Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

A Bayesian latent gaussian model with time-varying spatial weight matrices: Application to modeling the impact of multi-pollutant exposure on tuberculosis

I Gede Nyoman Mindra Jaya[a*], Yudhie Andriyana[a], Bertho Tantular[a] and Farah Kristiani[b]

[a]*Department of Statistics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia*
[b]*Department of Mathematics, Parahyangan Catholic University, Jl. Ciumbuleuit No. 94, Hegarmanah, Kec. Cidadap, Kota Bandung 40141, Indonesia*

C H R O N I C L E

Article history:
Received: May 3, 2024
Received in revised format: May 28, 2024
Accepted: July 16, 2024
Available online: July 16, 2024

Keywords:
Latent Gaussian model
Time-varying spatial weight matrices
Monte-Carlo
Air pollutants
Tuberculosis

A B S T R A C T

The main objective of spatiotemporal analysis is to offer precise predictions of outcomes. The objective of this study is to assess the accuracy of the Bayesian Latent Gaussian Model in predicting outcomes by utilizing both time-varying and fixed spatial weight matrices. The results of the Monte Carlo simulation suggest that when there is moderate spatial autocorrelation (between 0.3 and 0.7), it is strongly advised to use a time-varying spatial weight matrix. This approach yields the most precise predictions and minimizes any distortion in parameter estimates. Furthermore, we provide an illustrative case study where we simulate the effects of exposure to multiple pollutants on tuberculosis. The analysis revealed that particulate matter 10 (PM10), nitrogen oxides (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), have a positive influence on the risk of TB, with spatial effects that change over time. The model demonstrates that a rise of 1 mg/m³ in the levels of PM10, NO2, SO2, CO, and O3 is linked to corresponding increases in TB cases by 2.1%, 21.17%, 13.20%, 6.72%, and 6.59%, respectively. NO2 and SO2 have the most significant influence on the risk of tuberculosis (TB). These findings enhance our comprehension of the spatial correlation of TB over time and promote further investigation to determine the most efficacious strategies for mitigating the dissemination of TB.

1. Introduction

Spatiotemporal analysis is an appropriate method for describing the spatial and temporal relationship between TB risk factors (Carrasco-Escobar, Schwalb, Tello-Lizarraga, Vega-Guerovich, & Ugarte-Gil, 2020) (Li, Ge, & Zhang, 2022). The main goal of spatiotemporal modeling is to attain accurate predictions and identify relevant risk factors associated with the relative risk of diseases such as tuberculosis (TB). The utilization of data-driven modeling is crucial for achieving this objective. Defining spatial dependence is a crucial element of spatiotemporal data analysis. Temporal dependence progresses solely in a linear fashion, while spatial dependencies can expand in various directions. The precision of forecasts greatly depends on accurately identifying these spatial interdependencies. Usually, it is assumed that spatial dependence remains consistent over time. Nevertheless, this assumption is excessively inflexible, as spatial dependence is constantly changing and influenced by a multitude of factors (Dubé & Legros, 2013). Lately, there has been significant attention given to the modeling of temporal variations in spatial dependencies. The main topics under discussion are time-varying parameters and time-varying weight matrices related to spatial factors. This research specifically focuses on the latter. The study conducted by (Ou, Zhao, & Wang, 2015) demonstrated, through a Monte Carlo simulation, that spatiotemporal models with spatial weight matrices that vary over time have higher statistical power than models with fixed spatial weight matrices. The Moran's I statistic is employed to assess the spatial interdependence of variables of interest among adjacent regions, where a higher value of Moran's I signifies a more pronounced spatial dependence (Moran, 1950). Nevertheless, the scarcity of spatiotemporal research incorporating time-

* Corresponding author.
E-mail address mindra@unpad.ac.id (I G. N. M. Jaya)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.7.007

varying spatial dependence hampers the attainment of precise prediction outcomes. The objective of this study is to establish that the precision with which we determine evolving dependency structures has a substantial impact on the accuracy of spatiotemporal predictions.

The Bayesian Latent Gaussian Model (LGM) is a frequently employed approach for modeling spatiotemporal data (Blangiardo & Cameletti, 2015). The LGM is extensively utilized in diverse disciplines such as biostatistics, econometrics, epidemiology, Earth sciences, genetics, social sciences, engineering, economics, and regional science. The main objective of utilizing LGM is to simulate and analyze spatial or temporal interdependencies and their interactions. The LGM postulates that a collection of unobserved parameters, referred to as latent variables, conform to a Gaussian distribution (Hrafnkelsson & Bakka, 2023). Latent Gaussian models are a type of Bayesian additive models that have a structured additive predictor. These models have a hierarchical structure, where Gaussian prior distributions are applied to the latent parameters within the Bayesian framework (Rue, Martino, & Chopin, 2009; Hrafnkelsson & Bakka, 2023).

The Bayesian LGM framework operates under the assumption that the data adheres to a particular parametric distribution, such as the number of cases conforming to a Poisson distribution (Rustand, Niekerk, Krainski, Rue, & Proust-Lima, 2024). The unidentified variables are converted using a particular link function and subsequently represented at the hidden level, encompassing both predetermined and random influences that collectively conform to a Gaussian distribution. The Bayesian estimation method's hierarchical structure makes spatiotemporal modeling with fixed and random effects more manageable. The hierarchical Bayesian method is highly efficient in solving intricate models that have a larger number of parameters compared to the sample size (Lin, et al., 2022). This approach is in line with the Bayesian concept, as it considers all parameters as random variables, making it easier to deal with random elements. Recently, the use of Integrated Nested Laplace Approximation (INLA) in the R software platform has become popular for Bayesian LGM modeling (Rue, Martino, & Chopin, 2009; Jaya & Folmer, 2021). The Bayesian Integrated Nested Laplace Approximation (INLA) method is more straightforward to apply and offers significantly faster computational efficiency compared to the Markov Chain Monte Carlo (MCMC) approach.

In order to assess the predictive capability of the spatiotemporal model, we will compare models that utilize a time-varying spatial weight matrix with those that use a fixed spatial weight matrix. The comparison will be carried out using a Monte Carlo simulation, with a specific focus on evaluating the performance differences between the models based on mean square error prediction. Furthermore, we will demonstrate a practical use of the model by evaluating the influence of exposure to multiple pollutants on tuberculosis.

The structure of this paper is as follows: Section 2 presents the detailed Bayesian latent Gaussian model that incorporates spatial weight matrices that change over time. Section 3 of our study focuses on the Monte Carlo simulation, where we compare the predictive performance of the spatiotemporal model using time-varying spatial weight matrices to that using fixed spatial weight matrices. Section 4 utilizes the model to evaluate the influence of exposure to multiple pollutants on tuberculosis. Section 5 presents a detailed analysis, while Section 6 serves as the final section of the paper.

2. Bayesian Latent Gaussian Model with Time-Varying Spatial Weight Matrices

2.1 Bayesian Latent Gaussian Modeling

The Latent Gaussian Model (LGM) is a flexible and extensively utilized structured additive regression model in diverse application domains. The response variable in LGM is assumed to have an exponential family distribution, such as normal, Poisson, binomial, or others. The mean μ_{it} of the response variable is linked to the additive predictor structure η_{it} through a link function $g(.)$, where $g(\mu_{it}) = \eta_{it}$. The structure of the additive predictor η_{it} can include various effects from covariates, as well as structured and unstructured spatial and temporal random effects and their interactions (Rue, Martino, & Chopin, 2009):

$$\eta_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{itk} + \sum_{l=1}^{L} f_l(u_{itl}) + \in_{it}; i = 1, \dots n, \text{and } t = 1, \dots, \mathrm{T} \quad (1)$$

where β_0 represents the intercept or overall mean of the response variable. The terms $\{\beta_k\}$ represent the linear fixed effects of the covariate $\mathbf{x}$, while $\{f_l(.)\}$ represent the unknown smoothing functions of the covariate $\mathbf{u}$. The symbol $\in$ represents the unstructured component. The LGM, or Latent Gaussian Model, is a specific type of Bayesian additive model. It has an additive predictor structure, meaning that it combines multiple predictors in a linear fashion. In this model, the variables β_0, $\{\beta_k\}$, $\{f_l(.)\}$, and $\{\in_{it}\}$ are assumed to be random variables that follow a Gaussian prior distribution.

In order to provide a clear explanation of Bayesian latent models with time-varying spatial weight matrices, we will analyze a straightforward spatiotemporal model that includes K covariates, as well as spatially and temporally structured random effects.

$$\eta_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{itk} + \omega_{i|t} + \nu_t. \tag{2}$$

The function $f_1(.) = \omega_{it}$ represents the spatially structured effects that change over time, while the function $f_2(.) = v_t$ represents the temporally structured effects. The Conditional Autoregressive (CAR) model is frequently employed to represent the spatially structured effects ω. An appropriate model within this framework is the CAR Leroux model (Leroux, Lei, & Breslow, 1999):.

