A spatio-temporal Dirichlet process mixture model for coronavirus disease-19

Jaewoo Park1,2 | Seorim Yi2 | Won Chang3 | Jorge Mateu4

1Department of Applied Statistics, Yonsei University, Seoul, South Korea
2Department of Statistics and Data Science, Yonsei University, Seoul, South Korea
3Division of Statistics and Data Science, University of Cincinnati, Cincinnati, Ohio, USA
4Department of Mathematics, University Jaume I, Castellón de la Plana, Spain

Correspondence
Jaewoo Park, Department of Statistics and Data Science, Yonsei University, Seoul 03722, Korea.
Email: jwpark88@yonsei.ac.kr

Funding information
National Research Foundation of Korea, Grant/Award Numbers: 2020R1C1A0100386814, RS-2023-00217705; ICT Challenge and Advanced Network of HRD, Grant/Award Number: RS-2023-00259934; Spanish Ministry of Science, Grant/Award Number: PID2019-107392RB-I00

Understanding the spatio-temporal patterns of the coronavirus disease 2019 (COVID-19) is essential to construct public health interventions. Spatially referenced data can provide richer opportunities to understand the mechanism of the disease spread compared to the more often encountered aggregated count data. We propose a spatio-temporal Dirichlet process mixture model to analyze confirmed cases of COVID-19 in an urban environment. Our method can detect unobserved cluster centers of the epidemics, and estimate the space-time range of the clusters that are useful to construct a warning system. Furthermore, our model can measure the impact of different types of landmarks in the city, which provides an intuitive explanation of disease spreading sources from different time points. To efficiently capture the temporal dynamics of the disease patterns, we employ a sequential approach that uses the posterior distribution of the parameters for the previous time step as the prior information for the current time step. This approach enables us to incorporate time dependence into our model in a computationally efficient manner without complicating the model structure. We also develop a model assessment by comparing the data with theoretical densities, and outline the goodness-of-fit of our fitted model.

KEYWORDS
Bayesian hierarchical model, Dirichlet process Gaussian mixture, Infectious diseases, Markov chain Monte Carlo, spatio-temporal point patterns

1 INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has led to millions of deaths worldwide and presents an unprecedented challenge to economics, social life, and public health systems. Understanding the dynamics of the disease is essential to plan and implement effective intervention measures. Most of the disease modeling efforts in the literature have been focusing on modeling and predicting the total incidence, prevalence and mortality observed within certain regions or countries (cf. Reference 1; Fanelli and Piazza; Anastassopoulou et al; Barlow and Weinstein). While such analysis is helpful for regional or national agencies to manage and plan resources allocation over time, it may not be informative about how the disease is spread within the study area over time.

Considering that COVID-19 spreads quickly through close contact with infected people, detecting the disease hot spots and understanding their dynamics over time and space is important for an effective intervention. Indeed, understanding the spatial distribution of disease hot spots, along with their size, duration, and proximity to landmarks can provide insights into disease outbreaks within the local area. This can be done through analyzing geocoded epidemiological data at a fine spatio-temporal scale, which records the geographic location (home addresses) and time...
of confirmed cases. In this line, we analyze here the spatio-temporal point patterns of COVID-19 cases in the city of Cali (Colombia).

Our data provides every confirmed case’s exact location and time information, offering vital insights for the spatio-temporal interaction between individuals concerning the disease spread in a metropolis. In contrast, usual aggregated data lack precise information about individual cases and present a significant challenge in modeling the spatio-temporal dynamics of human-to-human disease transmission when capturing the fine spatial heterogeneity of case distribution in a small region. Aggregated data may lose fine-grained spatio-temporal information, which will lead the administrative officials to make biased decisions.

Our unique high-resolution dataset for individual cases of COVID-19 in Cali, Colombia, highlights the typical behavior of the spread of an infectious disease, that of forming clusters along a region with varying dimensions in space and time. This sort of triggering mechanism can be fully analyzed using point process methodology. Indeed, spatio-temporal point processes provide a natural framework for modeling COVID-19 cases by regarding the observations as a realization from the process.

Spatio-temporal analysis of COVID-19 plays a pivotal role in understanding the dynamics of the spread of COVID-19. Just a few studies attempt to model the dynamics of COVID-19 using point processes. Gajardo and Müller6 propose a point process regression framework of COVID-19 cases and deaths conditioned on mobility and economic covariates. Giudici et al7 focus on country-level case prediction in 27 European countries by augmenting spatio-temporal point process model with mobility network covariates. Li et al8 introduce a generative and intensity-free point process model based on an imitation learning framework to track the spread of COVID-19 and forecast county-level cases in the United States. Closer to our point process approach, Briz-Redón et al9 propose a mechanistic spatio-temporal point process model by accounting for both human mobility and geographical proximity. Chiang et al10 adapt Hawkes processes11 to explain a contagious disease pattern at an areal level. Although both approaches can describe the characteristics of epidemic outbreaks through the conditional intensity functions, it is challenging to specify the space-time location of disease-spreading centers. Recently, Park et al12 developed an interaction Neyman-Scott point process13 to detect unobserved cluster centers. However, the model cannot explain temporal dependencies. Furthermore, this method is of limited applicability to large samples (e.g., N = 10,000) because it requires birth-death Markov chain Monte Carlo (MCMC),14 which suffers from the slow mixing. This motivates the development of a new approach that allows researchers to study spatio-temporal patterns of the disease within a faster inferential framework.

In this manuscript, we develop a spatio-temporal point process model that can automatically select the number of disease spreading clusters and their location by adopting a Dirichlet stochastic process into our framework. Dirichlet process Gaussian mixture models have been widely used for various scientific problems.15-20 Examples include clustering for population genetics data18 and for market segmentation.20 From estimated space and time range parameters, we can diagnose the degree of COVID-19 risk. In addition, to quantify the impact of the outbreak on major landmarks in the city, we incorporate landmarks range parameters into our model. We can outline the time-varying effect of landmarks, which provide useful epidemiological information for constructing public health interventions.

Our method is computationally practical because the MCMC algorithm can be easily implemented from the full conditional distribution. This is one of the advantages of our method over the Neyman-Scott-based approaches12,21 which require expensive simulations for parameter estimation. Furthermore, we implement a sequential approach that uses the information from the previous time step, which can provide a quick and realistic analysis.

The remainder of this paper is organized as follows. In Section 2, we describe the different data sources analyzed in this study. In Section 3, we propose a spatio-temporal Dirichlet process mixture model and describe Bayesian inference for the model. We introduce our rolling-window strategies with implementation details. We also provide a model assessment strategy to validate our method. In Section 4, we apply our methods to COVID-19 datasets in Cali, Colombia. We show that our model can detect the spatio-temporal cluster centers of COVID-19 cases and provide the impact of the clusters across different times. Furthermore, we can quantify the time-varying effect of the landmarks on COVID-19 cases, which can provide important epidemiological interpretations. We conclude with a summary and discussion in Section 5.

