entropy

MDPI

Article

MCMCINLA Estimation of Missing Data and Its Application to Public Health Development in China in the Post-Epidemic Era

Jiaqi Teng, Shuzhen Ding, Xiaoping Shi, Huiguo Zhang and Xijian Hu *

College of Mathematics and System Science, Xinjiang University, Urumqi 830046, China; tengjq111@stu.xju.edu.cn (J.T.); dszspur@xju.edu.cn (S.D.); shixiaoping@xju.edu.cn (X.S.); zhanghg@xju.edu.cn (H.Z.)
* Correspondence: xijianhu@xju.edu.cn; Tel.: +86-130-7990-0717

Abstract: Medical data are often missing during epidemiological surveys and clinical trials. In this paper, we propose the MCMCINLA estimation method to account for missing data. We introduce a new latent class into the spatial lag model (SLM) and use a conditional autoregressive specification (CAR) spatial model-based approach to impute missing values, making the model fit into the integrated nested Laplace approximation (INLA) framework. Combining the advantages of both the Markov chain Monte Carlo (MCMC) and INLA frameworks, the MCMCINLA algorithm is used to implement imputation of the missing data and fit the model to derive estimates of the parameters from the posterior margins. Finally, the economic data and the hemorrhagic fever with renal syndrome (HFRS) disease data of mainland China from 2016–2018 are used as examples to explore the development of public health in China in the post-epidemic era. The results show that compared with expectation maximization (EM) and full information maximum likelihood estimation (FIML), the predicted values of the missing data obtained using our method are closer to the true values, and the spatial distribution of HFRS in China can be inferred from the imputation results with a southern-heavy and northern-light distribution. It can provide some references for the development of public health in China in the post-epidemic era.

Keywords: missing data; spatial lag model; MCMC; INLA; public health

Citation: Teng, J.; Ding, S.; Shi, X.; Zhang, H.; Hu, X. MCMCINLA Estimation of Missing Data and Its Application to Public Health Development in China in the Post-Epidemic Era. *Entropy* **2022**, *24*, 916. https://doi.org/10.3390/e24070916

Academic Editor: Pentti Nieminen

Received: 16 May 2022
Accepted: 26 June 2022
Published: 30 June 2022

1. Introduction

Missing data are common and unavoidable in daily life. For example, in engineering design, equipment failure may cause some data to fail to be collected normally, and during market research, there may also be situations where respondents refuse to answer relevant questions. Besides, missing data often occurs in the medical context. For instance, in epidemiological surveys, equipment limitations prevent access to complete information, and in medical databases, not all patients' clinical test results are available at a given time, leaving a portion of the attribute values vacant. In addition, data can be lost due to the failure of storage media and transmission media.

Information theory is an important discipline based on the methods of probability theory and mathematical statistics for the study of information entropy, data processing, and data transmission [1]. It has a wide range of applications in the effective processing and reliable transmission of information. In the era of big data, the data processing of complex information becomes an important part of research and analysis. Dealing with missing data has always been the first problem that researchers need to solve before conducting statistical analysis, as improper handling can lead to deviations in statistical inferences and even affect the final decision. Little and Rubin (2002) [2] gave a detailed and systematic introduction to the different missing mechanisms and imputation models of missing data. Buuren (2011) [3] supplemented and improved it on this basis and provided a demonstration of R code to make the imputation process more intuitive. At present, scholars have proposed a variety of imputation methods for the missing data problem, which can basically be divided into two

categories: statistical methods and machine learning methods. Statistical methods mostly make assumptions based on the dataset itself, and then use the original dataset to impute the missing data accordingly. Common methods include expectation maximization (EM) imputation [4], regression imputation [5], and multiple imputation [6]. Machine learning methods generally impute the missing dataset by clustering with K-nearest neighbor imputation [7] and K-means imputation [8], and Bayesian networks [9] are represented. In recent years, with the rise of the machine learning boom, Bayesian networks have become a frequently used method to deal with missing data. Mason (2009) [10] proposed the use of a Bayesian method to model nonrandom missing data through the Bayesian missing imputation framework to adjust the missing covariates in longitudinal studies. Erler et al. (2016) [11] suggested that missing values can be imputed in a joint estimation framework using Bayesian methods. Zhang et al. (2017) [12] proposed a missing data processing method based on plain Bayesian and EM algorithms for the software workload problem. Ding (2020) [13] conducted a comparative study of the missing data imputation problem in normal models using Bayesian and jackknife multiple imputation methods, respectively, and concluded that Bayesian imputation results are more accurate.

The Markov chain Monte Carlo (MCMC) algorithm has now become a standard method for parameter estimation in many models. Doğan and Taspinar (2018) [14] (hereinafter DT) performed parameter estimation for the spatial error sample selection model with nonrandom missing data using MCMC; Hajime Seya et al. (2020) [15] improved on the work done by DT, and in parallel, they proposed that MCMC can handle the parameter estimation problem of the spatial lag sample selection model with nonrandom missing data. Although the MCMC can solve Bayesian inference excellently, MCMC may be limited by the speed of convergence and numerical stability when faced with larger models or more data. To address this problem, Rue et al. (2009) [16] proposed an algorithm that combines Laplace approximation with modernized numerical integration under a Bayesian framework—the integrated nested Laplace approximation (INLA)—which can significantly reduce computation time while guaranteeing the accuracy of an MCMC estimation. G'omez-Rubio et al. (2017) [17] described the realization of a new class of latent model in INLA which can be used directly for fitting spatial econometric models, and G´omez-Rubio and Rue (2018) [18] created a new approach combining INLA and MCMC, namely, MCMCINLA, and used it to fit spatial econometric models, linear regression models with missing data in covariates, Bayesian Lasso models, and mixed models. Gomez-Rubio et al. (2019) [19] also redefined the problem of missing values in regression models covariates by latent Gaussian Markov random field (GMRF) for analysis and imputation of missing data, and they applied it to the spatial model and the multiple linear regression model to overcome the problem wherein INLA cannot handle a model with missing values in covariates.

This paper proposes a new MCMCINLA imputation method for missing data and uses the hemorrhagic fever with renal syndrome (HFRS) disease data with random missing in covariates to establish a spatial lag (SLM) latent model to explore the developmental inputs to public health in China before the COVID-19 outbreak, and to provide reference suggestions for the national financial inputs to public health in the post-epidemic era. In addition, the use of the imputation effects of EM, Full Information Maximum Likelihood estimation (FIML), and MCMCINLA method on the missing data are compared to illustrate the effectiveness of the imputation method proposed in this paper.

The paper is structured as follows: Section 2 reviews the different missing mechanisms for the missing data and introduces the SLM latent model with random missing data in covariates; Section 3 gives the proof procedure of the model GMRF structure and describes the process of implementing the MCMCINLA algorithm for the SLM latent model with random missing data in covariates; Section 4 conducts numerical simulations for the model and algorithm proposed in this paper to verify the correctness of the method; Section 5 presents an empirical analysis of the reform adjustment problem of public health development in China in the post-epidemic era as an example; and finally, the conclusions and discussions are given in Section 6.

2. Model Building

2.1. Missing Mechanism

The relationship between missing data and complete data is known as the missing mechanism, which is broadly classified into three types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR occurs when the missing data is random and unrelated to both observed and unobserved data, MAR occurs when the missing data is only related to observed data, and NMAR occurs when the missing data is related to both observed and unobserved data [20].