$$\omega_{i|t}|(\boldsymbol{\omega}_{-i|t}, \boldsymbol{W}_t) \sim N\left(\frac{\rho_t \sum_{j=1}^{n} w_{ij|t}\omega_{j|t}}{\rho_t \sum_{j=1}^{n} w_{ij|t} + 1 - \rho_t}, \frac{\sigma_{\omega_t}^2}{\left(\rho_t \sum_{j=1}^{n} w_{ij|t} + 1 - \rho_t\right)}\right) \tag{3}$$

where $\boldsymbol{W}_t$ represents a binary spatial weights matrix of size $n \times n$ at time t. This matrix defines the neighborhood structure of the areas at that specific time. It is commonly constructed as either an inverse distance matrix or a contiguity matrix $\boldsymbol{\omega_t}$. In addition, ρ_t represents the spatial autoregressive coefficient for the spatially structured random effects. One frequently used spatial weight matrix combines the queen and rook spatial weight matrix. The joint prior density function of for time t is:

$$p\left(\boldsymbol{\omega}_t \middle| \sigma_{\omega_t}^2\right) \propto \left(\sigma_{\omega_t}^2\right)^{-\frac{n-1}{2}} exp\left(-\frac{1}{2}\boldsymbol{\omega}_t' \boldsymbol{Q}_{\omega_t} \boldsymbol{\omega}_t\right) \text{ for t} = 1, \ldots, \text{T} \tag{4}$$

with $\mathbf{Q}_{\omega_t} = \left(\frac{1}{\sigma_{\omega_t}^2}\right) \boldsymbol{R}_{\omega_t}$ as the precision matrix with and $\boldsymbol{R}_{\omega_t}$ denotes the $n \times n$ spatial structure matrix at time t defined as:

$$\boldsymbol{R}_{\omega_t} = \begin{cases} \rho_t n_{\text{i}} + (1 - \rho_t), \text{if } i = j \\ -1, \qquad\quad \text{if } i \sim j \\ 0, \qquad \text{otherwise} \end{cases}, \tag{5}$$

where n_{i} represents the number of neighbors in region i, and the notation $i \sim j$ indicates that region i and j are adjacent to each other. The Leroux model is used to produce the spatially structured effect $\boldsymbol{\omega}_t$. The $\mathbf{Q}_{\omega_t}$ matrix can alternatively be defined as:

$$\mathbf{Q}_{\omega_t} = \frac{1}{\sigma_{\omega_t}^2}[\rho_t(\mathbf{D}_t - \mathbf{W}_t) + (1 - \rho_t)\mathbf{I}_n]; t = 1, \ldots, T \tag{6}$$

where $\mathbf{D}_t = diag\left[\sum_j w_{ij}\right]$ represents a diagonal matrix with elements equal to the sum of the weights $w_{ij|t}$ for time t. The symbol $\mathbf{I}_n$ represents an identity matrix with dimensions $n \times n$. The spatial weight matrices $\mathbf{W}_t$ can be created by considering customized neighboring dependencies that can change over time. The selection of the appropriate neighborhood structure is typically based on evaluating the highest level of spatial autocorrelation of the observed response variable using Moran's I (Ou, Zhao, & Wang, 2015). The temporal effects are represented by an autoregressive model.

$$\nu_t = \lambda \nu_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\nu^2) \text{ for every } i \text{ and } t = 2, .., T, \tag{7}$$

with $|\lambda| < 1$ the autoregressive parameter and $\nu \sim N\left(0, \frac{\sigma_\nu^2}{1-\lambda^2}\right)$. The joint prior distribution of $\boldsymbol{\nu} = (\nu_1, .., \nu_T)'$ is defined as:

$$p(\boldsymbol{\nu}|\sigma_\nu^2) \propto (\sigma_\nu^2)^{-\frac{(T-1)}{2}} exp\left(-\frac{1}{2}\boldsymbol{\nu}' \boldsymbol{Q}_{\boldsymbol{\nu}(T\times T)}^{(AR1)} \boldsymbol{\nu}\right), \forall\, i \tag{8}$$

with $\boldsymbol{Q}_{\boldsymbol{\nu}(T\times T)}^{(AR1)} = \left(\frac{1}{\sigma_\nu^2}\right) \boldsymbol{R}_{\nu(T\times T)}^{(AR1)}$ as the precision matrix and $\boldsymbol{R}_{\nu(T\times T)}^{(AR1)}$ as the $T \times T$ temporal trend structure matrix for the $AR1$ prior:

$$\boldsymbol{R}_{\nu(T\times T)}^{(AR1)} = \begin{bmatrix} 1 & -\lambda & \square & \square \\ -\lambda & (1+\lambda^2) & -\lambda & \square \\ \ddots & \ddots & \ddots & \square \\ \square & -\lambda & (1+\lambda^2) & -\lambda \\ \square & \square & -\lambda & 1 \end{bmatrix} \tag{9}$$

The Integrated Nested Laplace Approximation (INLA) is presently the prevailing technique for making inferences in Latent Gaussian Models (LGMs). $\boldsymbol{\Omega}$ represents the vector that includes all latent Gaussian variables, while $\boldsymbol{\Psi}$ represents the vector of hyperparameters, which may or may not be Gaussian. Bayesian Latent Gaussian Models (LGMs) are structured hierarchically, comprising of three distinct stages (Blangiardo & Cameletti, 2015):

1. The first stage involves defining the conditionally independent likelihood function $p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\psi})$.
2. The second stage involves specifying the prior distribution $p(\boldsymbol{\Omega}|\boldsymbol{\psi})$ for the vector parameters $\boldsymbol{\Omega}$, known as the latent field.
3. The final stage involves the distribution of the hyperparameters, known as the hyperprior distribution, that is $\boldsymbol{\Psi} \sim p(\boldsymbol{\Psi})$. Commonly used hyperprior distributions include the inverse Gamma, half-Cauchy, Penalized Complexity, and Uniform distributions.

Following the three stages outlined above, the parameter estimation process begins by computing the posterior distribution using Bayes' theorem.

$$p(\boldsymbol{\Omega}, \boldsymbol{\psi}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\psi})p(\boldsymbol{\Omega}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\mathbf{y}|\boldsymbol{\psi})}. \tag{10}$$

where $p(\boldsymbol{\Omega}, \boldsymbol{\psi}|\mathbf{y})$ represents the joint posterior distribution of the parameter vector $\boldsymbol{\Omega}$ and the hyperparameter vector $\boldsymbol{\psi}$. The term $p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\psi})$ denotes the likelihood function, $p(\boldsymbol{\Omega}|\boldsymbol{\psi})$ signifies the joint prior distribution of $\boldsymbol{\Omega}$ conditional on the hyperparameter $\boldsymbol{\psi}$, $p(\boldsymbol{\psi})$represents the joint hyperprior distribution of $\boldsymbol{\psi}$, and $p(\mathbf{y}|\boldsymbol{\psi})$ indicates the marginal distribution of the observations $\mathbf{y} = (y_{11}, \ldots, y_{nT})'$, given $\boldsymbol{\psi}$.

Because the marginal distribution of the observations $p(\mathbf{y}|\boldsymbol{\psi})$ does not include the parameter vector $\boldsymbol{\Omega}$ to be estimated, and to ensure that the joint posterior distribution is a proper density function, $p(\mathbf{y}|\boldsymbol{\psi})$is treated as a normalizing constant. Consequently, it can be ignored in the inference process. Therefore, the posterior density can be expressed as:

$$p(\boldsymbol{\Omega}, \boldsymbol{\psi}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\psi})p(\boldsymbol{\Omega}|\boldsymbol{\psi}) \times p(\boldsymbol{\psi}). \tag{11}$$

where $\propto$ is called as "proportional to". Assuming y is independent, the likelihood function $p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\psi})$ is:

$$p(\mathbf{y}|\boldsymbol{\Omega}, \boldsymbol{\psi}) = \prod_{l=1}^{n} \prod_{t=1}^{T} p(\mathrm{y}_{it}|\boldsymbol{\Omega}_{it}, \boldsymbol{\psi}) \tag{12}$$

where each data point y_{it} is connected to only one element $\boldsymbol{\Omega}_{it}$. The vector parameter $\boldsymbol{\Omega}$ conditional $\boldsymbol{\psi}$ is assumed to follow a multivariate normal distribution:

$$p(\boldsymbol{\Omega}|\boldsymbol{\psi}) \propto |\mathbf{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\Omega}'\mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\Omega}\right), \tag{13}$$

where $|\mathbf{Q}(\boldsymbol{\psi})|$ denotes the determinant of matrix precision matrix $\mathbf{Q}(\boldsymbol{\psi})$. Finally, the joint posterior distribution of $\boldsymbol{\Omega}$ and $\boldsymbol{\psi}$ for the described hierarchical LGM model is expressed as follows (Rue, Martino, & Chopin, 2009):

$$p(\boldsymbol{\Omega}, \boldsymbol{\psi}|\mathbf{y}) \propto p(\boldsymbol{\psi})|\mathbf{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\Omega}'\mathbf{Q}(\boldsymbol{\psi})\boldsymbol{\Omega} + \sum_{i=1}^{n}\sum_{t=1}^{T} \log(\mathrm{y}_{it}|\boldsymbol{\Omega}_{it}, \boldsymbol{\psi})\right) \tag{14}$$

To obtain the parameter estimate of the vector parameter, we need the posterior marginal distribution of each element of the parameter vector. The marginal posterior distribution of $\boldsymbol{\Omega}_{it}$ is (Rue, Martino, & Chopin, 2009):

$$p(\boldsymbol{\Omega}_{\mathrm{it}}|\mathbf{y}) = \int p(\boldsymbol{\Omega}_{\mathrm{it}}, \boldsymbol{\psi}|\mathbf{y})\mathrm{d}\boldsymbol{\psi} = \int p(\boldsymbol{\Omega}_{\mathrm{it}}|\boldsymbol{\psi}, \mathbf{y})p(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} \tag{15}$$

Following the INLA approach, the marginal posterior distribution is numerically solved through a finite weighted sum (Rue, Martino, & Chopin, 2009):

$$\tilde{p}_{LA}(\boldsymbol{\Omega}_{it}|\mathbf{y}) \approx \sum_{j}^{J} \tilde{p}_{LA}\left(\boldsymbol{\Omega}_{it}|\boldsymbol{\psi}^{(j)}, \mathbf{y}\right) \tilde{p}_{LA}\left(\boldsymbol{\psi}^{(j)}|\mathbf{y}\right)\Delta_j \tag{16}$$

where the index LA denotes the Laplace approximation. The set $\{\boldsymbol{\psi}^{(j)}\}$ represents a collection of values of $\boldsymbol{\psi}$ that are associated with the integration weights Δ_j.

