

A Bayesian spatially-clustered coefficient model with temporal structures for hepatitis A data in South Korea

Jaeseon Lee & Jungsoon Choi

To cite this article: Jaeseon Lee & Jungsoon Choi (2026) A Bayesian spatially-clustered coefficient model with temporal structures for hepatitis A data in South Korea, Journal of Applied Statistics, 53:7, 1253-1273, DOI: [10.1080/02664763.2025.2555593](https://doi.org/10.1080/02664763.2025.2555593)

To link to this article: <https://doi.org/10.1080/02664763.2025.2555593>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 17 Sep 2025.



Submit your article to this journal [↗](#)



Article views: 859



View related articles [↗](#)



View Crossmark data [↗](#)

A Bayesian spatially-clustered coefficient model with temporal structures for hepatitis A data in South Korea*

Jaeseon Lee^{a,b} and Jungsoon Choi^{c,d}

^aDepartment of Statistics, Texas A&M University, College Station, Texas, USA; ^bDepartment of Applied Statistics, Hanyang University, Seoul, South Korea; ^cDepartment of Mathematics, Hanyang University, Seoul, South Korea; ^dResearch Institute for Natural Sciences, Hanyang University, Seoul, South Korea

ABSTRACT

Hepatitis A, a highly contagious and perilous viral liver infection, is globally widespread, with its data collected across spatial and temporal domains. Also, demographic and socioeconomic covariates, such as population density and per capita income, are gathered over space and time units. As a result, the association between infectious disease outcomes and risk factors may differ across space and time. Some sub-regions may have a heterogeneous association with others, while a homogeneous temporal structure may exist within certain sub-regions. Acknowledging the potential variability in these associations, this study focused on comprehending the spatio-temporal dynamics of hepatitis A through a statistical model. In this paper, we analyzed monthly hepatitis A infection counts in the Republic of Korea from January 2020 to December 2021 using a Bayesian spatio-temporal model. Specifically, we employed a Bayesian spatially-clustered coefficient model with temporal structures to estimate sub-regions with the temporally varying risk effects associated with hepatitis A. Our focus lies in utilizing the Bayesian spatio-temporal model to uncover insights into the spatio-temporally varying relationships between covariates and hepatitis A outcomes. Furthermore, we addressed the spatial confounding bias prevalent in common spatial models with space-time random components by incorporating two-stage framework within our analysis.

ARTICLE HISTORY

Received 15 September 2024
Accepted 27 August 2025

KEYWORDS

Hepatitis A; spatio-temporal model; spatially-clustered coefficient; spatial confounding bias; Bayesian inference

1. Introduction

Hepatitis A, a viral infection affecting the liver, can lead to fatal outcomes. The disease is typically transmitted through contaminated food and water, or direct contact with an infected person. The disease is prevalent in areas with poor sanitation, and individuals who have not been vaccinated or previously infected are at risk. According to the World Health Organization, hepatitis A infections are not chronic, but they can result in acute liver failure and atypical complications. In the year 2015, hepatitis A claimed the lives of approximately

CONTACT Jungsoon Choi  jungsoonchoi@hanyang.ac.kr  Natural Sciences Building Room 726, Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, 04763, South Korea

*This paper is an extended version of master's thesis of Jaeseon Lee.

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

11,000 individuals, representing 0.8% the mortality attributed to viral hepatitis [32]. In addition, the fatality rate estimate is up to 1.8% among adults aged 50 or older [19].

The existence of variations in risk effects across space and time is plausible, considering that health data, along with covariates, is commonly gathered across both spatial and temporal dimensions. For example, the association between the outbreak of infectious disease and temperature may not be identical across regions or time. In some sites, higher temperatures might be related to an increase in counts, while in other areas, the opposite trend could appear [11,31]. Further, these associations can change over time, implying that a weak or inconsistent relationship in the early stages of an outbreak may become stronger and more clearly observed later on [31]. This suggests that there may exist spatio-temporal variations in risk effects, where both the strength and direction of risk effects vary. The evolving associations across regions and over time highlight the necessity for a tailored model in the regression analysis of spatio-temporal data, including hepatitis A infection.

Several studies have addressed a spatially and temporally varying regression coefficient, aiming to explain how risk effects manifest across space and time. Kang [24] proposed a spatially-clustered coefficient model with a Bayesian hierarchical framework. However, their model assumed that the estimates of the coefficient are fixed over time, thereby constraining the exploration of locally varying temporal patterns within each spatial cluster. Serban [33] suggested a space-time varying coefficient model, but coefficients were not clustered spatially based on temporal structures of covariate effects. Lee *et al.* [29] introduced a methodology for spatio-temporally varying coefficient regression, employing a hypothesis testing approach. Their methodology used circular cluster windows, necessitating spatial regions to be contiguous to be considered part of the sample spatial cluster. Choi *et al.* [8] developed a Bayesian spatio-temporal latent model to investigate spatial clusters with distinct time patterns, focusing specifically on low birth weight data from the state of Georgia, USA. However, their study did not address the problem of spatial confounding bias, which can significantly impact parameter estimation quality.

Existing spatio-temporal studies on hepatitis A primarily focused on the space-time variations in the disease infection, rather than exploring variations in risk effects. Guerrero-Vadillo *et al.* [17] described hepatitis A outbreaks in Spain from 2016 to 2017. In addition, they conducted space-time cluster analysis by sex on all hepatitis A cases in Spain during the same period, employing Kulldorff's space-time scan statistics. Zhu *et al.* [35] explored various types of viral hepatitis infection in China from 2003–2015, using visualization methods such as time series plots, Moran's I statistics, and space-time scan statistics. Brouwer *et al.* [5] studied the space-time spread patterns of hepatitis A in Southeast Michigan. They utilized a susceptible-infectious-recovered type model, reporting the influence of vaccination on infections. Despite the importance of considering space-time characteristics of hepatitis A, it is also crucial to understand the dynamic association between the disease and risk factors for developing effective prevention strategies.

Therefore, some researchers have studied the relationship between hepatitis A and covariates. Unfortunately, these studies assumed that risk effect is constant over all space-time domains, potentially overlooking variations in space and time of risk effects. Leal *et al.* [28] studied relationships between hepatitis A and environmental and sociodemographic risk factors in Para, Brazil, employing a variety of models, including generalized linear models with Poisson and Negative Binomial distributions. However, they did not account for potential variations in associations across space and time domains. Recently, Jeong

et al. [23] explored spatio-temporal variations and risk effects of diverse covariates for hepatitis A in South Korea using a Bayesian spatio-temporal model. They investigated the space-time distribution of the disease using diverse visualization and statistical methods, incorporating socioeconomic and weather-related factors in their model. However, their model assumed constant regression coefficients across all space-time domains, limiting its ability to grasp spatio-temporally varying risk effects on hepatitis A.

In this study, we investigated monthly hepatitis A infection cases in the Republic of Korea from 2020 to 2021 using a Bayesian spatio-temporal zero-inflated Poisson model. We focused on estimating the spatial groups exhibiting heterogeneous risk effects compared to other spatial clusters, simultaneously exploring a homogeneous temporal trend within each spatial group, within a Bayesian hierarchical framework. As a consequence, we proposed a Bayesian spatially-clustered coefficient model incorporating temporal structures to analyze hepatitis A incidence data in South Korea. The proposed model identified spatial clusters based on the temporal profiles of risk effects and characterized distinct temporal patterns. In addition, we addressed the issue of spatial confounding bias in space-time statistical modeling [1] by adopting a two-stage approach to reduce spatial collinearity [6,27]. In the first stage, space-time variations were independently estimated after fitting the covariates-only model. In the second stage, the spatially-varying time profiles of risk effects were re-estimated with the estimated space-time variations as the inputs. To our best knowledge, we first investigated spatio-temporal variations in risk effects of hepatitis A. The flexibility in our model fit aligns with diverse spread patterns of infectious disease in the real world. We expect that our study enhance our ability to predict hepatitis A and facilitate more efficient responses to the disease.

This study goes on as follows: In Section 2, we introduced the dataset containing hepatitis A case and risk factors. Our focus was on understanding space-time variations, and the relationship between disease infection and average temperature for each spatial unit. Section 3 presented the illustration of a Bayesian spatially-clustered coefficient model with temporal structure. The introduction of competing models and exploration of various model comparison criteria was covered in Section 4. The analysis results were presented in Section 5, and our study concluded with a discussion in Section 6.