Furthermore, missing data can be classified into four categories:(Y_{obs}, X_{obs}), (Y_{mis}, X_{obs}), (Y_{obs}, X_{mis}), and (Y_{mis}, X_{mis}), depending on where the missing values are located. For the first case, when neither X nor Y contains missing data, the fit can be performed directly using INLA or MCMCINLA (see, for example, Gómez-Rubio, V. et al. (2017, 2018) [17,18]); for the second case, when the missing data are in the response variables, INLA can use its own properties to predict the missing values by directly calculating the predicted distribution of all the missing data in the response variable (see, for example, Gómez-Rubio, V. et al. (2017) [17]); for the third and fourth case, when the missing data are in the covariates, it is necessary to first define the imputation submodel as latent effects with a GMRF structure to make it suitable for the INLA framework and include the imputation in the main model, and then further perform the model fitting with the help of the INLA, which is also the focus of this paper.

2.2. The SLM Latent Model

In spatial statistics, the SLM has received increasingly wide attention from many scholars. It is mainly used to study the impact of the behavior of adjacent regions on the behavior of other regions of the whole system, expressed formally as:

$$Y = \rho_{Lag} WY + X\beta + e, e \sim MVN(0, \sigma^2 I_n),$$

where Y represents the observation vectors of n different regions, ρ_{Lag} is the spatial autocorrelation parameter, W is the adjacency matrix, β is the coefficient vector of covariates, and the error term e obeys a multivariate normal distribution with the mean 0 and diagonal variance-covariance matrix $\sigma^2 I_n$. We can also shift the term for Y and rewrite the model as:

$$Y = (I_n - \rho_{Lag} W)^{-1}(X\beta + e), e \sim MVN(0, \sigma^2 I_n). \quad (1)$$

The key to enabling the SLM to be implemented under the MCMCINLA is whether the model can be implemented within INLA. Normally, the SLM cannot be fitted directly with INLA, and we need to construct latent classes and redefine the original SLM as a model with GMRF so as to conform to the INLA framework. We can construct a latent class for the SLM as follows:

$$x = (I_n - \rho W)^{-1}(X\beta + e), \quad (2)$$

where x denotes the vector of n random effects, ρ is the spatial autocorrelation parameter, W is the weight matrix, $X = (X_{mis}, X_{obs})$ are the covariates with random missing data, β is the coefficient vector of covariates, and the error term e obeys an independent Gaussian distribution with the mean 0 and precision matrix τI_n.

Using the constructed latent class of Equation (2) to rewrite the SLM model, we can obtain the SLM latent model:

$$Y = x + \varepsilon = (I_n - \rho W)^{-1}(X\beta + e) + \varepsilon, \quad (3)$$

where ε is a small error term that is used to fit the model.

3. Algorithm Description

The INLA algorithm mainly targets models with structured additive regression models with a latent random field of GMRF. The premise of using the MCMCINLA fitting model is that the model needs to conform to the INLA framework, i.e., it has a GMRF structure in order to be solved. Therefore, this section first explains the GMRF structure of the main model and the imputation model, and then describes the implementation process of the MCMCINLA algorithm.

3.1. Proof of GMRF Structure

3.1.1. GMRF Structure of the Main Model

If we assign a Gaussian prior with a zero mean and the precision matrix Q to β in Equation (2), the e obeys a Gaussian distribution with a zero mean and the precision matrix τI_n, with τ as a precision parameter. Then, INLA will want to obtain the joint distribution $\pi(x, \beta)$ of x and β. By Bayes' theorem, we have:

$$\pi(x, \beta) = \pi(x|\beta)\pi(\beta),$$

and, by definition,

$$\begin{gathered}
E(\beta) = 0, \\
Prec(\beta) = Q, \\
J = E(x|\beta) = (I_n - \rho W)^{-1} X\beta, \\
var(x|\beta) = var\left[(I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1} e \middle| \beta\right] \\
= (I_n - \rho W)^{-1} var(e|\beta)((I_n - \rho W)^{-1})' \\
= (I_n - \rho W)^{-1} \tfrac{1}{\tau} I_n ((I_n - \rho W)^{-1})' \\
= \tfrac{1}{\tau}(I_n - \rho W)^{-1}(I_n - \rho W')^{-1}, \text{ and} \\
K = Prec(x|\beta) = \frac{1}{var(x|\beta)} = \tau(I_n - \rho W')(I_n - \rho W).
\end{gathered}$$

Thus,

$$\begin{gathered}
\pi(x, \beta) = \pi(x|\beta)\pi(\beta) \\
\propto \exp\left\{-\tfrac{1}{2}(x - J)'K(x - J)\right\} \exp\left\{-\tfrac{1}{2}\beta' Q\beta\right\} \\
= \exp\left\{-\tfrac{1}{2}(x'Kx - x'KJ - J'Kx + J'KJ + \beta' Q\beta)\right\} \\
= \exp\left\{-\tfrac{1}{2}(x, \beta)' P(x, \beta)\right\},
\end{gathered} \tag{4}$$

where P is the precision matrix of (x, β) with the structure:

$$\begin{aligned}
P &= \begin{pmatrix} K & -K(I_n - \rho W)^{-1} X \\ -X'(I_n - \rho W')^{-1} K & Q + \tau X'X \end{pmatrix} \\
&= \begin{pmatrix} \tau(I_n - \rho W')(I_n - \rho W) & -\tau(I_n - \rho W')(I_n - \rho W)(I_n - \rho W)^{-1} X \\ -X'^{(I_n - \rho W')^{-1}} \tau(I_n - \rho W')(I_n - \rho W) & Q + \tau X'X \end{pmatrix} \\
&= \begin{pmatrix} \tau(I_n - \rho W')(I_n - \rho W) & -\tau(I_n - \rho W')X \\ -\tau X'^{(I_n - \rho W)} & Q + \tau X'X \end{pmatrix}.
\end{aligned} \tag{5}$$

This shows that for the given hyperparameters τ and ρ, the mean and precision matrix of (x, β) are 0 and P; that is, the constructed latent class x has a GMRF structure with the mean 0 and precision matrix P, which is consistent with the INLA framework, and thus can be implemented with the help of MCMCINLA.

3.1.2. GMRF Structure of the Imputation Model

For the covariates $X = (X_{mis}, X_{obs})$ containing missing data, X_{mis} denotes the part with the missing values and X_{obs} denotes the part that is observable. We define the imputation

submodel of the covariates as the latent effect $x' = (x'_{mis}, x'_{obs})$, and set different priors for x'_{mis} and x'_{obs}, respectively, where the observed term x'_{obs} in the latent effect is set to the mean equal to X_{obs} and a high precision matrix (e.g., here, we take it as $5^{10}I$) so that its variance is very small, i.e., it makes the observed term x'_{obs} in the latent effect as infinitely close to the observed covariate data X_{obs} as possible [19]. A spatial model with the mean μ_c and precision matrix Q_c is built for the missing term x'_{mis} in the latent effect to impute the missing covariate data, and the procedure is as shown below.

The imputation model, that is, given the observed data X_{obs} and the hyperparameter θ, provides the distribution of the missing values X_{mis}; hence, we have:

$$\pi(X_{mis}|X_{obs}) = \int \pi(X_{mis}, \theta|X_{obs})d\theta = \int \pi(X_{mis}|X_{obs}, \theta)\pi(\theta|X_{obs})d\theta. \tag{6}$$

Since $\pi(\theta|X_{obs})$ is only related to the observed data X_{obs}, it can be considered as a priori of θ, which can further rewrite $\pi(\theta|X_{obs})$ as:

$$\pi(\theta|X_{obs}) \propto \pi(X_{obs}|\theta)\pi(\theta). \tag{7}$$

In the general case, we assume that the covariates $X = (X_{mis}, X_{obs})$ follow a multivariate normal distribution of:

$$X|\theta \sim Normal\left(\begin{pmatrix}\mu_{mis}\\ \mu_{obs}\end{pmatrix}, \begin{pmatrix}Q_{mis,mis} & Q_{mis,obs}\\ Q_{obs,mis} & Q_{obs,obs}\end{pmatrix}\right),$$

and then it defines that its imputation model will obey:

$$X_{mis}|X_{obs}, \theta \sim Normal(\mu_c, Q_c),$$

where $\mu_c = \mu_{mis} - Q^{-1}_{mis,mis}Q_{mis,obs}(X_{obs} - \mu_{obs})$ and $Q_c = Q_{mis,mis}$.