In order to obtain the outcome prediction $\hat{\boldsymbol{y}}$, the predictive distribution $p(\hat{y}\,|\mathbf{y})$ is defined as follows:

$$p(\hat{\boldsymbol{y}}\,|\mathbf{y}) = \int p(\hat{\boldsymbol{y}}\,|\boldsymbol{\psi}, \mathbf{y}\,)p(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} \tag{17}$$

where the marginal posterior distribution for $p(\hat{\boldsymbol{y}}\,|\boldsymbol{\psi},\mathbf{y})$ is obtained by integrating the posterior conditional distribution $p(\hat{\boldsymbol{y}}\,,\boldsymbol{\Omega}|\boldsymbol{\psi},\mathbf{y})$ over $\boldsymbol{\Omega}$:

$$p(\hat{\boldsymbol{y}}|\boldsymbol{\psi},\mathbf{y}) = \int p(\hat{\boldsymbol{y}}\,,\boldsymbol{\Omega}|\boldsymbol{\psi},\mathbf{y})d\boldsymbol{\Omega} = \int p(\hat{\boldsymbol{y}}\,|\boldsymbol{\Omega},\boldsymbol{\psi})p(\boldsymbol{\Omega}|\boldsymbol{\psi},\mathbf{y})d\boldsymbol{\Omega} \tag{18}$$

The numerical method outlined in Eq. (12) can be used to approximate the integrals in Eq. (17) and Eq. (18).

The structured matrix input for the spatiotemporal model with time-varying spatial weight matrix in R-INLA is as follows:

$$\begin{matrix} \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{it} \\ \vdots \\ y_{nt} \end{bmatrix} \\ y_{it} \end{matrix} = \begin{matrix} \begin{bmatrix} 1 & x_{111} & \dots & x_{11K} \\ 1 & x_{121} & \dots & x_{11K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{it1} & x_{itk} & x_{itK} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{nt1} & \dots & x_{ntK} \end{bmatrix} \\ \beta_0 \quad \beta_1 \quad \dots \quad \beta_5 \end{matrix} + \begin{matrix} \begin{bmatrix} 1 & \dots & NA \\ 2 & \dots & NA \\ \vdots & \vdots & \vdots \\ NA & \ddots & NA \\ \vdots & \dots & 1 \\ NA & \dots & 2 \end{bmatrix} \\ id1 \quad \dots \quad idn \end{matrix} + \begin{matrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ t \\ \vdots \\ T \end{bmatrix} \\ \nu_t \end{matrix} \tag{19}$$

3. Monte Carlo Simulation

This section will evaluate the predictive accuracy of models using a time-varying spatial weight matrix and a fixed spatial weight matrix. Specifically, we aim to demonstrate that inaccuracies in determining the spatial weight matrix or treating it as equal across all time periods can lead to significant prediction errors, especially when the level of spatial autocorrelation is higher. We employed Monte Carlo simulation for this assessment.

3.1 Experimental Design

To evaluate the predictive accuracy of the spatiotemporal model with time-varying coefficients, we generated Poisson data using model (Eq. 2), deliberately omitting the temporal effect ν_t to ensure that the model's predictive capability is solely determined by spatial dependencies and a single predictor variable. Let's assume y_{it} follows a Poisson distribution with a linear predictor η_{it} expressed as follows:

$$\eta_{it} = \beta_0 + \beta_1 x_{it1} + \omega_{i|t} \tag{20}$$

We considered nine regular grid areas. To assess spatial dependencies, we employed nine types of spatial weight matrices based on rook contiguity, each representing a different time period (Figure 1).

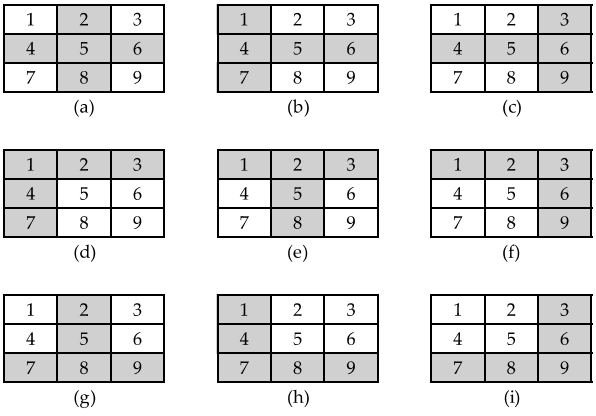

Fig. 1. Nine types of rook spatial contiguity (grey color represents the rook dependencies).

A. Data generation stages:

1. Defining the time-varying rook spatial weight matrices according to Fig. 1

The time-varying spatial weight matrices $\boldsymbol{W}_t$ for $t = 1, \dots, T$ $(T = 9)$ are generated during this phase to identify the dynamic spatial dependencies that exist within the data over time. When considering a fixed spatial weight matrix, we denote it as $\boldsymbol{W}1$ (Fig. 2(a)), a commonly employed matrix typically used for rook contiguity.

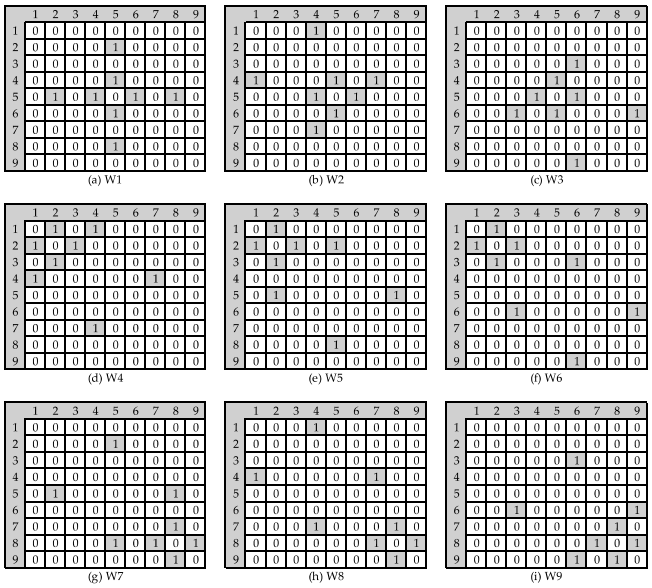

Fig. 2. Tme-varying rook spatial weight matrices

2. Generating spatially structured random effects with the Leroux model. The first step in this process involves calculating the precision matrix, represented as $\mathbf{Q}_{\omega_t}$ (Eq. 6), for $t = 1, .., T$ ($T = 9$). The spatially random effects $\boldsymbol{\omega}_t$ are drawn from a multivariate normal distribution (MVN) with mean 0 and precision matrix $\mathbf{Q}_{\omega_t}$, that is, $\boldsymbol{\omega}_t \sim MVN\left(\mathbf{0}, \mathbf{Q}_{\omega_t}^{-1}\right)$ and $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_T)'$. We define $\rho = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\sigma_{\omega_t}^2 = 1$ for $t = 1, \dots, T$.
3. Generating predictor variables $\mathbf{x}$ of size nT. Predictor variables $\mathbf{x}$ of size nT are generated from a standard normal distribution, denoted as $\mathbf{x} \sim N(0,1)$
4. Setting parameters β_0 and β_1. The parameters β_0 and β_1 are set as $\beta_0 = 1$ and $\beta_1 = 0.5$, respectively.
5. Defining the Linear Predictor η_{it}. The linear predictor η_{it} is defined in (19) :
6. Drawing response variable y_{it} from Poisson distribution with $E(y_{it}) = Var(y_{it}) = \eta_{it}$, that is $y_{it}|\eta_{it} \sim Poisson(\eta_{it})$

B. Model estimation using R-INLA

In starting the modeling process with R-INLA, it is crucial to define hyperpriors for the hyperparameters $\sigma_{\beta_0}^2$, $\sigma_{\beta_1}^2$, and $\left\{\sigma_{\omega_1}^2, \dots \sigma_{\omega_T}^2\right\}$. It is presumed that σ_{α}^2 and $\sigma_{\beta_1}^2$ are constant, with both values set at 10^6. Regarding the hyperparameter set $\left\{\sigma_{\omega_1}^2, \dots \sigma_{\omega_T}^2\right\}$, we assume that the standard deviation $\left\{\sigma_{\omega_1}, \dots, \sigma_{\omega_T}\right\}$ is distributed according to a Half Cauchy distribution. The value of the scale parameter γ is assigned as 25 (Gelman, 2006):

$$p_{HC}(\sigma|\gamma) = \frac{2}{\pi\gamma\left(1 + \left(\frac{\sigma}{\gamma}\right)^2\right)} \text{ for } \sigma = \left\{\sigma_{\omega_1}, \dots, \sigma_{\omega_T}\right\} \tag{21}$$

C. Model evaluation criteria

To assess the predictive performance of the model across time-varying and fixed spatial weight matrices, we employ the Mean Square Error (MSE) criteria, defined as follows:

$$MSE_j = \frac{1}{nTM}\sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{itm} - \exp\left(\hat{\eta}_{itmj}\right)\right)^2; j = 1,2 \tag{22}$$

where $j = 1$ signifies a time-varying spatial weight matrix model, whereas $j = 2$ represents the fixed spatial weight matrix model. Additionally, M denotes the total number of iterations in the Monte Carlo simulation. $M = 1{,}000$ replications are carried out. We ran the model on an Apple M1 Pro with 16 GB of memory using R version 4.3.3 and the INLA package (INLA_24.02.09). The R-code can be accessed at https://github.com/mindra-bit/Time-varying

3.2 Simulation result

Prediction performance

The predictive performance of spatiotemporal Poisson regression models, evaluated using MSE, is depicted in Fig. 3 for both fixed and time-varying spatial weight matrices. A comprehensive summary for each spatial autocorrelation coefficient value is presented in Table 1.