2 DATA: COVID-19 CASES IN CALI, COLOMBIA

The COVID-19 data that motivates our approach is provided by the Municipal Public Health Secretary of Cali (Colombia)*, institution that documents individual-level confirmed cases of acute respiratory infections due to the COVID-19. This is part of a running program from the National Surveillance System in Public Health (SIVIGILA). Cali represents one of the
major cities in Colombia, the capital of Valle del Cauca department, and the most populated city in southwest Colombia, with about 2.2 million residents according to the 2018 census. The city spans 560.3 square kilometers (216.3 square miles) with 120.9 square kilometers (46.7 square miles) of urban area. As the only major Colombian city with access to the Pacific coast, Cali is the leading industrial and economic center in the country’s south, with one of Colombia’s fastest-growing economies. See Figure 1 to depict the location of the city of Cali within Colombia.

Cali is geographically diverse, with an altitude descent from west to east. The city is mainly flat with a small part of mountainous areas on the western city border, leading to a higher population density in the southeastern compared to the northwestern of the city. Owing to its proximity to the equator, there are no significant seasonal variations in Cali.

As shown in Figure 2a, more than half of the population concentrates in neighborhoods of low socioeconomic strata located mainly in the east, northeast, and west. Almost a tenth of the population under the line of poverty agglomerates in the city’s eastern neighborhoods. The population with higher socioeconomic strata distributes in the other city areas, concentrating the wealthiest population in the city’s south.

Our dataset consists of 16,309 records of cases from March 02 to July 25 of 2020. Specifically, a COVID-19 case was recorded once confirmed, with the diagnosed date of the patient and the geographical location (measured in longitude and latitude) of their residence. The testing procedures were carried out across the entire urban area, with similar testing rates in each local community. See Cuartas et al for a general description and exploratory analysis of COVID-19 cases in Cali.
Unlike other commonly-seen COVID-19 datasets that only report the aggregated number of cases or deaths at a state or county level, this dataset records the exact location and time information of each single confirmed case, which opens the possibility to look at the data from the perspective of point patterns. To have an overall picture of the evolution, Figure 3 presents the spatial distribution of confirmed cases at three particular weeks in Cali. Note that although the original dataset belonging to the corresponding authorities does have individual information of infected patients, including names, all this sort of information was totally removed from the files made available in the analysis. From our standpoint, we can not cross the information between locations and individuals as we also do not have access to the names of where people are living. Thus, the privacy issue is not a problem here.

We note that the first confirmed case of COVID-19 in Colombia appeared on March 6, 2020. On March 12, the country soon declared a state of emergency. On March 15, Cali reported the first positive person. Then the authorities announced the mandatory isolation for the entire city for just 8 days. However, the disease quickly spread and concentrated in the most vulnerable areas with low socioeconomic strata. After early efforts of the government to contain the pandemic, inevitably, the virus spread throughout the city, affecting a large part of the population. The overall public health decisions were known to not significantly affect the dynamics of the virus spreading. Thus an initial lockdown of (only) 8 days was not a strong measure, leading to as if no in-home long-staying period was implemented. Also, recall the time interval here considered goes from March 2 to July 25, a period in which the academic year is running, and primary, secondary schools, and universities are totally open (summer holidays in Colombia run from the end of December to the end of February). This all meant people were moving around and attending to their daily obligations.

Besides COVID-19 events, we also collected the location of a number of distinct landmarks in Cali, including town halls, churches, schools, shopping centers, bus stations, and parks, from the Administrative Department of Municipal Planning †, as these locations play an important role in understanding the wide and rapid spreading of the virus. The landmark dataset has three town halls, 48 small and large churches, 79 schools, 143 shopping centers, 65 bus stations, and 71 parks. Figure 2b shows the exact locations of these collected landmarks.
3 | A SPATIO-TEMPORAL DIRICHLET PROCESS MIXTURE MODEL

In this section, we propose a spatio-temporal Dirichlet process Gaussian mixture model (STDPG) that can capture and describe unobserved spatio-temporal cluster events of COVID-19. We additionally obtain maps of COVID-19 risk.

3.1 | Model framework

Consider the realization of a point pattern \( \{ (x_i, t_i) \}_{i=1}^{N} \) from the spatio-temporal point process \( X \), defined over the bounded domain \( S \times T \). Here, \( S \in \mathbb{R}^2 \), \( T \in \mathbb{R}^+ \) denotes the spatial region of interest and the study period, respectively. We consider latitude and longitude coordinates with a normalized time domain \( T = [0, 1] \). In our context, each \( (x_i, t_i) \) represents the location and time of a confirmed individual. Let \( C = \{ (c_j^x, c_j^t) \}_{j=1}^{M} \in S \times T \) be the unobserved space-time cluster centers. For the \( i \)th observation, we introduce the cluster membership variable \( g_i \in \{ 1, \ldots, M \} \). We model the membership distribution via a Dirichlet process prior\(^{25} \) to automatically choose the number of clusters. Especially, we consider the distribution of \( g_i \) as

\[
g_i \sim \text{Categorical}(q_1, \ldots, q_M),
\]

where \( q_1, \ldots, q_M \) are cluster probabilities. We use the stick breaking prior\(^ {26} \) for cluster probabilities as

\[
q_j = \begin{cases} U_1, & \text{for } j = 1 \\ U_j \prod_{k=1}^{j-1} (1 - U_k), & \text{for } j = 2, \ldots, M, \end{cases}
\]

where \( U_1, \ldots, U_k \sim_{\text{iid}} \text{Beta}(1, b_u) \) with a rate parameter \( b_u \). Following the previous works\(^ {18,20} \), we use a hyperprior for \( b_u \) defined as \( b_u \sim \text{Gamma}(1, 1/4) \), which encourages the numbers of the non-empty clusters to be between 1 and 15. This type of regularization can reduce the possibility of overfitting. In (1), the first mixture probability \( q_1 \) is modeled as a beta random variable \( U_1 \). For \( j = 2, \ldots, M \), the subsequent mixture probabilities \( q_j \) can be obtained through multiplication between the remaining probability \( 1 - \sum_{k=1}^{j-1} q_k \) and the proportion assigned to the \( j \)th cluster component \( (U_j) \). Although the stick breaking prior considers infinite \( M \) in theory, we can approximate it with a reasonably large \( M < \infty \) in practice. In this case, the remaining probability \( 1 - \sum_{k=1}^{j-1} q_k \) becomes close to 0 for large \( j \) (i.e., no observations come from the \( j \)th membership). Therefore, the finite representation in (1) can automatically estimate the number of non-empty space-time clusters \( M^* \) out of \( M \) possible clusters.
We can also incorporate the landmark information into STDPG. Consider we have $\ell = 1, \ldots, p$ different types of landmarks and let $\{z_i^\ell\}_{i=1}^N \in S$ be the location of the closest landmark $\ell$ from the $i$th observation. Given the cluster membership $g_i$, the space-time locations given by confirmed individuals can be modeled as

$$
f((x_i, t_i)|g_i = j) \propto \frac{1}{\omega_{s\ell} \omega_{t\ell}} \exp \left( -\frac{||x_i - c_{i\ell}||^2}{2\omega_{s\ell}^2} - \frac{(t_i - c_{t\ell})^2}{2\omega_{t\ell}^2} \right) \times \prod_{\ell=1}^p \frac{1}{\omega_{s\ell} \omega_{t\ell}} \exp \left( -\frac{||z_i^\ell - c_{i\ell}||^2}{2\omega_{s\ell}^2} \right),$$

(2)

where $\omega_{s\ell}, \omega_{t\ell}, \{\omega_{s\ell}\}_{\ell=1}^p$ denote space, time, and landmarks range parameters, respectively. The first term in (2) comes from the 3-dimensional multivariate normal distribution with a $3 \times 3$ diagonal covariance matrix whose elements are $\omega_{s\ell}^2, \omega_{t\ell}^2, \omega_{s\ell}^2; \frac{1}{\omega_{s\ell} \omega_{t\ell}}$ is calculated from the determinant of the covariance matrix. In our context, $\omega_{s\ell}$ and $\omega_{t\ell}$ control the width of space-time cluster events, and $\omega_{s\ell}$ can measure the effect of the closest landmark $\ell$. Smaller $\omega_{s\ell}$ implies that the $\ell$th landmark is located nearby space-time cluster centers; therefore, the $\ell$th landmark can be regarded as important in determining disease hot spots. We set non-informative priors for range parameters and update them sequentially based on results from previous time windows. We describe the details of sequential update procedures in Section 3.2.