Considering that the covariates in the SLM are spatially correlated, the missing data in the covariates are imputed using a conditional autoregressive specification (CAR) spatial model-based approach. Under CAR, the mean of the model is set as $\mu = \alpha^T$ and the precision matrix as $Q = \tau(I - \rho W)$, where α is the intercept of the linear predictor, τ is the precision parameter, ρ is the spatial autocorrelation parameter, W denotes the adjacency matrix, and the hyperparameter θ at this time consists of τ, ρ, and α.

Substituting the values of μ and Q, the covariates X follow a multivariate normal distribution of:

$$\begin{aligned} X|\theta &\sim Normal\left(\begin{pmatrix}\mu_{mis}\\ \mu_{obs}\end{pmatrix}, \begin{pmatrix}Q_{mis,mis} & Q_{mis,obs}\\ Q_{obs,mis} & Q_{obs,obs}\end{pmatrix}\right)\\ &= Normal\left(\begin{pmatrix}\alpha^T_{mis}\\ \alpha^T_{obs}\end{pmatrix}, \begin{pmatrix}\tau(I_{mis} - \rho W_{mis,mis}) & -\tau\rho W_{mis,obs}\\ -\tau\rho W_{obs,mis} & \tau(I_{obs} - \rho W_{obs,obs})\end{pmatrix}\right), \end{aligned} \tag{8}$$

and the imputation model based on the spatial model will obey:

$$X_{mis}|X_{obs}, \theta \sim Normal(\mu_c, Q_c), \tag{9}$$

where $\mu_c = \alpha^T_{mis} - (I_{mis} - \rho W_{mis,mis})^{-1}(-\rho W_{mis,obs})(X^T_{obs} - \alpha^T_{obs})Q_c = \tau(I_{mis} - \rho W_{mis,mis})$. The hyperparameters $\tau, \rho,$ and α are obtained from $\pi(\tau, \rho, \alpha|X_{obs})$, which is proportional to $\pi(X_{obs}|\tau, \rho, \alpha)\pi(\tau, \rho, \alpha)$.

Thereby, the latent effect $x' = (x'_{mis}, x'_{obs})$ will obey the following multivariate normal distribution:

$$x'|\theta \sim Normal\left(\begin{pmatrix}\mu_c\\ X_{obs}\end{pmatrix}, \begin{pmatrix}Q_c & 0\\ 0 & 5^{10}I\end{pmatrix}\right), \tag{10}$$

where μ_c and Q_c are taken as shown in Equation (9). This shows that the latent effect x' that we defined for the imputation model has a GMRF structure that can be applied to the INLA framework, and thus it can be implemented with the help of MCMCINLA.

3.2. Implementation of the MCMCINLA Algorithm

When using MCMCINLA to deal with the SLM latent model with random missing data in covariates, it is first necessary to define an imputation submodel to impute the covariates with missing values so as to substitute the complete covariate data back into the SLM, and then use MCMCINLA to fit the model for estimation. The core of the MCMCINLA algorithm parameter estimation lies in dividing the estimated parameters into two groups: the first group is estimated using the Metropolis–Hastings (MH) algorithm in MCMC and the second group is estimated using the Bayesian model averaging (BMA) algorithm [21] in INLA. Therefore, the whole algorithmic process is carried out in three main steps:

1. Imputation of the missing covariates $X = (X_{mis}, X_{obs})$ using INLA In this paper, we chose to impute the missing covariates $X = (X_{mis}, X_{obs})$ using a CAR space model-based approach, which first requires the definition of the mean, precision, hyperparameters, and prior of each hyperparameter in the latent effect x'. The key codes are as follows:

$$\begin{array}{c} \textit{inla.rgeneric.micar.model = function(cmd = c(“graph”, “Q”, “mu”,} \\ \textit{“initial”,”log.norm.const”, “log.prior”, “quit”), theta = NULL),} \end{array}$$

which define the spatial weight matrix (“*graph*”), precision matrix (“*Q*”), mean (“*mu*”), hyperparameter prior (“*log.prior*”), etc. in the latent effect x' by the *inla.rgeneric.micar.model*, respectively, in preparation for defining the imputation model;

$$\textit{model = inla.rgeneric.define(inla.rgeneric.micar.model, debug = TRUE, n, x, idx.mis, W),}$$

which defines the imputation model via the *inla.rgeneric.define()*, where *n* denotes the total number of indices, *x* denotes the covariates containing the missing data, *idx.mis* denotes the index of each missing datum, and *W* is the spatial weight matrix; and

$$\textit{inla(x\sim0 + f(idx, model = model),data, \ldots),}$$

where, finally, INLA is used to complete the fit of the imputation model, and where *f(idx, model =model)* represents the spatial effect of the imputation model. The covariate X after imputation via INLA is incorporated into the SLM, at which time there are three parameters to be estimated in the model, namely, the spatial autocorrelation parameter ρ, the covariate coefficient β, and the error term precision τ. Here, we use MH estimation for the parameter ρ and BMA estimation for the remaining parameters.

2. Estimation of the spatial autocorrelation parameter ρ using MH The estimation of ρ using MH is carried out in three main steps, as follows:
   - Step 1: Assume that starting from the initial point $\rho^{(1)} = 0$, the model is fitted conditionally with $\rho^{(1)}$ to obtain $\pi(Y\left|\rho^{(1)}\right.)$, $\pi(\beta\left|Y,\rho^{(1)}\right.)$, and $\pi(\tau\left|Y,\rho^{(1)}\right.)$;
   - Step 2: Use the MH algorithm to sample from the posterior of ρ, propose a new point ρ^* for ρ by proposing the distribution $q(\cdot\left|\rho^{(j-1)}\right.)$, fit the model conditionally on ρ^* to obtain $\pi(Y|\rho^*)$, $\pi(\beta|Y,\rho^*)$, and $\pi(\tau|Y,\rho^*)$, and calculate $\pi(\rho^*)$, $q(\rho^*\left|\rho^{(j)}\right.)$, and $q(\rho^{(j)}\left|\rho^*\right.)$;
   - Step 3: Calculate the acceptance probability α of ρ^* and determine whether the proposal is acceptable (or not), where:

$$\alpha = \min\left\{1, \frac{\pi(\beta|Y,\rho^*)\pi(\tau|Y,\rho^*)\pi(Y|\rho^*)\pi(\rho^*)q(\rho^{(j)}\left|\rho^*\right.)}{\pi(\beta\left|Y,\rho^{(j)}\right.)\pi(\tau\left|Y,\rho^{(j)}\right.)\pi(Y\left|\rho^{(j)}\right.)\pi(\rho^{(j)})q(\rho^*\left|\rho^{(j)}\right.)}\right\}. \quad (11)$$

If the proposal is accepted, then $\rho^{(j+1)} = \rho^*$ with $\pi(\beta\left|Y,\rho^{(j+1)}\right.) = \pi(\beta|Y,\rho^*)$ and $\pi(\tau\left|Y,\rho^{(j+1)}\right.) = \pi(\tau|Y,\rho^*)$; otherwise, $\rho^{(j+1)} = \rho^{(j)}$, and $\pi(\beta\left|Y,\rho^{(j+1)}\right.) = \pi(\beta\left|Y,\rho^{(j)}\right.)$

and $\pi(\tau\Big|Y,\rho^{(j+1)}) = \pi(\tau\Big|Y,\rho^{(j)})$. This iterative process is executed until the end of the estimation.