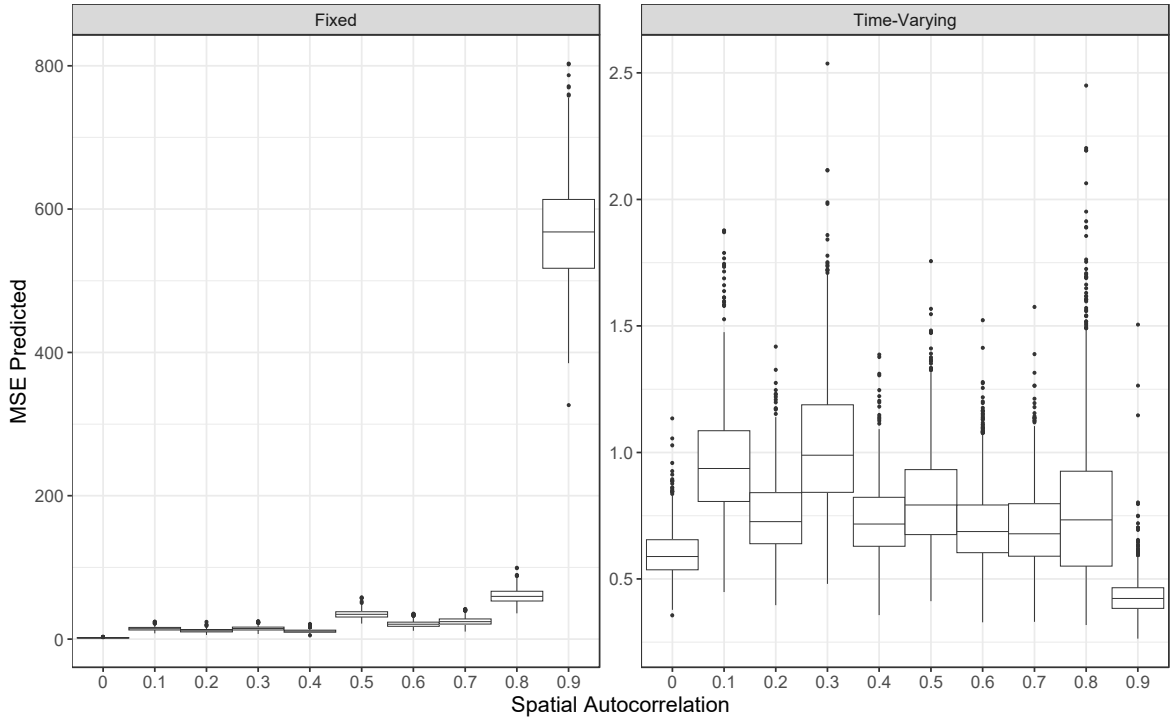

Fig. 3. Evaluating the predictive performance of spatiotemporal Poisson regression models for fixed and time-varying spatial weight matrices using MSE

Table 1
Summary of Predictive Performance for Poisson Spatiotemporal Models with Fixed and Time-Varying Spatial Weight Matrices

Rho	MSE	
	Fixed	Time-Varying
0	1.788	0.599
0.1	14.697	0.963
0.2	11.754	0.749
0.3	15.008	1.031
0.4	11.282	0.732
0.5	34.788	0.817
0.6	21.124	0.710
0.7	24.822	0.704
0.8	60.195	0.789
0.9	568.444	0.434

The results of the comparative analysis, as shown in Table 1 and Fig. 4, demonstrate a significant difference in predictive accuracy between models that use fixed and time-varying spatial weight matrices. More precisely, the model that uses a fixed spatial weight matrix has much higher Mean Squared Error (MSE) values compared to the model that adjusts to changes in the spatial structure over time. The significant differences in mean squared error (MSE) highlight the crucial need to consider changes over time in the relationships between spatially dependent variables when analyzing spatiotemporal data.

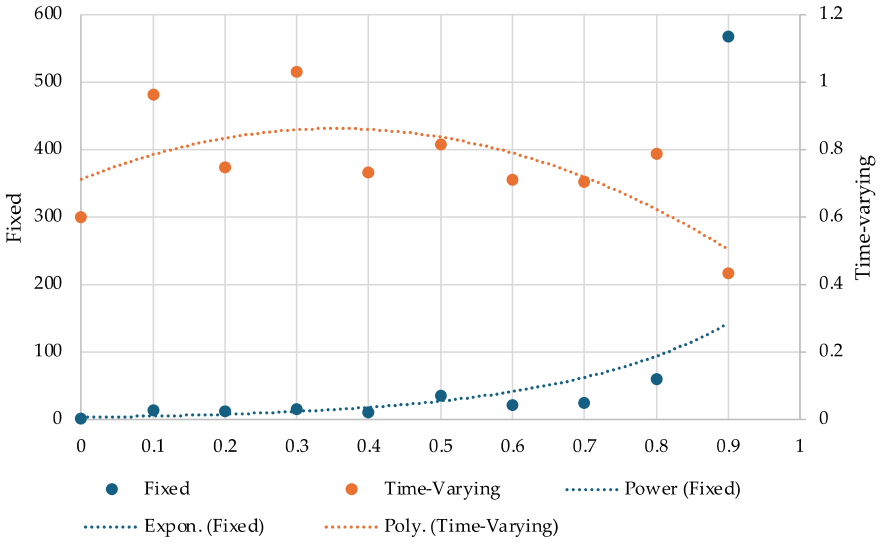

Fig. 4. Variation of MSE with Increasing Spatial Autocorrelation

Fig. 4 shows a clear pattern: as the level of spatial autocorrelation increases, the model's ability to predict outcomes improves significantly when using time-varying weight matrices. This improvement is particularly noticeable at an autocorrelation level of 0.90. In contrast, the model that uses a fixed spatial weight matrix shows a different pattern, where the accuracy of predictions decreases as the level of spatial autocorrelation increases. The observed phenomenon demonstrates a non-linear correlation between the increase in spatial autocorrelation and MSE for the model using time-varying spatial weight matrices. The fluctuating impact of the predictor variable on the response variability is attributed to the changing spatial autocorrelation. When spatial autocorrelation is low or close to zero, the predictor variable primarily explains the variability in response variable. Therefore, even if there is only a small amount of spatial correlation, a model that uses a fixed spatial weight matrix can still provide accurate prediction values with a relatively low MSE. On the other hand, when spatial autocorrelation increases to higher levels, specifically larger than 0.30, the influence of spatial structure becomes more significant in explaining the response variable. The variability of response variable in these cases is mainly attributed to the spatial random component, which represents the complex spatial relationships between years. Therefore, the effectiveness of using weight matrices that change over time becomes more noticeable in these situations. On the other hand, the utilization of fixed spatial weight matrices demonstrates a reciprocal correlation. If there is an increase in spatial autocorrelation but the specifications of the spatial matrix are inaccurate, the prediction errors may increase significantly. Hence, selecting the correct weight matrix specifications is vital in reducing prediction errors, especially when spatial autocorrelation fluctuates. Nevertheless, it is important to acknowledge that there is a potential for overfitting, particularly when dealing with significant spatial autocorrelation.

Biased estimate of the regression parameter

We also evaluated the estimator bias of the regression parameters, with the results displayed in Fig. 5 and Fig. 6.

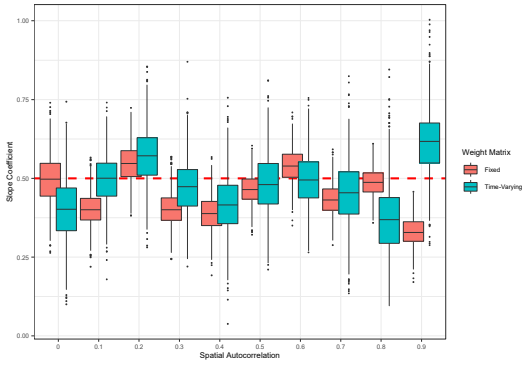

Fig. 5. The estimation of the regression parameter for 1000 sample sets

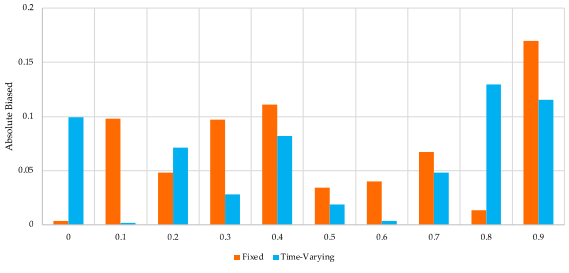

Fig. 6. Biased estimate of the regression parameter

The results of the Monte-Carlo simulation show that the fixed spatial weight matrix has the least bias compared to the time-varying spatial weight matrix when there is no spatial autocorrelation. Nevertheless, when the spatial autocorrelation falls within the range of 0.3 to 0.7, employing a spatial weight matrix that varies over time yields parameter estimates that are the least biased. Overfitting can occur when a time-varying spatial weight matrix is used at both low and high levels of autocorrelation. As a result, the differences in the response variable are primarily attributed to spatial effects rather than the covariates in the model.

4. Application to Modeling the Impact of Multi-Pollutant Exposure on Tuberculosis

4.1 Data

Tuberculosis (TB) is a highly infectious and potentially fatal disease caused by the bacterium Mycobacterium tuberculosis (M. tuberculosis) (Herrera, et al., 2022). TB is transmitted through the air when individuals with pulmonary tuberculosis cough, sneeze, or spit. This disease is categorized as a re-emerging disease due to its consistent annual occurrence. Based on data provided by the World Health Organization (WHO), approximately 10 million individuals receive a tuberculosis diagnosis annually, resulting in 1.5 million deaths each year (WHO, 2024). Indonesia is among the Asian nations grappling with a severe TB issue. Based on the Global TB Report 2022, Indonesia is the second highest country in terms of TB cases, following only India. The prevalence rate in Indonesia is 354 per 100,000 population (RI, 2023). Jakarta, the capital of Indonesia, is significantly impacted, primarily because of its poor air quality. TB is characterized by spatial clustering, often found in economically disadvantaged urban regions with low air quality, which may be linked to a higher likelihood of infection (Carrasco-Escobar, Schwalb, Tello-Lizarraga, Vega-Guerovich, & Ugarte-Gil, 2020). Over the past few years, Jakarta has consistently been listed as one of the top five cities with the most severe air pollution globally (Nurhaliza, 2024).

Multiple studies have investigated the correlation between air pollution and the likelihood of contracting TB and have determined that higher levels of air pollution can heighten the risk of developing TB (Yang, et al., 2020) (Feng, et al., 2022) (Lin, et al., 2019) (Dimala & Kadia, 2022). Studies have demonstrated that specific contaminants, including particulate matter 2.5 (PM2.5), particulate matter 10 (PM10), nitrogen oxides (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), heighten the vulnerability to TB infection. Furthermore, research has demonstrated that air pollution amplifies the likelihood of mortality in individuals with tuberculosis (Peng, Liu, Xu, Kan, & Wang, 2017). The data used in these studies typically have both spatial and temporal dimensions. The analysis techniques employed include correlation analysis and log-linear models. The utilization of spatiotemporal information is essential for making well-informed decisions regarding the timing and location of outbreaks, which allows for more effective allocation of resources. These techniques have been successfully employed in the management of infectious diseases to identify areas of high disease prevalence and outbreaks (Jaya & Folmer, 2021).