From the Gaussian mixture model described above, the membership variable for each $i$th observation is sampled from

$$g_i \sim \text{Categorical}(\tilde{q}_{ij}, \ldots, \tilde{q}_{iM}),$$

where

$$\tilde{q}_{ij} \propto q_{ij} \frac{1}{\omega_{s\ell} \omega_{t\ell}} \exp \left( -\frac{||x_i - c_{i\ell}||^2}{2\omega_{s\ell}^2} - \frac{(t_i - c_{t\ell})^2}{2\omega_{t\ell}^2} \right) \times \prod_{\ell=1}^p \frac{1}{\omega_{s\ell} \omega_{t\ell}} \exp \left( -\frac{||z_i^\ell - c_{i\ell}||^2}{2\omega_{s\ell}^2} \right),$$

(3)

for $j = 1, \ldots, M$. Intuitively, the first term considers the space-time clustering, and the second term measures the impact of the closest landmark on the membership probabilities. Note that we have different types of landmarks which are located nearby each other in Cali (Figure 2). Therefore, quantifying the impact of different landmarks is a challenging but very important epidemiological question. (3) implies that an unobserved cluster center is likely to be estimated nearby an event and different landmark variables. Especially, the impact of different landmarks is separately quantified through $\{\omega_{s\ell}\}_{\ell=1}^p$. At the same time, our model can quantify the burden of multiple disease sources through $\tilde{q}_{ij}$; we can calculate the probability of the $i$th event belonging to the $j$th cluster. Therefore, STDPG utilizes landmark information to (1) measure the impact of different landmarks and (2) quantify the burden of multiple disease sources. Note that the stick breaking prior for $q_i$ in (1) encourages only $M^* (< M)$ clusters to be non-empty (ie, at least a single observation comes from the cluster). From this model specification, the full likelihood function is

$$L(\theta|X, c) \propto \prod_{i=1}^N \sum_{\text{non-empty}} q_{ij} \frac{1}{\omega_{s\ell} \omega_{t\ell}} \exp \left( -\frac{||x_i - c_{i\ell}||^2}{2\omega_{s\ell}^2} - \frac{(t_i - c_{t\ell})^2}{2\omega_{t\ell}^2} \right) \times \prod_{\ell=1}^p \frac{1}{\omega_{s\ell} \omega_{t\ell}} \exp \left( -\frac{||z_i^\ell - c_{i\ell}||^2}{2\omega_{s\ell}^2} \right),$$

(4)

where $\theta = (\omega_{s\ell}, \omega_{t\ell}, \{\omega_{s\ell}\}_{\ell=1}^p)$ is a parameter vector. We have both COVID-19 events and landmark variables as observations. Given these observations, (3) implies that an unobserved cluster center $(c_{i\ell}', c_{t\ell}')$ is likely to be located nearby an event $(x, t)$ as well as the landmark location $z_i^\ell$. In our preliminary study, we observed that by including landmark variables, the model offers additional cluster centers compared to those from the model without landmarks. This implies that the cluster centers are also detected near landmarks, which is realistic in that the landmarks act as sources of infections. Therefore, with a different set of landmarks, the model can identify different clusters, which is natural. However, the cluster centers detected from the cases would be similar because the contribution of the cases to the likelihood function remains. Including other variables (eg, socioeconomic variables) can also affect cluster locations in DP processes. Considering that such variables are not provided with the point level, data aggregation method can be combined with our method to address misalignment issues. Note that in (4), both events-related terms and landmark-related terms are defined through the Gaussian kernels. This implies that $\{\omega_{s\ell}\}_{\ell=1}^p$ and $\omega_{s\ell}, \omega_{t\ell}$ play a similar role; the measurement of the spread between cluster centers and space-time events (or landmark variables). Based on this, we provide interpretations in Section 4. We provide the conditional distributions for all parameters in Appendix.
3.2 Rolling-window analysis

In this section, we describe our rolling-window strategies with their justification. In a Bayesian framework, it is natural to use the previous window’s information as a prior for the current model fitting. The procedure of rolling-window updates is as follows. For a given time window \( w \), we use a prior \( \pi^{(w)}(\theta) \) as \( N(\mu^{(w-1)}, \Sigma^{(w-1)}) \mathbf{I}(\theta > 0) \), where \( \mu^{(w-1)} \) is a posterior mean vector and \( \Sigma^{(w-1)} \) is a diagonal matrix whose elements are the posterior variance of parameters from window \( w - 1 \). Here, we use the truncated normal prior to ensure range parameters are well defined. We multiply a constant \( c = 2 \) to \( \Sigma^{(w-1)} \) to make prior wider. Furthermore, in our MCMC updates, we use the initial locations of cluster centers \( c^{(w)} \) from the last updated location of cluster centers \( c^{(w-1)} \). For \( w = 1 \), we use a positive non-informative prior (ie, \( \pi^{(1)}(\theta) = \mathbf{I}(\theta > 0) \)) and use randomly generated locations from landmarks as \( c^{(1)} \).

The first advantage is its interpretability compared to the full data analysis. For a spatio-temporal analysis of point processes, the observed events need to be separated into different time steps to capture the dynamics of event occurrence over time. The length of time step needs to be chosen based on domain knowledge so that the resulting spatio-temporal pattern provides a scientifically meaningful representation of the disease spread dynamics. Based on the previous studies of the early stages of the COVID-19 spread, two weeks is the average period of recovery from infection; \(^{1, 28}\) the impact of a confirmed case at a specific space-time location would become weaker as time goes by. Therefore, it would be most informative to utilize the dataset collected over the last several weeks to detect space-time cluster centers of disease spread.

Also, the rolling-window updates are computationally much more efficient than all-at-once updates. Although our model shows much faster mixing than interaction point process models, \(^{12, 29}\) fitting STDPG with the entire dataset \((N = 16, 309)\) would still be computationally demanding. Instead, we fit STDPG to the subset of the dataset from the considered time window while at the same time incorporating information from the previous time window into the current model fit. This would be particularly useful for constructing a real-time disease warning system because we can quickly conduct inference when new confirmed cases are collected.

Lastly, we can obtain reasonable initial values of an MCMC algorithm based on the previous window’s information. As we described, our method requires updating both model parameters and unobserved (latent) cluster centers. Although one can use arbitrary initial values for the update, this can lead to a slower mixing of the chain for the high-dimensional posterior.

