Received: 17 November 2020 | Revised: 30 July 2021 | Accepted: 20 September 2021

DOI: 10.1002/sim.9227

RESEARCH ARTICLE

Statistics in Medicine WILEY

Identification of geographic clusters for temporal heterogeneity with application to dengue surveillance

Pei-Sheng Lin[1,2]

[1]Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan

[2]Department of Mathematics, National Chung Cheng University, Minxiong, Taiwan

Correspondence
Pei-Sheng Lin, Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan.
Email: pslin@nhri.edu.tw

Identifying transmission of hot spots with temporal trends is important for reducing infectious disease propagation. Cluster analysis is a particularly useful tool to explore underlying stochastic processes between observations by grouping items into categories by their similarity. In a study of epidemic propagation, clustering geographic regions that have similar time series could help researchers track diffusion routes from a common source of an infectious disease. In this article, we propose a two-stage scan statistic to classify regions into various geographic clusters by their temporal heterogeneity. The proposed scan statistic is more flexible than traditional methods in that contiguous and nonproximate regions with similar temporal patterns can be identified simultaneously. A simulation study and data analysis for a dengue fever infection are also presented for illustration.

KEYWORDS

dengue infection, disease surveillance, scan statistics, spatial cluster, temporal heterogeneity

1 | INTRODUCTION

Cluster analysis is a useful tool to explore underlying structures of a stochastic process and relations between observations by grouping items into categories according to their similarity.[1] This analytical approach has long been of great importance in applications of epidemiology, including contagious disease surveillance. We are here particularly interested in clustering geographic units with similar time series of elevated risks while accounting for spatial-temporal correlation. A motivation for this article is the study and surveillance of dengue fever infection in Taiwan. Dengue fever is a disease caused by the flavivirus, which is transmitted by certain species of mosquito of the Aedes type, whose incubation period has a seasonal cycle in Taiwan. Due to climate change, dengue is now a global problem, and it has had outbreaks in more than 110 countries since the Second World War. Dengue fever infection also had consecutive outbreaks in Taiwan in 2014 and 2015, which caused 15,732 and 43,784 confirmed cases, respectively. Since there is no effective vaccine or specific medicine to treat dengue infection, dengue prevention usually puts more focus on vector control, such as chemical spraying and physical removal of breeding sites. A surveillance system that can map clusters of cases to explore epidemic propagation is thus essential to preventing a dengue outbreak.

Epidemic propagation is a process by which an infectious disease spreads from its origin. Taiwan, an island straddling the Tropic of Cancer, is an excellent place to investigate the epidemic propagation of dengue fever. Geographically, Taiwan is surrounded by China, Japan, and some southeast Asian counties. While dengue pandemics occurred often in southeast Asian counties, dengue outbreaks have been found in East Asian countries just in recent years. Since Taiwan forms

an isolated geographic environment and most dengue epidemics initiated from imported cases, a better understanding of the epidemic propagation could help other countries prevent their own dengue outbreaks. For dengue progression, the incubation period of mosquitos is a pivot factor. Since the incubation period depends on climate and environmental situations,[2] the epidemic diffusion would usually be related to an evolution of disease clusters that have space and time interaction.[3] An exploratory spatial analysis such as kernel density estimation may thus not be suitable to explore complex diffusion processes.[4] On the other hand, a disease mapping model that can characterize different temporal trends into their corresponding geographic clusters would be useful to evaluate epidemic progression in time and space for disease surveillance. For this purpose, we aim to develop an identification method to group regions into various geographic clusters by their temporal heterogeneity of incidence rates.

To cluster geographic regions that have similar time series, we note that these regions should also have similar annual incidences. So, in the first stage of the identification procedure, we modify the quasi-likelihood (QL) scan statistic by Lin et al[5] for the dengue data to find geographic clusters that are in proximity and have similar elevated (annual) disease rates. Then, proceeding from the geographic clusters identified by the spatial scan statistic, in the second stage we use a combination procedure based on a chi-squared test for comparison of temporal risks to regroup the clusters. Instead of considering candidate clusters with sizes up to the same order of the sampling regions, the proposed identification procedure uses a concept similar to local regression with restriction for sizes of candidate clusters in the first stage. By knitting the identified geographic clusters into temporal-heterogeneity (TH) clusters with suitable selection procedures in the second stage, the proposed method can sense regions that are distant from each other but have similar temporal trends. For convenience, we refer to the proposed method as a TH scan statistic.

When an infectious disease can be transmitted by persons in transportation, identification of nonproximate geographic clusters that share temporal similarity would provide useful information to track multiple routes of propagation that come from a common source, for example, as shown in the data analysis of Section 5. However, traditional scan statistics are mainly designed to cluster contiguous regions with similar mean infection rates.[6] On the other hand, although some classification methods, such as the k-means method, can be used to group nonproximate spatial-temporal units with similar event rates, there is no guarantee that temporal patterns can be observed from the grouped units since the k-means method does not supervise the known spatial structure. Additionally, the above method generally does not take correlation between observations into account. The TH scan statistic, on the other hand, can bridge clustering and classification methods. For example, in the first stage of the identification procedure, our method bears a similarity to the traditional clustering approaches by using the spatial scan statistic to identify clusters with contiguous regions. And, in the second stage of the TH scan statistic, we merge geographic clusters that may not be in proximity but have the same temporal patterns into TH clusters, which results in an identification result similar to the classification method. The TH scan statistic can thus not only identify geographic clusters in arbitrary shapes but also summarize temporal waves of outbreaks over the whole time period.