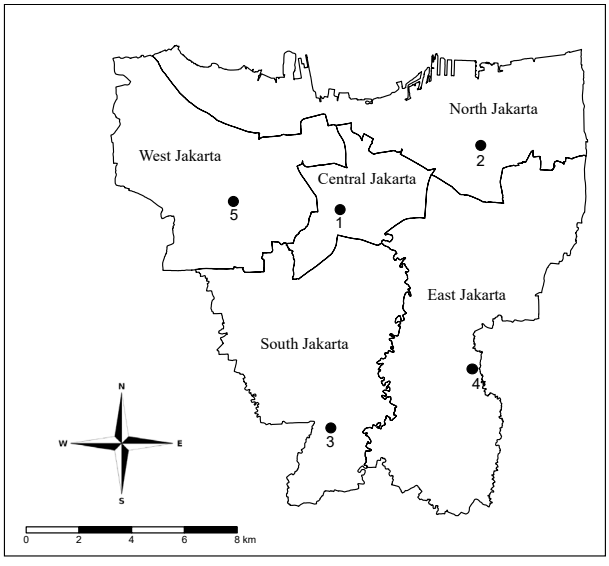

Fig. 7. Jakarta Region with Five Observation Stations

Jakarta covers an area of 661.52 square kilometers and has a population of 10.956 million people, resulting in a population density of 16,562.25 individuals per square kilometer. The city is divided into five administrative regions: Central Jakarta, South Jakarta, North Jakarta, East Jakarta, and West Jakarta. The Jakarta health profile report (2014-2020) provides information on the yearly occurrence of tuberculosis cases in these administrative divisions. The Jakarta Environmental Monitoring

Center provided data on the yearly average levels of different air pollutants, including PM10, NO2, SO2, CO, and O3. The data was collected from five separate monitoring stations, each located in a different administrative zone (Fig. 7). It covers the identical period as the records of pneumonia and TB cases.

We refrained from using data from 2021 onwards because of the COVID-19 pandemic, especially during its peak in Indonesia in 2021. A significant number of TB cases were not reported during this time, and the levels of pollutants decreased due to limitations on movement. Therefore, examining the influence of air pollutants on TB during this period would be filled with difficulties.

4.2 Statistical Method

Time-Varying Spatial Weight Matrices

Spatial weight matrices are typically created using contiguity measures, such as queen's or rook's contiguity, or distance measures. The weight matrix in this study is established by considering the cluster areas that exhibit either high or low risk. High-risk areas are characterized by an incidence rate that exceeds the average, whereas low-risk areas are characterized by an incidence rate that falls below the average. The areas should exhibit spatial clustering, whereby all areas within high-risk or low-risk clusters should be interconnected.

1. Calculate the incidence rate using the formula:
$$IR_{it} = \frac{y_{it}}{N_{it}} \times 100{,}000 \tag{23}$$
where y_{it} represents the number of TB cases and N_{it} represents the population at risk at the i-th location and t-th period.
2. A spatial cluster is defined as a high-risk area with an incidence rate greater than the average or a low-risk area with an incidence rate lower than the average. Each area within a spatial cluster must be connected to at least one other area within the same cluster.
3. Create a weight matrix based on high or low spatial cluster areas by identifying clusters with many connected areas.
4. Evaluate spatial dependency using spatial autocorrelation measures such as Moran's I.
Moran's I is a statistical measure commonly used to evaluate spatial autocorrelation, which is the degree to which a set of spatial data points is correlated with itself over a given area. It helps in determining whether similar values occur near each other or if there is a random distribution of values. The Moran's I statistic for a given weight matrix $\boldsymbol{W}_t$ can be formulated as follows:
$$Moran'sI_t = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{tij}} \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{tij} \left(IR_{it} - \overline{IR}_t\right)\left(IR_{jt} - \overline{IR}_t\right)}{\sum_{i=1}^{n}(IR_{it} - \overline{IR}_t)^2} \right) \tag{24}$$
where n denotes the total number of areas, i and j represent basic units, w_{tij} is the spatial weight matrix that varies over time t, and IR_{it} and IR_{jt} represent the TB incidence rate in units i and j, respectively. $\overline{IR}_t$ is the average TB incidence rate at time t. Moran's I index ranges from -1 to 1, with I>0 indicating positive spatial correlation. The larger the value, the higher the degree of spatial clustering.

Latent Gaussian Model

This study employed a latent Gaussian model (LGM) (Rue, Martino, & Chopin, 2009)to analyze the impact of air pollutants on the incidence of Pneumonia and TB cases in Jakarta, Indonesia. The mathematical representation of the LGM is outlined as follows. Assume the number of TB cases at area-i and time -t follows a Poisson distribution with mean and variance equal to $\lambda_{it} = E_{it}\theta_{it}$. This expressed as:

This study utilized a latent Gaussian model (LGM) to examine the influence of air pollutants on the occurrence of Pneumonia and TB cases in Jakarta, Indonesia. The mathematical formulation of the LGM is presented as follows. Let's assume that the number of tuberculosis (TB) cases in area-i and at time-t is distributed according to a Poisson distribution. The mean and variance of this distribution are both equal to λ_{it}, which is calculated as the product of E_{it}. and θ_{it}. This can be stated as:

$$y_{it}|E_{it}\theta_{it} \sim Poisson(E_{it}\theta_{it}); \text{for i} = 1, \dots, \text{n and t} = 1, \dots, \text{T} \tag{25}$$

where E_{it} and θ_{it} denote expected count and the relative risk at location i and time -t respectively. Here $n = 5$ and $T = 7$ denote the total number areas and period respectively. The expected count is formulated as (Jaya & Folmer, 2021):

$$E_{it} = N_{it} \times \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T} N_{it}} \tag{26}$$

The relative risk is modeled (θ_{it}) by a log-linear model as follows:

$$\log(\theta_{it}) = \eta_{it} = \beta_0 + \mathrm{x}_{1it}\beta_1 + \cdots + \mathrm{x}_{5it}\beta_5 + \omega_{i|t} + \nu_t \tag{27}$$

where β_0 represents the intercept, which explains the average TB risk across different locations and time periods. The variables $\mathrm{x}_1, \ldots, \mathrm{x}_5$ represent the air pollutants PM10, NO2, SO2, CO, and O3, respectively. The effects of these pollutants are denoted by $\beta_1, \ldots, \beta_5$. The random effects component represents the spatially structured effects $\omega_{i|t}$ that exhibit temporal variation, while v_t represents the effects that vary over time. We omitted the consideration of spatiotemporal interaction effects due to their potential to render the impact of air pollution on TB risk insignificant. Not all components may be included in the final model. We examine model selection criteria to ensure the inclusion of air pollutants' significant effects. If including the random effects components leads to an insignificant impact of air pollutants on TB risk, they are not considered.

A crucial goal in disease mitigation, such as tuberculosis, is to accurately pinpoint regions with a high prevalence of the disease, commonly known as hotspots. By utilizing a Bayesian approach, it is possible to calculate the posterior probability exceedance value of the relative risk. The probability $\widehat{\Pr}(\theta_{it} > c|\mathbf{y})$ denotes the likelihood that the estimated posterior mean of the relative risk θ_{it} for area i at time t exceeds a specified threshold value c. The estimation is defined as:

$$\widehat{\Pr}(\theta_{it} > c|\mathbf{y}) = 1 - \int_{-\infty}^{c} p(\theta_{it}|\mathbf{y})\, d\theta_{it} \tag{28}$$

The expression $\int_{-\infty}^{c} p(\theta_{it}|\mathbf{y})\, d\theta_{it}$ is the cumulative probability of θ_{it} with threshold value c. The initial parameter to consider is the threshold value, denoted as c, for θ_{it}. Typical threshold values include 1, 2, and 3. A value of 1 signifies that a region has a moderate relative risk, while values of 2 or 3 indicate a high and very high risk, respectively. The second parameter represents the cut-off value $\gamma = (1 - \alpha)$, of the exceedance probability. Typical values for γ are 0.90, 0.95, and 0.99.

To determine the best model for explaining the impact of air pollutants on TB, we considered the following four models:

1. Model 1: Uncorrelated spatiotemporal dependencies model, which assumes that different regions have no association with each other ($\omega_{i|t}$) and no temporal trend (ν_t). We only consider the overall level of RR (β_0) and the effects of air pollutants ($\beta_1, \ldots, \beta_5$) (M1).
2. Model 2: The overall level of RR (β_0) and the effects of air pollutants ($\beta_1, \ldots, \beta_5$) with autoregressive time effects (ν_t) (M2)
3. Model 3: The overall level of RR (β_0) and the effects of air pollutants ($\beta_1, \ldots, \beta_5$) are modeled along with a spatially structured effect using a fixed spatial weight matrix (ω_i) (M3)
4. Model 4: The overall level of RR (β_0) and the effects of air pollutants ($\beta_1, \ldots, \beta_5$) are modeled along with a spatially structured effect using a time-varying spatial weight matrix ($\omega_{i|t}$) (M4)

Various Bayesian model selection criteria, such as the Deviance Information Criterion (DIC), Watanabe Akaike Information Criterion, and Log-Marginal-Likelihood (LML), are commonly employed to determine the best model. Additionally, classical model selection criteria like Mean Absolute Deviance (MAE), Mean Absolute Prediction Error (MAPE), Root Mean Square Error (RMSE), and Pearson correlation (R) are also utilized. The model with the lowest DIC, WAIC, MAE, MAPE, RMSE, and largest LML and R is selected as the best model with superior predicted performance (Blangiardo & Cameletti, 2015).

4.3 Result

Descriptive analysis

Between 2014 and 2020, a total of 84,651 new cases of tuberculosis (TB) were recorded in the five administrative cities of Jakarta. Table 1 presents the incidence rate of TB (per 100,000 inhabitants) across these cities during this period, and Table 2 shows the concentration of air pollutants (μg/m³) from 2014 to 2020.