2. Data

2.1. Hepatitis A data

The 2019 hepatitis A outbreak in South Korea was the largest since 2009, with over 16,000 cases, primarily influencing adults in their 30s and 40s due to low immunity, lack of childhood vaccination, and a cohort effect [25]. Given these factors, the years of 2020–2021 served as critical periods for evaluating post-outbreak trends and understanding the subsequent spatio-temporal dynamics of hepatitis A risk. In this study, we analyzed monthly hepatitis A incidence data in the Republic of Korea from 2020 to 2021, obtained from [the Infectious Disease Portal of Korea Disease Control and Prevention Agency](#). The study covered a spatial domain of 250 districts in South Korea, and a temporal domain of 24 months. Notably, 44.55 % of monthly hepatitis A infections on each space was zero, requiring the use of a zero-inflated model. Table 1 shows the summary statistics for monthly hepatitis A counts in the Republic of Korea from 2020 to 2021.

Table 1. Summary of the monthly hepatitis A data.

Variable	Mean	SD	Median	Q1	Q3	Min.	Max.
Hep. A infected cases	1.762	2.771	1.000	0.000	2.000	0.000	46.000
Non-zero Hep. A infected cases	3.178	3.057	2.000	1.000	4.000	1.000	46.000

Note: ^aSD, standard deviation; Q1, first quartile; Q3, third quartile.

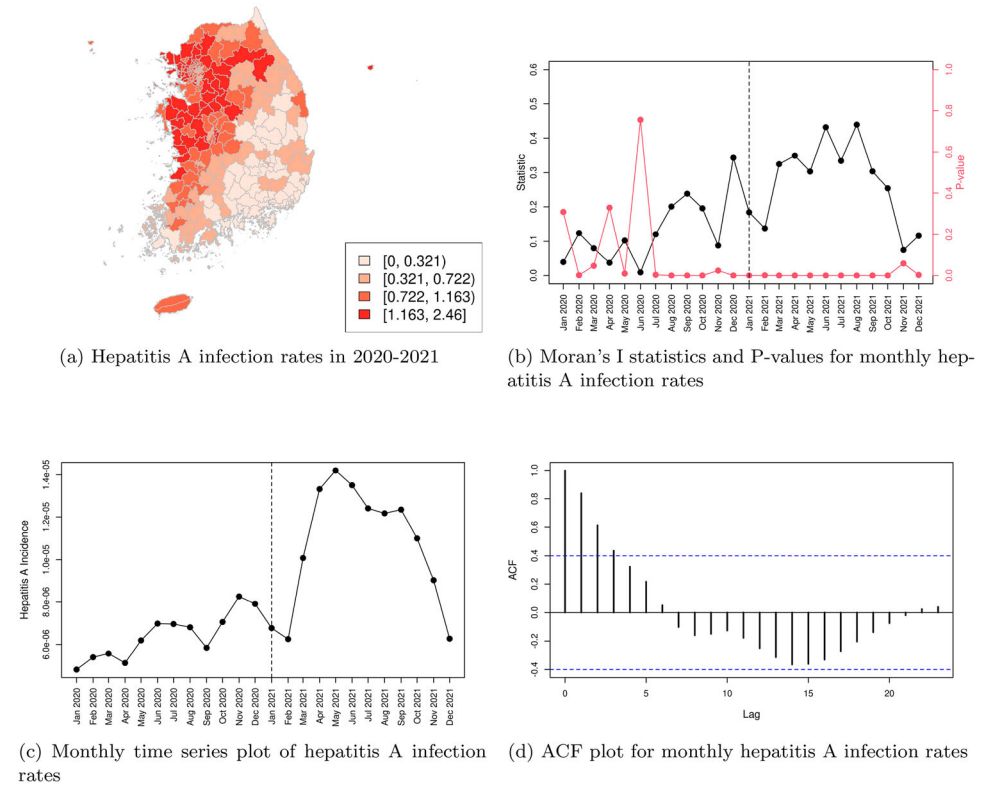


Figure 1. Spatial and temporal analysis Hepatitis A infection rates (infected cases per 10,000 population). (a) Hepatitis A infection rates in 2020–2021. (b) Moran's I statistics and P-values for monthly hepatitis A infection rates. (c) Monthly time series plot of hepatitis A infection rates and (d) ACF plot for monthly hepatitis A infection rates.

Figure 1(a) presents a spatial distribution map of cumulative hepatitis A cases per 10,000 population in the Republic of Korea from 2020 to 2021. The incidence rates are divided into four groups using quartiles. The northwestern areas, particularly metropolitan areas, as well as western regions, show the highest rates. In contrast, the southern and eastern regions exhibit relatively lower case rates. Figure 1(b) illustrates Moran's I statistics [30] and corresponding p-values for the period from January 2020 to December 2021 to explore the spatial association of monthly hepatitis A infections per 10,000 people. The statistics show positive values with the highest 0.439 in August 2021. Starting July 2020, the p-values were less than 0.001, indicating significant positive spatial dependencies in hepatitis A infection in South Korea.

To elucidate temporal dynamics, a time series plot of monthly hepatitis A cases per 10,000 population in South Korea is depicted in Figure 1(c). In 2020, infection counts per 10,000 people were high from June to August and again from October to December. A decreasing trend was observed until February 2021, followed by a significant surge that peaked in May 2021. Subsequently, the number of cases declined steadily. In addition, autocorrelation function (ACF; [4]) was applied. The ACF plot, presented in Figure 1(d), indicates the potential presence of temporal association in the data. The positive values of ACF exceeded the threshold of 0.4 up to a lag of three, suggesting a significant temporal correlation in the Korea’s hepatitis A data.

2.2. Covariates

As covariates of hepatitis A cases, we considered three yearly demographic factors: yearly elderly ratio (population aged 65 and above to total population), yearly sex ratio (males to females), and yearly population density (total population to total land area), sourced from the Open Government Data Portal. In addition, epidemiological research on hepatitis A risk factors has shown that climate-related risk factors, such as precipitation and temperature, are highly related to hepatitis A incidence [2,18,23,28,34]. Thus, we included two other monthly covariates, monthly total precipitation and monthly average temperature, obtained from the Korea National Climate Data Center. Table 2 shows summary statistics for covariates.

To explore the spatial and temporal variations in covariates, we analyzed covariate maps. For example, Figure 2 displays the quartile-stratified map of average temperature from 2020 to 2021 by seasons, revealing a common trend of warmer southern areas. Figure 2(a) shows relatively higher average temperatures in the eastern regions during winter, while Figure 2(b) indicates hotter weather in the western areas during summer. Given observed space-time variations in hepatitis A counts and covariate maps, it is plausible that the effects of covariates on hepatitis A incidence differ spatially and temporally. Supporting this, Figure 3 illustrates the map of Pearson correlation between hepatitis A infections per 10,000 people and average temperature in spatial regions, categorized into quartiles. The correlation exhibited variations across regions, suggesting that the association between the disease and average temperature differs among areas. Spatial grouping dependencies were also evident, suggesting spatial variations in risk factors’ impacts on

Table 2. Summary of covariates.