The remainder of this article is organized as follows. In Section 2, we establish cluster models and QL estimation methods for the proposed models. In Section 3, the QL function is adapted to develop a spatial scan statistic for geographic clusters, and then a chi-squared test is built to regroup geographic clusters. A simulation study is conducted to compare the TH scan statistic with other existing approaches in Section 4. Data analysis and discussion are given in Sections 5 and 6, respectively.

2 | CLUSTER MODELS AND ESTIMATES

2.1 | Model settings

Assume that we have a spatial-temporal data set with n geographic regions and T time periods. Let $(\boldsymbol{s}_i, t)$, $i = 1, \dots, n; t = 1, \dots, T$, denote a spatial-temporal coordinate for region i at time t, where $\boldsymbol{s}_i$ denotes a geographic location for the centroid of region i. Let $N_T = nT$ denote the total number of spatial-temporal observations. Also, let $O_{i,t}$ and $N_{i,t}$ denote numbers of observed cases and people at risk, respectively, in region $\boldsymbol{s}_i$ at time t. Assume that $O_{i,t}$ is affected by a spatial-temporal noise $\epsilon_{i,t}$ from a zero-mean Gaussian random field with variance σ^2 and correlation $\rho_{i,t;j,t'}$ for $\epsilon_{i,t}$ and $\epsilon_{j,t'}$. With consideration of regional noises, we are interested to know whether TH clusters for elevated infection rates exist. Let H denote a null hypothesis that the observations do not have a geographic cluster (but could have spatial-temporal correlation). Under H, an expected number of cases $E_{i,t}$ in region $\boldsymbol{s}_i$ at time t is given by $E_{i,t} = N_{i,t} \sum_{j,t'} O_{j,t'} / \sum_{j,t'} N_{j,t'}$. A spatial-temporal (standardized) incidence rate $Y_{i,t}$ is therefore given by $Y_{i,t} = O_{i,t}/E_{i,t}$.

We now fit a spatial-temporal cluster model for $Y_{i,t}$. In this article, each TH cluster is assumed to have its own temporal pattern, which is determined mainly by a risk coefficient within the given cluster in each time t. Let $C_k, k = 1, \ldots, K$, denote disjoint TH clusters, and let $\delta_{C_k}(\boldsymbol{s}_i)$ denote an indicator variable such that $\delta_{C_k}(\boldsymbol{s}_i) = 1$ if $\boldsymbol{s}_i \in C_k$. Let $\xi_{k,t}$ denote a log risk associated with cluster C_k at time t. Conditional on $\epsilon_{i,t}$, C_k, and $\xi_{k,t}$, an integrated model is given by

$$\log(Y_{i,t}|\epsilon_{i,t}, C_k, \xi_{k,t}) = \mu_0 + \sum_{k=1}^{K} \xi_{k,t}\delta_{C_k}(\boldsymbol{s}_i) + \epsilon_{i,t}, \tag{1}$$

where μ_0 denotes an intercept parameter. Note that model (1) is different from a traditional generalized linear model (GLM) since both $\xi_{k,t}$ and C_k are unknown. Thus, a traditional GLM method cannot be directly applied for estimation, and therefore the concept of scan statistics is later used to address the related estimation problem in Section 2.2.

To find geographic clusters by scan statistics, we next develop a spatially marginal model associated with C_k from (1). Let $O_{i,+} = \sum_{t=1}^{T} O_{i,t}$ and $E_{i,+} = \sum_{t=1}^{T} E_{i,t}$ denote sums of observed and expected cases, respectively, in region $\boldsymbol{s}_i$ over all time periods. A spatial (standardized) incidence rate in region $\boldsymbol{s}_i$ can thus be defined as $Y_{i,+} = O_{i,+}/E_{i,+}$. When $N_{i,t} \equiv N_i$ for all $t \in T$, the spatial incidence rate can be expressed as $Y_{i,+} = \sum_{t=1}^{T} Y_{i,t}$. Let $\xi_{k,+} = \sum_{t=1}^{T} \xi_{k,t}$ denote a yearly log-risk associated with cluster C_k and let $\epsilon_i^s = \sum_{t=1}^{T} \epsilon_{i,t}$ denote a spatial noise associated with region $\boldsymbol{s}_i$. It then follows from (1) that $Y_{i,+} = b_{i,+}\exp(\mu_s + \epsilon_i^s)$, where μ_s denotes an intercept and $b_{i,+} = \exp(\sum_{k=1}^{K} \xi_{k,+}\delta_{C_k}(\boldsymbol{s}_i))$, as derivation can be seen in the Appendix. So, conditional on C_k and ϵ_i^s, we can model $Y_{i,+}$ by

$$\log(Y_{i,+}|C_k, \xi_{i,+}, \epsilon_i^s) = \mu_s + \sum_{k=1}^{K} \xi_{k,+}\delta_{C_k}(\boldsymbol{s}_i) + \epsilon_i^s, \tag{2}$$

where ϵ_i^s is from a Gaussian process with mean zero, variance σ_s^2, and correlation $\rho_{i,j}^S$ for ϵ_i^s and ϵ_j^s.

Finally, we set a covariance structure for the responses by assuming that $\rho_{i,t;j,t'}$ is spatial-temporally separable. That is, $\rho_{i,t;j,t'} = \rho_{i,j}^S(h)\rho_{t,t'}^T(l)$, where $\rho_{i,j}^S(h)$ and $\rho_{t,t'}^T(l)$ denote the spatial and temporal correlations, respectively, with a Euclidean distance $h = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and a time lag $l = |t - t'|$ between t and t'. Also, let $o(\cdot)$ denote a little o function. We make the following assumptions for the dependence structure.

