Received: 4 November 2022 | Accepted: 7 May 2023

DOI: 10.1002/sta4.585

ORIGINAL ARTICLE

WILEY

Spatial dynamic panel models with missing data

Jin Liu[1] | Jing Zhou[2] | Wei Lan[3] | Hansheng Wang[4]

[1]School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, Tianjin, China

[2]Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

[3]The Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China

[4]Guanghua School of Management, Peking University, Beijing, China

Correspondence
Jing Zhou, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China.
Email: jing.zhou@ruc.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 12201316, 72171226, 11971504, 71532001, 11931014, 12171395, 71991472, 12271012, 11831008; Beijing Municipal Social Science Foundation, Grant/Award Number: 19GLC052; Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China, Grant/Award Number: 21XNA027; Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics; Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Grant/Award Number: KLATASDS-MOE-ECNU-KLATASDS2101

Missing data are a common problem that researchers face in practice. In this article, we focus on the missing response problem for a spatial dynamic panel data (SDPD) model, which allows for both spatial and temporal dependencies. A logistic regression with a set of prespecified covariates is used to model the missingness mechanism, which is assumed to be missing at random (MAR). A weighted maximum likelihood estimator (WMLE) is proposed for parameter estimation in the presence of incomplete data. The associated asymptotic properties are investigated. Thereafter, we develop a novel imputation method, which makes use of the information from spatial dependence, temporal dependence and exogenous regression covariates. Lastly, the performance of WMLE and the proposed imputation method are demonstrated by both simulation studies and a real data example.

KEYWORDS
imputation, missing at random, spatial dynamic panel data, weighted maximum likelihood estimator

1 | INTRODUCTION

This research is motivated by a real application. China Engineering Cost Network (CECN, https://www.cecn.org.cn/) is a state-owned institute responsible for collecting price information for various construction materials, for example, price index compiling, references for price setting of materials and salary setting of employee. As a consequence, CECN needs to compile and publish price indices in a monthly manner. To this end, CECN collects price information for various building materials at different locations and time points. However, due to many practical reasons, the collected price information is seldom complete. For example, the missing rate of price information in the proposed CECN dataset is around 25%. This becomes a serious challenge for price index compiling. Therefore, how to handle these incomplete price information becomes a problem of great interest.

It is remarkable that the price information is collected by CECN at regular time points and from a fixed set of locations. This leads to a dataset of spatial panel structure. To conduct appropriate modelling approach for this kind of data, two types of dependencies should be considered. The first type is spatial dependence, which means the price collected from neighbouring locations, should be correlated. The second type is temporal dependence; that is, the current price value should be correlated with the historical ones. Then, how to develop a model allowing for both spatial and temporal dependencies becomes a problem of great importance.

There has been a large body of literature dealing with the above two types of dependencies. For temporal dependence, time series models have been developed (Brockwell & Davis, 1991; Fuller, 1996). For spatial dependence, spatial autoregressive (SAR) models have been widely adopted (Anselin, 1980; Lee et al., 2013; Ord, 1975). To allow for both spatial and temporal dependencies, Yu et al. (2008) proposed a spatial dynamic panel data (SDPD) model and suggested a quasi-maximum likelihood method for estimation. Later on, Lee and Yu (2014) developed a generalized method of moments (GMM) for estimating a SDPD model with multiple spatial lags. Li (2017) further extended the model to allow for multiple spatial time lags. Su and Yang (2015) proposed the Quasi Maximum Likelihood Estimator for dynamic panel models with spatial errors and random/fixed effects. Yang (2018) studied the SDPD models with spatial error. Li and Yang (2021) developed SDPD models with correlated random effects. Feng et al. (2022) considered spatial–temporal model with heterogeneous random effects. More discussions on models with both spatial and temporal dependencies can be found in the above studies and the references therein. It is noteworthy that all these pioneered researches are conducted based on complete datasets. The problem of missing data seems not well studied, and no imputation method has been developed.

However, as we point out in the beginning, missing data are a common issue in practice. This is particularly true for SDPD as we can see in the CECN case. So it is necessary to develop statistical methods to handle this challenge. Typically, there are three types of assumptions about the missingness mechanism (Rubin, 1976). They are *missing completely at random* (MCAR), *missing at random* (MAR) and *nonignorable missing* (NM); see Little and Rubin (2002) for a more detailed discussion. Under a regression setup with observed covariates and incomplete responses, MCAR implies that the missingness is completely independent of both covariates and responses. In contrast, MAR means that the missingness could depend on the observed covariates. However, conditional on the observed covariates, the missingness should be independent of the responses. Lastly, NM suggests that the missingness depends on the responses even after controlling for the covariate effect. It is remarkable that MCAR is an assumption too restrictive to hold in many practical applications. On the other hand, NM is an assumption often leading to identification issue. This makes MAR an assumption widely used for statistical research. Accordingly, MAR is adopted throughout the rest of this article (Rao & Shao, 1992; Rubin, 1987; Sun & Wang, 2020).

Under the MAR assumption, estimators based on complete data might lead to seriously biassed results (Nakai & Ke, 2011; Shao & Wang, 2002). To solve this problem, various modelling and imputation methods have been developed for missing data. For example, Shao and Wang (2002) investigated a sample correlation coefficients-based regression method. Semiparametric of this type was also developed (Liang et al., 2007; Wang & Dai, 2008; Wang et al., 2004, 2016; Zhao & Tang, 2016). For other more details, see Little and Rubin (2002) for an excellent overview. Despite of their usefulness, most of the existing estimation and imputation methods are developed for independent data (Schafer, 1997; Qin et al., 2008; Miao et al., 2016). Recently, a few researchers show interests in exploring estimation methods with either spatial or temporal correlation. Rahman et al. (2015) proposed an estimation method in time series using lagged correlation. Wang and Lee (2013a) used three methods to estimate a SAR model with incomplete responses. Sun and Wang (2020) proposed an estimation and imputation method for a SAR model. Zhou et al. (2022) developed an autoregressive model with SAR error for missing data without covariates. It seems to us no estimation method exists for spatial panel data so that both spatial and temporal dependencies, together with exogenous covariates, can be accommodated simultaneously.