The key code for the process is as follows:

fit.inla < −*slm.inla (formula, d, W, rho, . . .).*

in which we fit the SLM by defining *fit.inla()* to prepare for subsequent sampling, where *formula* is a formula with the response variable and the fixed effects, *d* represents the complete data set after imputation, *W* is the spatial weight matrix as above, and *rho* is the spatial autocorrelation parameter; and

INLAMH (d.mis, fit.inla, d.init, rq, dq, prior, n.sim = 200, n.burnin = 100, n.thin = 1, . . .),

in which we fit the parameters ρ by *INLAMH()*, where *fit.inla* is the model fitted earlier, and here, namely, the SLM, and *d.init*, *rq*, *dq*, and *prior* are all basic settings in the MH algorithm which denote the initial value of sampling, the proposed distribution, the density function of the proposed distribution, and the prior distribution, respectively.

3. Estimation of the coefficient β and precision τ using BMA For the parameter β and hyperparameter τ, the conditional margins $\pi(\beta\Big|Y,\rho^{(j)})$ and $\pi(\tau\Big|Y,\rho^{(j)})$ generated by the MH algorithm in each iteration can be obtained using BMA, and further, the posterior margins of β and τ are derived by integrating over ρ, namely:

$$\pi(\beta|Y) = \int \pi(\beta|\rho,Y)\pi(\rho|Y)d\rho = \frac{1}{N}\sum_{j=1}^{N}\pi(\beta\Big|Y,\rho^{(j)}) \text{ and} \tag{12}$$

$$\pi(\tau|Y) = \int \pi(\tau|\rho,Y)\pi(\rho|Y)d\rho = \frac{1}{N}\sum_{j=1}^{N}\pi(\tau\Big|Y,\rho^{(j)}). \tag{13}$$

The key code for the process is as follows:

INLABMA:::fitmatrixBMA(l.models, ws, "summary.fixed")

INLABMA:::fitmatrixBMA(l.models, ws, "summary.hyperpar")

which calculate the posterior margins for fixed effects and hyperparameters by the *fitmatrixBMA()* function in the INLABMA package and where *l.models* are the INLA models to be averaged and *ws* is the weight vector.

4. Simulation Study

4.1. Data Generation

Assuming that $(I_n - \rho_{Lag}W)$ is invertible, we consider the numerical simulation process of the SLM latent model for covariates with random missing data as follows:

$$Y = (I_n - \rho_{Lag}W)^{-1}(X_1\beta_1 + X_2\beta_2 + e) + \varepsilon, \tag{14}$$

where Y is the response variable, X_1 and X_2 are the covariates, and β_1 and β_2 are the coefficients corresponding to X_1 and X_2, respectively, I_n is an $n \times n$ unit matrix, W is an $n \times n$ spatial weight matrix, e is a random error term with $e \sim N(0,\sigma^2)$, and ε is a fitting error term.

The specific simulation data are taken as follows:

- In the main model, $X_1 \sim U(0,1)$ and $X_2 \sim U(0,1)$, $\beta_1 = 0.3$ and $\beta_2 = 0.5$, $\rho_{Lag} = 0.9$, and Y is generated by Equation (14) and we randomly remove 15% data as the missing values in X_1 using the MAR mechanism;
- In the imputation model, $\alpha = 0.5$ and $\rho = 0.2$; and

- The spatial weight matrices W in both the main model and the imputation model are selected as Queen-type adjoints, created in the regular lattice, and all error terms in the model obey the normal distribution $N(0, 0.5^2)$, taking the sample size n = 250, pre-burn simulation of 20 times, interval rejection after pre-burn to keep one of the 5 iterations, and, finally, a total of 80 iterations of simulation are completed.

4.2. Fitting Effect Evaluation Indicators

The mean square error (MSE) value is used to reflect the difference between estimates and estimates, the deviance information criterion (DIC) value is used to measure the fit of the Bayesian model, and the accuracy is used to measure the prediction of the missing data. The MSE can evaluate the degree of change of the data, and the smaller the value of the MSE, the higher the accuracy of the data; the DIC criterion can weigh the complexity of the estimated model and the goodness of the model fit, and the smaller the value, the better the model fit; the accuracy is the ratio of the predicted value to the true value, which is used to reflect the similarity between the true value and the predicted value of the missing data and capture the missing information, and the larger the value, the more accurate the prediction result.

4.3. Results Analysis

Using MCMCINLA to first impute the missing data in the covariates by the method based on the CAR spatial model, and then fit the estimates to the SLM latent model, the results of each simulation are shown in Tables 1 and 2, the fitted curves of ρ_{Lag}, β_1, and β_2 and τ and the prediction comparison plots of the missing data are shown in Figures 1–3.

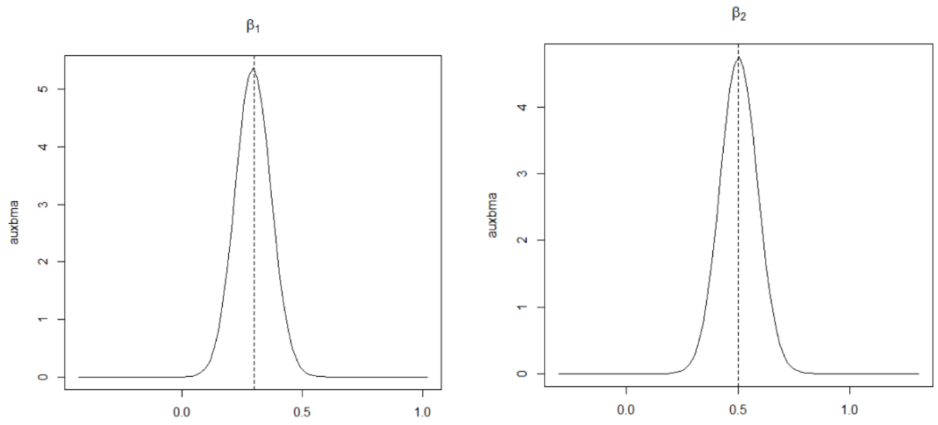
Table 1. Parameter estimation results and model fitting results in the simulation.

	β_1	β_2	ρ_{Lag}	τ
Mean	0.2970	0.5040	0.9153	4.1133
Standard Deviation	0.0744	0.0840	0.0210	0.3036
95% Credible interval	(0.2463, 0.3466)	(0.4468, 0.5599)	(0.8711, 0.9432)	(3.9036, 4.3127)
MSE	2.25×10^{-8}	4×10^{-8}	5.929×10^{-7}	3.209×10^{-5}
DIC	12.7841			

According to the results in Table 1, it can be found that the mean of each parameter β_1, β_2, ρ_{Lag}, and τ estimated in the SLM latent model are very close to the true values and the MSE and the DIC are small, indicating that the parameter estimation and model fitting of this estimation method are good. In addition, Table 2 shows the imputation results for each missing datum in the simulation, where "Mean" represents the mean of the predicted value of the missing data, "95%CI" denotes the 95% credible interval of the predicted value of the missing data, "True value" represents the true value of the missing data, and "Accuracy" is the ratio of "Mean" to "True value". As shown in Table 2, except for the prediction accuracy of V7 and V34, which is around 85%, the prediction accuracy of each missing datum is basically above 90%, indicating that the imputation accuracy of our method is high and the information contained in the missing data can be effectively mined for research with the help of this method.