Assumption 1. (a) $\rho_{i,j}^S(h)$ is a positive-definite correlation function satisfying $\rho_{i,j}^S(h) = o(h)$ as $h \to \infty$. (b) The temporal correlation $\rho_{t,t'}^T(l)$ decays exponentially in lag $|t - t'|$.

We note that in model (1), $\xi_{k,t}$ is associated with time t, and therefore the temporal structure of $Y_{i,t}$ within a TH cluster is almost included in $(\xi_{k,1}, \ldots, \xi_{k,T})'$. Thus, it is reasonable to assume that the temporal correlation of $\epsilon_{i,t}$ is negligible in the sense that $\sum\sum_{t\neq t'} \rho_{t,t'}^T = o(T)$ as $T \to \infty$, as described in Assumption 1(b). It then follows from the definition of ϵ_i^s that $\text{cov}(\epsilon_i^s, \epsilon_j^s) = \sum_{t=1}^{T}\text{cov}(\epsilon_{i,t}, \epsilon_{j,t}) + \sum\sum_{t\neq t'} \text{cov}(\epsilon_{i,t}, \epsilon_{j,t'}) \approx T\sigma^2\rho_{i,j}^S$. This implies that $\text{corr}(\epsilon_i^s, \epsilon_j^s) = \text{corr}(\epsilon_{i,t}, \epsilon_{j,t}) = \rho_{i,j}^S$ and $\text{var}(\epsilon_i^s)(\equiv\sigma_s^2) = T\sigma^2$. This relationship will be used to connect spatial and spatial-temporal variations in later sections.

2.2 | QL estimating equations for scan statistics

For estimation of unknown parameters, we rely on QL estimating equations, which involve first-order and second-order moments. To develop a spatial estimating equation for model (2), let $\theta_{i,+} = E(Y_{i,+})$ denote the corresponding marginal mean. By moment generating functions for a normal distribution, we can get $\theta_{i,+} = \exp\left\{0.5\sigma_s^2 + \mu_s + \sum_{k=1}^{K} \xi_{k,+}\delta_{C_k}(\boldsymbol{s}_i)\right\}$ and $\text{cov}(Y_{i,+}, Y_{j,+}) = \theta_{i,+}\theta_{j,+}\{\exp(\sigma_s^2\rho_{i,j}^S) - 1\}$. Let $\boldsymbol{\Sigma}_s$ denote an $n \times n$ matrix with the (i,j) element to be

$$(\boldsymbol{\Sigma}_s)_{i,j} = \exp\left(\sigma_s^2\rho_{i,j}^S\right) - 1. \tag{3}$$

Let $\boldsymbol{Y}_s = (Y_{1,+}, \ldots, Y_{n,+})'$, let $\boldsymbol{\theta}_s = E(\boldsymbol{Y}_s) \equiv (\theta_{1,+}, \ldots, \theta_{n,+})'$, and let $\boldsymbol{\Theta}_s = \text{diag}(\boldsymbol{\theta}_s)$ denote a diagonal matrix formed by $\boldsymbol{\theta}_s$. The covariance matrix of $\boldsymbol{Y}_s$ can then be expressed as $\boldsymbol{V}_s = \boldsymbol{\Theta}_s\boldsymbol{\Sigma}_s\boldsymbol{\Theta}_s$. Let $\boldsymbol{D}_s = \partial\boldsymbol{\theta}_s/\partial(\mu_s, \xi_{1,+}, \ldots, \xi_{K,+})$ denote a derivative matrix of $\boldsymbol{\theta}_s$ with respect to $(\mu_s, \xi_{1,+}, \ldots, \xi_{K,+})$. The QL estimating equation for the spatial model (2) is given by

$$\boldsymbol{D}_s'\boldsymbol{V}_s^{-1}(\boldsymbol{Y}_s - \boldsymbol{\theta}_s)\big|_{(\mu_s=\hat{\mu}_s, \xi_{1,+}=\hat{\xi}_{1,+}, \ldots, \xi_{K,+}=\hat{\xi}_{K,+})} = \boldsymbol{0}, \tag{4}$$

where $\boldsymbol{V}_s^{-1}$ denotes an inverse matrix of $\boldsymbol{V}_s$. We refer to (4) as the spatial estimating equation, and to the estimates $\hat{\mu}_s$ and $\hat{\xi}_{k,+}$ as the QL estimates for μ_s and $\xi_{k,+}$, $k = 1, \ldots, K$, respectively. It is easy to see that the derivative matrix of $\boldsymbol{V}_s$ with respect to θ_s is symmetric, and therefore the QL estimating equation has a unique root in this setting.[7]