The model developed in this work for missing data makes full use of both spatial and temporal dependencies, together with exogenous covariates. Specifically, we borrowed the classical SDPD model of Yu et al. (2008) as our model foundation. This model allows cross-sectional spatial dependence by a SAR structure. It accommodates temporal dependence by a vector autoregressive (VAR) component. It also considers the exogenous covariate effects by a linear regression model. To reflect the missingness mechanism, a logistic regression model is used. Since we adopt the MAR assumption, the model allows the exogenous covariates to be correlated with missing probability. Throughout the rest of this article, we assume that these covariates are fully observed and complete. To estimate the unknown parameters, a novel weighted log-likelihood function is developed, and this leads to a weighted maximum likelihood estimator (WMLE). Under appropriate regularity conditions, we show theoretically that WMLE is consistent and asymptotically normal. With the help of WMLE, a novel imputation method is developed. The imputed values take both the spatial and temporal dependencies into consideration, together with the exogenous covariate effects. This leads to much improved imputation results. Accordingly, the missing price in the CECN dataset can be appropriately imputed. As a result, a principled construction material price index can be compiled.

The rest of this article is organized as follows. Section 2 introduces the model and notations. The WMLE is also developed, and its asymptotic properties are studied. A novel imputation method is proposed in this section. Section 3 demonstrates numerical studies, including simulation experiments and a real data example. Lastly, the article is concluded with a brief discussion in Section 4. All technical details are relegated in the supporting information Appendix.

2 | MODEL AND METHODOLOGY

2.1 | Model and notations

Let $Y_{it} \in \mathbb{R}^1$ be a continuous response collected from the ith ($1 \le i \le N$) location at time point t ($1 \le t \le T$). To model Y_{it}, we adopt a SDPD model from Yu et al. (2008) as follows:

$$\mathbb{Y}_t = \lambda W\mathbb{Y}_t + \gamma\mathbb{Y}_{t-1} + \rho W\mathbb{Y}_{t-1} + \mathbb{X}_t\beta + \mathcal{E}_t, \tag{1}$$

where $\mathbb{Y}_t = (Y_{1t},...,Y_{Nt})^\top \in \mathbb{R}^N$ is the response vector collected at time point t and $\mathbb{X}_t = (X_{1t}^\top,...,X_{Nt}^\top)^\top \in \mathbb{R}^{N\times p}$ is the associated covariate matrix. Here, $X_{it} = (X_{it1},...,X_{itp})^\top$ is a p-dimensional exogenous covariate. The matrix $W \in \mathbb{R}^{N\times N}$ is a row normalized weight matrix, which is used to capture the spatial dependence among different locations. For example, assume that there exists an adjacency matrix as $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, where $a_{ij} = 1$ if location i is bordered with location j and $a_{ij} = 0$ otherwise. Then $W = (w_{ij}) \in \mathbb{R}^{N\times N}$ can be defined as $w_{ij} = a_{ij}/d_i$, where $d_i = \sum_{j=1}^N a_{ij}$ is the total number of locations that i is bordered with. Lastly, $\mathcal{E}_t = (\varepsilon_{1t},...,\varepsilon_{Nt})^\top \in \mathbb{R}^N$ is residual vector, which is assumed to follow a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I \in \mathbb{R}^{N\times N}$. Here, I stands for an identity matrix with an appropriate dimension.

Write $S = I - \lambda W$. Then by Lee (2004), we know that S is always invertible as long as $|\lambda| < 1$. Thus, throughout the rest of this article, we assume that $|\lambda| < 1$. Define $M = S^{-1}(\gamma I + \rho W)$. Then, model (1) can be rewritten as follows:

$$\mathbb{Y}_t = M\mathbb{Y}_{t-1} + S^{-1}\mathbb{X}_t\beta + S^{-1}\mathcal{E}_t. \tag{2}$$

In practice, the response Y_{it} could be incomplete. We then use a binary indicator $Z_{it} \in \{0,1\}$ to indicate whether Y_{it} is observed or not. Specifically, define $Z_{it} = 1$ if Y_{it} is observed and define $Z_{it} = 0$ otherwise. Next, assume

$$P(Z_{it} = 1|\mathcal{F}) = p_{it} = \frac{\exp(\varsigma^\top X_{it})}{1 + \exp(\varsigma^\top X_{it})}, \tag{3}$$

where $\mathcal{F}$ is the σ-field generated by $\{(Y_{it}, X_{it}) : 1 \le t \le T, 1 \le i \le N\}$ and $\varsigma = (\varsigma_1,...,\varsigma_p)^\top \in \mathbb{R}^p$ is the associated regression coefficient vector. By model (3), it suggests that a MAR missingness mechanism is adopted. This is because conditional on the observed covariates X_{it}, the missingness of response Y_{it} is independent of the response Y_{it} itself. It is also worth noting that the exogenous covariates in (3) could be different from those in model (1). They may contain the variables on regions or time.

2.2 | Weighted maximum likelihood

We next consider how to estimate the parameters in model (1). Define $\theta = (\delta^\top, \lambda, \sigma^2)^\top \in \mathbb{R}^{p+4}$, where $\delta = (\gamma, \rho, \beta^\top)^\top \in \mathbb{R}^{p+2}$. Recall that $\mathcal{E}_t$ is assumed to follow a multivariate normal distribution with mean 0 and covariance $\sigma^2 I$. Then, by model (2), we have the following full data log-likelihood function (omitting some constants),

$$\ell_1(\theta) = (T-1)\log|S| - \frac{N(T-1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^T \mathcal{E}_t^\top \mathcal{E}_t, \tag{4}$$

where $S = I - \lambda W$ and $\mathcal{E}_t = S\mathbb{Y}_t - \gamma\mathbb{Y}_{t-1} - \rho W\mathbb{Y}_{t-1} - \mathbb{X}_t\beta = \tilde{\mathbb{Y}}_t - \tilde{\mathbb{X}}_t\delta$. Here, $\tilde{\mathbb{X}}_t = S\mathbb{Y}_t \in \mathbb{R}^N$, $\tilde{\mathbb{X}}_t = (\mathbb{Y}_{t-1}, W\mathbb{Y}_{t-1}, \mathbb{X}_t) \in \mathbb{R}^{N\times(p+2)}$.