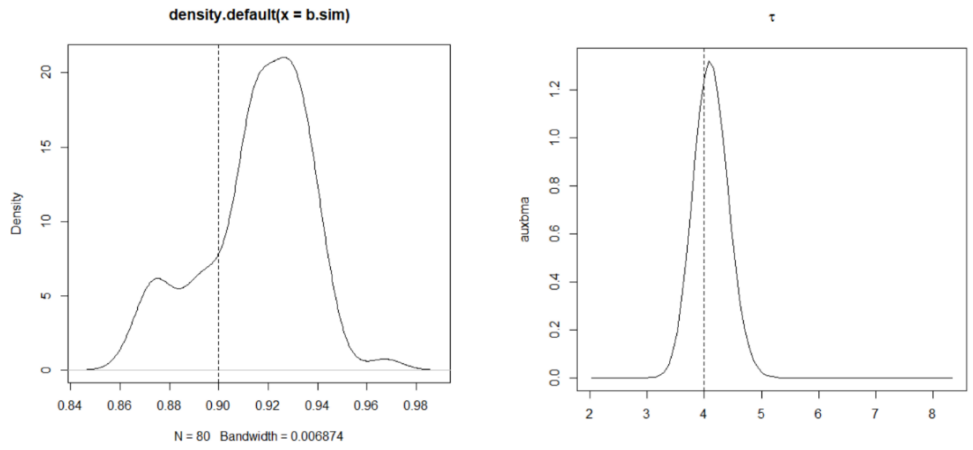
Table 2. Imputation results for each missing datum in the simulation.

	Mean	95%CI	True Value	Accuracy
V1	0.1732	(0.1010, 0.2454)	0.1848	0.9372
V2	0.6821	(0.6502, 0.7140)	0.7023	0.9712
V3	0.5500	(0.4928, 0.6073)	0.5733	0.9593
V4	0.1720	(0.1405, 0.2033)	0.1680	0.9791
V5	0.9578	(0.9284, 0.9872)	0.9438	0.9853
V6	0.9276	(0.9020, 0.9534)	0.9434	0.9832
V7	0.1101	(0.0118, 0.2084)	0.1291	0.8528
V8	0.8169	(0.7875, 0.8461)	0.8334	0.9802
V9	0.4790	(0.4479, 0.5105)	0.4680	0.9770
V10	0.5222	(0.4633, 0.5813)	0.5499	0.9496
V11	0.5312	(0.4862, 0.5763)	0.5526	0.9612
V12	0.2510	(0.1945, 0.3055)	0.2388	0.9513
V13	0.7798	(0.7504, 0.8092)	0.7605	0.9752
V14	0.1673	(0.0911, 0.2435)	0.1808	0.9253
V15	0.4223	(0.3773, 0.4673)	0.4052	0.9595
V16	0.8467	(0.8292, 0.8644)	0.8535	0.9920
V17	0.9601	(0.9390, 0.9810)	0.9763	0.9834
V18	0.2333	(0.2039,0.2627)	0.2258	0.9678
V19	0.4298	(0.4006,0.4594)	0.4448	0.9662
V20	0.0800	(0.0075,0.1558)	0.0749	0.9362
V21	0.6472	(0.6215, 0.6729)	0.6618	0.9779
V22	0.3664	(0.3075, 0.4253)	0.3875	0.9455
V23	0.8468	(0.8269, 0.8667)	0.8368	0.9881
V24	0.1436	(0.0908, 0.1967)	0.1505	0.9541
V25	0.3280	(0.2684, 0.3872)	0.3472	0.9447
V26	0.5998	(0.5814, 0.6186)	0.6114	0.9810
V27	0.3776	(0.3254, 0.4230)	0.3949	0.9561
V28	0.5401	(0.5300, 0.5500)	0.5396	0.9990
V29	0.8579	(0.8455, 0.8703)	0.8616	0.9957
V30	0.7492	(0.7491, 0.7618)	0.7524	0.9957
V31	0.7376	(0.7083, 0.7695)	0.7582	0.9728
V32	0.9165	(0.9066, 0.9264)	0.9238	0.9920
V33	0.3179	(0.2753, 0.3605)	0.3289	0.9665
V34	0.1323	(0.0294, 0.2355)	0.1563	0.8464
V35	0.2290	(0.1604, 0.2976)	0.2151	0.9393
V36	0.1075	(0.0586, 0.1564)	0.1128	0.9530
V37	0.2001	(0.1712, 0.2232)	0.1957	0.9780
V38	0.8436	(0.8235, 0.8640)	0.8576	0.9836
V39	0.2179	(0.1712, 0.2646)	0.2277	0.9569
V40	0.7447	(0.7214, 0.7683)	0.7670	0.9709



(**a**) (**b**)

Figure 1. The fitted plots of β_1 and β_2 in the simulation. (**a**) Posterior density plot for the regression coefficient β_1 in the simulation. (**b**) Posterior density plot for the regression coefficient β_2 in the simulation.



(**a**) (**b**)

Figure 2. Density functional plot of ρ_{Lag} and the fitted plot of τ in the simulation. (**a**) Density functional plot of the spatial autocorrelation parameter ρ_{Lag} in the simulation. (**b**) Posterior density plot for the error term precision τ in the simulation.

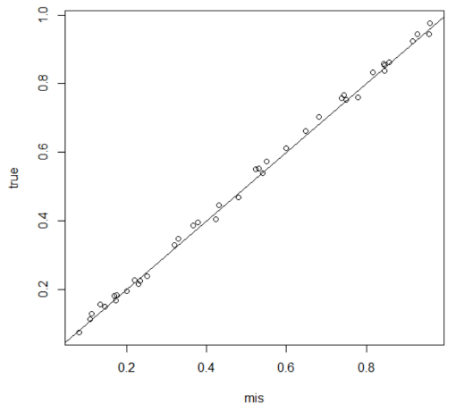


Figure 3. Predicted versus true values for the missing data in the simulation.

Figures 1 and 2 show the fitted curves of the covariate coefficients β_1 and β_2, the error term precision τ, and the density function curve of the spatial autocorrelation parameter ρ_{Lag}, respectively, where the black solid line is the fitted value and the black dashed line perpendicular to the x-axis is the true value. Figure 3 shows the predicted versus true values of the missing data, where the x-axis represents the predicted values of the missing data and the y-axis represents the true values of these missing data that are artificially removed. As can be seen from Figures 1 and 2, except for the peak of ρ_{Lag}, which deviates slightly from the true value, the peaks of all the other parameters are very close to the true value, and all points in Figure 3 basically fall on the line $y = x$. This indicates that the SLM latent model with the covariates containing missing data can be estimated and predicted relatively well using MCMCINLA.

5. Empirical Analysis

5.1. Subject Presents

In the post-epidemic era, outbreaks of various infectious diseases, especially the outbreak of COVID-19, have raised thoughts about the restructuring of the public health system and the corresponding financial investment reforms [22]. How to effectively adjust to the various problems revealed under the epidemic to ensure the continued health of the country's public health has become a key and urgent issue for current research. In the case of medical data, data are often missing during collection and transmission due to clinical trials or equipment failures [23]. It is important to make full use of the large amount of data to mine the important information and make correct predictions of the missing data to make a comprehensive analysis of the research subject.