Similarly, we develop a spatial-temporal estimating equation for model (1). Let $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})'$ denote the temporally marginal data and let $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_T')'$ denote the ST data. Also, let $\theta_{i,t} = E(Y_{i,t})$. Then, it follows from the moment generating function that $\theta_{i,t} = \exp\{\mu_0 + \sum_{k=1}^{K} \xi_{k,t}\delta_{C_k}(\boldsymbol{s}_i) + 0.5\sigma^2\}$. Let $\boldsymbol{\theta}_t = (\theta_{1,t}, \ldots, \theta_{n,t})'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_T')'$. Let $\boldsymbol{\Sigma}$ denote an $n \times n$ matrix with the (i,j) element to be $(\boldsymbol{\Sigma})_{i,j} = \exp(\sigma_s^2 \rho_{i,j}^S / T) - 1$. By Assumption 1, the covariance matrix of $\boldsymbol{Y}$ can be approximated by

$$\boldsymbol{V} = \boldsymbol{\Theta}(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma})\boldsymbol{\Theta}, \tag{5}$$

where $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta})$, $\boldsymbol{I}_T$ is an identity matrix of size T, and $\otimes$ denotes a Kronecker product. Let $\boldsymbol{\xi}_k = (\xi_{k,1}, \ldots, \xi_{k,T})'$ denote the time series in cluster C_k, and let $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_K'\}'$. Also, let $\boldsymbol{D} = \partial\boldsymbol{\theta}/\partial(\mu_0, \boldsymbol{\xi})$ denote a derivative matrix of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\xi}$. The spatial-temporal estimating equation,

$$\boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\theta})\big|_{(\mu_0,\boldsymbol{\xi})=(\hat{\mu}_0,\hat{\boldsymbol{\xi}})} = \boldsymbol{0}, \tag{6}$$

is then used to estimate the parameters $(\mu_0, \boldsymbol{\xi})$, where $\boldsymbol{V}^{-1}$ denotes an inverse matrix of $\boldsymbol{V}$. We refer to $\hat{\mu}_0$ and $\hat{\boldsymbol{\xi}}$ as the QL estimates for μ_0 and $\boldsymbol{\xi}$, respectively.

In practice, locations of the true clusters are usually unknown; therefore, the scan statistic is commonly used to identify the clusters. Let Λ_m denote a candidate geographic cluster, and let $\boldsymbol{\Lambda} = \{\Lambda_m : m = 1, \ldots, M_n\}$ denote a collection of the candidate geographic clusters, where M_n denotes the total number of candidate clusters. Let $|A|$ denote a cardinality for set A. To ensure asymptotic properties for the QL estimating equation on Λ_m, the size of each candidate cluster should be restricted.[5] We make an assumption for the size of Λ_m.

Assumption 2. For each candidate cluster Λ_m, $m = 1, \ldots, M_n$, we require $|\Lambda_m| \leq Mn^{1/2}$ for some $M > 0$.

The idea of Assumption 2 for the spatial scan statistic is similar to the concept of local regression. In spatial epidemiology, identification of localized clusters is important for exploration of disease transmission. However, in some areas with uneven terrains, close regions may still have different disease transmission patterns. For this reason, restriction for the sizes of candidate clusters with local models may not only be useful to obtain statistical inference for the proposed scan statistic, but also provide ability to detect regions that have abrupt changes in disease rates. Thus, by applying the concept of local regressions in the identification process, the proposed scan statistic could more accurately identify temporal trends by piecewisely regrouping clusters with suitable criteria. In Section 3.2, we propose a test statistic to combine geographic clusters with similar patterns into a larger cluster. On the other hand, although the size of candidate clusters is limited, the scan statistic used in this article can allow the candidate clusters to be arbitrary shapes.

By Assumptions 1 and 2, it is reasonable to assume that $n^{-1}\boldsymbol{D}_s'\boldsymbol{V}_s^{-1}\boldsymbol{D}_s$ converges to a positive-definite matrix $\boldsymbol{\Upsilon}_s^0$ as $n \to \infty$. Let $\tilde{\boldsymbol{\Upsilon}}_s^0$ denote $\boldsymbol{\Upsilon}_s^0$ evaluated at the estimated parameters. Thus, under suitable conditions, we have

$$n^{1/2}\left(\hat{\mu}_s - \mu_s, \hat{\xi}_{1,+} - \xi_{1,+}, \ldots, \hat{\xi}_{K,+} - \xi_{K,+}\right)' \to N\left\{\boldsymbol{0}, \left(\tilde{\boldsymbol{\Upsilon}}_s^0\right)^{-1}\right\}, \tag{7}$$

in distribution as $n \to \infty$, where $(\tilde{\boldsymbol{\Upsilon}}_s^0)^{-1}$ denotes an inverse matrix of $\tilde{\boldsymbol{\Upsilon}}_s^0$. Similarly, we can assume that $n^{-1}\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}$ converges to a positive-definite matrix as $n \to \infty$. Under Assumption 2 and suitable conditions, we have asymptotic properties for the spatial-temporal estimating Equation (6). Details can be seen in the Appendix.

3 | IDENTIFICATION FOR TH CLUSTERS

3.1 | Scan statistics for geographic clusters

To identify local spatial clusters associated with $\boldsymbol{Y}_s$, we first use an independent scan statistic to search a collection of single clusters. Given the single cluster Λ_m, we fit a regression model for $Y_{i,+}$ by

$$\log\left(Y_{i,+}|\Lambda_m,\epsilon_i^s\right)=\mu_m+\xi_m\delta_{\Lambda_m}(\boldsymbol{s}_i)+\epsilon_i^s. \tag{8}$$

Under an independence assumption (ie, $\rho_{i,j}^S=0$), we estimate μ_m and ξ_m for model (8) by the QL estimating Equation (4), $m=1,\ldots,M_n$. Since we have M_n estimated cluster coefficients to be evaluated under the null hypothesis H, the Benjamini-Hochberg procedure[8] is used to control a false discovery rate (FDR) at level α. However, after the Benjamini-Hochberg procedure, some of the initial estimated clusters could have overlapping regions, which would cause some problems in statistical inference for a multiple-cluster model. To address this issue, a partition procedure for the initial estimated clusters is proposed to make these clusters disjoint. Details of the Benjamini-Hochberg and partition procedures can be seen in Section 5.2.