Table 2
Incidence Rate of TB Across Five Cities from 2014 to 2020 (per 100,000 Inhabitants)

City	2014	2015	2016	2017	2018	2019	2020	Average
Central Jakarta	47.648	101.753	53.457	164.434	103.386	239.519	244.920	136.445
East Jakarta	88.955	12.367	87.532	144.359	172.701	167.060	232.278	129.322
North Jakarta	27.881	27.428	23.915	100.768	146.888	138.065	162.170	89.588
South Jakarta	89.593	79.425	92.615	110.920	111.792	143.286	182.911	115.792
West Jakarta	67.839	101.973	105.590	112.972	155.039	152.552	153.942	121.415
Average	64.383	64.589	72.622	126.691	137.961	168.096	195.244	118.512

Table 3
The concentration of air pollutants (μg/m³) from 2014 to 2020

Pollutant	2014	2015	2016	2017	2018	2019	2020	Average
PM10	50.610	52.796	50.818	47.680	52.144	57.146	51.480	51.811
NO2	12.918	12.714	8.228	8.902	8.420	10.422	14.318	10.846
SO2	9.764	15.606	20.242	26.364	20.678	17.210	23.462	19.047
CO	23.526	23.426	26.312	19.266	17.628	18.138	17.990	20.898
O3	64.478	54.508	54.620	63.516	75.116	78.364	57.570	64.025

Table 2 shows that Central and East Jakarta had the highest occurrence rates of TB, with five-year averages of 136.44 and 129.33 per 100,000 residents, respectively. Notably, each successive year witnessed a rise in reported incidents compared to the preceding year. One of the primary factors thought to contribute to this rising trend is the air quality in Jakarta, which has a substantial influence. Table 3 indicates a notable rise in various air pollutants from 2018 to 2020. Among these pollutants, O3 and PM10 exhibit the highest concentration levels, surpassing NO2, SO2, and CO.

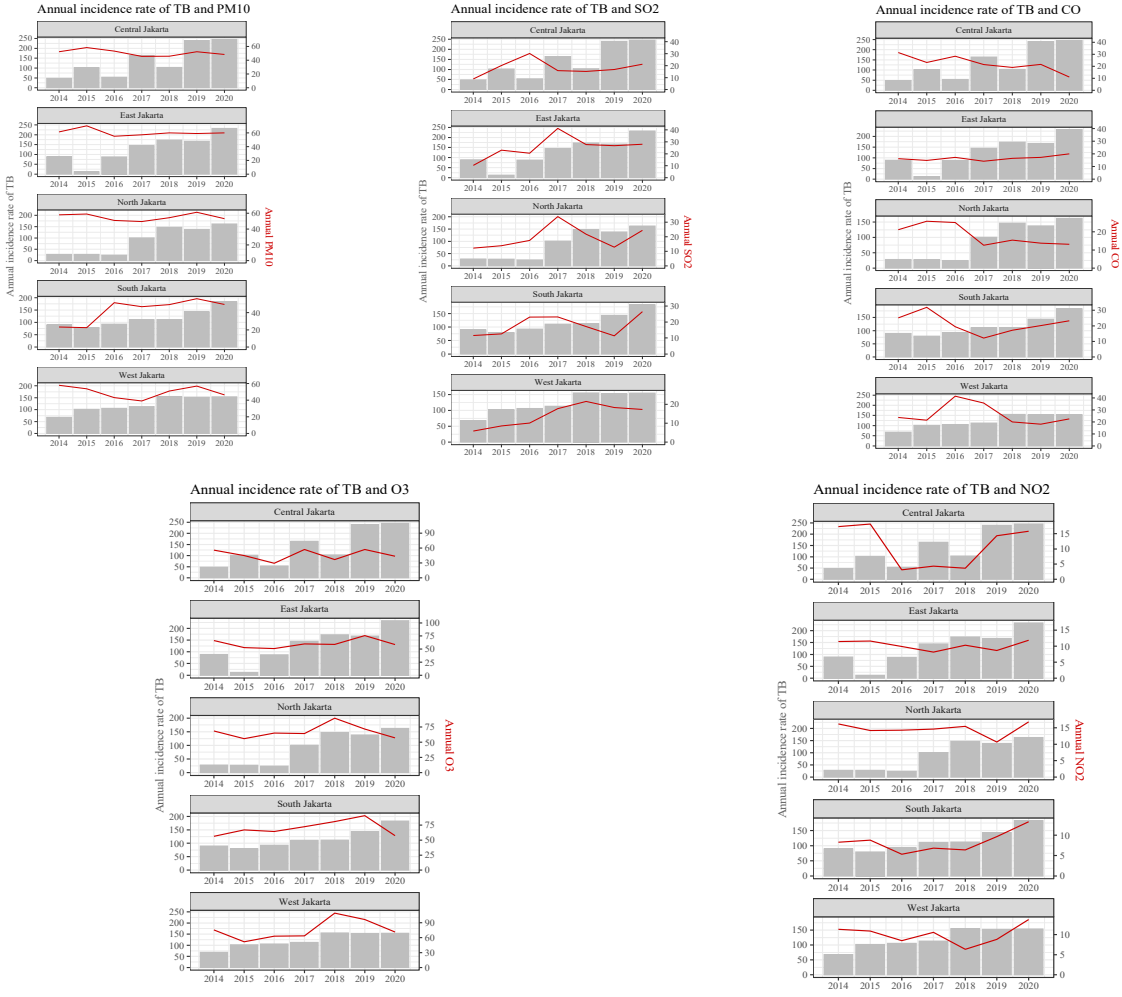

Fig. 8. Trends in TB Incidence Rates (Bar Chart) and Air Pollutant Levels (Line Chart)

During the past five years, the mean concentration of PM10 was 51,811 µg/m³, while the mean concentration of O3 was 64,025 µg/m³. The extensive data reveals a steady and continuous rise in air pollution levels annually, with some fluctuations depicted in the accompanying line chart of Fig. 8. Fig. 9 shows the spatial and temporal distribution of TB incidence rates throughout the research period. The data exhibits variations in both space and time. Temporally, there is a clear and consistent upward trend observed each year. Spatially, the data shows the existence of spatial clusters among administrative districts. In general, East Jakarta consistently demonstrates a high rate of occurrence each year. Fig. 10 displays the spatial and temporal distributions of air pollutant levels, including PM10, NO2, SO2, CO, and O3. Although there have been variations in space and time, pollution levels in 2020 have consistently increased.

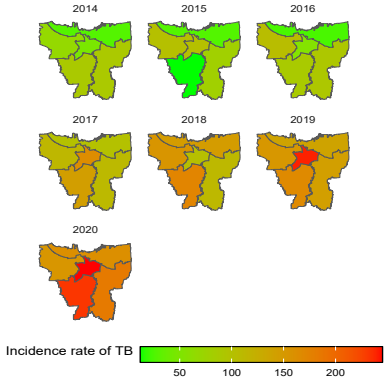

Fig. 9. Incidence rate of TB from 2014 to 2020

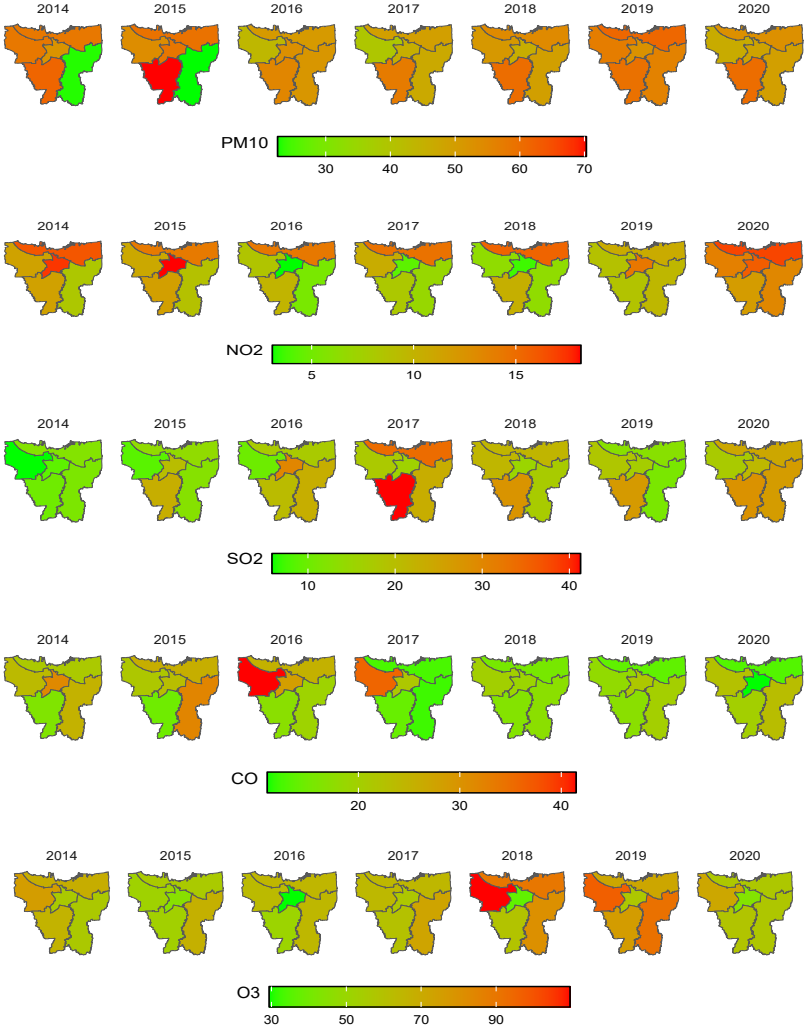

Fig. 10. Spatiotemporal distribution of air pollutants

Associations between TB and air pollutants

To create a dependable model for evaluating the influence of exposure to multiple pollutants on tuberculosis, it is essential to include spatial and temporal interdependencies. Fig. 9 illustrates that certain areas exhibit comparable TB incidence rates, creating spatial clusters. Fig. 11 provides a more distinct representation of these clusters, which are defined by similar incidence rate values across different areas, whether consistently low or consistently high. Nevertheless, these clusters display temporal variability, which requires modifying the modeling approach to account for these annual fluctuations.