Variable	Description	Mean	SD	Median	Q1	Q3	Min	Max
Elderly Ratio	Yearly	21.836	8.561	19.450	14.600	28.550	8.200	43.200
Sex Ratio	Yearly	100.552	5.645	99.725	96.956	103.183	88.794	132.640
Population Density	Yearly	3872.577	5811.424	608.192	104.082	5897.552	19.254	26097.810
Total Precipitation (unit: mm)	Monthly	118.416	134.571	75.500	27.703	151.000	0.000	992.357
Average Temperature (unit: °C)	Monthly	13.091	8.835	13.175	5.900	21.503	−7.350	29.200

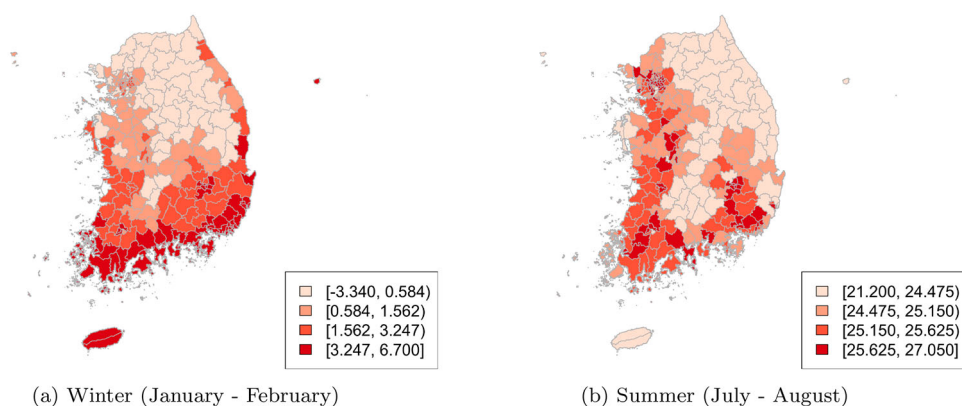


Figure 2. Average temperature in 2020–2021 by seasons. (a) Winter (January – February) and (b) Summer (July – August).

hepatitis A incidence in the Republic of Korea. Furthermore, the correlation may have distinct temporal properties, indicating varying risk effects across both space and time. To illustrate, Figure 4 shows the time series plot of correlation between hepatitis A infections per 10,000 people and average temperature across seven provinces (Seoul, Gyeonggi-do, Incheon, Chungcheong-do, Gangwon-do, Gyeongsang-do, Jeolla-do) in South Korea. See Appendix 1 for specific locations of these provinces. Remarkably, some regions display similar temporal patterns, indicating the possibility of spatial clustering. Conversely, other regions exhibit distinct time structures, suggesting that each group may have a unique association over time. Consequently, considering the temporal structures of coefficients in spatially-clustered regression model is crucial for capturing these spatio-temporal dynamics.

3. Statistical model

3.1. Statistical model

A common approach to analyzing the risk effects of covariates on health outcomes assumes a Poisson regression model. In particular, a log link function is specified as a function of risk factors and space-time random effects if data is spatial-temporal count health data. However, challenges arise in parameter estimation due to spatial confounding bias. Addressing this spatial collinearity issues, Lawson *et al.* [27] suggested a two-stage modeling approach for space-time health data. To be specific, space-time variations were independently estimated after fitting the covariates-only model in the first stage. The covariate regression coefficients were re-estimated with the previously estimated space-time variations incorporated as inputs. Following their methodology, we employed a two-stage framework to effectively reduce the influence of spatial confounding bias on parameter estimation.

However, in many cases like our hepatitis A infection data, where excess zeros may lead to over-dispersion in health outcomes, the Poisson distribution becomes less suitable for statistical modeling. To tackle this issue, we utilized a zero-inflated Poisson (ZIP) model.

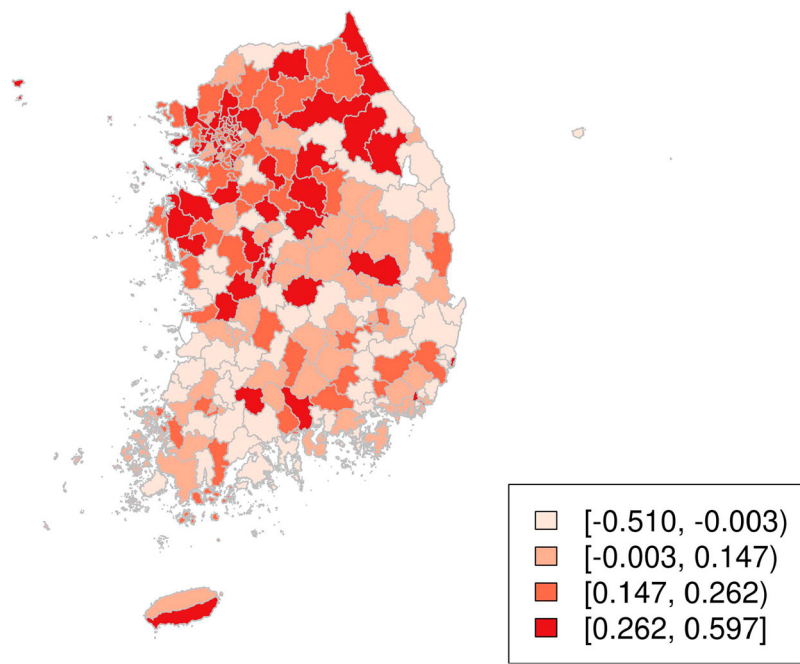


Figure 3. Correlation between hepatitis A infection rates and average temperature.

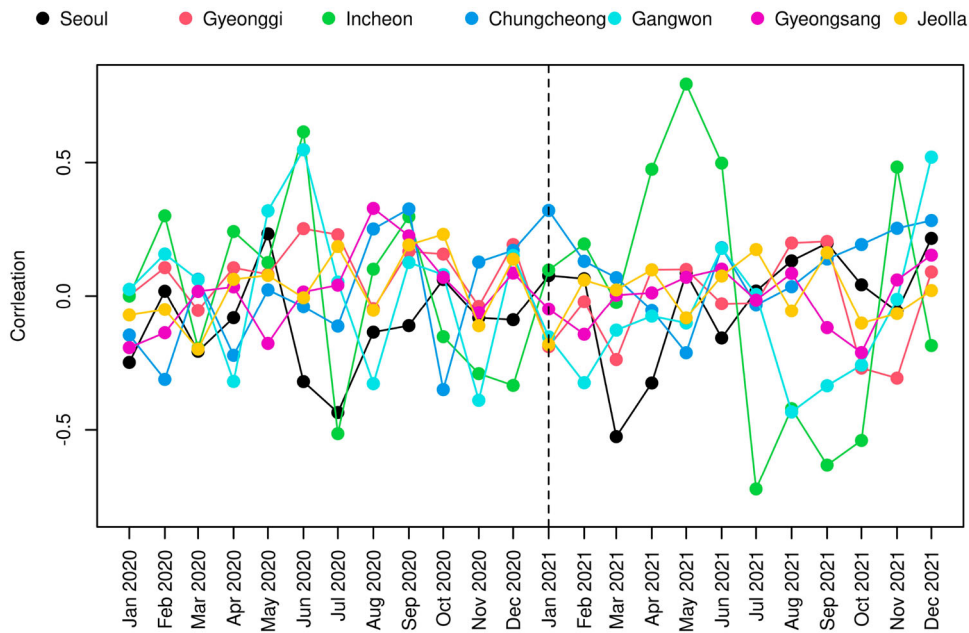


Figure 4. Temporal plot of correlation between hepatitis A infection rates and average temperature by provinces.

Let y_{it} denote the observed count in the i th space at the t th time for space-time health outcome data, where $i = 1, 2, \dots, I$ and $t = 1, 2, \dots, T$. Then, we can model y_{it} as follows:

$$\begin{aligned} y_{it} \mid z_{it} &\sim \text{Poisson}(\mu_{it} = (1 - z_{it}) \times \lambda_{it}), \\ z_{it} &\sim \text{Bernoulli}(p_{it}), \end{aligned}$$

where μ_{it} and λ_{it} indicate the mean of Poisson distribution with and without structure zeros, respectively. An indicator variable z_{it} is one if a zero appear from structure zeros with probability p_{it} , and zero otherwise. This hierarchical structure of the ZIP model can be simply written as

$$y_{it} \sim \text{ZIP}(p_{it}, \lambda_{it}).$$

Following Lawson *et al.* [27], the first stage log link function includes only covariates, thus termed as the covariates-only model.

$$\begin{aligned} \log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^{\top} \boldsymbol{\beta}_{it} + \log(N_{it}), \\ \text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^{\top} \boldsymbol{\alpha}, \end{aligned}$$

where $\mathbf{X}_{it}$ is a vector of p_1 covariates, and $\mathbf{Z}_{it}$ is a vector of p_2 covariates. Also, $\boldsymbol{\beta}_{it} = (\beta_{1it}, \beta_{2it}, \dots, \beta_{p_1it})^{\top}$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{p_2})^{\top}$ are vectors of regression coefficients, while β_0 and α_0 are intercept terms and N_{it} is the population as the offset variable.