In this paper, we fit STDPG with two different windows (last two weeks and last four weeks) at any given time point. Figure 4 shows a time series plot of confirmed cases of COVID-19 with different time windows. The main goal of this study is to detect space-time cluster centers of confirmed cases as well as to examine the time-varying effects of parameters across different times of the disease spread.
3.3 Model structure and implementation details

Here, we summarize the hierarchical structure for STDGP as follows.

\[
\text{Likelihood: } \prod_{i=1}^{N} \sum_{\{x\}_{\text{non-empty}}} q_i \frac{1}{\omega_i^2 \omega_t} \exp \left( -\frac{\|x_i - c_j^t\|^2}{2\omega_t^2} - \frac{(t_i - c_j^t)^2}{2\omega_t^2} \right) \\
\times \prod_{i=1}^{p} \frac{1}{\omega_i^2} \exp \left( -\frac{\|z_i' - c_j^t\|^2}{2\omega_t^2} \right)
\]

**Dirichlet process prior:** \(g_t \sim \text{Categorical}(q_1, \ldots, q_M)\)

\(q_j = U_1\) for \(j = 1\) and \(q_j = U_j \prod_{k=1}^{j-1} (1 - U_k)\) for \(j = 2, \ldots, M\)

\(U_1, \ldots, U_k \sim_{\text{iid}} \text{Beta}(1, b_u)\)

\(b_u \sim \text{Gamma}(1, 1/4)\)

**Parameter prior:** \(\theta \sim N(\mu^{(w-1)}, \Sigma^{(w-1)}) I(\theta > 0)\)

**Cluster location prior:** \(c \sim \frac{1}{|S| T}\)

From the above specification, we can construct the joint posterior distribution \(\pi(\theta, c, \{g_t\}_{i=1}^{N}, \{U_j\}_{j=1}^{M}, b_u|X)\). Here, \(\theta = (\omega_t, \omega_t, \{\omega_t\}_{\text{grid}})\) is the range parameters, \(c = \{(c_j^t, c_j^s)\}_{j=1}^{M}\) is the unobserved (latent) space-time cluster centers, and \((\{g_t\}_{i=1}^{N}, \{U_j\}_{j=1}^{M}, b_u)\) is a set of Dirichlet process parameters. For \(\theta\), we use the truncated normal prior at each time window \(w\) as described in Section 3.2. For \(c\), we use the uniform prior defined through the area of the space and the length of the time. The conditional distributions for all components are described in the Appendix. An MCMC algorithm for STDGP is summarized in Algorithm 1.

3.4 Model assessment

To validate our models, we compare the observed proportion of data points and the theoretical proportion of density from the likelihood function. First, we construct \(G = n_x \times n_y \times n_t\) number of rectangular cubes covering our study domain \(S \times T\). Here, we use \(n_x = 8\) and \(n_y = 13\) numbers of grids for the longitudinal and latitudinal axes, respectively, to make rectangular cubes (ie, longitudinal range/8 \(\approx\) latitudinal range/13). For the time axis, we use \(n_t = 6\) for 2 weeks window and \(n_t = 6\) for 4 weeks window so that each cube covers approximately 4.6 days. One may use different choices of \(n_x\), \(n_y\), and \(n_t\) for model assessment. However, if we use extremely small or large size cubes, it would be challenging to compare the observed and theoretical proportions. For instance, if we use \(G = 1 \times 1 \times 1\) (ie, too large size cube), we would only have a single pair of \((p_{k^t}^g, p_{k^s}^g)\), which is hard to compare. On the other hand, if we use \(G = 100 \times 100 \times 100\) (ie, too small size cubes), we would have a lot of zeros for \((p_{k^t}^g, p_{k^s}^g)\) because there will be no data points (or mass) in these small cubes. In our preliminary studies, we observe that the results are robust for moderate grid sizes (ie, \(n_x \times n_y \times n_t \in [6, 10] \times [10, 15] \times [2, 10]\)). From (4), we can obtain the theoretical proportion of density over a cube \([a_g, b_g]^3\) as

\[
p_{k^t}^g = \int_{[a_g, b_g]^3 \in S \times T} \sum_{\{x\}_{\text{non-empty}}} q_i \frac{1}{\omega_i^2 \omega_t} \exp \left( -\frac{\|x - c_j^t\|^2}{2\omega_t^2} - \frac{(t - c_j^t)^2}{2\omega_t^2} \right) dX,
\]

(5)

where \(X \in S \times T\) is the spatio-temporal point process. In practice, we obtain \(\widehat{p}_{k^t}^g\) via numerical integration using \texttt{pmvnorm} function in R. Similarly, we can calculate the observed proportion of data points located in a cube \([a_g, b_g]^3\) as

\[
p_{k^t}^o = \frac{\text{the number of data points in } [a_g, b_g]^3}{\text{total number of observations}}.
\]
Algorithm 1. STDGP MCMC algorithm

Given \{\theta^{(m)}, c^{(m)}, \{g^{(m)}_i\}_{i=1}^N, \{U^{(m)}_j\}_{j=1}^M, b^{(m)}_u\} at the \(m\)th iteration

Update the range parameters \(\theta^{(m+1)}\):
Propose \(\theta^* \sim q(\cdot | \theta^{(m)})\) and accept \(\theta^{(m+1)} = \theta^*\) with probability

\[
\alpha = \min \left\{ \frac{\pi(\theta^* | c^{(m)}, \{g^{(m)}_i\}_{i=1}^N, \{U^{(m)}_j\}_{j=1}^M, b^{(m)}_u)}{\pi(\theta^{(m)} | c^{(m)}, \{g^{(m)}_i\}_{i=1}^N, \{U^{(m)}_j\}_{j=1}^M, b^{(m)}_u)} \right\}.
\]

Update space-time cluster centers \(c^{(m+1)}\):
Propose \((c^{(m+1)}_j, \hat{c}^{(m+1)}_j) \sim \frac{1}{|S_j|}\) and accept \((c^{(m+1)}_j, \hat{c}^{(m+1)}_j) = (c^{(m+1)}_j, \hat{c}^{(m+1)}_j)\) with probability

\[
\alpha = \min \left\{ \frac{\pi(c^{(m+1)}, \hat{c}^{(m+1)} | \theta^{(m)}, \{g^{(m)}_i\}_{i=1}^N, \{U^{(m)}_j\}_{j=1}^M, b^{(m)}_u)}{\pi(c^{(m)}, \hat{c}^{(m)} | \theta^{(m)}, \{g^{(m)}_i\}_{i=1}^N, \{U^{(m)}_j\}_{j=1}^M, b^{(m)}_u)} \right\}
\]

for \(j = 1, \ldots, M\).

Update Dirichlet process parameters
Generate \(g^{(m+1)}_i \sim \text{Categorical}(\tilde{q}^1_i, \ldots, \tilde{q}^M_i)\), where

\[
\tilde{q}^i_j \propto q_j \frac{1}{\omega_j^{2(m+1)} \omega_i^{2(m+1)}} \exp \left( - \frac{|x_i - c^{(m+1)}_j|^2}{2 \omega_j^{2(m+1)}} - \frac{(t_i - \hat{c}^{(m+1)}_j)^2}{2 \omega_i^{2(m+1)}} \right) \prod_{j' \neq j} \frac{1}{\omega_j^{2(m+1)}} \exp \left( - \frac{|z_i - c^{(m+1)}_{j'}|^2}{2 \omega_j^{2(m+1)}} \right)
\]

for \(i = 1, \ldots, N\) and \(j = 1, \ldots, M\).