Next, consider how to handle incomplete observations. Note that $E(Z_{it}Z_{i(t-1)}|\mathcal{F}) = p_{it}p_{i(t-1)}$. This suggests that different weights (e.g. $p_{it}, p_{i(t-1)}$) are expected for different sample pairs $(Y_{it}, Y_{i(t-1)})$. In other words, each sample pair is no longer treated equally in the estimation process. This might make θ less efficient (Zhou et al., 2022). This inspires us to consider the following weighted log-likelihood function as follows:

$$\ell_2(\theta) = (T-1)\log|S| - \frac{N(T-1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^T\sum_{i=1}^N \frac{Z_{it}Z_{i(t-1)}}{p_{it}p_{i(t-1)}}\varepsilon_{it}^2. \tag{5}$$

One can easily verify that $E\{\ell_2(\theta)|\mathcal{F}\} = \ell_1(\theta)$. This suggests that the weighted log-likelihood function (5) is an unbiassed estimator for the full data log-likelihood function (4). Accordingly, it might lead to a sensible estimator for θ.

Nevertheless, the weighted log-likelihood function (5) cannot be directly used for parameter estimation. This is because p_{it}s are unknown parameters. To fix the problem, we can replace p_{it} by its consistent estimator $\hat{p}_{it} = \{\exp(\hat{\varsigma}^\top X_{it})\}/\{1 + \exp(\hat{\varsigma}^\top X_{it})\}$, where $\hat{\varsigma}$ is maximum likelihood estimator from the logistic regression model (3). This leads to the following practically feasible weighted log-likelihood function as

$$\ell_3(\theta) = (T-1)\log|S| - \frac{N(T-1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^T\sum_{i=1}^N \frac{Z_{it}Z_{i(t-1)}}{\hat{p}_{it}\hat{p}_{i(t-1)}}\varepsilon_{it}^2. \tag{6}$$

Then, a feasible WMLE can be obtained as $\hat{\theta} = \arg\max_\theta \ell_3(\theta)$. Its asymptotic properties are to be carefully studied in the following subsection.

2.3 | Theoretical results

We first introduce some notations. For an arbitrary $N \times N$ matrix $C = (c_{ij}) \in \mathbb{R}^{N\times N}$, define $\|C\|_\infty = \max_{1\le i\le N}\sum_{j=1}^{N}|c_{ij}|$, $\|C\|_1 = \max_{1\le j\le N}\sum_{i=1}^{N}|c_{ij}|$, and $\mathrm{abs}(C) = (|c_{ij}|) \in \mathbb{R}^{N\times N}$. Denote $G = WS^{-1} = (G_{ij}) \in \mathbb{R}^{N\times N}$, $\mathcal{P}_t = \mathrm{diag}\{p_{it}p_{i(t-1)}\} \in \mathbb{R}^{N\times N}$. Moreover, define $\Delta_{NT} = [\Delta_{NT,11}, \Delta_{NT,12}, 0; \Delta_{NT,12}^\top, \Delta_{NT,22}, \Delta_{NT,23}; 0, \Delta_{NT,23}^\top, \Delta_{NT,33}] \in \mathbb{R}^{(p+4)\times(p+4)}$, where

$$\Delta_{NT,11} = \frac{1}{NT}\frac{1}{\sigma^2}\sum_{t=2}^{T}\tilde{\mathbb{X}}_t^\top \mathcal{P}_t^{-1}\tilde{\mathbb{X}}_t, \qquad \Delta_{NT,12} = -\frac{1}{NT}\frac{1}{\sigma^2}\sum_{t=2}^{T}\tilde{\mathbb{X}}_t^\top \mathcal{P}_t^{-1}(G\tilde{\mathbb{X}}_t\delta),$$

$$\Delta_{NT,22} = \Delta_{NT,22,1} + \Delta_{NT,22,2}, \qquad \Delta_{NT,22,1} = \frac{1}{NT}\frac{1}{\sigma^2}\sum_{t=2}^{T}(G\tilde{\mathbb{X}}_t\delta)^\top \mathcal{P}_t^{-1}(G\tilde{\mathbb{X}}_t\delta),$$

$$\Delta_{NT,22,2} = \frac{1}{NT}\sum_{t=2}^{T}\left\{2\sum_{i=1}^{N}G_{ii}^2(p_{it}^{-1}p_{i(t-1)}^{-1} - 1) + \mathrm{tr}(GG^\top \mathcal{P}_t^{-1}) + \mathrm{tr}(G^2)\right\},$$

$$\Delta_{NT,23} = \frac{1}{NT}\frac{1}{2\sigma^2}\sum_{t=2}^{T}\{3\mathrm{tr}(G\mathcal{P}_t^{-1}) - \mathrm{tr}(G)\}, \Delta_{NT,33} = \frac{1}{4NT\sigma^4}\sum_{t=2}^{T}\{3\mathrm{tr}(\mathcal{P}_t^{-1}) - N\}.$$

To investigate the asymptotic properties of the proposed WMLE of θ, we consider the following technical conditions.

(C1) (WEIGHT MATRIX) The spatial weight matrix W satisfies that $\|W\|_\infty < \infty$.

(C2) (SPATIAL DEPENDENCE) Assume $\lambda \in (-1,1)$. In addition, assume $\mathcal{M} = \sum_{k=1}^{\infty}\mathrm{abs}(M^k)$ exists and satisfies that $\|\mathcal{M}\|_\infty < \infty$ and $\|\mathcal{M}\|_1 < \infty$.