Let $I_1,\ldots,I_{K_0}$ denote the estimated clusters by the independent scan statistic after the partition procedure. We then evaluate whether the independence assumption is suitable. Based on $I_1,\ldots,I_{K_0}$, a multiple-cluster model for spatial data is given by $\log\left(Y_{i,+}|I_1,\ldots,I_{K_0}\right)=\mu_s+\sum_{k=1}^{K_0}\xi_{k,+}\delta_{I_k}(\boldsymbol{s}_i)+\epsilon_i^s$. The estimating equation (4) under the independence assumption is used again to obtain estimates $(\tilde{\mu}_s,\tilde{\xi}_{1,+},\ldots,\tilde{\xi}_{K_0,+})$ for $(\mu_s,\xi_{1,+},\ldots,\xi_{K_0,+})$. Residuals can then be computed by $\hat{\epsilon}_i^s=Y_{i,+}-\exp\left\{\tilde{\mu}_s+\sum_{k=1}^{K_0}\tilde{\xi}_{k,+}\delta_{I_k}(\boldsymbol{s}_i)\right\}$. For spatially correlated data, a variogram, $\gamma(h)=\mathrm{var}(\epsilon_i^s-\epsilon_j^s)$, is commonly used to measure spatial dependence. Under suitable conditions, $\gamma(h)$ can be estimated by an empirical variogram $\hat{\gamma}(h)=\sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in N(h)}\left(\hat{\epsilon}_i^s-\hat{\epsilon}_j^s\right)^2/|N(h)|$, where $N(h)$ denotes a collection of pairs $(\boldsymbol{s}_i,\boldsymbol{s}_j)$ with distance h apart. We obtain estimates $(\hat{\sigma}_s,\hat{\rho}_{i,j}^S)$ for $(\sigma_s,\rho_{i,j}^S)$ by a weighted least squares estimation[9]

$$\left(\hat{\sigma}_s,\hat{\rho}_{i,j}^S\right)=\arg\min_{\sigma,\rho_{i,j}^s}\sum_h|N(h)|\left\{1-\frac{\hat{\gamma}(h)}{\gamma(h;\sigma_s,\rho_{i,j}^S)}\right\}^2, \tag{9}$$

where $\gamma(h;\sigma_s,\rho_{i,j}^S)$ denotes a specific variogram model. In (9), the used weights are the numbers of pairs $N(h)$.

If $\hat{\rho}_{i,j}^S$ is *not* significantly away from zero, then $I_1,\ldots,I_{K_0}$ are the estimated clusters for the spatial data $\boldsymbol{Y}_s$, and the identification procedure jumps to the combination procedure shown in Section 3.2. When the correlation estimate $\hat{\rho}_{i,j}^S$ from (9) is significantly away from zero, it provides evidence for the existence of spatial correlation. In this situation, a spatial scan statistic by adding empirical estimates $\hat{\sigma}_s$ and $\hat{\rho}_{i,j}$ into the estimating equation is developed below to calibrate estimated clusters from the independent scan statistic. Specifically, let $\theta_{i,+}^*=\exp\left\{0.5\hat{\sigma}_s^2+\mu_s+\sum_{k=1}^{K_0}\xi_{k,+}\delta_{I_k}(\boldsymbol{s}_i)\right\}$ and $\boldsymbol{\theta}_s=\left(\theta_{1,+}^*,\ldots,\theta_{n,+}^*\right)'$. We then use the estimating equation (4) to estimate $(\mu_s,\xi_{1,+},\ldots,\xi_{K_0,+})$. However, in practice, the true covariance parameters in $\boldsymbol{V}_s$ are usually unknown. We therefore implement the estimates $\hat{\sigma}_s$ and $\hat{\rho}_{i,j}^S$ from (9) into σ_s and $\rho_{i,j}^S$, respectively, for $\boldsymbol{\Sigma}_s$ of (3). Let $\hat{\boldsymbol{\Sigma}}_s$ denote the estimate of $\boldsymbol{\Sigma}_s$ from the variogram method, and let $\hat{\boldsymbol{V}}_s=\boldsymbol{\Theta}_s\hat{\boldsymbol{\Sigma}}_s\boldsymbol{\Theta}_s$. The spatial estimating equation for the multiple-cluster model associated with $I_1,\ldots,I_{K_0}$ now becomes $\boldsymbol{D}_s'\hat{\boldsymbol{V}}_s^{-1}(\boldsymbol{Y}_s-\boldsymbol{\theta}_s)=\boldsymbol{0}$, where $\hat{\boldsymbol{V}}_s^{-1}$ is an inverse matrix of $\hat{\boldsymbol{V}}_s$.

Let $(\hat{\mu}_s,\hat{\xi}_{1,+},\ldots,\hat{\xi}_{K_0,+})$ denote the corresponding QL estimates for $(\mu_s,\xi_{1,+},\ldots,\xi_{K_0,+})$. By the limiting distribution (7), we can obtain the P-values of $\hat{\xi}_k$. The significance of each estimated cluster I_k, $k=1,\ldots,K_0$, is evaluated again by the corresponding P-value. For those estimated clusters with nonsignificant P-values, we remove them from the collection of estimated clusters. Let $G_1,\ldots,G_{K_1}$, $K_1\le K_0$, denote the significant identified clusters by the spatial scan statistic.