Generate

\[
U^{(m+1)}_j \sim \text{Beta}(1 + \sum_{i=1}^N I(g^{(m+1)}_i = j), b^{(m)}_u + \sum_{i=1}^N I(g^{(m+1)}_i > j))
\]

for \(j = 1, \ldots, M\).

Generate

\[
b^{(m+1)}_u \sim \text{Gamma}(M, 1/4 - \sum_{j=1}^{M-1} \log(1 - U^{(m+1)}_j))
\]

for \(j = 1, \ldots, M\).

Then, we draw Q-Q plots using \(p^{(m)}_g, \hat{p}^{(m)}_g\), calculated from \(g = 1, \ldots, G\). If our model fits well, Q-Q plots would follow a straight line. Furthermore, we compute the mean-squared error between the observed and theoretical proportions as

\[
\text{MSE} = \frac{1}{G} \sum_{g=1}^{G} (\hat{p}^{(m)}_g - p^{(m)}_g)^2.
\]

Naturally, the smaller MSE values are, the better the model fits.

Such model validation is closely related to the residual analysis\(^{30}\) in that we also compare discrepancies between the observed and theoretical quantities. The difference comes from whether we use the number of points\(^{30}\) or the proportions of the points (our method). We also apply the residual analysis to diagnose spatial trends after model fittings. For each grid, we calculate the residuals through \(\hat{p}_g^m - p_g^0\). The model fits well if the residual values are small and do not show spatial trends.
4 | APPLICATION

Here, we apply our method to COVID-19 cases in Cali, Colombia. The proposed method is implemented through R and C++ using the Rcpp and RcppArmadillo packages. All codes were run on AMD Ryzen Threadripper 2990WX processors. Implementing the MCMC algorithm for the entire dataset takes about 39 and 35 h for 2 weeks intervals and 4 weeks intervals, respectively. For each time window, we run the MCMC algorithm for 20,000 iterations, with the first 10,000 iterations discarded for burn-in. In our analysis, we set the maximum number of clusters $M = 120$ for the 2 weeks window and $M = 200$ for the 4 weeks window, which can ensure the number of non-empty clusters $M^*$ is smaller than $M$. For parameter interpretation, we convert spatial range parameters $\omega_s, \{\omega_{t}\}_{\ell=1}^{p}$ into kilometers scale using the Haversine formula (see Appendix for details). We also convert the scale of temporal range parameter $\omega_t$ into days by multiplying it by 28.

4.1 | Real data analysis

4.1.1 | Visualization of COVID-19 hot spots

Figure 5 (with 4 weeks intervals) shows the posterior mean surfaces of the spatio-temporal densities for the observed cases. To illustrate the mean surfaces, we first construct space-time grids over our study domain $S \times T$. For each grid,
we calculate the density values from (4) for the given posterior mean of the parameters and the estimated cluster centers. To illustrate densities over 2-dimensional figures, we summed the density values across different time grids that belong to the same spatial grid. We repeat this for each time window to obtain posterior mean surfaces separately. Lastly, we normalize the density values across different time windows to the [0, 1] range for comparison. It shows how the detected hot spots vary and evolve with time. In the Appendix (Figure C1), we also provide a density map with 2 weeks intervals; both maps are sort of similar, indicating the way the disease was spreading. It started in the central and northeastern part of Cali, then showed generalized outbreaks later in May and June with increasing densities. Overall infected cases decreased later in July. To show how well these densities describe the spatio-temporal distribution of the observed cases, we created Q-Q plots for $p^o_k$ and $\hat{p}_k$ as previously described. The plots in Figure 6 show that the estimated spatio-temporal densities well capture the observed cases, as the estimated proportion $\hat{p}_k$ is in good agreement with the observed proportion $p^o_k$ in both rolling windows. We observe that MSE from 4 weeks window is smaller than that of 2 weeks window. Naturally, the more information is used, the more likely we can provide accurate estimates; the 4 weeks window uses longer cumulative data points than the 2 weeks window. However, MSEs are fairly small for both cases, indicating the models fit well and the results are robust across different window settings. We observe that there are no significant spatial patterns from the residual plots (Figure 7), and the residual values are distributed around zero. This implies that our model fits the observed data well for both 2 and 4 weeks windows. Considering that 2 weeks is the average period of recovery from COVID-19 infection, using 2–4 weeks as a time window can be a practical choice in terms of both statistical and epidemiological perspectives.

4.1.2 Interpretation of landmark parameters

To describe how the properties of case clusters change over time in detail, we plot the time series of posterior means of the range parameters $(\omega_s, \omega_n, \{\omega_r^p\}_{r=1}^5)$ in Figure 8. We fit our model for each window and obtain posterior distributions of the range parameters. For example, with the 4 weeks window, we can obtain five sets of range parameters across windows. Note that our rolling-window strategy naturally incorporates the information from the previous window through the prior, as we described in the previous section. We also provide posterior means and 95% HPD intervals for all types of range parameters in the Appendix; see Tables C1, C2, C3, and C4. As discussed in Section 3.1, a smaller range parameter value for a certain type of landmark means the clusters tend to show up closer to those landmarks. In other words, a smaller range value can be interpreted as a stronger influence of a landmark on the disease spread for the given period, whereas a larger value indicates a weaker influence.

As shown in Dong et al., more than half of the population concentrates in neighborhoods of low socioeconomic strata located mainly in the east, northeast, and west. People in these places mainly move around the neighborhood attending the closest church, going to the closest town hall to take care of their administrative needs with the city, and the students go to the closest school. This goes in line with what is reported by James et al., who highlight a high COVID-19 positive rate among attendees to events at places, such as churches or schools, as we use here. Also, these authors reported that among 92 attendees at a rural Arkansas church during March 6–11, 35 (38%) developed laboratory-confirmed COVID-19, and three persons died. Dong et al. performed a preliminary study of the same dataset and observed a strong spatial
correlation in the vicinity of landmark locations, while the correlation between two locations weakens with their distance. This was a motivation to develop a spatio-temporal point process model that can consider the impact of the closest landmarks.

The range parameter for school decreases as the cases increase for the first 14 weeks and then slightly increases as the cases decrease for the last 6 weeks. This indicates that schools may be more important in determining the disease hot spots when the number of cases is higher. The range parameter for the shopping center, on the other hand, moves in the opposite direction, showing that shopping centers play less important roles in detecting the hot spots when the cases are
more prevalent. This is perhaps because the higher number of COVID-19 cases did not much discourage people to visit schools, while certainly discouraged people to go to shopping centers. This suggests that the disease control effort should focus more on schools when the prevalence of disease is high, and more on shopping centers when the prevalence is low.

The range parameter for churches and bus stations did not change much over the study period, being the values always quite low (around 0.8 for churches and 0.7 for bus stations). This implies that public transportation is always important in detecting hot spots regardless of the phase of the disease spread. The range parameter for parks shows an interesting pattern. At the beginning, parks seemed to play an important role in detecting disease clusters, and then they became less important in the following weeks, becoming more important again in the last 12 weeks. This shows that people decided to avoid parks on the onset of the pandemic, but then changed their behavior later and visited parks again. This is perhaps not surprising given that outdoor activities were viewed as a safer alternative to indoor activities during the pandemic period.