To account for this variability, we suggest modifying the spatial weight matrix on a yearly basis. We conducted a spatial autocorrelation analysis using Moran's I, comparing fixed and time-varying spatial weight matrices. The results of this analysis are summarized in Table 4. The spatial weight matrix, which was created using queen contiguity, produced a negative Moran's I value. This indicates that there is a lack of significant spatial dependencies and a negative autocorrelation of TB incidence rates. On the other hand, the custom time-varying spatial weight matrix, which considers changes in both high and low spatial clusters, showed a positive spatial correlation, indicating significant statistical spatial dependencies.

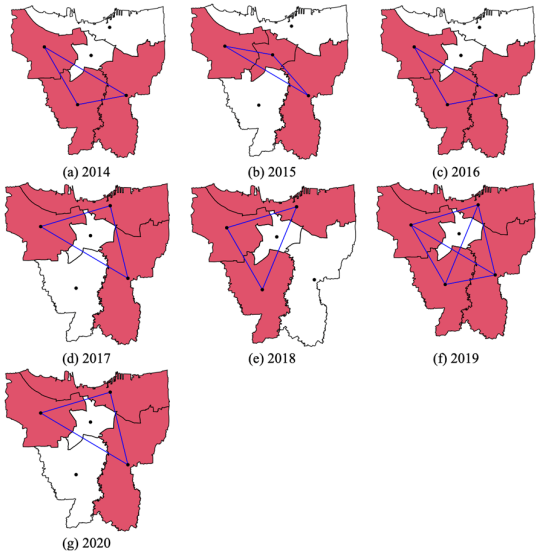

Fig. 11. Defining Varying Spatial Weight Matrices for Clustered Incidence Risk Areas

Table 4
Comparison of Moran's I for Fixed and Varying Spatial Weight Matrices from 2014 to 2020

Period	Fixed Spatial Weight		Varying Spatial Weight	
	Moran's I	p-value	Moran's I	p-value
2014	-0.137	0.533	0.461	0.041
2015	-0.551	0.125	0.585	0.013
2016	-0.316	0.697	0.543	0.026
2017	-0.235	0.933	0.574	0.018
2018	-0.278	0.882	0.509	0.026
2019	-0.312	0.501	0.203	0.048
2020	-0.233	0.930	0.560	0.016

Table 5
Bayesian spatiotemporal model comparison

Model	WAIC	MPL	MAE	RMSE	MAPE	R
M1: $\eta_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta}$	6289.631	-8462.083	0.347	0.449	70.648	0.394
M2: $\eta_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \nu_t$	5319.353	-3402.907	0.221	0.268	39.764	0.835
M3: $\eta_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \omega_i$	7812.061	-6734.343	0.301	0.368	60.184	0.654
M4: $\eta_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \omega_{i\|t}$	1657.759	-652.828	0.067	0.079	10.095	0.987

By applying a wide range of model selection criteria, including WAIC, MAE, RMSE, MAPE, and R, we determined that Model 4 from Table 5 is the most suitable option. This model, characterized by its use of different spatial weight matrices, outperformed Model 3, which depends on a static spatial weight matrix. Initially, our evaluation included the effects of time and space-time interactions. However, these factors made the effects of air pollution statistically insignificant. As a result, we narrowed down our focus to highlight the impact of air pollution, specifically excluding these elements.

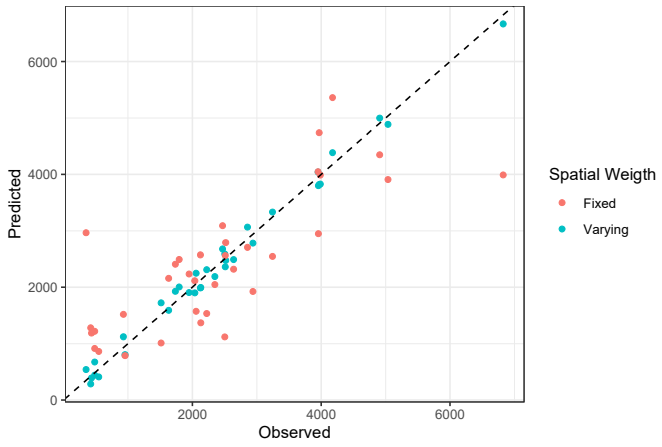

Fig. 12. Comparison of Observed and Predicted Values Using Fixed (M3) and Varying (M4) Spatial Weight Matrices

Fig. 12 clearly demonstrates that Model 4 outperforms Model 3 in terms of fit. The predictive values produced by Model 4 demonstrate a more pronounced correlation with observed values, as indicated by the closer alignment of blue points with the black linear line in comparison to red points. Model 4 incorporates the influence of air pollutants (PM10, NO2, SO2, CO, and O3) and spatially organized effects, providing the most precise understanding of the influence of multiple air pollutants on TB.

Table 6
Estimations of the associated factors in Bayesian model of total tuberculosis

Parameters	Mean	SD	q(0.025)	q(0.975)	RR	Δ (%)
(Intercept)	-11.042	0.425	-11.885	-10.218	0.000	
PM10	0.021	0.002	0.016	0.025	1.021	2.12
NO2	0.191	0.007	0.177	0.205	1.210	21.05
SO2	0.123	0.004	0.116	0.130	1.131	13.09
CO	0.064	0.007	0.050	0.079	1.066	6.61
O3	0.063	0.003	0.058	0.068	1.065	6.50

Table 7
Estimations of the spatial random effects

Hyperparameter	Mean	Sd	q(0.025)	Q(0.975)	Percentage Variance (%)
SD spatial effect 2014 (σ_{2014})	2.021	0.909	0.850	4.351	16.397
SD spatial effect 2015 (σ_{2015})	1.986	0.871	0.854	4.211	15.833
SD spatial effect 2016 (σ_{2016})	1.304	1.088	0.236	4.251	6.823
SD spatial effect 2017 (σ_{2017})	2.334	1.026	0.994	4.949	21.862
SD spatial effect 2018 (σ_{2018})	2.701	1.271	1.058	5.950	29.287
SD spatial effect 2019 (σ_{2019})	1.389	0.659	0.511	3.050	7.740
SD spatial effect 2020 (σ_{2020})	0.716	0.714	0.085	2.658	2.059
Spatial autocorrelation 2014 (ρ_{2014})	0.326	0.265	0.008	0.891	
Spatial autocorrelation 2015 (ρ_{2015})	0.241	0.236	0.003	0.822	
Spatial autocorrelation 2016 (ρ_{2016})	0.625	0.296	0.050	0.990	
Spatial autocorrelation 2017 (ρ_{2017})	0.316	0.261	0.008	0.883	
Spatial autocorrelation 2018 (ρ_{2018})	0.374	0.272	0.013	0.908	
Spatial autocorrelation 2019 (ρ_{2019})	0.287	0.260	0.005	0.881	
Spatial autocorrelation 2020 (ρ_{2020})	0.689	0.289	0.065	0.996	

Table 6 and Table 7 display the calculated values for the posterior estimates of fixed and random effects, respectively. All fixed effects in the study are statistically significant at a significant level of 5%. The estimators' 95% credible intervals do not include the value 0. These findings suggest that higher levels of air pollutants (PM10, NO2, SO2, CO, and O3) are associated with an increase in the number of tuberculosis cases. Significantly, nitrogen dioxide (NO2) and sulfur dioxide (SO2) are identified as the most significant components of air pollution, with the risk of TB increasing by more than 10% for every 1 μg/m³ increase in their concentrations. In addition, Table 7 shows that there is a positive spatial autocorrelation $\rho_{2014} - \rho_{2020}$, indicating that a higher risk of TB in one location increases the risk of TB in nearby areas.

In Figure 13, we present the projected relative risk and exceedance probability by considering both the fixed and random components in Model 4. Green areas represent low-risk locations, whereas red areas indicate high-risk areas (RR > 1). A

probability of exceeding close to 1 indicates a significantly elevated level of risk. The analysis findings reveal that during the years 2019 and 2020, all locations were classified as high-risk areas.

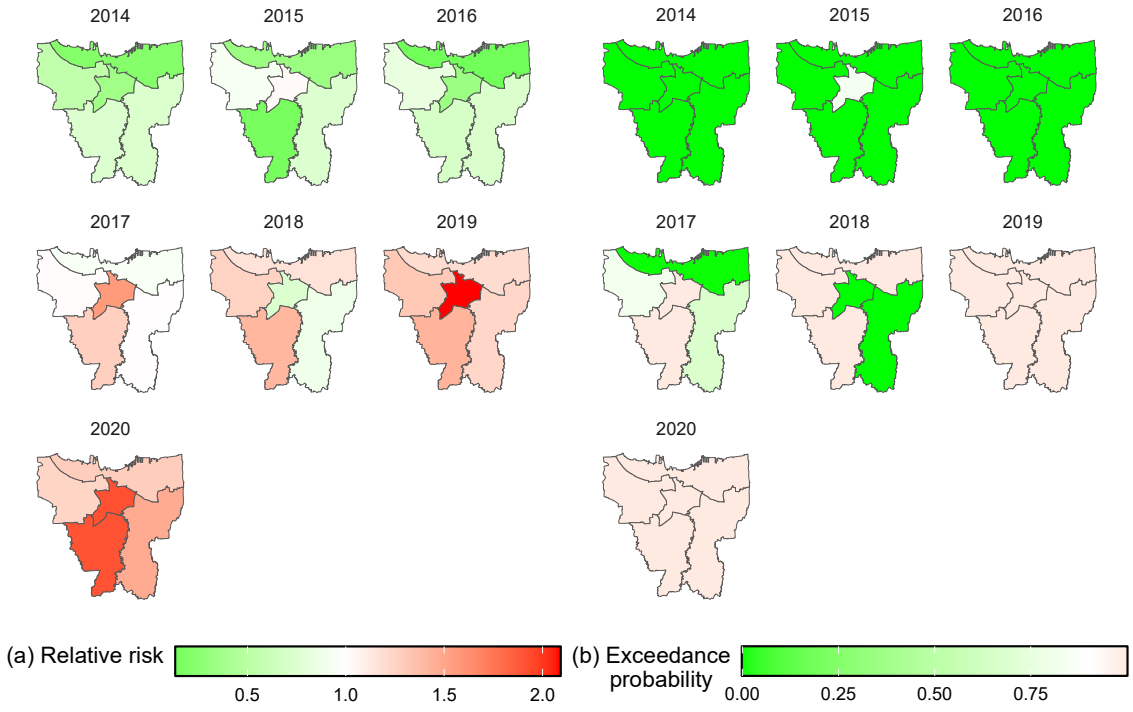

Fig. 13. Posterior Mean of (a) Relative Risk and (b) Exceedance Probability of TB

