}\Big[\sum_{k=1}^{T}\{\widehat{\theta}_{\mathrm{MLE}}(t_k) - \theta(t)\} \\ & K_h(t_k - t)\Big]\Big\{T^{-1}\sum_{k=1}^{T}K_h(t_k - t)\Big\}^{-1} \\ =& \frac{1}{T}\Big[\sum_{k=1}^{T}\{\widehat{\theta}_{\mathrm{MLE}}(t_k) - \theta(t_k) + \theta(t_k) - \\ & \theta(t)\}K_h(t_k - t)\Big]\Big\{\frac{1}{T}\sum_{k=1}^{T}K_h(t_k - t)\Big\}^{-1} \\ =& \{\widehat{m}_1(t) + \widehat{m}_2(t)\}/\widehat{f}(t),\end{aligned}$$

where $\widehat{m}_1(t) = T^{-1}\sum_{k=1}^{T}\{\widehat{\theta}_{\mathrm{MLE}}(t_k) - \theta(t_k)\}K_h(t_k - t)$, $\widehat{m}_2(t) = T^{-1}\sum_{k=1}^{T}\{\theta(t_k) - \theta(t)\}K_h(t_k - t)$, and $\widehat{f}(t) = T^{-1}\sum_{k=1}^{T}K_h(t_k - t)$. We then study the orders of $\widehat{m}_1(t), \widehat{m}_2(t)$, and $\widehat{f}(t)$, respectively.

First, we examine $\widehat{m}_1(t)$. Define $\mathbf{u}_k = \widehat{\theta}_{\mathrm{MLE}}(t_k) - \theta(t_k)$, then $\widehat{m}_1(t) = \sum_{k=1}^{T}\mathbf{u}_k K_h(t_k - t)/T$, where $\mathbf{u}_k$ can be derived from the Taylor expansion of $\dot{l}\{\widehat{\theta}_{\mathrm{MLE}}(t_k)\}$ at $\theta(t_k)$. For convenience, we denote $\gamma_k = N^{-1}\ddot{\ell}\{\theta(t_k)\}$, $\gamma = E(\gamma_T)$, and $\eta_k = \sqrt{N}^{-1}\dot{\ell}\{\theta(t_k)\}$. Assume conditions (C1)-(C4) hold, then we have (9). It leads to $\mathbf{u}_k = -(N\gamma_k)^{-1}\sqrt{N}\eta_k + O_p(N^{-1})$, which can be verified to be equivalent to $\mathbf{u}_k = -\gamma^{-1}N^{-1/2}\eta_k + R_k$, where R_k is the remainder term that can be ignored. Therefore, we get

$$\widehat{m}_1(t) = \frac{-1}{\sqrt{NTh}}\frac{1}{\sqrt{Th}}\sum_{k=1}^{T}\gamma^{-1}\eta_k K(\frac{t_k - t}{h}). \tag{C.1}$$

According to [12], we can verify that $\sqrt{Th}^{-1}\sum_{k=1}^{T}\gamma^{-1}\eta_k K((t_k - t)/h) = O_p(1)$. Then, (C.1) can be finally expressed as $\widehat{m}_1(t) = O_p(NTh^{-\frac{1}{2}})$.

Second, we study $\widehat{m}_2(t)$. Using the Taylor expansion of $\theta(t_k)$ at t, we get $\theta(t_k) - \theta(t) = \dot{\theta}(t)(t_k - t) + \ddot{\theta}(t)(t_k - t)^2/2 + o((t_k - t)^2)$. Therefore, $\widehat{m}_2(t)$ can be rewritten as

$$\begin{aligned}\widehat{m}_2(t) =& \frac{1}{Th}\sum_{k=1}^{T}\Big\{\dot{\theta}(t)(t_k - t) \\ & + \frac{1}{2}\ddot{\theta}(t)(t_k - t)^2 + o((t_k - t)^2)\Big\}K(\frac{t_k - t}{h}).\end{aligned} \tag{C.2}$$

Assume $t_k - t = hu$, and then as $T \to \infty$, we have

$$\begin{aligned}&\frac{1}{Th}\sum_{k=1}^{T}K(\frac{t_k - t}{h}) \to \int K(u)du = 1, \\ &\frac{1}{Th}\sum_{k=1}^{T}(t_k - t)K(\frac{t_k - t}{h}) \to h\int uK(u)du = 0, \\ &\frac{1}{Th}\sum_{k=1}^{T}(t_k - t)^2K(\frac{t_k - t}{h}) \to h^2\int u^2K(u)du = h^2\mu_2,\end{aligned}$$

where μ_2 is the second moment of the kernel $K(\cdot)$. Then (C.2) can be written as $\widehat{m}_2(t) = \ddot{\theta}(t)h^2\mu_2/2 + o_p(h^2) = O_p(h^2)$.