(C3) (LAW OF LARGE NUMBER) There exists a positive definite matrix Δ, which satisfies that $\Delta_{NT} \to_p \Delta$ as $\min\{N,T\} \to \infty$.

These conditions are commonly used in literature. Condition (C1) is a standard assumption in SAR literature (Yu et al., 2008; Wang & Lee, 2013a, 2013b). Condition (C2) is an assumption on the absolute summability of M and its power. This controls the dependence between time series and between cross-sectional locations. It is trivially satisfied if $\gamma = \rho = 0$. More detailed discussion can be found in Yu et al. (2008) and Li (2017). Condition (C3) is a law of large number type assumption. To guarantee the positive definiteness of Δ, it is worth noting that the missing rate should not be too large. It is trivially satisfied if there is no missing data, such as, $\mathcal{P}_t = I$. As a result, we can obtain the associated positive matrix Λ. In this case, similar assumption can be found in Theorem 3 of Yu et al. (2008). With the help of above conditions, we then have the following theorem.

Theorem 1. *Assume conditions (C1)–(C3) as given above. Further assume that* $\min\{N,T\} \to \infty$, *we then have* $\sqrt{NT}(\hat{\theta} - \theta) \to_d N(0, \Lambda^{-1}\Delta\Lambda^{-1})$, *where* Λ *is a special case of* Δ *when all* p_{it}*s are fixed to be 1, that is,* $\mathcal{P}_t = I$.

The proof of Theorem 1 is given in the supporting information Appendix B. This theorem establishes the consistency and asymptotic normality of the WMLE of $\hat{\theta}$. It is worth noting that the convergence rate also depends on the missing rate, which is included in Δ. As for a valid statistical inference, we can consistently estimate Δ by $\hat{\Delta}_{NT}$, which is obtained by replacing θ and p_{it} in Δ_{NT} by $\hat{\theta}$ and $\hat{p}_{it}$, respectively. We can also consistently estimate Λ by $\hat{\Lambda} = \hat{\Lambda}_{NT}$. In this case, all of $\hat{p}_{it}$ in $\hat{\Delta}_{NT}$ are 1.

2.4 | Imputation method

With the estimated $\hat{\theta}$, we next consider how to conduct imputation for $\mathbb{Y}_t$. For simplicity, denote $\mathbb{Y}_t = (\mathbb{Y}_t^{(1)}, \mathbb{Y}_t^{(2)})^\top$, where $\mathbb{Y}_t^{(1)}$ is observed vector and $\mathbb{Y}_t^{(2)}$ is the unobserved one. It is remarkable that $\mathbb{Y}_t^{(1)}$ could be associated with different locations for different time points t. In order to impute $\mathbb{Y}_t^{(2)}$, we first study $E(\mathbb{Y}_t^{(2)}|\mathbb{Y}_t^{(1)}, \mathbb{Y}_{t-1}, \mathbb{X}_t)$. We then have the following proposition.

Proposition 1. *Given* $\mathbb{Y}_{t-1}$ *and* $\mathbb{X}_t$, *we can obtain that* $\mathbb{Y}_t = (\mathbb{Y}_t^{(1)}, \mathbb{Y}_t^{(2)})^\top$ *follows multivariate normal distribution with conditional mean* $\mu_t = (\mu_t^{(1)}, \mu_t^{(2)})^\top = M\mathbb{Y}_{t-1} + S^{-1}\mathbb{X}_t\beta$ *and conditional covariance* $\Sigma = [\Sigma^{(11)}, \Sigma^{(12)}; \Sigma^{(21)}, \Sigma^{(22)}] = \sigma^2(S^\top S)^{-1}$. *Then, we have* $E(\mathbb{Y}_t^{(2)}|\mathbb{Y}_t^{(1)}, \mathbb{Y}_{t-1}, \mathbb{X}_t) = \mu_t^{(2)} - \Sigma^{(21)}(\Sigma^{(11)})^{-1}(\mathbb{Y}_t^{(1)} - \mu_t^{(1)})$.

Proposition 1 suggests an interesting and recursive imputation method. Specifically, we start with $\mathbb{Y}_0$. In this case, the whole response vector $\mathbb{Y}_0$ is not observed at all. We then impute it by a relatively simple estimator $\mathbb{Y}_0^\star = \overline{\mathbb{Y}}^c$, where $\overline{\mathbb{Y}}^c = (Y_1^c, \ldots, Y_N^c)^\top \in \mathbb{R}^N$ and $Y_i^c = \left(\sum_{t=1}^{T}Z_{it}Y_{it}\right)/\left(\sum_{t=1}^{T}Z_{it}\right)$ is the simple averaging over observed responses. Obviously, $\overline{\mathbb{Y}}^c$ is a very crude estimator for $\mathbb{Y}_0$. We can try other alternative (e.g. $\mathbb{Y}_0 = 0$). We find that the overall results are fairly similar as long as T is sufficiently large. Once $\mathbb{Y}_{t-1}^\star$ is obtained, we then treat it as

$\mathbb{Y}_{t-1}$. Then by Proposition 1, we know that $\mathbb{Y}_t^\star$ can be replaced by $\mathbb{Y}_t^\star = (\mathbb{Y}_t^{(1)}, \mathbb{Y}_t^{\star(2)})^\top$. Here, $\mathbb{Y}_t^{\star(2)} = \hat{\mu}_t^{(2)} - \hat{\Sigma}^{(21)}(\hat{\Sigma}^{(11)})^{-1}(\mathbb{Y}_t^{(1)} - \hat{\mu}_t^{(1)})$, where $(\hat{\mu}_t^{(1)}, \hat{\mu}_t^{(2)})^\top = \hat{\mu}_t = \hat{M}\mathbb{Y}_{t-1}^\star + \hat{S}^{-1}\mathbb{X}_t\hat{\beta}$ and $\hat{\Sigma} = [\hat{\Sigma}^{(11)}, \hat{\Sigma}^{(12)}; \hat{\Sigma}^{(21)}, \hat{\Sigma}^{(22)}] = \hat{\sigma}^2(\hat{S}^\top \hat{S})^{-1}$. Here, $\hat{M}$ is obtained by replacing θ in M by $\hat{\theta}$, and $\hat{S}$ is computed in a similar way. Then repeating the above procedure, this leads to the entire imputed responses sequence $\{\mathbb{Y}_t^\star\}_{t=1}^T$. Thereafter, standard statistical analysis (e.g. computing sample mean at different time points) can be conducted based on $\{\mathbb{Y}_t^\star\}_{t=1}^T$.