3.2 | A combination procedure for geographic clusters

To regroup geographic clusters $G_1,\ldots,G_{K_1}$ such that temporal patterns are different between TH clusters but similar within the clusters, we first estimate the temporal pattern in each geographic cluster G_k, $k=1,\ldots,K_1$. For the given cluster G_k, a spatial-temporal model from (1) for the marginal mean is given as

$$\theta_{i,t}^{(k)}=\exp\left\{0.5\sigma^2+\mu_0+\xi_{k,t}\delta_{G_k}(\boldsymbol{s}_i)\right\}, \tag{10}$$

where $\xi_{k,t}$ denotes a cluster coefficient associated with G_k at time t. Note that in (10), the intercept parameter μ_0 is set to be the same for all k, $k=1,\ldots,K_1$, so that the temporal patterns can be compared under the same baseline. The spatial-temporal estimating equation (6) is then applied to estimate the parameters $\boldsymbol{\xi}_k=(\xi_{k,1},\ldots,\xi_{k,T})'$. Since the true

covariance matrix $\boldsymbol{V}$ of (5) is unknown, we use $\hat{\boldsymbol{V}} = \boldsymbol{\Theta}(\boldsymbol{I}_T \otimes \hat{\boldsymbol{\Sigma}})\boldsymbol{\Theta}$ to replace $\boldsymbol{V}$ in the spatial-temporal estimating equation, where $\hat{\boldsymbol{\Sigma}}$ is the estimate of $\boldsymbol{\Sigma}$ with σ_s^2 and $\rho_{i,j}^S$ being estimated by the variogram method. Let $\hat{\boldsymbol{\xi}}_k = (\hat{\xi}_{k,1}, \ldots, \hat{\xi}_{k,T})'$ denote the estimated temporal pattern by the QL estimating equation for $\boldsymbol{\xi}_k$, $k = 1, \ldots, K_1$.

Based on the estimated temporal patterns, we now propose a chi-squared test to evaluate whether two geographic clusters G_k and $G_{k'}$ (that may not be adjacent) should be combined. Let $\boldsymbol{\theta}^{(k)} = \left(\theta_{1,1}^{(k)}, \ldots, \theta_{n,1}^{(k)}, \ldots, \theta_{1,T}^{(k)}, \ldots, \theta_{n,T}^{(k)}\right)'$ and $\boldsymbol{\theta}^{(k')} = \left(\theta_{1,1}^{(k')}, \ldots, \theta_{n,1}^{(k')}, \ldots, \theta_{1,T}^{(k')}, \ldots, \theta_{n,T}^{(k')}\right)'$. Let $\boldsymbol{D}_{\xi_k} = \partial\boldsymbol{\theta}^{(k)}/\partial\boldsymbol{\xi}_k$ and $\boldsymbol{D}_{\xi_{k'}} = \partial\boldsymbol{\theta}^{(k')}/\partial\boldsymbol{\xi}_{k'}$ denote derivative matrices of $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\theta}^{(k')}$ with respect to $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_{k'}$, respectively. By Assumptions 1 and 2, it is reasonable to assume that $n^{-1}\boldsymbol{D}'_{\xi_k}\boldsymbol{V}^{-1}\boldsymbol{D}_{\xi_k} \to \boldsymbol{\Upsilon}_{\xi_k}$ as $n \to \infty$. The chi-squared test statistic,

$$U_{k,k'} = n\left(\hat{\boldsymbol{\xi}}_k - \hat{\boldsymbol{\xi}}_{k'}\right)'\left(\hat{\boldsymbol{\Upsilon}}_{\xi_k}^{-1} + \hat{\boldsymbol{\Upsilon}}_{\xi_{k'}}^{-1}\right)^{-1}\left(\hat{\boldsymbol{\xi}}_k - \hat{\boldsymbol{\xi}}_{k'}\right), \tag{11}$$

is proposed to evaluate temporal heterogeneity between the two geographic clusters G_k and $G_{k'}$, where $\hat{\boldsymbol{\Upsilon}}_{\xi_k}$ denotes $\boldsymbol{\Upsilon}_{\xi_k}$ evaluated at the estimated parameters. When the two geographic clusters G_k and $G_{k'}$ have the same temporal pattern, $U_{k,k'}$ follows a chi-square distribution with degrees of freedom T. Thus, G_k and $G_{k'}$ are combined if $U_{k,k'} \le \chi^2_{T,1-\alpha}$, where $\chi^2_{T,1-\alpha}$ denotes a $100(1-\alpha)$-percentile of the chi-squared distribution with degrees of freedom T. In the data analysis, we present a regrouping procedure based on a forward selection procedure associated with the chi-squared statistic $U_{k,k'}$. Details can be seen in Section 5.3. Note that this procedure can also be applied to further partition larger clusters into smaller ones.

Let $C_1, \ldots, C_K$ denote the final TH cluster from the regrouping procedure. For the spatial-temporal model (1) associated with $C_1, \ldots, C_K$, we apply the spatial-temporal estimating Equation (6) again to estimate $\boldsymbol{\xi}_k = (\xi_{k,1}, \ldots, \xi_{k,T})'$ for $k = 1, \ldots, K$. The expected standardized incidence rate can thus be estimated from the TH cluster model by $\hat{\theta}_{i,t} = \exp\left\{\hat{\mu}_0 + \sum_{k=1}^K \hat{\xi}_{k,t}\delta_{C_k}(\boldsymbol{s}_i)\right\}$. The whole identification procedure shown in Section 3 is therefore called the TH scan statistic method.