The range parameter for the town hall shows much larger values, showing that the town halls are much weaker attractors of the local hot spots than the other types of landmarks. Its value was smaller in the first 4 weeks, and then became larger for the rest of the study period. This perhaps indicates that people decided to avoid the town center after seeing the rapid increase of the number of cases at the beginning, and then the behavior did not change even after the cases has decreased.

4.1.3 Interpretation of space and time range parameters

The estimated posterior means and HPD intervals (Appendix) indicate a significant spatio-temporal effect in the dataset, which coincides with a preliminary study in Dong et al. The range parameter for space takes values between 0.4 and 0.7 km (Figure 8), meaning about 95% of the observed points are located within $2 \times 0.5 = 1$ km radius from the cluster center on average, because normal kernels are used. The range parameter for time starts with a value of 5 days and then increases to around 8 days in the 2 weeks window results, and starts with a value of 6 days and then increases to around 7 days in the 4 weeks window results (Figure 8). This means that the 95% of the observed values are within a 2 weeks range or slightly longer from the cluster center.

Our approach can also visualize “risk boundaries” from parameter estimates; such boundaries would be useful for diagnosing the degree of COVID-19 risk. Figure 9 illustrates cluster centers and their spatial risk boundaries. Specifically, cluster centers $c$ are sampled from the conditional posterior distribution (see Appendix for details), and the radii are obtained by $2 \times \omega_s$. The range parameter for space seems to be negatively correlated to the number of total cases. As shown in Figure 9, the number of kernels got larger and their sizes became smaller as the number of cases increased. This pattern coincides with what has been observed by Park et al. in their similar but simpler purely spatial model. Individual clusters became smaller as the spatial distribution of events became denser, which led to a smaller value of the range parameter for spatial kernels. Figure 9 shows that the risk boundaries of the clusters are somewhat overlapped. Note that our model provides cluster membership probabilities for each point through $\tilde{q}_{ij}$ (i.e., probability that a point $(x_i, t_j)$ belongs to the $j$th cluster) as a by-product. Then the point is assigned to the cluster having a maximum membership probability. This implies that a single point can have positive membership probabilities for the multiple clusters. This is realistic in that multiple disease sources can affect a single disease event. Our method can quantify the burden of multiple disease sources through $\tilde{q}_{ij}$, which can be beneficial for constructing effective interventions. For example, we can raise the social distance level for the regions in clusters having the top three largest $\tilde{q}_{ij}$ to prevent the further spread of disease around the $i$th point. Risk boundary maps with two weeks intervals are also illustrated in Appendix (Figure C2). The range parameter for time shows a small discrepancy between the results based on 2 weeks intervals and 4 weeks intervals. The slightly shorter range (hence the longer time dependence) in the 4 weeks interval results seems to be a consequence of stronger smoothing effects due to the longer time window.

4.2 Simulation studies

To validate the proposed method, we conduct a simulation study. First, we simulate the space-time cluster centers $c$ randomly over the Cali region. Then for each cluster center, we simulate observations from a multivariate normal distribution centered at the cluster center with the range parameters $\omega_s$ and $\omega_t$. Similarly, we simulate six different types of landmarks $\{z^l\}_{l=1}^6$ from a multivariate normal distribution centered at the cluster center with the range parameters
Illustration of spatio-temporal clusters with 4 weeks intervals. Black dots indicate the location of cluster centers, and grey circles indicate their risk boundaries. The number of clusters (in order of time) is estimated 43, 59, 88, 112, 120, respectively.

Table 1: Inference results for simulated data sets.

| Parameter | True | Mean | SE  | Coverage |
|-----------|------|------|-----|----------|
| $\omega_s$ | 0.555 | 0.551 | 0.011 | 0.905    |
| $\omega_t$ | 5.600 | 5.598 | 0.208 | 0.975    |
| $\omega_1$ | 0.777 | 0.777 | 0.015 | 0.945    |
| $\omega_2$ | 0.666 | 0.668 | 0.012 | 0.985    |
| $\omega_3$ | 3.330 | 3.296 | 0.056 | 0.990    |
| $\omega_4$ | 0.888 | 0.888 | 0.017 | 0.955    |
| $\omega_5$ | 0.666 | 0.669 | 0.013 | 0.960    |
| $\omega_6$ | 0.666 | 0.669 | 0.011 | 0.975    |

Note: The mean of estimates, SE, and coverage are obtained from 200 simulations. As in the real data analysis, we covert spatial range parameters $w_s$, $\{\omega_f\}_{f=1}^6$ into kilometers scale and convert the temporal range parameter $\omega_t$ into days scale.
Note that this procedure mimics our data-generating mechanism in (2). For realistic simulations, we set the true parameters \((\omega_s, \omega_t, \left\{\omega_r \right\}_{r=1}^6)\) having similar values from the real data application, resulting in about 800 observations.

We repeat the simulation 200 times to study whether the true parameter values are well estimated. Table 1 indicates that our method can accurately estimate the simulated truth and provide reasonable empirical coverage close to the 95% nominal rate. Furthermore, we compare the true and estimated cluster centers to investigate how well the model detects the disease-spreading sources. For illustrative purposes, Figure 10 provides six cases from 200 repeated simulations. We obtain the estimated location of the cluster centers and the membership variables from the full conditional posterior distribution (see Appendix for details). Figure 10 indicates good agreements between the true and estimated cluster centers.
DISCUSSION

In this manuscript, we have proposed a spatio-temporal point process method based on the Dirichlet process Gaussian mixture. Our model can detect spatio-temporal clusters of disease with pragmatic parameter interpretations, which is one of the advantages over existing mechanistic point process model frameworks.\textsuperscript{35,36} To provide a realistic and on-time inference, we have provided a rolling-window scheme by incorporating prior information from the posteriors from the previous time window. We have also provided a spatio-temporal density map from the fitted model, which gives an intuitive epidemiological interpretation. We have validated our model by comparing the expected densities and the observations. The results indicate that our model captures well the spatio-temporal pattern of observed events. The ideas and method proposed in this paper can generally be applicable to model other clustering processes such as distribution of rat sightings\textsuperscript{37} and dynamics of crime.\textsuperscript{38}

From a computational perspective, the use of Dirichlet process allows us to detect disease spreading centers quickly with much lower computational cost, compared to the Neyman–Scott process approaches;\textsuperscript{12,21} fitting the Neyman–Scott process model is computationally prohibitive for complex spatio-temporal processes due to the slow mixing of birth-death MCMC. In terms of modeling and application aspects, our method provides insights on disease hot spots, which are useful for planning public health policy. For example, posterior means of $\omega_s$, $\omega_t$ provide the spatio-temporal range of disease spreading hot spots. From $\{\omega_{\ell}\}$, we can estimate the impact of major landmarks during disease outbreaks.