3 | NUMERICAL STUDIES

3.1 | Simulation models

To demonstrate the finite sample performance of the proposed method, we present some simulation studies. First, we generate N independent and identically distributed random variables according to an exponential distribution with mean 10. Denote these variables by U_i with $1 \le i \le N$. For each location i, we randomly select a sample size of $[U_i]$ from $\mathcal{S}_F = \{1,2,\ldots,N\}$ without replacement, where $[U_i]$ stands for the smallest integer no less than U_i. Denote the sample by $\mathcal{S}_i$. Define $a_{ij} = 1$ if $j \in \mathcal{S}_i$ and $a_{ij} = 0$ otherwise. Lastly, let $a_{ii} = 0$ for every $1 \le i \le N$. We next force A to be symmetric by replacing a_{ij} by a_{ji} for $i < j$. Subsequently, W can be obtained by normalizing each row of A. Then W is fixed across different time points.

Once W and N are given, the response variable $\mathbb{Y}_t$ is generated according to $\mathbb{Y}_t = M\mathbb{Y}_{t-1} + S^{-1}\mathbb{X}_t\beta + S^{-1}\mathcal{E}_t$, where $\mathcal{E}_t$ is simulated from a multivariate normal distribution with mean 0 and covariance $\sigma^2 I$. The true value of $(\gamma,\rho,\beta,\lambda,\sigma^2)^\top$ is set to be $(0.3,0.2,2,0.5,1)^\top$. To simulate $\mathbb{Y}_t$ sequence, we first generate $\mathbb{Y}_0$ from a multivariate standard normal distribution. Then we generate $\mathbb{Y}_t$ sequentially according to Equation (2) for $t = 1,\ldots,T_0+T$, where T_0 is a prespecified integer. For example, in this work, we assume $T_0 = 1000$. We then redefine $\mathbb{Y}_t = \mathbb{Y}_{t-T_0}$, for $t = T_0+1,\ldots,T+T_0$. This leads to the final sequence of $\{\mathbb{Y}_t : 1 \le t \le T\}$.

For illustration purpose, we consider $p = 2$ and $X_{it} = (X_{it1}, X_{it2})^\top \in \mathbb{R}^2$. We fix $X_{it1} = 1$ to be the intercept, define $X_{it2} = Y_{it}Y_{i(t-1)} + e_{it1}$ and generate e_{it1} from a standard normal distribution. In this way, X_{it2} has an effect on $\mathbb{Y}_t$ and it is fully observed. This satisfies our MAR assumption. For simplicity, we only consider the temporal correlation between different X_{it2}s. We generate Z_{it} according to model (3). Recall $\varsigma = (\varsigma_0, \varsigma_1)^\top \in \mathbb{R}^2$. Fix $\varsigma_1 = 0.1$ but allow different values of ς_0, so that the overall missing rate is controlled between 25% ($\varsigma_0 = 1$) and 35% ($\varsigma_0 = 0.5$).

3.2 | Performance measures and simulation results

We consider different network sizes (N = 100, 200, 500) and different number of time points (T = 100, 200, 500). For a reliable evaluation, the experiment is randomly replicated $R = 500$ times for every (N,T) combination. For a given (N,T) combination, we use $\hat{\alpha}^{(r)}$ to repeat one particular estimator (e.g. $\hat{\gamma}$) obtained in the rth replication. We further assume that the estimating target is α. Then the root mean square error (RMSE) is evaluated by $\text{RMSE} = \{R^{-1}\sum_{r=1}^R(\hat{\alpha}^{(r)} - \alpha)^2\}^{1/2}$. In addition to that, a 95% confidence interval is constructed as $\text{CI}^{(r)} = (\hat{\alpha}^{(r)} - z_{0.975}\widehat{\text{SE}}^{(r)}, \hat{\alpha}^{(r)} + z_{0.975}\widehat{\text{SE}}^{(r)})$, where $\widehat{\text{SE}}^{(r)}$ is computed according to the asymptotic covariance in Theorem 1 by plugging in the resulting estimator respectively. Here, z_α is the αth quantile of a standard normal distribution. Consequently, the empirical coverage probability (ECP) is computed as $\text{ECP} = R^{-1}\sum_{r=1}^R I(\alpha \in \text{CI}^{(r)})$, where $I(\cdot)$ is the indicator function. Detailed simulation results are summarized in Tables 1 and 2.

We can draw the following conclusions from Tables 1 and 2. For example, Table 1 presents the case with a missing rate around 25%. We can find that the WMLE of θ are consistent, with RMSE decreasing towards 0 as $\min\{N,T\} \to \infty$. Additionally, the ECP is fairly close to their nominal level 95%. This suggests that the resulting estimator is asymptotically normal, and the estimated standard error (i.e. $\widehat{\text{SE}}$) can approximate the true SE well. Table 2 reports the case with a missing rate around 35%. The findings are quantitatively similar.