This paper obtains the economic and disease data of 31 regions in mainland China from 2016 to 2018 through the China Statistical Yearbook (http://www.stats.gov.cn, accessed on 31 December 2020) and the Public Health Science Data Center (http://www.phsciencedata.cn, accessed on 31 December 2018), and uses the national financial investment in public health as the response variable Y and the number of infectious disease cases, economic development, and scientific and technological development as the covariates X to construct the SLM latent model to explore how public health in China can further be developed in the post-epidemic era. The national financial investment in public health is expressed by the indicator "health expenditure in general public budget expenditure by region". Since China has become the most seriously affected country in the world by HFRS [24], the infectious disease studied here is HFRS, and the number of infectious disease cases is expressed by the indicator "number of HFRS cases by region". The economic development is expressed by the indicator "regional gross product". The scientific and technological development is expressed by the indicator "number of research and experimental development R&D projects by region". Using the MAR mechanism to randomly remove 15% of the data from the number of infectious disease cases to construct the SLM latent model for covariates with random missing data, we have:

$$Y = (I_n - \rho_{Lag} W)^{-1}(X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + e) + \varepsilon. \tag{15}$$

If we set the prior distribution for the hyperparameters, the ρ_{Lag} of the main model is assigned a uniform distribution that satisfies $(-2, 1)$, and at this time, the minimum eigenvalue of the adjacency matrix W is -0.5. The prior of the coefficient vector β_i and τ are set as the default values in R-INLA. Then, we assign a Gaussian prior with a zero mean and 0.00005 precision to the intercept α in the imputation model, set in the imputation model as above using the default settings in R-INLA, and assign $logit(\rho)$ to a Gaussian prior with a zero mean and 0.001 precision.

5.2. Exploring the Development of Public Health in China in the Post-Epidemic Era

Using MCMCINLA, the missing data are imputed with the help of *inla.rgeneric.define()*, and the model is fitted using *fit.inla*( ) and the estimated values of the spatial autocorrelation

parameters and the posterior estimates of the regression coefficients of the influencing factors are obtained as shown in Table 3. The regression coefficients of the influencing factors are plotted in Figure 4.

Table 3. Posterior estimates of the regression coefficients of each influencing factor of financial investment in public health.

Variables	Mean	Standard Deviation	95% Credible Interval
The number of infectious disease cases	−0.0421	0.0429	(−0.0711, −0.0135)
Economic development	1.2719	0.1001	(1.2043, 1.3385)
Scientific and technological development	−0.3777	0.0925	(−0.4399, −0.3160)

As estimated by MCMCINLA, the spatial autocorrelation parameter is ρ_{Lag} = 0.6933, which indicates that there is a significant spatial correlation between the national financial investment in public health between regions. To a certain extent, the public health development of a region also has some influence on the public health development of the surrounding regions. From the results of Table 3, it can be found that the posterior mean of economic development is 1.2719, which indicates that the increase of economic level will increase the national financial investment in public health, which can provide the power financial source for the development of public health. However, the posterior mean of the number of infectious disease cases and scientific and technological development are −0.0421 and −0.3777, respectively, indicating that the incidence of infectious diseases and the country's scientific and technological development during 2016–2018 have the opposite effect on the amount of financial investment in public health, which, to some extent, also reveals the shortcomings and inadequacies of the public health management system and the institutional mechanism for epidemic prevention and control. The increase in the number of incidences of general infectious diseases did not attract sufficient attention to the refinement of the CDC structural system, and the rising level of science and technology did not contribute to the problem of updating and replenishing the equipment of specialized public health institutions, leading to the sudden outbreak of the COVID-19 resulting in 2019 a shortage of medical resources and medical personnel and insufficient reserves of premises and materials [25]. Therefore, in the post-epidemic era, it is more important for the government to learn from previous experiences and lessons, establish a reserve mechanism of materials for public health emergencies, build a modern epidemic prevention material reserve system [26], increase the construction of professional public health institutions and the procurement of professional equipment updates, and improve the ability to face public health emergencies in order to quickly and effectively control sudden major infectious epidemics.

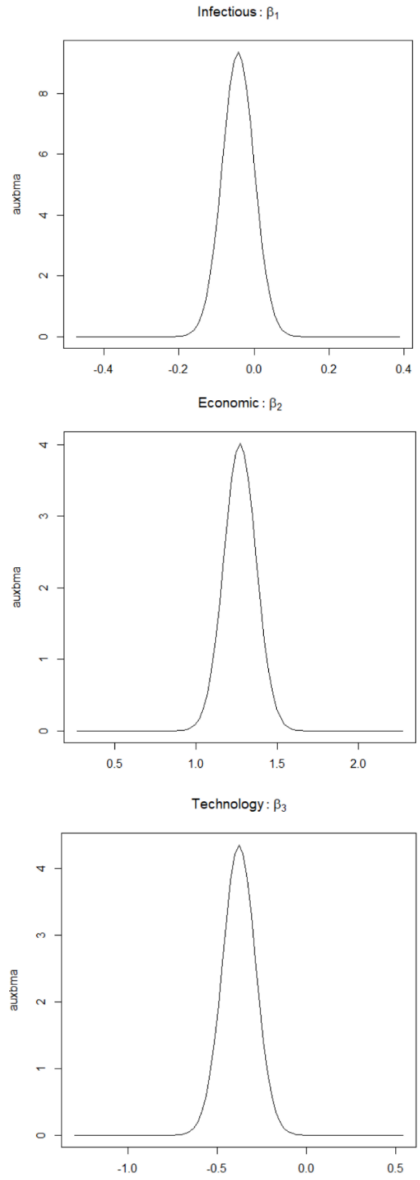


Figure 4. Estimated figures of the regression coefficients of each influencing factor of financial investment in public health.

5.3. Imputation of Missing Predictor Values

Using the CAR spatial model-based approach to impute the SLM latent model with the missing data in X_1, we obtain the prediction information of the missing data and comparing the imputed predicted value with the true value to determine the imputation accuracy.

Based on the prediction results of the missing data in Table 4, it can be seen that the predicted means of the missing values are very close to the true values, and the accuracy rates are basically above 90%, indicating that the imputation of the missing covariates using the CAR space model-based approach with the help of MCMCINLA works well. Meanwhile, observing the data information in the table, we can find that the number of HFRS cases in different regions varies significantly, with a certain spatial heterogeneity and a spatial distribution trend of south-heavy and north-light. The areas with more incidences are Guangdong and Fujian, mostly in the mountainous areas in the south, which have richer vegetation and more precipitation; the areas with fewer incidences are Xinjiang, Tibet, Ningxia, and Gansu, which are mostly in the drier plains. To some extent, increased precipitation promotes the growth of vegetation and crops and provides a more suitable environment for rodents, such as rats, in mountainous areas, creating an increased risk

of HFRS transmission [27]. Therefore, for HFRS, the relevant CDC departments should focus on the prevention and control of the epidemic in the southern region in the future to effectively control the further spread of the epidemic.

Table 4. Missing data prediction results in the number of infectious disease cases (MCMCINLA).

	Mean	95%CI	True Value	Accuracy
Ningxia (2016)	4.3678	(2.8478, 5.8889)	4	0.9157
Tibet (2016)	1.1752	(0.1642, 2.1894)	1	0.8509
Hubei (2016)	248.4683	(226.5048, 270.4318)	236	0.9498
Guangdong (2016)	436.1809	(391.1957, 481.1663)	410	0.9399
Gansu (2016)	20.4501	(11.3075, 29.5927)	19	0.9290
Hebei (2016)	456.7421	(406.4227, 510.0615)	434	0.9502
Guangxi (2017)	14.9736	(9.7006, 20.2477)	14	0.9349
Beijing (2017)	7.7038	(4.8404, 10.5663)	7	0.9086
Xinjiang (2018)	1.1431	(0.0471, 2.2392)	1	0.8748
Shanxi (2018)	22.5906	(14.3455, 30.8357)	21	0.9295
Tibet (2018)	1.1620	(0.1601, 2.1639)	1	0.8605
Shandong (2018)	1268.9431	(1025.2239, 1512.6623)	1218	0.9598
Fujian (2018)	447.8379	(408.5648, 487.1110)	430	0.9601
Tianjin (2018)	21.0547	(12.3284, 29.7810)	20	0.9499
Chongqing (2018)	13.0454	(7.7685, 18.3223)	12	0.9198

5.4. Comparison of Different Imputation Methods

With the development of statistical techniques, a series of model-based missing data processing methods, such as the maximum likelihood estimation, have received increasing attention from academics [28]. They mainly include EM and FIML, which have the advantages of convenient operation and more applicable models.