4 | SIMULATION

In this section, we conduct a simulation study by using the geographic structure of Kaohsiung City, which consists of 891 villages. The purpose of the simulation study is to evaluate whether the proposed method can cluster hot-spot villages whose temporal patterns of elevated incidence rates are the same. Let Ω denote a collection of the 891 villages, and let $\boldsymbol{s}_i \in \Omega$ denote the administrative centroid for the ith village. Since Kaohsiung is partitioned by rivers and hills, using nearest neighbors to construct candidate clusters would probably be better than traditional distance methods. We thus use the concept of nearest neighbors to define candidate clusters. Specifically, for a given village $\boldsymbol{s}_i$, we define its lth-order neighbors $B_i^{(l)}$ to be a collection of villages that share a common border with $B_i^{(l-1)}$, $l = 1, 2, \ldots$, with $B_i^{(0)} \equiv \{\boldsymbol{s}_i\}$. Similar to the procedure used in Section 5, candidate clusters associated with the centroid $\boldsymbol{s}_i$ are unions of its neighbors up to the third order. That is, the candidate clusters associated with $\boldsymbol{s}_i$ are given by $\left\{\cup_{l=0}^L B_i^{(l)}, L = 0, \ldots, 3\right\}$. More details on how to choose centroids to make candidate clusters can be seen in the data analysis of Section 5. In total, there are 855 candidate clusters. Besides using the TH scan statistic, we also employ the *SaTScan* method with two different settings to analyze the simulated data for comparisons.

In the simulation study, we consider two scenarios. The first mainly follows the dengue data analysis result shown in Section 5. Five geographic clusters identified by the spatial scan statistic in the data analysis, $G_1, \ldots, G_5$, are chosen as hot-spot regions. Figure 1A shows the locations of $G_1, \ldots, G_5$ on a map of downtown Kaohsiung with $|G_1| = 34$, $|G_2| = 21$, $|G_3| = 10$, $|G_4| = 23$, and $|G_5| = 23$. (Downtown Kaohsiung consists of 528 villages with most dengue cases happening there.) Four of the five geographic clusters, $G_2, \ldots, G_5$, are set to have the same temporal pattern, while geographic cluster G_1 has its own temporal pattern (Figure 1B). Let $C_1 \equiv G_1$ and $C_2 \equiv G_2 \cup \cdots \cup G_5$ denote the TH clusters. The total numbers of villages in C_1 and C_2 are 34 and 77, respectively. As can be seen in Figure 1B, some villages in the TH cluster C_2 are in proximity, but others are not contiguous. Also, some villages (cluster G_5) in the TH cluster C_2 are next to the TH cluster C_1, while clusters G_5 and C_1 have similar annual incidence rates but different temporal patterns. These characteristics may make traditional scan statistics difficult to accurately identify the TH clusters C_1 and C_2.

In the second simulation scenario, we choose two proximate geographic clusters $G_1 \equiv C_1$ and G_5 (representing C_2) with various sizes $a|G_1|$ and $a|G_5|$, $a = 0.5$ or 1. The temporal patterns set in G_1 and G_5 are also the same as

FIGURE 1 In the dengue data analysis, locations of the estimated clusters by the (A) spatial scan statistic and (B) regrouping procedure for temporal heterogeneity

those set in the first simulation scenario. Note that the geographic clusters G_1 and G_5 belong to the TH clusters C_1 and C_2, respectively, and therefore the temporal patterns in G_1 and G_5 are different. Figure A1 in the Supplementary Material depicts locations of the half-size geographic clusters, say, $0.5G_1$ and $0.5G_2$. We use the second simulation scenario to evaluate finite sample properties for the TH scan statistic. To simulate responses for both scenarios, we first generate a spatial-temporal noise $\epsilon_{i,t}$ from a Gaussian random field with mean zero, variance $\sigma^2 = 0.01$, 0.02, or 0.05, and correlation $\text{corr}(\epsilon_{i,t}, \epsilon_{j,t'}) = \rho^S_{i,j}(h)\rho^T_{t,t'}(l)$. In the simulation, the spatial correlation function is set to be $\rho^S_{i,j}(h) = 0.55 - 0.83(h/1.23) + 0.28(h/1.23)^3$, which is the same as that given in (13) from the data analysis. Also, due to seasonal dengue infection in Taiwan, we set $\rho^T_{t,t'}(l) = \kappa^{|t-t'|}$ if t and t' are both in the interval [30, 51], and 0, otherwise. In the simulation setting, we consider $\kappa = 0$, 0.1, or 0.3. Let $\epsilon^r_{i,t}$ denote the simulated spatial-temporal noise for the rth simulation run, $r = 1, \ldots, R$, where R denotes a total number of simulation runs. Given $\epsilon^r_{i,t}$, a temporal pattern of responses for village $\boldsymbol{s}_i$, $i = 1, \ldots, 894$, is simulated by $Y^r_{i,t} = \exp\{-0.077 + \xi_{1,t}\delta_{C_1}(\boldsymbol{s}_i) + \xi_{2,t}\delta_{C_2}(\boldsymbol{s}_i) + \epsilon^r_{i,t}\}$, $t = 1, \ldots, 52$, where $\delta_{C_k}(\boldsymbol{s}_i)$ denotes an indicator variable for whether $\boldsymbol{s}_i$ is in C_k, and values of $\xi_{k,t}$, $k = 1, 2$, are the same as those given in model (16). (Figure 2B also depicts $\xi_{k,t}$, $k = 1, 2$, $t = 1, \ldots, 52$.) In the simulation, we simulate $R = 200$ replicates for each setting.