3.3 | Imputation results

As we mentioned before, this work is motivated by a real data application, which is the CECN price index composition problem. In case of no missing responses (i.e. the price information), a price index can be easily constructed by simply averaging the price values collected from different locations but at the same time point. Statistically, this amounts to compute $\hat{\mu}_t = N^{-1}\sum_{i=1}^N Y_{it}$. Unfortunately, this simple statistic is not computable if a significant portion of the responses is missing. Then, imputation becomes a natural choice. More specifically, for a given (i,t), let Y_{it}^* be the imputed value for Y_{it} by one particular type of imputation method (e.g. the proposed imputation method in Section 2.4). Once Y_{it}^*s is obtained, the price index can be computed based on the imputed responses as $\hat{\mu}_t^* = N^{-1}\sum_{i=1}^N\{Z_{it}Y_{it} + (1-Z_{it})Y_{it}^*\}$. Recall that $Z_{it} = 1$ if Y_{it} is not missing, and $Z_{it} = 0$ otherwise. Under a simulation setup, the imputation accuracy can be measured by RMSE as $\text{RMSE} = \left\{T^{-1}\sum_{t=1}^T(\hat{\mu}_t - \hat{\mu}_t^*)^2\right\}^{1/2}$. Because for each (N,T) combination, the simulation experiments, as given in the previous subsection, are randomly replicated for a total of 500 times. This leads to 500 RMSE values for each (N,T) combination, and then they are further averaged and reported in Tables 3 and 4.

TABLE 1 Simulation results with missing rate of 25% ($\varsigma_0 = 1$).

N	T	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\beta}$	$\hat{\lambda}$	$\widehat{\sigma^2}$
100	100	0.59(94.8)	1.67(94.4)	1.37(95.6)	1.52(94.0)	2.05(96.8)
	200	0.42(95.0)	1.26(95.6)	0.95(96.0)	1.19(94.4)	1.44(96.8)
	500	0.27(94.4)	0.85(94.6)	0.60(95.2)	0.76(94.0)	0.92(96.2)
200	100	0.42(95.6)	1.14(95.2)	1.01(93.4)	1.09(94.4)	1.47(96.4)
	200	0.30(95.2)	0.89(95.8)	0.68(96.4)	0.79(94.8)	1.06(95.0)
	500	0.19(94.6)	0.55(97.0)	0.41(96.0)	0.53(94.2)	0.64(96.2)
500	100	0.27(94.6)	0.73(94.6)	0.60(95.8)	0.67(93.8)	0.89(97.4)
	200	0.20(94.2)	0.58(94.4)	0.44(94.4)	0.51(95.2)	0.60(97.2)
	500	0.12(95.0)	0.36(95.2)	0.29(94.2)	0.32(94.8)	0.40(97.6)

Note: The RMSE values ($\times 10^{-2}$) are reported for every (N,T) combination and estimator. The ECP (in %) is given in parentheses.

TABLE 2 Simulation results with missing rate of 35% ($\varsigma_0 = 0.5$).

N	T	$\hat{\gamma}$	$\hat{\rho}$	$\hat{\beta}$	$\hat{\lambda}$	$\widehat{\sigma^2}$
100	100	0.70(95.6)	1.87(95.8)	1.62(94.6)	1.71(94.4)	2.36(96.4)
	200	0.49(95.8)	1.43(96.2)	1.11(96.2)	1.33(95.0)	1.69(96.8)
	500	0.31(95.6)	0.99(93.8)	0.69(96.2)	0.87(94.2)	1.08(97.2)
200	100	0.48(95.4)	1.29(97.0)	1.22(92.4)	1.29(93.2)	1.72(95.4)
	200	0.34(95.0)	1.01(95.8)	0.79(96.6)	0.92(95.0)	1.24(96.2)
	500	0.22(96.0)	0.64(96.4)	0.52(94.6)	0.60(95.8)	0.73(96.6)
500	100	0.32(94.4)	0.84(96.2)	0.72(95.8)	0.78(95.0)	1.05(95.6)
	200	0.23(94.0)	0.67(93.0)	0.49(95.8)	0.59(95.0)	0.71(98.4)
	500	0.14(93.2)	0.43(95.4)	0.34(94.6)	0.38(94.4)	0.48(96.6)

Note: The RMSE values ($\times 10^{-2}$) are reported for every (N,T) combination and estimator. The ECP (in %) is given in parentheses.

For comparison purpose, we consider the following competitive methods for imputation. The first imputation method is our proposed imputation method. It is a method based on the spatial dynamic panel data model of Yu et al. (2008). We thus refer to it as SDPD method for short. On the other hand, in practice, without a SDPD type model to support, one can only consider some simple imputation methods. For example, one can consider the method of mean imputation based on complete cases (MIBC). That is to impute Y_{it} by $\hat{\mu}_t^c = n_{ct}^{-1}\sum_{i=1}^N Z_{it}Y_{it}$ and $n_{ct} = \sum_{i=1}^N Z_{it}$. Second, one can also consider a complete case-based regression (CCBR) imputation method. That is to impute Y_{it} by $\hat{Y}_{it}^c = X_{it}^\top \hat{\beta}^c$, where $\hat{\beta}^c = \left(\sum_{i=1}^N\sum_{t=1}^T Z_{it}X_{it}X_{it}^\top\right)^{-1}\left(\sum_{i=1}^N\sum_{t=1}^T Z_{it}X_{it}Y_{it}\right)$ is the ordinary least squares estimator obtained based on complete cases. Third, one can also consider to impute Y_{it} by its observed spatial neighbourhood average (OSNA). That is to impute Y_{it} by $Y_{it}^s = n_{ts}^{-1}\sum_{j=1}^N a_{ij}Y_{jt}Z_{jt}$, where $n_{ts} = \sum_{j=1}^N a_{ij}Z_{jt}$. Recall that $A = (a_{ij})$ is the associated adjacency matrix. Lastly, one can also consider to impute Y_{it} by carrying forward its nearest historical observation. This is a method that has been referred to as the last observation carrying forward (LOCF, Shao & Zhong, 2003). More specifically, for a given Y_{it}, define $t_{\max} = \max\{s : s < t, Z_{is} = 1\}$. We then impute Y_{it} by $Y_{it_{\max}}$. This leads to a total of five different imputation methods (our proposed SDPD method and four competing methods). They are all evaluated in our simulation experiments, and their detailed RMSE values are summarized in Tables 3 and 4.