Third, we study $\widehat{f}(t)$. For any time t, according to [6], we can easily know that $\widehat{f}(t) - 1 = O_p(h^2)$.

For any $t \in (0, 1)$, we finally have $\widehat{\theta}_{KSE}(t) - \theta(t) = \{\widehat{m}_1(x) + \widehat{m}_2(t)\}/\widehat{f}(t) \longrightarrow_p 0$, as $\min\{N, T\} \to \infty$, $h \to 0$, and $NTh \to \infty$. This completes the proof of Theorem 1.

APPENDIX D. DETAILS ABOUT THE CONDITIONAL EXPECTATION OF $Y_I(T)$

Define the distribution of $\mathbb{Y}(t)$ to be $N(\mu(t), \Sigma(t))$, where $\mu(t) \in \mathbb{R}^N$ and $\Sigma(t) \in \mathbb{R}^{N\times N}$. Let $\mathbb{X}_{i_1,\cdot}(t) \in \mathbb{R}^{1\times p}$ be the i_1th row of $\mathbb{X}(t)$ and $\mathbb{X}_{(-i_1),\cdot}(t) \in \mathbb{R}^{(N-1)\times p}$ be $\mathbb{X}(t)$ without the i_1th row. Define $E[\{Y_{i_1}(t), Y_{(-i_1)}^\top(t)\}^\top] = [\mu_{i_1}(t), \mu_{(-i_1)}^\top(t)]^\top$ and $\mathrm{var}[\{Y_{i_1}(t), Y_{(-i_1)}^\top(t)\}^\top] = \Sigma_{i_1} = [\Sigma_{i_1,11}(t), \Sigma_{i_1,12}(t); \Sigma_{i_1,21}(t), \Sigma_{i_1,22}(t)]$, where $\mu_{i_1}(t) \in \mathbb{R}$, $\mu_{(-i_1)}(t) \in \mathbb{R}^{N-1}$, $\Sigma_{i_1,11}(t) \in \mathbb{R}$, $\Sigma_{i_1,12}(t) \in \mathbb{R}^{1\times(N-1)}$, $\Sigma_{i_1,21}(t) \in \mathbb{R}^{(N-1)\times 1}$, and $\Sigma_{i_1,22}(t) \in \mathbb{R}^{(N-1)\times(N-1)}$. Assume $\Sigma_{i_1,11}(t) \neq 0$ and recall $W = (w_{i_1 i_2}) \in \mathbb{R}^{N\times N}$. We can derive that $E\{Y_{i_1}(t)|Y_{(-i_1)}(t), \mathbb{X}(t)\} = \mu_{i_1}(t) + \sum_{i_3\neq i_1}\alpha_{i_1 i_3}(t)\{Y_{i_3}(t) - \mu_{(-i_1)}(t)\}$, where

$$\alpha_{i_1 i_3}(t) = \frac{\rho(t)\{w_{i_1 i_3} + w_{i_3 i_1} - \rho(t)\sum_{i_2}w_{i_2 i_1}w_{i_2 i_3}\}}{1 + \rho^2(t)\sum_{i_2}w_{i_2,i_1}^2},$$

$[\mu_{i_1}(t); \mu_{(-i_1)}(t)] = [\{A\mathbb{X}_{i_1,\cdot}(t) + B\mathbb{X}_{(-i_1),\cdot}(t)\}\beta(t); \{C\mathbb{X}_{i_1,\cdot}(t) + D\mathbb{X}_{(-i_1),\cdot}(t)\}\beta(t)]$, and

$$A = \frac{1}{1-\rho(t)w_{i_1i_1}} - \frac{\rho(t)W_{i_1,(-i_1)}}{1-\rho(t)w_{i_1i_1}}\Big\{I_{N-1} - \rho W_{(-i_1),(-i_1)} - \frac{\rho^2(t)}{1-\rho(t)w_{i_1i_1}}W_{(-i_1),i_1}W_{i_1,(-i_1)}\Big\}^{-1},$$

$$B = \frac{\rho(t)}{1-\rho(t)w_{i_1i_1}}W_{i_1,(-i_1)}\Big\{I_{N-1} - \rho(t)W_{(-i_1),(-i_1)} - \frac{\rho^2(t)}{1-\rho(t)w_{i_1i_1}}W_{(-i_1),i_1}W_{i_1,(-i_1)}\Big\}^{-1},$$

$$C = \frac{\rho(t)}{1-\rho(t)w_{i_1i_1}}\Big\{I_{N-1} - \rho(t)W_{(-i_1),(-i_1)} - \frac{\rho^2(t)}{1-\rho(t)w_{i_1i_1}} \times W_{(-i_1),i_1}W_{i_1,(-i_1)}\Big\}^{-1}W_{(-i_1),i_1},$$

$$D = \Big\{I_{N-1} - \rho(t)W_{(-i_1),(-i_1)} - \frac{\rho^2(t)}{1-\rho(t)w_{i_1i_1}}W_{(-i_1),i_1} \times W_{i_1,(-i_1)}\Big\}^{-1}.$$