Note that our models assume (1) conditional independence for $X$ given $c$ and (2) independence for $c$, and still can capture a suitable dependence structure. Such assumptions are common under log Gaussian Cox process class models\textsuperscript{39} such as the Neyman–Scott processes.\textsuperscript{13} Our STDPG also assumes conditional independence among the observed points in $X$ given the unobserved cluster centers in $c$. The clustering behavior is captured through the dependence between $c$ and $X$, which is adequate for data from a pandemic situation where most observations form disease hot spots with many incidences. For a rare disease with a low number of incidences, considering direct interaction among observed points might be more helpful. Although our prior on the cluster centers in $c$ assumes their independence, the data model given by the STDPG in (4) will place them near the observed points $X$ and their corresponding landmarks $\{z_{\ell}^i\}_{i=1}^N$ as long as there are enough observations. This naturally induces clustering behavior for $c$. For low occurrence data; however, including some interactions among the cluster centers might be necessary. In such a case, inferential and computational issues caused by complex interactions need to be considered.

As per the future directions, one might consider adding interactions among spatio-temporal clusters through more advanced mixture models. For example, one might be able to formulate a model with a two-level hierarchy of clusters to introduce clustering behaviors of clusters. Another possibility is to construct a dependent covariance structure between Gaussian mixture components at the expense of computational costs. Our method utilizes Dirichlet processes,\textsuperscript{18,20} which have been developed for unsupervised learning such as clustering. Therefore, the direct application of STDPG to forecasting spatio-temporal events is challenging. We can consider clusters right after a short time period (e.g., 1 day), assuming that the distribution within that period can be well approximated based on the previous 2 (or 4) weeks; however, this is not a formal prediction. Lastly, including covariates describing the properties of different subareas (such as the socio-economic status of the areas) in the analysis might enable a more epidemiologically insightful analysis. Developing such extensions of our model may provide an interesting avenue for future research.

ACKNOWLEDGEMENTS
The authors are grateful to anonymous reviewers for their careful reading and valuable comments. This research was supported by the National Research Foundation of Korea (2020R1C1C1A0100386814, RS-2023-00217705), ICAN (ICT Challenge and Advanced Network of HRD) support program (RS-2023-00259934) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and National Project from the Spanish Ministry of Science (PID2019-107392RBI00/AEI/10.13039/501100011033).

DATA AVAILABILITY STATEMENT
The source code and data can be downloaded from https://github.com/jwpark88/STDP.

ENDNOTES
\textsuperscript{*}https://www.cali.gov.co/salud/
\textsuperscript{†}https://www.cali.gov.co/planeacion/
REFERENCES

1. Choi S, Ki M. Estimating the reproductive number and the outbreak size of COVID-19 in Korea. *Epidemiol Health*. 2020;42:e2020011.
2. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos Solitons Fract*. 2020;134:109761.
3. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS One*. 2020;15(3):e0230405.
4. Barlow NS, Weinstein SJ. Accurate closed-form solution of the SIRepidemic model. *PhysD:NonlinearPhenomena*. 2020;408:132540.
5. CDC. *How COVID-19 Spreads*. 2020 https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html.
6. Gajardo Á, Müller HG. Point process models for COVID-19 cases and deaths. *J Appl Stat*. 2023;50(11-12):2287-2293.
7. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS One*. 2020;15(3):e0230405.
How to cite this article: Park J, Yi S, Chang W, Mateu J. A spatio-temporal Dirichlet process mixture model for coronavirus disease-19. *Statistics in Medicine*. 2023;42(30):5555-5576. doi: 10.1002/sim.9925

APPENDIX A. FULL CONDITIONALS FOR MARKOV CHAIN MONTE CARLO

- The conditional distribution of range parameters \((\omega_s, \omega_t, \{\omega_\ell\}_{\ell=1}^p)\) takes the form:

\[
\pi(\omega_s, \omega_t, \{\omega_\ell\}_{\ell=1}^p | \text{others}) \propto \prod_{i=1}^N \int f(\{x_i, t_i\} | \omega_s, \omega_t, \{\omega_\ell\}_{\ell=1}^p) \pi(\omega_s) \pi(\omega_t) \prod_{\ell=1}^p \pi(\omega_\ell) \times \prod_{i=1}^N \frac{1}{f(\{x_i, t_i\} | \omega_s, \omega_t, \{\omega_\ell\}_{\ell=1}^p)} \left[ \exp \left( -\frac{||x_i - c_j'||^2}{2\omega_s^2} - \frac{(t_i - c_j')^2}{2\omega_t^2} \right) \right] \times \prod_{\ell=1}^p \frac{1}{\omega_\ell} \exp \left( -\frac{||z_\ell' - c_j'||^2}{2\omega_\ell^2} \right) \times \pi(\omega_s) \pi(\omega_t) \pi(\omega_1) \cdots \pi(\omega_p).
\]

- The conditional distribution of membership variable \(g_i\) is of the form:

\[
\pi(g_i | \text{others}) = \text{Categorical}(\tilde{q}_{i1}, \ldots, \tilde{q}_{iM}),
\]

where

\[
\tilde{q}_{ij} \propto q_{ij} \frac{1}{\omega_t^2}\omega_t \exp \left( -\frac{||x_i - c_j'||^2}{2\omega_s^2} - \frac{(t_i - c_j')^2}{2\omega_t^2} \right) \times \prod_{\ell=1}^p \frac{1}{\omega_\ell} \exp \left( -\frac{||z_\ell' - c_j'||^2}{2\omega_\ell^2} \right)
\]

for \(j = 1, \ldots, M\).

- The conditional distribution of space-time cluster centers \((c_j', c_j')\) becomes:

\[
\pi((c_j', c_j') | \text{others}) \propto \prod_{i: g_i = j} \left[ q_{ij} \frac{1}{\omega_t^2}\omega_t \exp \left( -\frac{||x_i - c_j'||^2}{2\omega_s^2} - \frac{(t_i - c_j')^2}{2\omega_t^2} \right) \times \prod_{\ell=1}^p \frac{1}{\omega_\ell} \exp \left( -\frac{||z_\ell' - c_j'||^2}{2\omega_\ell^2} \right) \right] \frac{1}{|S|}
\]

for \(j = 1, \ldots, M\).

- The conditional distribution of \(U_j\):

\[
\pi(U_j | \text{others}) = \text{Beta}(1 + \sum_{i=1}^N I(g_i = j), b_u + \sum_{i=1}^N I(g_i > j)).
\]

- The conditional distribution of \(b_u\) can be written as:

\[
\pi(b_u | \text{others}) = \text{Gamma}(M - 1 + a, b - \sum_{j=1}^{M-1} \log(1 - U_j)),
\]

where \(a = 1, b = 1/4\).
APPENDIX B. HAVERSINE FORMULA

Consider latitude and longitude coordinates \((x_1, y_1), (x_2, y_2)\). Then we can calculate the great-circle distance \(d\) (\(km\)) between two coordinates using the Haversine formula, as follows:

\[
\begin{align*}
dx &= \frac{\pi(x_2 - x_1)}{180}, \\
\frac{\pi(y_2 - y_1)}{180} \\
\sin^2 \left( \frac{dx}{2} \right) + \cos \left( \frac{\pi x_1}{180} \right) \cos \left( \frac{\pi x_2}{180} \right) \sin^2 \left( \frac{dy}{2} \right) \\
d &= 6371 \times 2 \arctan \left( \frac{\sqrt{a}}{\sqrt{1 - a}} \right).
\end{align*}
\]