Suppose covariates have spatially-clustered coefficients with temporal structures, $\boldsymbol{\beta}_{it} = \boldsymbol{\beta}_{c(i),t}$, where $c(i)$ serves as a spatial cluster indicator. Specifically, we assume heterogeneous relationships among sub-regions and a homogeneous temporal pattern within each sub-region. Consequently, we expect risk effects on space-time health data to vary across spatial groups, with each cluster having its distinct temporal structure. Let C denote the number of spatial clusters; then, we model $c(i)$ as

$$c(i) \sim \text{categorical}(q_{i1}, q_{i2}, \dots, q_{iC}; 1),$$

where $q_{i,c(i)}$ is a probability that i th area is included in spatial cluster $c(i)$ for $c(i) \in \{1, 2, \dots, C\}$. We model these probabilities using the unnormalized structure

$$q_{i,c(i)} = \frac{\gamma_{i,c(i)}}{\sum_{c=1}^C \gamma_{i,c}},$$

where $\gamma_{i,c}$ is non-negative parameter. In order to explain spatial dependency in the unnormalized weight, we use the intrinsic conditional autoregressive (ICAR) model proposed by Besag [3], expressed as:

$$\begin{aligned} \log(\gamma_{i,c(i)}) &\sim N(\eta_{i,c(i)}, \sigma_{c(i)}^2), \\ \eta_{i,c(i)} \mid \eta_{i',c(i)}, \quad i' \neq i &\sim N\left(\frac{1}{n_i} \sum_{i' \sim i} \eta_{i',c(i)}, \frac{\sigma_{\eta_{c(i)}}^2}{n_i}\right), \end{aligned}$$

where $\eta_{i,c(i)}$ represents spatially correlated mean, and $\sigma_{c(i)}^2$ is the variance. The notation $i' \sim i$ indicates that region i' is contiguous to region i , and n_i is the number of

neighboring areas of region i . Additionally, $\sigma_{\eta_{c(i)}}^2$ accounts for the magnitude of spatial variation. To elucidate temporal pattern within the cluster $c(i)$, we model $\boldsymbol{\beta}_{c(i),t} = (\beta_{1,c(i),t}, \beta_{2,c(i),t}, \dots, \beta_{P,c(i),t})^\top$ using an autoregressive (AR) model or a random walk (RW) process. For instance, assuming $\beta_{p,c(i),t}$ follows AR(1) model and $p \in \{1, 2, \dots, P\}$:

$$\beta_{p,c(i),t} \sim N(\rho_{c(i)}\beta_{p,c(i),t-1}, \sigma_{\beta_{c(i)}}^2),$$

where $\rho_{c(i)}$ denotes the spatial weight of cluster $c(i)$ with $\rho_{c(i)} \in (0, 1)$, and $\sigma_{\beta_{c(i)}}^2$ is the variance. Finally, the continuous-type residual in i th area at t th time in the first stage is estimated as following.

$$\widehat{r}_{it} = \log(y_{it} + 0.1) - \log(\widehat{\mu}_{it} + 0.1),$$

where an additional value of 0.1 is added into the residual calculation due to the zero values of y_{it} .

After fitting the covariates-only model, we assume the conditional distribution of $\widehat{r}_{it}$ given space-time random component ST_{it} , as there may be unexplained space-time variation

$$\begin{aligned}\widehat{r}_{it} | ST_{it} &\sim N(ST_{it}, \sigma_r^2), \\ ST_{it} &= \nu_i + \xi_i + \tau_t + \psi_t + \phi_{it}.\end{aligned}$$

Here, σ_r^2 is the variance of the residuals. We assume that both the spatially uncorrelated component ν_i and the temporally uncorrelated component τ_t follow a normal distribution, i.e. $\nu_i \sim N(0, \sigma_\nu^2)$ and $\tau_t \sim N(0, \sigma_\tau^2)$. The correlated spatial component $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_I)^\top$ is assumed to follow the ICAR distribution, i.e. $\boldsymbol{\xi} \sim \text{ICAR}(\sigma_\xi^2)$. The AR model or RW process is commonly considered for the temporally correlated component ψ_t . For the space-time interaction term ϕ_{it} , a matrix $\boldsymbol{\phi}$ with its element ϕ_{it} follows a multivariate normal distribution as:

$$\boldsymbol{\phi} \sim \text{MVN}(\mathbf{0}, \sigma_\phi^2 \boldsymbol{\Sigma}_\phi),$$

where σ_ϕ^2 represents the overall variation, and Knorr-Held [26] proposed four types of space-time structures for $\boldsymbol{\Sigma}_\phi$.

In the second stage, the estimated space-time variation $\widehat{ST}_{it}$ is incorporated into the model with other covariates to estimate the locally varying time profiles.

$$\begin{aligned}y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\ \log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_{c(i),t} + \log(N_{it}) + \widehat{ST}_{it} + \epsilon_{it}, \\ \text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha},\end{aligned}$$

where ϵ_{it} is the space-time uncorrelated interaction term unexplained by $\widehat{ST}_{it}$, and follows a normal distribution, i.e. $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$. The second stage is termed the restricted zero-inflated Poisson regression model, as the estimated space-time variation is included as an input in the model.

3.2. Bayesian computation

Based on the exploratory data analysis, we concluded that the covariate ‘average temperature’ has spatial groups where each has homogeneous time-varying coefficients. We further assumed a constant coefficient for other $p_1 - 1$ covariates. For simplicity, let $\beta_{1,c(i),t}$ be the corresponding coefficient to average temperature, thus $\beta_{c(i),t} = (\beta_{1,c(i),t}, \beta_2, \dots, \beta_{p_1})^\top$.

The likelihood functions for the covariate-only model in the first stage and the restricted ZIP model in the second stage are defined as below:

$$L(\Theta_1 | Y) = \prod_{i=1}^I \prod_{t=1}^T \text{Poisson}(\mu_{it} | X_{it}, Z_{it}, \Theta_1),$$

$$L(\Theta_3 | Y) = \prod_{i=1}^I \prod_{t=1}^T \text{Poisson}(\mu_{it} | X_{it}, Z_{it}, \Theta_3),$$

where Y indicates the observed case, and Θ_1 is the parameter sets in the covariate-only model. Also, Θ_3 denotes the parameter sets in the restricted ZIP model. In this study, $\beta_{1,c(i),t}$ is assumed to follow RW process. In the residual model after the first stage, the likelihood function is derived as:

$$L(\Theta_2 | Y) = \prod_{i=1}^I \prod_{t=1}^T \text{Normal}(\widehat{r}_{it} | \nu_i, \xi_i, \tau_t, \psi_t, \phi_{it}, \Theta_2),$$

where Θ_2 is the set of parameters in the residual model. Finally, the posterior distribution of $\Theta \in \{\Theta_1, \Theta_2, \Theta_3\}$, $p(\Theta | Y)$, is derived as:

$$p(\Theta | Y) = L(\Theta | Y)p(\Theta),$$

where $p(\Theta)$ is the prior distribution.

For parameter estimation, we assumed non-informative priors of the coefficients:

$$\alpha_0, \alpha_1, \dots, \alpha_{p_2} \sim N(0, 10^2),$$

$$\beta_0, \beta_2, \dots, \beta_{p_1} \sim N(0, 10^2).$$

In addition, the prior distributions for all the standard deviations were assumed to follow a uniform distribution, Uniform (0, 10), as suggested by Gelman [14]. Estimation was performed by Markov Chain Monte Carlo (MCMC) sampling via NIMBLE package in statistical software R. The spatial cluster indicator $c(i)$ was estimated with its posterior mode, since $c(i)$ has the nominal value. Meanwhile, the posterior mean was used for all the other parameters. A single long chain was employed to tackle the issue of label switching problem during a posterior sampling [7,22]. The first 150,000 samples were abandoned as a burn-in, and 5000 posterior samples were collected with a thinning interval of 50. The MCMC convergence of posterior samples was identified with several diagnostics, such as traceplot and Geweke diagnostic (1992) [15].

Table 3. Model summary.