Table 3 displays the imputation results for the case with missing rate around 25%. We can see that the imputation accuracy of SDPD method is considerably better than the other four methods in terms of the averaged RMSE values. Additionally, the imputation accuracy improves as $\{N,T\}$ increases. Table 4 is the case with a missing rate of 35%. The findings are quantitatively similar.

3.4 | A real data example

As our last example, we apply the proposed method to the CECN dataset. As we mentioned before, this is a dataset about price information for construction material. Specifically, the response is the price change in logarithm which is defined as $Y_{it} = \log(P_{it}) - \log(P_{i,t-1})$, where P_{it} is the price of one kind of building materials collected at location i (one particular province) and time point t (one particular month). For this particular case, a

TABLE 3 Imputation results with missing rate of 25% ($\varsigma_0 = 1$).

N	T	SDPD	MIBC	CCBR	OSNA	LOCF
100	100	0.1018	0.2749	0.2602	0.2927	0.1937
	200	0.0716	0.1972	0.1819	0.2071	0.1404
	500	0.0450	0.1301	0.1137	0.1325	0.0989
200	100	0.0977	0.2857	0.2615	0.2936	0.1920
	200	0.0684	0.2037	0.1842	0.2078	0.1407
	500	0.0436	0.1343	0.1150	0.1346	0.0985
500	100	0.0957	0.2935	0.2638	0.2953	0.1919
	200	0.0671	0.2083	0.1837	0.2086	0.1404
	500	0.0425	0.1362	0.1150	0.1342	0.0986

Note: The RMSE values ($\times 10^{-2}$) are reported for every (N,T) combination and imputation method.
Abbreviations: CCBR, complete case-based regression; LOCF, last observation carrying forward; MIBC, mean imputation based on complete cases; OSNA, observed spatial neighbourhood average; SDPD, spatial dynamic panel data.

TABLE 4 Imputation results with missing rate of 35% ($\varsigma_0 = 0.5$).

N	T	SDPD	MIBC	CCBR	OSNA	LOCF
100	100	0.1334	0.3731	0.3596	0.3963	0.2518
	200	0.0937	0.2671	0.2511	0.2797	0.1844
	500	0.0588	0.1758	0.1575	0.1784	0.1316
200	100	0.1276	0.3880	0.3615	0.3980	0.2519
	200	0.0893	0.2764	0.2548	0.2816	0.1844
	500	0.0569	0.1813	0.1591	0.1810	0.1309
500	100	0.1246	0.3981	0.3643	0.4002	0.2505
	200	0.0873	0.2823	0.2541	0.2822	0.1837
	500	0.0553	0.1842	0.1592	0.1807	0.1311

Note: The RMSE values ($\times 10^{-2}$) are reported for every (N,T) combination and imputation method.
Abbreviations: CCBR, complete case-based regression; LOCF, last observation carrying forward; MIBC, mean imputation based on complete cases; OSNA, observed spatial neighbourhood average; SDPD, spatial dynamic panel data.

TABLE 5 The real data estimation result.

Parameter	Coefficient	Std. err	*P*-value
γ	−0.3238	0.0900	<0.001
λ	0.5799	0.0569	<0.001
ρ	0.3734	0.1233	<0.05
ς_0	1.8066	0.1960	<0.001
ς_1	−1.2206	0.2465	<0.001
ς_2	−0.0177	0.0089	<0.05
β_1	−0.0025	0.0070	0.721
β_2	0.0003	0.0002	0.134
σ^2	0.0023	0.0003	<0.001

total of 26 locations (i.e. selected provinces) are studied and a total of $T = 15$ monthly data are collected. The objective of this study is to form an index to reflect the overall price dynamics (i.e. changes). Statistically, this amounts to compute $\hat{\mu}_t = N^{-1}\sum_{i=1}^{N} Y_{it}$ for each time point t. Unfortunately, due to various practical reasons, around 25% of the price information Y_{it} are missing. The missing rate seems to be clearly depends on months (e.g. the Spring Festival happens in February). This inspires us to collect for each Y_{it} a covariate of X_{it1}, where $X_{it1} = 1$ if t happens to be January, February or March. Those are the months with missing rate substantially higher than other months due to some practical reason.

Additionally, we consider another covariate of X_{it2}, which is the growth rate of fixed asset investment in construction industry. Obviously, X_{it1} and X_{it2} are the covariates always observed.

We then try to apply the proposed estimation methods to the dataset. The detailed results are given in Table 5. First, we find that ς_1 is estimated to be $\hat{\varsigma}_1 = -1.2206$, which is negatively significant at 0.1% level. This confirms our experience that the missing rate in those months (January, February or March) is substantially higher than other months. Second, we find that the other parameters (e.g. γ, λ, ρ) are significant at 5% level. This implies that both spatial and temporal dependencies do exist in price change dynamics. Specifically, γ is negatively estimated to be $\hat{\gamma} = -0.3238$. This suggests that higher (lower) price change from the previous time points might lead to lower (higher) price change of the current time point for the same location. In the meanwhile, both ρ and λ are positively estimated to be $\hat{\rho} = 0.3734$ and $\hat{\lambda} = 0.5799$, respectively. This suggests that the price dynamics of neighbouring locations should be positively correlated across different time points.

4 | CONCLUSION

In this article, we develop a novel imputation method to analyse the missing response problem in spatial dynamic panel data. The SDPD model of Yu et al. (2008) is used as the model foundation. A logistic regression model is used to reflect the missingness mechanism. The WMLE is proposed for parameter estimation in the presence of incomplete data. The associated asymptotic properties are investigated. Moreover, a novel regression-based imputation method is proposed. The proposed method makes use of the information from spatial dependence, temporal dependence and exogenous regression covariates. Finally, the performance of WMLE and imputation methods is demonstrated by both simulation studies and a real data example.