EM is an iterative algorithm that processes the missing data by calculating the maximum likelihood, and its imputation of the missing values can be achieved by continuous iteration once the initial values of the estimated parameters are given [29]. The principle of FIML is to model the available data using a "one-step" operation and estimate the parameters using a likelihood function; thus, the missing data imputation and parameter estimation processes are implemented simultaneously [30]. The missing data imputation process of the EM and FIML methods is done with the help of the TestDataImputation package and GDINA packages, respectively. The EM and FIML are imputed separately for the missing infectious disease case data and the results are obtained as shown in Tables 5 and 6. The comparison shows that all three imputation methods can obtain more accurate estimates, and the results obtained by the different methods do not differ much. Relatively speaking, the imputation accuracy of MCMCINLA is slightly better than EM, and the FIML method ranks last; however, in terms of computational speed, both MCMCINLA and EM need to perform imputation before completing the estimation, and the estimation of ρ when MCMCINLA needs to be performed with the help of MH sampling, which takes a long time, takes about 0.5 h, while the FIML method can obtain both the imputed

and estimated values at the same time, which is more efficient. In addition to using the accuracy to evaluate the imputation effect of each missing datum from the individual point of view, we can also select three evaluation indicators: MSE, mean absolute percentage error (MAPE) and Pearson correlation coefficient (r) to compare the imputation performance of the three methods from the global point of view [31]. The results are shown in Table 7. MSE can be used to measure the absolute deviation between the imputed value of the missing data and the true value, MAPE can be used to measure the relative error between the imputed value and the true value, both of which are as small as possible, and r can be used to illustrate the correlation between the imputed value and the true value, and the larger the r value, the better the fitting effect. From the results in Table 7, it can be found that the MSE and MAPE values of the three algorithms of MCMCINLA, EM, and FIML are all small, and the r values are all large, indicating that the three algorithms can obtain ideal imputation effects for missing data, and in comparison, the imputation performance of the MCMCINLA proposed in this paper is more prominent.

Table 5. Missing data prediction results in the number of infectious disease cases (EM).

	Mean	95%CI	True Value	Accuracy
Ningxia (2016)	4.3922	(2.8294, 5.9552)	4	0.9107
Tibet (2016)	1.1610	(0.1573, 2.1647)	1	0.8613
Hubei (2016)	255.4665	(212.1427, 298.7903)	236	0.9238
Guangdong (2016)	444.9267	(372.6629, 517.1905)	410	0.9215
Gansu (2016)	20.6387	(10.9749, 30.3025)	19	0.9206
Hebei (2016)	457.8542	(401.1974, 514.5110)	434	0.9479
Guangxi (2017)	15.3475	(9.0223, 21.6727)	14	0.9122
Beijing (2017)	7.9113	(4.1923, 11.6303)	7	0.8848
Xinjiang (2018)	1.1326	(0.0012, 2.2640)	1	0.8829
Shanxi (2018)	22.9332	(13.5426, 32.3238)	21	0.9157
Tibet (2018)	1.1372	(0.1477, 2.1266)	1	0.8793
Shandong (2018)	1292.1705	(1009.3498, 1574.9912)	1218	0.9426
Fujian (2018)	453.3473	(396.4419, 510.2527)	430	0.9485
Tianjin (2018)	21.4799	(11.3123, 31.6475)	20	0.9311
Chongqing (2018)	13.3288	(7.0392, 19.6184)	12	0.9003

Table 6. Missing data prediction results in the number of infectious disease cases (FIML).

	Mean	95%CI	True Value	Accuracy
Ningxia (2016)	4.4400	(2.7943, 6.0857)	4	0.9009
Tibet (2016)	1.1219	(0.1621, 2.0817)	1	0.8913
Hubei (2016)	256.9686	(210.1379, 303.7993)	236	0.9184
Guangdong (2016)	450.1042	(368.4492, 531.7592)	410	0.9109
Gansu (2016)	20.9320	(10.8787, 30.9853)	19	0.9077
Hebei (2016)	468.7837	(392.1163, 545.4583)	434	0.9258
Guangxi (2017)	15.5936	(8.9372, 22.2501)	14	0.8978
Beijing (2017)	7.8431	(4.3958, 11.2904)	7	0.8925
Xinjiang (2018)	1.1282	(0.0297, 2.2267)	1	0.8863
Shanxi (2018)	23.2970	(12.4424, 34.1517)	21	0.9014
Tibet (2018)	1.1462	(0.1392, 2.1532)	1	0.8724
Shandong (2018)	1312.6414	(1004.2167, 1621.0661)	1218	0.9279
Fujian (2018)	457.3981	(391.2553, 523.5409)	430	0.9401
Tianjin (2018)	21.7936	(11.0744, 32.5128)	20	0.9177
Chongqing (2018)	13.5379	(7.0102, 20.0656)	12	0.8864

Table 7. Imputation performance comparison of MCMCINLA, EM, and FIML.

Evaluation Indicators	MCMCINLA	EM	FIML
MSE	4.27×10^{-5}	9.44×10^{-5}	2.153×10^{-4}
MAPE	3.498×10^{-5}	7.2×10^{-5}	1.058×10^{-4}
r	0.9222	0.9122	0.9051

6. Conclusions and Discussion

Medical data are often missing during collection and transmission due to clinical trials or equipment failures. In this paper, we investigate the problem of parameter estimation for SLM when the covariates contain random missing data, and we propose a new imputation method that uses MCMCINLA to get not only accurate parameter estimates for the model, but also good imputation results for the missing data. Taking the economic and HFRS disease data of mainland China from 2016–2018 as an example for empirical analysis, the study found that the HFRS epidemic in China had obvious spatial heterogeneity and a south-heavy and north-light distribution trend. Before the outbreak of COVID-19 in 2019, China's public health management system had certain problems, and the state's financial investment in public health did not receive certain attention. Compared with EM and FIML, the predicted values of the missing data obtained using our method are closer to the true values. Therefore, in the future, the relevant CDC departments should focus their attention on the south or areas with a high incidence of epidemics in wetter climatic conditions and do a good job of the research and diagnosis of HFRS epidemics. In the post-epidemic era, the government should play a leading role, actively learn from previous experiences and

lessons, establish a mechanism for stockpiling materials for public health emergencies, build a modern epidemic prevention material reserve system, and increase the construction of professional public health institutions and the procurement of professional equipment updates in order to quickly and effectively control sudden major infectious epidemics.

Since this paper mainly focuses on model imputation, fitting, and estimation with random missing data, the other two mechanisms are not explored in depth. It can continue to be extended in the future to study how to use the method to deal with different imputation models and missing mechanisms.

One of the fundamental aspects of imputation, in addition to the missing pattern (MAR, MCAR, and NMAR), is the percentage of missing data. It would be more intriguing to study how the new algorithm is affected by the percentage of missing data and compare its performance with other algorithms. In this paper, we take the 15% missing data percentage as an example to conduct intensive research. In the future, different levels of missing rates (such as 5%, 10%, 15%, 30%, and 50%) can be set to discuss the performance of the algorithm more comprehensively.