Model	Two-stage	Spatial clustering of coefficient	Temporal Structure
M1			
M2	0		
M3			0
M4	0		0
M5		0	M5
M6	0	0	
M7		0	0
M8 (Proposed)	0	0	0

4. Model comparison

4.1. Competing models

To assess the proposed model's performance, we evaluated seven competing models, as outlined in Table 3, which provides an overview of the fundamental format of the considered eight models. Initially, the models were categorized based on whether they integrated the two-stage framework proposed by Lawson *et al.* [27]. In specific, M2, M4, M6, and M8 employed the two-stage framework, while M1, M3, M5, and M7 did not. Consequently, these odd-numbered models can be regarded as full models. Also, the models can be differentiated based on the structure of their coefficients. Specifically, M1 and M2 featured constant coefficient β , while M3 and M4 included only temporal profiles for a coefficient without spatial clustering, denoted as β_t . In contrast, M5 and M6 considered a spatial grouping of a coefficient $\beta_{c(i)}$ without any time patterns, representing an approach opposite to that of M3 and M4. Finally, M7 and M8 adopted a spatially-clustered coefficient model with temporal profiles, represented as $\beta_{c(i),t}$. The detailed equations for the competing models are presented in Appendix 2.

4.2. Criteria

We evaluated all eight models using a range of goodness-of-fit criteria. Firstly, for a comparison of models in terms of predictive performance, we assessed Mean Absolute Error (MAE), Mean Squared Prediction Error (MSPE).

$$\text{MAE} = \frac{1}{IT} \sum_{i=1}^I \sum_{t=1}^T |y_{it} - \hat{y}_{it}|, \quad \text{and} \quad \text{MSPE} = \frac{1}{IT} \sum_{i=1}^I \sum_{t=1}^T (y_{it} - \hat{y}_{it})^2,$$

where $\hat{y}_{it}$ is the predictive value of observed value y_{it} in space i at time t . Smaller values of MAE and MSPE indicate better predictive performance, as they reflect a closer match between the predicted and observed values.

Also, we evaluated the overall Bayesian predictive measure using the marginal predictive likelihood (MPL) [9,21], which is derived from the conditional predictive ordinate (CPO) [12,13]. The estimated CPO is approximated through MCMC algorithms [9], and the MPL provides a penalized measure of the model's predictive performance.

$$\widehat{\text{CPO}}_{it} = \left[\frac{1}{S} \sum_{s=1}^S \frac{1}{f(y_{it} | \mathbf{X}, \boldsymbol{\theta}^s)} \right]^{-1},$$

where S denotes the number of posterior samples after burn-in, and θ^s , $s = 1, 2, \dots, S$, is a MCMC sample from the posterior distribution. A function $f(\cdot)$ represent a likelihood function, so $f(y_{it} \mid \mathbf{X}, \theta^s)$ stands for the likelihood for an observation y_{it} given the covariate matrix $\mathbf{X}$ and θ^s . As a result, MPL is calculated as follows:

$$\text{MPL} = \sum_{i=1}^I \sum_{t=1}^T \log(\widehat{\text{CPO}}_{it}).$$

The larger value of MPL indicates better model fit.

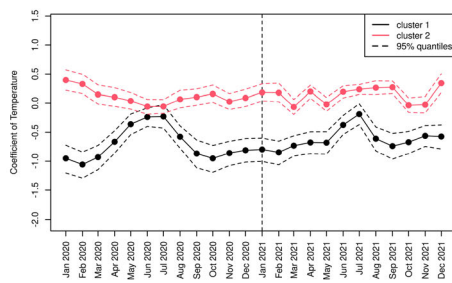
5. Data analysis results

Table 4 shows goodness-of-fit criteria results, including MAE, MSPE, and MPL, for models with all risk factors considered in Section 2. All models were fitted with various numbers of spatial clusters, and the estimated number of clusters was determined using the goodness-of-fit criteria results. Since the even-numbered models (M2, M4, M6, M8) included the two-stage framework, their performance measures were based on the second stage. These models outperformed the mixed-effect models without two-stage approach (M1, M3, M5, M7) in terms of MPL. Also, the even-numbered models except M8 with 4 clusters showed the lower MAE than the odd-numbered models. This featured the significance of introducing multi-level framework in spatio-temporal modeling of monthly hepatitis A count data. Notably, the proposed M8 with 2 spatial clusters consistently surpassed models without a spatially-clustered coefficient (M1, M2) across all measures, emphasizing the importance of accounting for spatial variations in hepatitis A infection data through the inclusion of spatial clustering in regression coefficients.

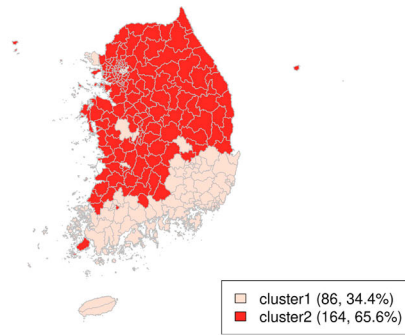
Furthermore, M8 with 2 spatial clusters outperformed temporal-only models (M3, M4), with the two-stage temporal-only model (M4) showing notable improvement over M3 across most criteria, except for MSPE. In addition, M8 with 2 spatial clusters surpassed

Table 4. Goodness-of-fit results for models with all covariates.

Number of Clusters	Model	MAE	MSPE	MPL
1	M1	0.760	1.230	−7918.548
	M2	0.745	1.609	−6987.478
	M3	0.761	1.234	−7923.736
	M4	0.720	1.489	−6889.486
2	M5	0.763	1.244	−7919.979
	M6	0.627	1.150	−6431.890
	M7	0.761	1.239	−7910.197
	M8	0.505	0.752	−6054.964
3	M5	0.760	1.232	−7917.405
	M6	0.730	1.525	−6846.661
	M7	0.761	1.241	−7909.244
	M8	0.652	1.232	−6555.156
4	M5	0.759	1.223	−7921.227
	M6	0.753	1.603	−6956.652
	M7	0.763	1.248	−7909.252
	M8	0.769	1.578	−7207.567
5	M5	0.761	1.232	−7924.331
	M6	0.623	1.133	−6448.468
	M7	0.760	1.242	−7898.710
	M8	0.740	1.560	−6958.027



(a) Estimated Temporal coefficient of average temperature



(b) Estimated cluster map

Figure 5. (a) Estimated temporal coefficient of average temperature and (b) estimated cluster map.

spatially-clustered models without temporal structures (M5, M6) in all the measures, highlighting the necessity of incorporating temporal trends within each cluster. In summary, the proposed model M8 with 2 spatial clusters demonstrated superiority in this hepatitis A data when compared to competing models with varying numbers of spatial groups. These findings imply the importance of considering both spatial clustering of coefficients and incorporating temporal structures within each cluster in the model.

Based on the best model, M8 with 2 spatial clusters, a variable selection was performed to retain only significant risk factors. As a result, monthly total precipitation and monthly average temperature were included as risk factors in the logit component of the ZIP model, $\text{logit}(p_{it})$, which captures the probability of excess zeros. Meanwhile, all covariates considered in Section 2, were incorporated into the log-linear component, $\text{log}(\lambda_{it})$, to model the expected count. Notably, the goodness-of-fit measures for M8 with 2 spatial clusters improved across all criteria, implying that the variable selection process enhanced the model's performance. Specifically, MAE and MSPE were reduced to 0.491 and 0.741, whereas MPL increased to -5955.325 . Therefore, the following results are based on M8 with two spatial clusters and their significant risk factors.

While studies by Baek *et al.* [2] and Jeong *et al.* [23] offered different perspectives on the relationship between temperature and hepatitis A in the Republic of Korea (with the note that Baek *et al.* [2] focused solely on Seoul), our study revealed distinct correlations between clusters, providing a finer understanding of the spatio-temporal dynamics of hepatitis A. Figure 5(a) presents the estimated temporal dependency structures of a coefficient for each cluster in M8. The first cluster consistently exhibited a negative coefficient value throughout the entire period, while the second cluster displayed a consistently positive coefficient. The temporal patterns for each group were clearly distinct, as evidenced by the non-overlapping 95 % credible intervals except on June and July in 2020. For example, the first cluster showed a convex shape from January 2020 to June 2020, whereas the second cluster displayed a steady decline during the same months. Following July 2021, the estimated coefficients for the first cluster exhibited a decreasing trend followed by an increasing trend, while those for the second cluster fluctuated, reaching the highest value in December 2021.