5. Discussion

The main objective of spatiotemporal data modeling is to attain precise and dependable predictions of outcomes, while minimizing any bias in the estimation of model parameters. Nevertheless, this form of modeling is intricate because it incorporates both spatial and temporal dependencies. The prediction accuracy of the spatiotemporal model is greatly influenced by the structure of these spatial dependencies. The level of intricacy is further intensified by the fact that the spatial arrangement of dependencies can change over time due to multiple factors. Precisely identifying the dependency structure, as indicated by the spatial weight matrix, is of utmost importance. Inaccurate prediction values and overfitting issues can arise from errors in defining the spatial structure of dependencies (Mingione, et al., 2022). Latent Gaussian models (LGM) are frequently employed for the modeling of spatiotemporal data (Blangiardo & Cameletti, 2015; Hrafnkelsson & Bakka, 2023; Hazra, Huser, & Jóhannesson, 2023). LGMs can incorporate covariate fixed effects and different spatial and temporal characteristics, such as structured and unstructured effects and their interactions. This study presents concrete evidence that highlights the significance of considering a spatial weight matrix that changes over time. The results of the Monte Carlo simulation show that using a time-varying spatial matrix, which adapts to the spatial clusters in the data, leads to more precise prediction results compared to assuming a fixed spatial weight matrix. This is particularly true when the level of autocorrelation is moderate (0.3-0.7). The simulation results demonstrate a quadratic trend in the Mean Squared Error (MSE) for prediction. When the autocorrelation levels are below 0.30 or above 0.70, using the time-varying spatial weight matrix still provides the most accurate prediction results. However, there is a significant increase in the likelihood of overfitting. The reason for this is that most of the variation in the response variable can be attributed to the spatial effect variable, which results in significant distortion in the estimates of the regression parameters at these levels of autocorrelation. Additionally, we offer a concrete demonstration of employing time-varying spatial weight matrices to model the influence of various air pollutants on tuberculosis risk in Jakarta. The dataset comprises spatiotemporal observations, with 5 spatial units and 7 temporal units. We created a spatial weight matrix by considering the spatial groupings of areas with low and high tuberculosis (TB) risk. This matrix was designed to maximize the values of Moran's I, which varied from 0.20 to 0.60. The evaluation of the model using various criteria such as DIC, WAIC, MAPE, MAE, RMSE, and R indicates that the spatiotemporal latent Gaussian model with a time-varying spatial weight matrix offers the most precise prediction results. Additionally, the regression coefficients demonstrate the anticipated effects. The analysis revealed that all pollutants (PM10, NO2, SO2, CO, and O3) exert a positive and substantial influence on the escalation of TB risk in Jakarta.

6. Conclusion

The results of the Monte-Carlo simulation show that utilizing a weight matrix that changes over time yields the most precise prediction values and the least amount of bias, especially for spatial autocorrelation levels within the medium range of 0.3 to 0.7. However, if the autocorrelation coefficient is extremely high or low, it can lead to overfitting issues. This occurs when the spatial effects of time-varying spatial weight matrices overpower the predictor variables. This is evidenced by a rise in the bias of the regression parameter estimates. Using time-varying spatial weight matrices in the application to model the impact of multi-pollutant exposure on TB results in more precise estimates of fixed and random effect parameters. This is demonstrated by the significantly narrower credible intervals. Elevated levels of air pollutants, including PM10, NO2, SO2, CO, and O3, significantly raise the risk of contracting tuberculosis. Furthermore, the study found that the impact of unquantified variables on the risk of tuberculosis differs based on the specific location and time, with a noticeable pattern suggesting an increase in tuberculosis risk over time. Nevertheless, this study is limited by its dependence on annual TB data, which may fail to consider monthly patterns that could provide more profound insights. Remarkably, there has been no prior investigation into the spatial and temporal impacts of air pollutants on tuberculosis (TB). Future research should focus on determining the most effective methodologies for managing tuberculosis, particularly in Jakarta, Indonesia.

Acknowledgement

Thanks to the Rector and the Directorate of Research, Community Service, and Innovation of Universitas Padjadjaran for providing the research grant program. This research was supported by Padjadjaran University through the Directorate of Research, Community Service, and Innovation (Grant No: 1893/UN6.3.1/PT.00/2024).

References

Blangiardo, M., & Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Chennai: John Wiley & Sons.

Carrasco-Escobar, G., Schwalb, A., Tello-Lizarraga, K., Vega-Guerovich, P., & Ugarte-Gil, C. (2020). Spatio-temporal co-occurrence of hotspots of tuberculosis, poverty and air pollution in Lima, Peru. *Infectious Diseases of Poverty*, *9*(02), 84-89.

Dimala, C., & Kadia, B. (2022). A systematic review and meta-analysis on the association between ambient air pollution and pulmonary tuberculosis. *Scientific Reports, 12*(11282), 1-13.

Dubé, J., & Legros, D. (2013). A spatio-temporal measure of spatial dependence: An example using real estate data. *Papers in Regional Science, 92*(1), 19-30.

Feng, Y., Wei, J., Hu, M., Xu, C., Li, T., Wang, J., & Chen, W. (2022). Lagged effects of exposure to air pollutants on the risk of pulmonary tuberculosis in a highly polluted region. *International Journal of Environmental Research and Public Health*, *19*(9), 5752.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis, 1*(3), 515–534.

Hazra, A., Huser, R., & Jóhannesson, Á. (2023). *Bayesian Latent Gaussian Models for High-Dimensional Spatial Extremes. In B. Hrafnkelsson, Statistical Modeling Using Bayesian Latent Gaussian Models* (pp. 219-251). Switzerland: Springer.

Herrera, M., Guzmán-Beltrán, S., Bobadilla, K., Santos-Mendoza, T., Flores-Valdez, M., Gutiérrez-González, L., & González, Y. (2022). Human Pulmonary Tuberculosis: Understanding the Immune Response in the Bronchoalveolar System. *Biomolecules, 12*(1148), 1-20.

Hrafnkelsson, B., & Bakka, H. (2023). *Bayesian Latent Gaussian Models. In B. Hrafnkelsson, Statistical Modeling Using Bayesian Latent Gaussian Models With Applications in Geophysics and Environmental Sciences* (pp. 1-80). Switzerland: Springer.

Jaya, I. G. N. M., & Folmer, H. (2021). Bayesian spatiotemporal forecasting and mapping of COVID-19 risk with application to West Java Province, Indonesia. *Journal of Regional Science*, *61*(4), 849-881.

Leroux, B., Lei, X., & Breslow, N. (1999). *Estimation of disease rates in small areas: a new mixed model for spatial dependence.* In M. Halloran, & D. Berry, Statistical Models in Epidemiology, the Environment and Clinical Trials (pp. 135–178). New York: Springer-Verlag.

Li, H., Ge, M., & Zhang, M. (2022). Spatio-temporal distribution of tuberculosis and the effects of environmental factors in China. *BMC Infectious Diseases, 22*(565), 1-13.

Lin, R., Shi, H., Yin, G., Thall, P. F., Yuan, Y., & Flowers, C. R. (2022). Bayesian hierarchical random-effects meta-analysis and design of phase I clinical trials. *The annals of applied statistics*, *16*(4), 2481.

Lin, Y. J., Lin, H. C., Yang, Y. F., Chen, C. Y., Ling, M. P., Chen, S. C., ... & Liao, C. M. (2019). Association between ambient air pollution and elevated risk of tuberculosis development. *Infection and drug resistance*, *12*, 3835-3847.

Mingione, M., Loro, P., Farcomeni, A., Divino, F., Lovison, G., Maruotti , A., & Lasinio, G. (2022). Spatio-temporal modelling of COVID-19 incident cases using Richards' curve: An application to the Italian regions. *Spatial Statistics, 49*(100544), 1-23.

Moran, P. (1950). A Test for the Serial Independence of Residuals. *Biometrika, 37*(1/2), 178-181.

Nurhaliza, S. (2024, March 30). Kualitas udara Jakarta urutan lima besar terburuk di dunia. Retrieved May 27, 2024, from ANTARA: https://www.antaranews.com/berita/4035441/kualitas-udara-jakarta-urutan-lima-besar-terburuk-di-dunia

Ou, B., Zhao, X., & Wang, M. (2015). Power of Moran's I Test for Spatial Dependence in Panel Data Models with Time Varying Spatial Weights Matrices. *Journal of Systems Science and Information, 3*(5), 463–471.

Peng, Z., Liu, C., Xu, B., Kan, H., & Wang, W. (2017). Long-term exposure to ambient air pollution and mortality in a Chinese tuberculosis cohort. *Science of The Total Environment, 580*, 1483-1488.

RI, M. o. (2023). Laporan Program Penanggulangan Tuberkulosis Tahun 2022. Jakarta: Ministry of Health RI.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B, 71*(2), 319–392.

Rustand, D., Niekerk, J., Krainski, E., Rue, H., & Proust-Lima, C. (2024). Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations. *Biostatistics, 25*(2), 429–448.

WHO. (2024, April 20). WHO. Retrieved from Tuberculosis: https://www.who.int/health-topics/tuberculosis#tab=tab_1

Yang, J., Zhang, M., Chen, Y., Ma, L., Yadikaer, R., Lud, Y., . . . Rui, B. (2020). A study on the relationship between air pollution and pulmonary tuberculosis based on the general additive model in Wulumuqi, China. *International Journal of Infectious Diseases, 96*, 43-47.