Finally, since in this paper we only design a set of simulation experiments to test the algorithm, objectively speaking, the effectiveness of the algorithm is questionable when it is extended to other situations (for example, if X takes different probability distributions, ρ_{Lag} takes different spatial autocorrelation degrees). Therefore, the test effect of the imputation algorithm in other cases can be further explored in the future.

Author Contributions: Conceptualization, J.T., S.D. and X.H.; methodology, J.T., S.D. and X.S.; software, J.T.; validation, S.D., X.S., H.Z. and X.H.; formal analysis, J.T., H.Z. and X.H.; investigation, J.T.; resources, X.H.; data curation, J.T.; writing—original draft preparation, J.T.; writing—review and editing, J.T., S.D., X.S., H.Z. and X.H.; visualization, J.T., S.D. and X.H.; supervision, X.S., H.Z. and X.H.; project administration, X.S., H.Z. and X.H.; funding acquisition, H.Z. and X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 11961065, and partially supported by the Natural Science Foundation of Xinjiang, grant number 2019D01C045, and the Ministry of education of Humanities and Social Science project, grant number 19YJA910007.

Data Availability Statement: All the data used in this paper can be obtained from the China Statistical Yearbook (http://www.stats.gov.cn, accessed on 31 December 2020) and the Public Health Science Data Center (http://www.phsciencedata.cn, accessed on 31 December 2018).

Acknowledgments: The authors sincerely thank the editor and two reviewers for their valuable comments and suggestions which led to significant improvement of the manuscript. Thanks for the support of the National Natural Science Foundation of China, the Natural Science Foundation of Xinjiang and the Ministry of education of Humanities and Social Science project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Natal, J.; Ávila, I.; Tsukahara, V.B.; Pinheiro, M.; Maciel, C.D. Entropy: From Thermodynamics to Information Processing. *Entropy* **2021**, *23*, 1340. [CrossRef]
2. Little, R.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
3. Buuren, S.V.; Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in r. *J. Stat. Softw.* **2011**, *3*, 1–67. [CrossRef]
4. Ghahramani, Z. Supervised learning from incomplete data via an EM approach. *Adv. Neural Inf. Process. Syst.* **1984**, *6*, 120–127.
5. Annas, S.; Kartikasari, P.; Arisandi, R. Handling Incomplete Data with Regression Imputation. *J. Phys. Conf. Ser.* **2021**, *1752*, 012049. [CrossRef]
6. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: Hoboken, NJ, USA, 1987.
7. Razavi-Far, R.; Cheng, B.; Saif, M.; Ahmadi, M. Similarity-learning information-fusion schemes for missing data imputation. *Knowl. Based Syst.* **2020**, *187*, 104805.1–104805.13. [CrossRef]
8. Raja, P.; Thangavel, K. Soft Clustering Based Missing Value Imputation. In *Convention of the Computer Society of India*; Springer: Singapore, 2016; pp. 119–133.

9. Ye, C.; Wang, H.; Lu, W.; Li, J. Effective Bayesian-network-based missing value imputation enhanced by crowdsourcing. *Knowl. Based Syst.* **2019**, *190*, 105199. [CrossRef]
10. Mason, A.J. *Bayesian Methods for Modelling Non-Random Missing Data Mechanisms in Longitudinal Studies*; Imperial College London: London, UK, 2009.
11. Erler, N.S.; Rizopoulos, D.; Rosmalen, J.V.; Jaddoe, V.W.V.; Franco, O.H.; Lesaffre, E.M.E.H. Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full bayesian approach. *Stat. Med.* **2016**, *17*, 2955–2974. [CrossRef]
12. Zhang, W.; Jiang, Y.; Yin, G.; Yu, L. A software workload missing data processing method based on plain Bayesian and EM algorithms. *Syst. Eng. Theory Pract.* **2017**, *37*, 2965–2974. (In Chinese)
13. Ding, M. A comparison of Bayesian and Jackknife multiple imputation methods for missing data of normal models. *Comput. Technol. Autom.* **2020**, *39*, 119–123. (In Chinese)
14. Doğan, O.; Taşpinar, S. Bayesian Inference in Spatial Sample Selection Models. *Oxf. Bull. Econ. Stat.* **2018**, *1*, 90–121. [CrossRef]
15. Seya, H.; Tomari, M.; Uno, S.; Phillips, A. Parameter estimation in spatial econometric models with non-random missing data. *Appl. Econ. Lett.* **2020**, *28*, 440–446. [CrossRef]
16. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* **2009**, *2*, 319–392. [CrossRef]
17. Gómez-Rubio, V.; Bivand, R.S.; Rue, H. Estimating spatial econometrics models with integrated nested Laplace approximation. *Mathematics* **2017**, *17*, 2044. [CrossRef]
18. Gómez-Rubio, V.; Rue, H. Markov Chain Monte Carlo with the Integrated Nested Laplace Approximation. *Stat. Comput.* **2018**, *28*, 1033–1051. [CrossRef]
19. Gómez-Rubio, V.; Cameletti, M.; Blangiardo, M. Missing Data Analysis and Imputation via Latent Gaussian Markov Random Fields. *arXiv* **2019**, arXiv:1912.10981.
20. Xiong, Z.; Guo, H.; Wu, Y. A review of missing data processing methods. *Comput. Eng. Appl.* **2021**, *14*, 12. (In Chinese)
21. Hoeting, J.; David Madigan, A.R.; Volinsky, C. Bayesian model averaging: A tutorial. *Stat. Sci.* **2001**, *14*, 382–401.
22. Zhong, R.; Chen, L. Study on the reform of public health financial investment in Hunan Province in the context of the COVID-19. *Econ. Res. Ref.* **2021**, *20*, 99–112. (In Chinese)
23. Zhang, H.; Mu, G.; Dang, J.; Li, C.; Liu, J. Missing data filling method in cardiac diagnosis system. *J. Adv. Sci.* **2021**, *41*, 44–49. (In Chinese)
24. Hjertqvist, M.; Klein, S.L.; Ahlm, C. Mortality rate patterns for hemorrhagic fever with renal syndrome caused by Puumala virus. *Emerg. Infect. Dis.* **2010**, *16*, 1584–1586. [CrossRef]
25. Shan, Y.; Kong, F.; Li, J.; Nie, W.; Li, S. Research progress of public health institutional mechanism reform in the late epidemic period. *China Rural. Health Manag.* **2021**, *41*, 581–585. (In Chinese)
26. Hao, S.J.; Xiong, L.P. Normalization of public health system for major infectious disease epidemics. *J. PLA Hosp. Manag.* **2021**, *28*, 1104–1106. (In Chinese)
27. Xiao, H.; Tian, H.; Cazelles, B. Atmospheric moisture variability and transmission of hemorrhagic fever with renal syndrome in Changsha City, Mainland China, 1991–2010. *PLoS Negl. Trop. Dis.* **2013**, *7*, e2260. [CrossRef] [PubMed]
28. Song, Z.; Guo, L.; Zheng, T. Comparison of data processing methods for cognitive diagnostic deficits: Zero replacement, multiple imputation and great likelihood estimation. *J. Psychol.* **2022**, *54*, 426–444. (In Chinese)
29. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Proc. R. Stat. Soc.* **1977**, *39*, 1–22.
30. Graham, W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **2009**, *60*, 549. [CrossRef] [PubMed]
31. Zhang, B.; Song, G. An Empirical Study on the Methods of Missing Data in Large-scale Air Quality Monitoring. *China Environ. Sci.* **2022**, *42*, 2078–2087. (In Chinese)