Table 5. Parameter estimates of α and β .

Parameter	Mean	Median	95% Lower	95% Upper	SD
α_0	-5.547	-5.437	-6.968	-4.635	0.769
α_1 (Temperature)	-0.813	-0.834	-1.479	-0.095	0.393
α_2 (Precipitation)	0.517	0.541	0.090	0.826	0.224
β_0	-11.343	-11.343	-11.374	-11.311	0.016
β_2 (Sex ratio)	0.107	0.108	0.084	0.130	0.012
β_3 (Elderly ratio)	-0.039	-0.039	-0.073	-0.005	0.017
β_4 (Precipitation)	0.043	0.043	0.021	0.065	0.011

The map illustrating the spatial groups of the estimated regression coefficient is presented in Figure 5(b). Cluster 1 encompassed 86 areas, while cluster 2 included 164 regions. The figure highlighted that the majority of central and northern inland regions were assigned to cluster 2. In contrast, many southern areas predominantly belonged to cluster 1. Appendix 3 provides a summary of the estimates for the parameter $\beta_{1,c(i),t}$. Notably, all estimates for $\beta_{1,1,t}$ show significantly negative values, while the estimates for $\beta_{1,2,t}$ are generally positive, though some are not statistically significant. In February 2020, a 1°C increase in average temperature resulted in a 65.3% decrease in the average number of hepatitis A cases in cluster 1 regions, while cluster 2 regions experienced a 38.7% increase. In February 2021, a 1°C increase in average temperature led to a 57.1% decrease in the average number of hepatitis A cases in the first cluster areas, while the second cluster regions was a 19.5% increase.

The estimated values of the regression coefficient, excluding $\beta_{1,c(i),t}$, are detailed in the Table 5. Regions with a higher proportion of males exhibited increased vulnerability to hepatitis A infection. Moreover, areas with a lower proportion of elderly individuals experienced elevated hepatitis A infection rates. These findings are consistent with the exploratory data analysis presented in Section 2. Metropolitan areas with relatively low elderly population ratios exhibited higher hepatitis A counts per 10,000 people, as shown in Figure 1(a). Regions with substantial precipitation also exhibited increased susceptibility to the disease.

6. Discussion

6.1. Conclusion

In this study, we examined monthly hepatitis A counts in the Republic of Korea from 2020 to 2021. We introduced a Bayesian spatially-clustered coefficient model with temporal structures to estimate sub-regions with temporally varying risk effects. Our focus was on capturing spatial clusters where each risk effect exhibits time variations and investigating their profiles. To address spatial confounding issue, we adopted a two-stage approach, inspired by Lawson *et al.* [27]. Comparing our model against alternatives, the results demonstrated its superiority over models with constant, temporally varying, and spatially varying coefficients. It significantly improved our ability to capture the complex space-time variations in hepatitis A data and its risk effects, underscoring the efficacy of our proposed model in analyzing infectious epidemiological data.

Our study significantly diverges from Jeong *et al.* [23] in both methodology and temporal period. While we both included the multi-level framework for reducing spatial

collinearity, our model uniquely incorporated clustering on regression coefficients, allowing us to capture how risk effects vary over space and time, an aspect that their model (M2) does not address. This distinct approach enabled our study to reveal more diverse correlations between disease clusters. Additionally, we used hepatitis A data in South Korea for period 2020–2021 to investigate post-outbreak dynamics, in contrast to previous studies that focused on the 2016–2019 period, which included the major outbreak in 2019. As a result, these differences provide more various insights into the evolving space-time dynamics of hepatitis A in the post-outbreak period in South Korea.

In epidemiology, risk effects vary across both space and time, as infectious diseases and associated risk factors exhibit space-time dependencies. Notably, hepatitis A incidence is known to be related to weather-related covariates [2,18,23,28]. In Section 2, we also examined the spatial-temporal variations in monthly hepatitis A counts and average temperature using visualization methods. We conducted a pioneering investigation into monthly hepatitis A counts in the Republic of Korea, with a specific focus on capturing spatial-temporally varying risk effects of average temperature. In contrast to existing studies concluding average temperature is positively correlated with hepatitis A infection in South Korea [23], our study expanded horizons and revealed distinct correlations between spatial clusters, indicating a negative association in some regions in South Korea. In addition, our study revealed that Seoul areas were also divided into two clusters, suggesting extended perspective compared to Baek *et al.* [2], who showed only negative risk effects in Seoul. This deeper understanding of risk effects provides valuable insights for developing more successful prevention and control strategies for hepatitis A.

For the novelty of this study, our proposed model not only allows for illustrating time profiles within each spatial group of risk effects on hepatitis A data in South Korea, but also addresses spatial confounding bias during the estimation process. This enabled us to better fit spatio-temporal model, since spatial collinearity leads to significant bias in estimates [10]. While our primary focus was on monthly hepatitis A cases in the Republic of Korea, the versatility of our model extends to other infectious diseases, including COVID-19 and scrub typhus, within a comprehensive space-time framework. The flexibility and practicality of our analysis model make it well-suited for analyzing diverse infectious disease data, contributing to the advancement of statistical and epidemiological research.

6.2. Future work

Despite of novelties, there are potential spaces for further research. Firstly, accurate estimation of the number of clusters is crucial, and the number of spatial clusters could be treated as an estimable parameter. In our approach, we predefined a candidate set of cluster numbers based on the analysis of small spatial domains. We assessed models' performance across different numbers of clusters within this predetermined range and selected the optimal number based on performance metrics. However, in larger spatial domains, the true cluster number might fall outside our predefined range. One potential solution is to estimate in within our process using Reversible Jump MCMC [16,20], which comes with increased computational costs, and this is beyond the scope of our study. Secondly, we assumed that only average temperature is spatially-clustered with time structures. However, other covariates may exhibit distinct risk effects in spatial domains. Future work

could directly expand our model to include multiple covariates with different spatial and temporal characteristics.

Acknowledgments

The authors acknowledge the Infectious Disease Portal of Korea Disease Control and Prevention Agency, the Open Government Data Portal, and the Korea National Climate Data Center for providing access to the data used in this analysis. Portions of this research were conducted with the advanced computing resources provided by Texas A&M Department of Statistics Arseven Computing Cluster.

Data availability statement

Data for this paper can be found at <https://github.com/jaeseonlee-stat/>.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C1010595).

References

- [1] A. Adin, T. Goicoa, J.S. Hodges, P.M. Schnell, and M.D. Ugarte, *Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India*, Stat. Modell. 23 (2023), pp. 9–30.
- [2] K. Baek, J. Choi, J.T. Park, and K. Kwak, *Influence of temperature and precipitation on the incidence of hepatitis a in Seoul, Republic of Korea: A time series analysis using distributed lag linear and non-linear model*, Int. J. Biometeorol. 66 (2022), pp. 1725–1736.
- [3] J. Besag, *Spatial interaction and the statistical analysis of lattice systems*, J. R. Stat. Soc. B 36 (1974), pp. 192–225.
- [4] G.E. Box and G.M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, CA, 1976.
- [5] A.F. Brouwer, J.L. Zelner, M.C. Eisenberg, L. Kimmins, M. Ladisky, J. Collins, and J.N. Eisenberg, *The impact of vaccination efforts on the spatiotemporal patterns of the hepatitis a outbreak in Michigan, 2016–2018*, Epidemiology 31 (2020), pp. 628–635.
- [6] J. Choi and A.B. Lawson, *A bayesian two-stage spatially dependent variable selection model for space–time health data*, Stat. Methods Med. Res. 28 (2019), pp. 2570–2582.
- [7] J. Choi, A.B. Lawson, B. Cai, and M.M. Hossain, *Evaluation of bayesian spatiotemporal latent models in small area health data*, Environmetrics 22 (2011), pp. 1008–1022.
- [8] J. Choi, A.B. Lawson, B. Cai, M.M. Hossain, R.S. Kirby, and J. Liu, *A bayesian latent model with spatio-temporally varying coefficients in low birth weight incidence data*, Stat. Methods Med. Res. 21 (2012), pp. 445–456.
- [9] P. Congdon, *Bayesian Models for Categorical Data*, John Wiley & Sons, New York, 2005.
- [10] E. Dupont, S.N. Wood, and N.H. Augustin, *Spatial+: A novel approach to spatial confounding*, Biometrics 78 (2022), pp. 1279–1290.
- [11] A. Gasparrini, Y. Guo, M. Hashizume, E. Lavigne, A. Zanobetti, J. Schwartz, A. Tobias, S. Tong, J. Rocklöv, B. Forsberg, M. Leone, M. De Sario, M. L. Bell, Yue-Liang L. Guo, Chang-fu Wu, H. Kan, Seung-Muk Yi, M. de Sousa Zanotti Stagliorio Coelho, P. H. N. Saldiva, Y. Honda, H. Kim, and B. Armstrong, *Mortality risk attributable to high and low ambient temperature: A multicountry observational study*, The Lancet 386 (2015), pp. 369–375.