APPENDIX C. ADDITIONAL RESULTS

**TABLE C1**  Posterior means and 95% HPD intervals (parenthesis) of space, time parameters for 2 weeks window.

|       | Space       | Time       |
|-------|-------------|------------|
| 3/16  | 0.67        | 4.96       |
|       | (0.59, 0.74)| (4.46, 5.54)|
| 3/30  | 0.66        | 7.84       |
|       | (0.61, 0.7 )| (7.36, 8.36)|
| 4/13  | 0.61        | 8.84       |
|       | (0.57, 0.65)| (8.28, 9.48)|
| 4/27  | 0.57        | 8.24       |
|       | (0.54, 0.6 )| (7.68, 8.78)|
| 5/11  | 0.45        | 8.16       |
|       | (0.43, 0.47)| (7.74, 8.56)|
| 5/25  | 0.47        | 8.14       |
|       | (0.46, 0.48)| (7.8, 8.48 )|
| 6/8   | 0.45        | 8.24       |
|       | (0.44, 0.46)| (7.98, 8.5 )|
| 6/22  | 0.5         | 7.98       |
|       | (0.49, 0.51)| (7.74, 8.18)|
| 7/6   | 0.53        | 8.28       |
|       | (0.52, 0.54)| (8.04, 8.5 )|
| 7/20  | 0.56        | 7.78       |
|       | (0.55, 0.57)| (7.54, 8)  |

**TABLE C2**  Posterior means and 95% HPD intervals (parenthesis) of landmark parameters for 2 weeks window.

|       | Church | School | Town hall | Shopping center | Bus | Park |
|-------|--------|--------|----------|-----------------|-----|------|
| 3/16  | 0.85   | 0.94   | 2.53     | 0.73            | 0.61| 0.67 |
|       | (0.75, 0.94)| (0.85, 1.04)| (2.29, 2.77)| (0.64, 0.81)| (0.53, 0.7)| (0.59, 0.77)|
| 3/30  | 0.9    | 0.8    | 2.74     | 0.82            | 0.73| 0.57 |
|       | (0.86, 0.96)| (0.76, 0.85)| (2.6, 2.89)| (0.77, 0.87)| (0.69, 0.77)| (0.53, 0.62)|
| 4/13  | 0.85   | 0.72   | 2.97     | 0.9             | 0.71| 0.77 |
|       | (0.8, 0.89)| (0.67, 0.76)| (2.84, 3.1)| (0.86, 0.95)| (0.67, 0.75)| (0.71, 0.82)|
### TABLE C2 (Continued)

|       | Church | School | Town hall | Shopping center | Bus | Park |
|-------|--------|--------|-----------|----------------|-----|------|
| 4/27  | 0.81   | 0.64   | 3.04      | 1              | 0.69| 0.76 |
|       | (0.78, 0.85) | (0.61, 0.68) | (2.92, 3.16) | (0.95, 1.05) | (0.65, 0.72) | (0.72, 0.8) |
| 5/11  | 0.76   | 0.55   | 2.88      | 1              | 0.72| 0.66 |
|       | (0.74, 0.79) | (0.53, 0.57) | (2.78, 2.97) | (0.96, 1.03) | (0.69, 0.74) | (0.63, 0.68) |
| 5/25  | 0.8    | 0.51   | 2.98      | 1.04           | 0.7 | 0.65 |
|       | (0.78, 0.82) | (0.49, 0.53) | (2.91, 3.05) | (1.01, 1.06) | (0.68, 0.73) | (0.63, 0.67) |
| 6/8   | 0.73   | 0.47   | 2.97      | 1.04           | 0.67| 0.6  |
|       | (0.72, 0.75) | (0.46, 0.49) | (2.9, 3.03) | (1.02, 1.06) | (0.65, 0.68) | (0.58, 0.61) |
| 6/22  | 0.81   | 0.56   | 2.97      | 0.96           | 0.71| 0.65 |
|       | (0.8, 0.82) | (0.55, 0.57) | (2.92, 3.02) | (0.95, 0.98) | (0.69, 0.72) | (0.64, 0.67) |
| 7/6   | 0.85   | 0.61   | 3.01      | 0.9            | 0.72| 0.69 |
|       | (0.83, 0.86) | (0.6, 0.62) | (2.97, 3.05) | (0.89, 0.91) | (0.71, 0.73) | (0.68, 0.7) |
| 7/20  | 0.81   | 0.67   | 2.9       | 0.85           | 0.69| 0.7  |
|       | (0.8, 0.82) | (0.66, 0.68) | (2.86, 2.94) | (0.84, 0.87) | (0.67, 0.7) | (0.69, 0.71) |

### TABLE C3 Posterior means and 95% HPD intervals (parenthesis) of space, time parameters for 4 weeks window.

|       | Space | Time |
|-------|-------|------|
| 3/30  | 0.64  | 5.52 |
|       | (0.61, 0.68) | (5.13, 5.95) |
| 4/27  | 0.58  | 7.54 |
|       | (0.56, 0.6) | (7.18, 7.9) |
| 5/25  | 0.46  | 7.37 |
|       | (0.45, 0.47) | (7.14, 7.6) |
| 6/22  | 0.48  | 6.99 |
|       | (0.47, 0.48) | (6.78, 7.21) |
| 7/20  | 0.55  | 7.05 |
|       | (0.54, 0.56) | (6.92, 7.19) |

### TABLE C4 Posterior means and 95% HPD intervals (parenthesis) of landmark parameters for 4 weeks window.

|       | Church | School | Town hall | Shopping center | Bus | Park |
|-------|--------|--------|-----------|----------------|-----|------|
| 3/30  | 0.89   | 0.82   | 2.7       | 0.8            | 0.7 | 0.56 |
|       | (0.84, 0.93) | (0.78, 0.87) | (2.59, 2.84) | (0.76, 0.85) | (0.66, 0.75) | (0.53, 0.6) |
| 4/27  | 0.82   | 0.67   | 3.05      | 0.94           | 0.67| 0.76 |
|       | (0.78, 0.85) | (0.64, 0.7) | (2.94, 3.15) | (0.91, 0.97) | (0.65, 0.71) | (0.74, 0.79) |
| 5/25  | 0.79   | 0.52   | 2.95      | 1.02           | 0.71| 0.64 |
|       | (0.77, 0.8) | (0.51, 0.55) | (2.9, 3.01) | (1, 1.04) | (0.69, 0.72) | (0.63, 0.66) |
| 6/22  | 0.79   | 0.52   | 2.96      | 1              | 0.69| 0.62 |
|       | (0.77, 0.8) | (0.51, 0.53) | (2.92, 3) | (0.98, 1.01) | (0.68, 0.71) | (0.61, 0.63) |
| 7/20  | 0.82   | 0.65   | 2.94      | 0.89           | 0.7 | 0.69 |
|       | (0.81, 0.83) | (0.64, 0.65) | (2.91, 2.98) | (0.87, 0.9) | (0.69, 0.71) | (0.68, 0.7) |
Posterior mean of density map for the observed cases with 2 weeks intervals.
FIGURE C2  Illustration of spatio-temporal clusters with 2 weeks intervals. Black dots indicate the location of cluster centers, and grey circles indicate their risk boundaries. The number of clusters (in order of time) is estimated 24, 41, 42, 49, 57, 71, 78, 89, 102, 86, respectively.