In the above equations, $I_{N-1} \in \mathbb{R}^{(N-1)\times(N-1)}$ is the identity matrix, $W_{i_1,(-i_1)} \in \mathbb{R}^{1\times(N-1)}$ is the i_1th row of W (i.e., a vector) with the i_1th element removed, $W_{(-i_1),i_1} \in \mathbb{R}^{(N-1)\times 1}$ is the i_1th column of W (i.e., a vector) with the i_1th element removed, and $W_{(-i_1),(-i_1)} \in \mathbb{R}^{(N-1)\times(N-1)}$ represents W with both the i_1th column and the i_1th row removed.

Received 9 October 2018

REFERENCES

[1] Anselin, L. (2013). *Spatial econometrics: methods and models*, Volume 4. Springer Science & Business Media.

[2] Baltagi, B. H. and L. Liu (2016). Random effects, fixed effects and hausman's test for the generalized mixed regressive spatial autoregressive panel data model. *Econometric Reviews 35*(4), 638–658. MR3464351

[3] Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press. MR3362184

[4] Bell, K. P. and N. E. Bockstael (2000). Applying the generalized-moments estimation approach to spatial problems involving micro-level data. *Review of Economics and Statistics 82*(1), 72–82.

[5] Case, A. C. (1991). Spatial patterns in household demand. *Econometrica: Journal of the Econometric Society 953–965*.

[6] Chen, J., J. Gao, and D. Li (2012). Semiparametric trending panel data models with cross-sectional dependence. *Journal of Econometrics 171*(1), 71–85. MR2970337

[7] Dong, C., J. Gao, and B. Peng (2015). Semiparametric single-index panel data models with cross-sectional dependence. *Journal of Econometrics 188*(1), 301–312. MR3371674

[8] Freedman, D. A. (2009). *Statistical models: theory and practice*. Cambridge University Press. MR2489600

[9] Huang, D., W. Lan, H. H. Zhang, H. Wang, et al. (2019). Least squares estimation of spatial autoregressive models for large-scale social networks. *Electronic Journal of Statistics 13*(1), 1135–1165. MR3935846

[10] Lee, L.-F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric theory 18*(2), 252–277. MR1891824

[11] Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica 72*(6), 1899–1925. MR2095537

[12] Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press. MR2283034

[13] Lin, X. and L.-f. Lee (2010). Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics 157*(1), 34–52. MR2652277

[14] Malikov, E. and Y. Sun (2017). Semiparametric estimation and testing of smooth coefficient spatial autoregressive models. *Journal of Econometrics 199*(1), 12–34. MR3652402

[15] Qu, X. and L.-f. Lee (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics 184*(2), 209–232. MR3290999

[16] Schabenberger, O. and C. A. Gotway (2017). *Statistical methods for spatial data analysis*. CRC Press. MR2134116

[17] Seber, G. A. and A. J. Lee (2012). *Linear regression analysis*, Volume 329. John Wiley & Sons. MR0436482

[18] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge. MR0848134

[19] Su, L. and S. Jin (2010). Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics 157*(1), 18–33. MR2652276

[20] Sun, Y., H. Yan, W. Zhang, Z. Lu, et al. (2014). A semiparametric spatial dynamic model. *The Annals of Statistics 42*(2), 700–727. MR3210984

[21] Wei, C., S. Guo, and S. Zhai (2017). Statistical inference of partially linear varying coefficient spatial autoregressive models. *Economic Modelling 64*, 553–559. MR3809255

[22] Xu, X. and L.-f. Lee (2015). A spatial autoregressive model with a nonlinear transformation of the dependent variable. *Journal of Econometrics 186*(1), 1–18. MR3321522

[23] Xu, X. and L.-f. Lee (2018). Sieve maximum likelihood estimation of the spatial autoregressive tobit model. *Journal of Econometrics 203*(1), 96–112. MR3758330

[24] Yang, K. and L.-f. Lee (2017). Identification and qml estimation of multivariate and simultaneous equations spatial autoregressive models. *Journal of Econometrics 196*(1), 196–214. MR3572822

Ke Xu
School of Statistics
University of International Business and Economics
Beijing
China

Luping Sun
Business School
Central University of Finance and Economics
Beijing
China

Jin Liu
Guanghua School of Management
Peking University
Beijing
China

Xuening Zhu
School of Data Science
Fudan University
Shanghai
China
E-mail address: xueningzhu@fudan.edu.cn

Hansheng Wang
Guanghua School of Management
Peking University
Beijing
China