- [12] S. Geisser and W.F. Eddy, *A predictive approach to model selection*, J. Am. Stat. Assoc. 74 (1979), pp. 153–160.
- [13] A.E. Gelfand and D.K. Dey, *Bayesian model choice: Asymptotics and exact calculations*, J. R. Stat. Soc. B 56 (1994), pp. 501–514.
- [14] A. Gelman, *Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)*, Bayesian Anal. 1 (2006), pp. 515–534, 20.
- [15] J. Geweke, *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Tech. Rep., Federal Reserve Bank of Minneapolis, 1991.
- [16] P.J. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika 82 (1995), pp. 711–732.
- [17] M. Guerrero-Vadillo, M. Peñuelas, Á. Domínguez, P. Godoy, D. Gómez-Barroso, N. Soldevila, C. Izquierdo, A. Martínez, N. Torner, A. Avellón, C. Rius, and C. Varela, *Epidemiological characteristics and spatio-temporal distribution of hepatitis a in Spain in the context of the 2016/2017 European outbreak*, Int. J. Environ. Res. Public Health 19 (2022), p. 16775.
- [18] P. Gullón, C. Varela, E.V. Martínez, and D. Gomez-Barroso, *Association between meteorological factors and hepatitis a in Spain 2010–2014*, Environ. Int. 102 (2017), pp. 230–235.
- [19] E. Hall, A. Wodi, J. Hamborsky, V. Morelli, and S. Schillie, *Epidemiology and Prevention of Vaccine-preventable Diseases*, US Department of Health and Human Services, Centers for Disease Control and Prevention, Washington, D.C., 2021.
- [20] D.I. Hastie and P.J. Green, *Model choice using reversible jump Markov chain Monte Carlo*, Stat. Neerl. 66 (2012), pp. 309–338.
- [21] J.G. Ibrahim, M.H. Chen, D. Sinha, J. Ibrahim, and M. Chen, *Bayesian Survival Analysis*, Vol. 2, Springer, New York, 2001.
- [22] A. Jasra, C.C. Holmes, and D.A. Stephens, *Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling*, Stat. Sci. 20 (2005), pp. 50–67, 18.
- [23] J. Jeong, M. Kim, and J. Choi, *Investigating the spatio-temporal variation of hepatitis a in korea using a Bayesian model*, Front. Public Health 10 (2023), p. 1085077.
- [24] D. Kang, *Bayesian spatial clustering of a regression coefficient for epidemiological data*, Ph.D. diss., Hanyang University, 2021.
- [25] M. Ki, H. Son, and B.Y. Choi, *Causes and countermeasures for repeated outbreaks of hepatitis a among adults in Korea*, Epidemiol. Health 41 (2019), pp. e2019038.
- [26] L. Knorr-Held, *Bayesian modelling of inseparable space-time variation in disease risk*, Stat. Med. 19 (2000), pp. 2555–2567.
- [27] A.B. Lawson, J. Choi, B. Cai, M. Hossain, R.S. Kirby, and J. Liu, *Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data*, J. Agric. Biol. Environ. Stat. 17 (2012), pp. 417–441.
- [28] P.R. Leal, R.J.D.P.S.E. Guimarães, and M. Kampel, *Associations between environmental and sociodemographic data and hepatitis-a transmission in pará state (Brazil)*, GeoHealth 5 (2021), p. e2020GH000327.
- [29] J. Lee, M.E. Kamenetsky, R.E. Gangnon, and J. Zhu, *Clustered spatio-temporal varying coefficient regression model*, Stat. Med. 40 (2021), pp. 465–480.
- [30] P.A. Moran, *Notes on continuous stochastic phenomena*, Biometrika 37 (1950), pp. 17–23.
- [31] L. Nottmeyer, B. Armstrong, R. Lowe, S. Abbott, S. Meakin, K.M. O'Reilly, R. von Borries, R. Schneider, D. Royé, M. Hashizume, M. Pascal, A. Tobias, A. M. Vicedo-Cabrera, E. Lavigne, P.M. Correa, N.V. Ortega, J. Kynčl, A. Urban, H. Orru, N. Rytí, J. Jaakkola, M. Dallavalle, A. Schneider, Y. Honda, C.F.S. Ng, B. Alahmad, G. Carrasco-Escobar, I.H. Holobác, H. Kim, W. Lee, C. Íñiguez, M.L. Bell, A. Zanobetti, J. Schwartz, N. Scovronick, M.S. Zanotti, S. Coêlho, P.H.N. Saldiva, M.H. Diaz, A. Gasparrini, and F. Sera, *The association of COVID-19 incidence with temperature, humidity, and UV radiation—A global multi-city analysis*, Sci. Total Environ. 854 (2023), p. 158636.
- [32] W.H. Organization, *Global Hepatitis Report 2017*, World Health Organization, Geneva, Switzerland, 2017.
- [33] N. Serban, *A space–time varying coefficient model: The equity of service accessibility*, Ann. Appl. Stat. 5 (2011), pp. 2024–2051, 28.

- [34] X. Zhao, M. Li, N. Haihambo, X. Wang, B. Wang, M. Sun, M. Guo, C. Han, *Periodic characteristics of hepatitis virus infections from 2013 to 2020 and their association with meteorological factors in Guangdong, China: Surveillance study*, JMIR Public Health Surveillance 9 (2023), p. e45199.
- [35] B. Zhu, J. Liu, Y. Fu, B. Zhang, and Y. Mao, *Spatio-temporal epidemiology of viral hepatitis in china (2003–2015): implications for prevention and control policies*, Int. J. Environ. Res. Public Health 15 (2018), article number 661.

Appendices

Appendix 1. Map of provinces in South Korea

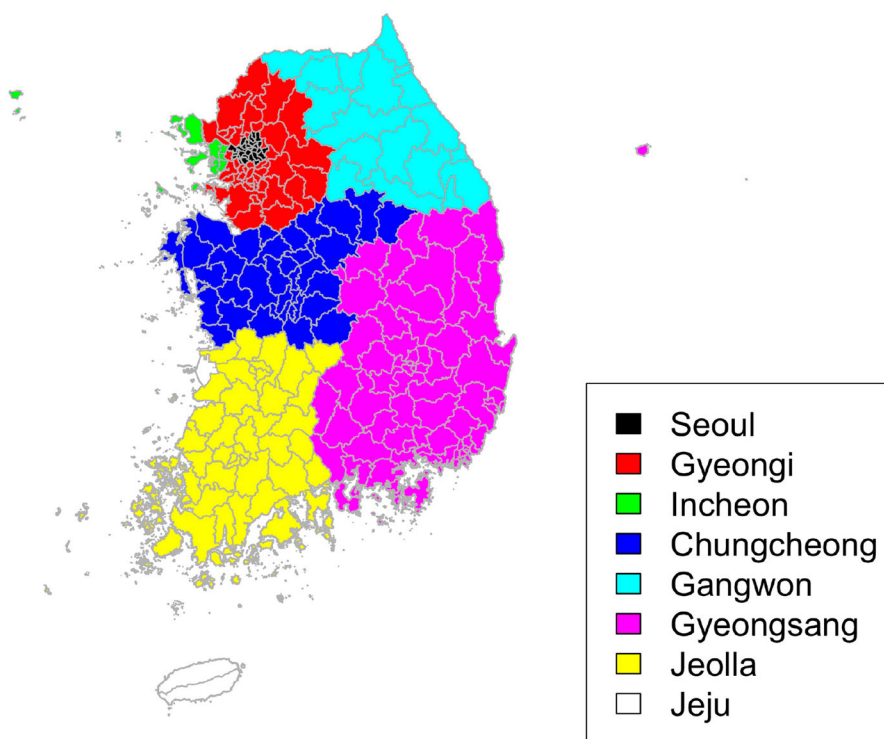


Figure A1. Provinces in South Korea mentioned in Figure 4.