To conclude this article, we discuss here several interesting topics for future study. First, the time and individual fixed effects can be considered in SDPD model (Lee & Yu, 2010). Second, more flexible spatial lags (e.g. different spatial weight matrices) and temporal lags could be considered (Li, 2017). Third, we can extend the error term to be spatially dependent (Yang, 2018). Fourth, the SDPD model used in this context assumes both the spatial and temporal dependencies to be reflected by scalar parameters. More flexible dependency parameters can be considered (Dou et al., 2016; Zhu et al., 2019). Lastly, the weight matrix in this work is predetermined. However, in many cases, the weight matrix is endogenously determined, because locations sharing similar features are more likely to be connected. Thus, how to model this endogenous phenomenon is an important topic worth future study.

ACKNOWLEDGEMENTS

The authors sincerely thank the co-editor, associate editor and anonymous referees for their helpful and insightful comments. Jin Liu's research is supported in part by the National Natural Science Foundation of China (no. 12201316). Jing Zhou's research is supported in part by the National Natural Science Foundation of China (nos. 72171226 and 11971504), the Beijing Municipal Social Science Foundation (no. 19GLC052), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China, no. 21XNA027. Wei Lan's research was supported by the National Natural Science Foundation of China (no. 71532001, 11931014, 12171395 and 71991472) and the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics. Hansheng Wang's research is partially supported by National Natural Science Foundation of China (nos. 12271012 and 11831008) and also partially supported by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (KLATASDS-MOE-ECNU-KLATASDS2101).

CONFLICT OF INTEREST STATEMENT

On behalf of all the authors, the corresponding author states that all of the authors agree to the submission, and there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

Anselin, L. (1980). Estimation methods for spatial autoregressive structures, *Regional science dissertation and monograph series 8*: Cornell University, Ithaca, NY.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: theory and methods*. Berlin, New York: Springer.

Dou, B., Parrella, M. L., & Yao, Q. (2016). Generalized Yule-Walker estimation for spatio-temporal models with unknown diagonal coefficients. *Journal of Econometrics*, *194*(2), 369–382.

Feng, X., Li, W., & Zhu, Q. (2022). Spatial-temporal model with heterogeneous random effects. *Statistica Sinica*. Forthcoming.

Fuller, W. A. (1996). *Introduction to statistical time series*: John Wiley and Sons.

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

Lee, L. F., Li, J., & Lin, X. (2013). Specification and estimation of social interaction models with network structure. *The Econometrics Journal*, *13*, 145–176.

Lee, L. F., & Yu, J. (2010). A spatial dynamic panel data model with both time and individual fixed effects. *Econometric Theory*, *26*(2), 564–597.
Lee, L. F., & Yu, J. (2014). Efficient GMM estimation of spatial dynamic panel data models with fixed effects. *Journal of Econometrics*, *180*(2), 174–197.
Li, K. (2017). Fixed-effects dynamic spatial panel data models and impulse response analysis. *Journal of Econometrics*, *198*(1), 102–121.
Li, L., & Yang, Z. (2021). Spatial dynamic panel data models with correlated random effects. *Journal of Econometrics*, *221*, 424–454.
Liang, H., Wang, S., & Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika*, *94*, 185–198.
Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken: John Wiley and Sons, Inc.
Miao, W., Deng, P., & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, *111*, 1673–1683.
Nakai, M., & Ke, W. (2011). Review of methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*, *5*(1), 1–13.
Ord, J. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, *70*, 120–126.
Qin, J., Shao, J., & Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing response. *Journal of the American Statistical Association*, *103*, 797–810.
Rahman, S. A., Huang, Y., Claassen, J., & Kleinberg, S. (2015). Imputation of missing values in time series with lagged correlations. In *2014 IEEE International Conference on Data Mining Workshop*.
Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*, 811–822.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. B. (1987). *Multiple imputation for nonrespondents in surveys*. New York: Wiley.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.
Shao, J., & Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association*, *97*, 544–552.
Shao, J., & Zhong, B. (2003). Last observation carry-forward and last observation analysis. *Statistics in Medicine*, *22*(15), 2429–2441.
Su, L., & Yang, Z. (2015). QML estimation of dynamic panel data models with spatial errors. *Journal of Econometrics*, *185*, 230–258.
Sun, Z., & Wang, H. (2020). Network imputation for a spatial autoregression model with incomplete data. *Statistica Sinica*, *30*(3), 1419–1436.
Wang, Q., & Dai, P. (2008). Semiparametric model-based inference in the presence of missing responses. *Biometrika*, *95*(3), 721–734.
Wang, Q., Zhang, T., & Härdle, W. K. (2016). An extended single-index model with missing response at random. *Scandinavian Journal of Statistics*, *43*(4), 1140–1152.
Wang, Q. H., Linton, O., & Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, *99*, 334–345.
Wang, W., & Lee, L. F. (2013a). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *Econometrics Journal*, *43*(3), 521–538.
Wang, W., & Lee, L. F. (2013b). Estimation of spatial panel data models with randomly missing data in the dependent variable. *Regional Science and Urban Economics*, *43*(3), 521–538.
Yang, Z. (2018). Unified M-estimation of fixed-effects spatial dynamic models with short panels. *Journal of Econometrics*, *205*(2), 423–447.
Yu, J., Jong, R., & Lee, L. F. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics*, *146*(2), 118–134.
Zhao, P. X., & Tang, X. R. (2016). Imputation based statistical inference for partially linear quantile regression models with missing responses. *Metrika*, *79*(8), 991–1009.
Zhou, J., Liu, J., Wang, F., & Wang, H. (2022). Autoregressive model with spatial dependence and missing data. *Journal of Business & Economic Statistics*, *40*(1), 28–34.
Zhu, X., Chang, X., Li, R., & Wang, H. (2019). Portal nodes screening for large scale social networks. *Journal of Econometrics*, *209*(2), 145–157.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, J., Zhou, J., Lan, W., & Wang, H. (2023). Spatial dynamic panel models with missing data. *Stat*, *12*(1), e585. https://doi.org/10.1002/sta4.585