Appendix 2. Equations for competing model

With the same notations illustrated in Section 3, the competing models in Section 4 are described as follows.

M1:

$$\begin{aligned}
 y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
 \log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta} + \log(N_{it}) + v_i + \zeta_i + \tau_t + \psi_t + \phi_{it} + \epsilon_{it}, \\
 \text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}.
 \end{aligned}$$

M2:

In the first stage,

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta} + \log(N_{it}), \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}, \\
\widehat{r}_{it} &= \log(y_{it} + 0.1) - \log(\widehat{\mu}_{it} + 0.1), \\
\widehat{r}_{it} \mid ST_{it} &\sim \text{N}(ST_{it}, \sigma_r^2), \\
ST_{it} &= v_i + \zeta_i + \tau_t + \psi_t + \phi_{it}.
\end{aligned}$$

In the second stage,

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta} + \log(N_{it}) + \widehat{ST}_{it} + \epsilon_{it}, \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}.
\end{aligned}$$

M3:

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_t + \log(N_{it}) + v_i + \zeta_i + \tau_t + \psi_t + \phi_{it} + \epsilon_{it}, \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha},
\end{aligned}$$

M4:

In the first stage,

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_t + \log(N_{it}), \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}, \\
\widehat{r}_{it} &= \log(y_{it} + 0.1) - \log(\widehat{\mu}_{it} + 0.1), \\
\widehat{r}_{it} \mid ST_{it} &\sim \text{N}(ST_{it}, \sigma_r^2), \\
ST_{it} &= v_i + \zeta_i + \tau_t + \psi_t + \phi_{it}.
\end{aligned}$$

In the second stage,

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_t + \log(N_{it}) + \widehat{ST}_{it} + \epsilon_{it}, \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}.
\end{aligned}$$

M5:

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_{c(i)} + \log(N_{it}) + v_i + \zeta_i + \tau_t + \psi_t + \phi_{it} + \epsilon_{it}, \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha},
\end{aligned}$$

M6:

In the first stage,

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_{c(i)} + \log(N_{it}), \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}, \\
\widehat{r}_{it} &= \log(y_{it} + 0.1) - \log(\widehat{\mu}_{it} + 0.1), \\
\widehat{r}_{it} \mid ST_{it} &\sim N(ST_{it}, \sigma_r^2), \\
ST_{it} &= v_i + \xi_i + \tau_t + \psi_t + \phi_{it}.
\end{aligned}$$

In the second stage,

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_{c(i)} + \log(N_{it}) + \widehat{ST}_{it} + \epsilon_{it}, \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha}.
\end{aligned}$$

M7:

$$\begin{aligned}
y_{it} &\sim \text{ZIP}(p_{it}, \lambda_{it}), \\
\log(\lambda_{it}) &= \beta_0 + \mathbf{X}_{it}^\top \boldsymbol{\beta}_{c(i),t} + \log(N_{it}) + v_i + \xi_i + \tau_t + \psi_t + \phi_{it} + \epsilon_{it}, \\
\text{logit}(p_{it}) &= \alpha_0 + \mathbf{Z}_{it}^\top \boldsymbol{\alpha},
\end{aligned}$$

Appendix 3. Estimates for $\beta_{1,c(i),t}$ of M8 with two clusters

Table A1. Parameter estimates of $\beta_{1,c(i),t}$ of M8 with two clusters.

Parameter	Mean	Median	95% Lower	95% Upper	SD
$\beta_{1,1,1}$	−0.951	−0.946	−1.203	−0.724	0.122
$\beta_{1,1,2}$	−1.058	−1.054	−1.293	−0.842	0.115
$\beta_{1,1,3}$	−0.930	−0.926	−1.146	−0.728	0.106
$\beta_{1,1,4}$	−0.664	−0.663	−0.861	−0.476	0.098
$\beta_{1,1,5}$	−0.363	−0.362	−0.541	−0.189	0.090
$\beta_{1,1,6}$	−0.239	−0.237	−0.401	−0.082	0.082
$\beta_{1,1,7}$	−0.229	−0.229	−0.430	−0.028	0.103
$\beta_{1,1,8}$	−0.578	−0.577	−0.776	−0.392	0.098
$\beta_{1,1,9}$	−0.868	−0.865	−1.115	−0.636	0.121
$\beta_{1,1,10}$	−0.951	−0.949	−1.191	−0.729	0.116
$\beta_{1,1,11}$	−0.860	−0.858	−1.076	−0.658	0.107
$\beta_{1,1,12}$	−0.811	−0.811	−1.016	−0.610	0.104
$\beta_{1,1,13}$	−0.798	−0.794	−1.004	−0.602	0.104
$\beta_{1,1,14}$	−0.847	−0.844	−1.060	−0.654	0.104
$\beta_{1,1,15}$	−0.732	−0.730	−0.908	−0.563	0.089
$\beta_{1,1,16}$	−0.676	−0.673	−0.868	−0.493	0.095
$\beta_{1,1,17}$	−0.680	−0.678	−0.877	−0.492	0.097
$\beta_{1,1,18}$	−0.376	−0.376	−0.540	−0.213	0.084
$\beta_{1,1,19}$	−0.189	−0.189	−0.370	−0.012	0.090
$\beta_{1,1,20}$	−0.611	−0.608	−0.823	−0.412	0.104
$\beta_{1,1,21}$	−0.739	−0.737	−0.963	−0.522	0.111
$\beta_{1,1,22}$	−0.672	−0.669	−0.869	−0.488	0.097
$\beta_{1,1,23}$	−0.562	−0.559	−0.745	−0.387	0.093
$\beta_{1,1,24}$	−0.574	−0.571	−0.789	−0.376	0.107
$\beta_{1,2,1}$	0.394	0.394	0.219	0.573	0.089
$\beta_{1,2,2}$	0.327	0.326	0.164	0.496	0.084
$\beta_{1,2,3}$	0.148	0.148	−0.016	0.312	0.083
$\beta_{1,2,4}$	0.102	0.102	−0.056	0.263	0.083
$\beta_{1,2,5}$	0.038	0.038	−0.102	0.178	0.072
$\beta_{1,2,6}$	−0.059	−0.059	−0.190	0.063	0.064
$\beta_{1,2,7}$	−0.058	−0.057	−0.175	0.061	0.059
$\beta_{1,2,8}$	0.065	0.064	−0.080	0.218	0.076
$\beta_{1,2,9}$	0.103	0.104	−0.036	0.242	0.071
$\beta_{1,2,10}$	0.158	0.158	0.012	0.308	0.076
$\beta_{1,2,11}$	0.026	0.026	−0.109	0.161	0.068
$\beta_{1,2,12}$	0.088	0.088	−0.059	0.236	0.075
$\beta_{1,2,13}$	0.181	0.179	0.036	0.330	0.075
$\beta_{1,2,14}$	0.178	0.177	0.024	0.342	0.080
$\beta_{1,2,15}$	−0.067	−0.067	−0.195	0.062	0.066
$\beta_{1,2,16}$	0.197	0.198	0.086	0.306	0.055
$\beta_{1,2,17}$	−0.022	−0.022	−0.140	0.097	0.060
$\beta_{1,2,18}$	0.193	0.193	0.089	0.295	0.053
$\beta_{1,2,19}$	0.236	0.236	0.148	0.318	0.043
$\beta_{1,2,20}$	0.263	0.263	0.146	0.382	0.060
$\beta_{1,2,21}$	0.271	0.271	0.163	0.377	0.055
$\beta_{1,2,22}$	−0.037	−0.036	−0.159	0.088	0.064
$\beta_{1,2,23}$	−0.028	−0.028	−0.168	0.114	0.071
$\beta_{1,2,24}$	0.341	0.340	0.181	0.502	0.082