In applying the TH scan statistic for the simulated data, we conduct an identification procedure similar to the one used for the data analysis of Section 5. Recall that the collection of candidate clusters used for the TH scan statistic is the same as that in the data analysis (Section 5.1), which has 855 candidate clusters. The level of significance for the FDR is controlled at $\alpha = 0.05$. In use of the *SaTScan* software, the first setting for *SaTScan*, say LR_1, is to use the default setting, which means each candidate cluster including up to 50% of the population. On the other hand, in the second setting for

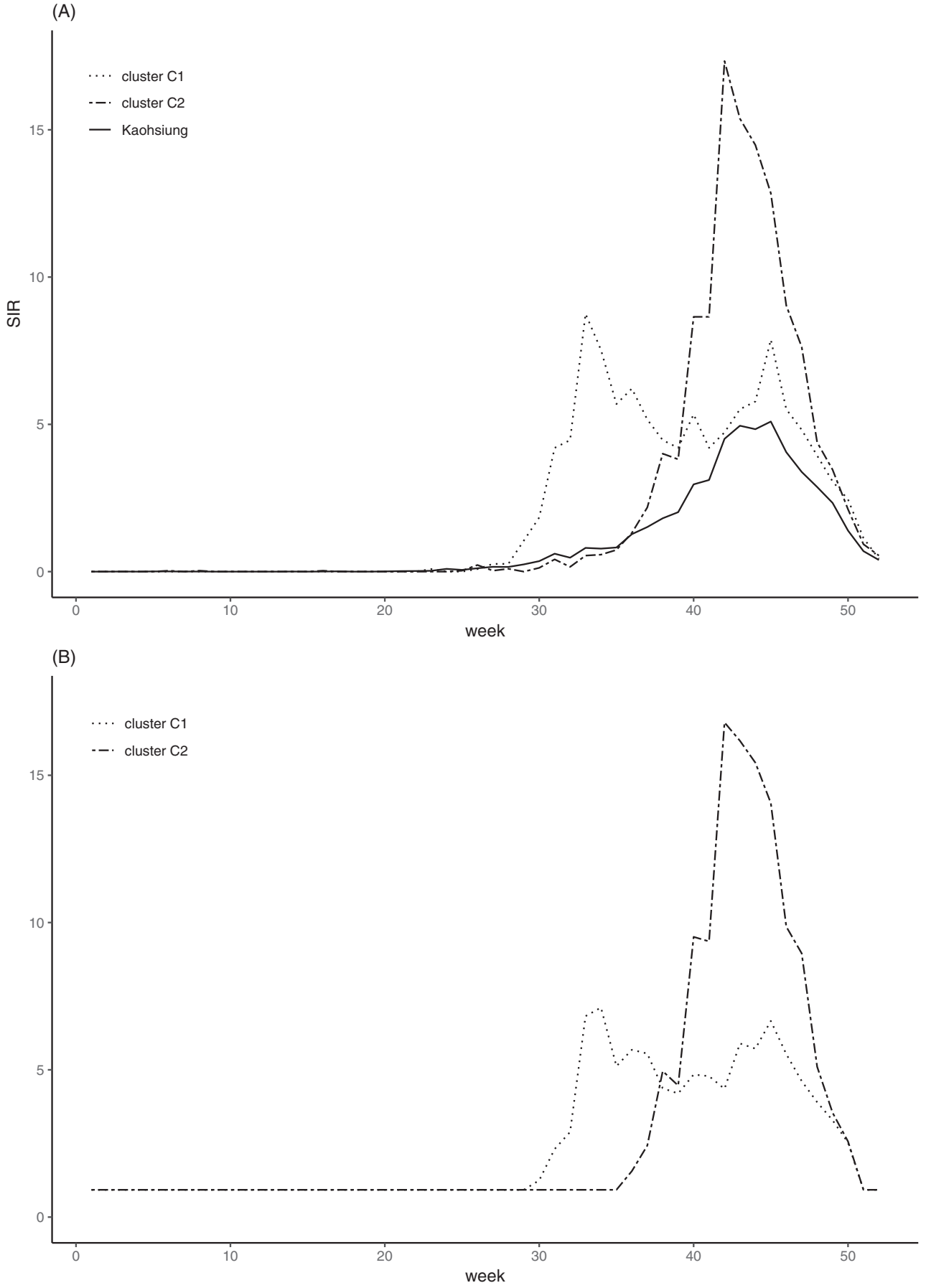

FIGURE 2 Temporal patterns of dengue infections for the 2014 Kaohsiung data. (A) Averages of observed incidence rates in the identified clusters (clusters C_1 and C_2) and whole study area (Kaohsiung City). (B) Estimated incidence rates by the spatial-temporal estimating equation in the identified clusters

TABLE 1 The average number of regions in the true cluster C_k, $k = 1, 2$, that are classified into $\hat{C}_{k'}$, $k' = 1, 2, 3$, by the TH scan statistic and two SaTScan methods (LR_1 and LR_2) for $\sigma^2 = 0.01$ and temporal correlation $\kappa = 0$

		C_1	C_2	$\overline{C}$
		$\lvert C_1\rvert = 34$	$\lvert C_2\rvert = 77$	$\lvert\overline{C}\rvert = 780$
TH	$\hat{C}_1$	30.0	0.0	0.0
	$\hat{C}_2$	4.0	73.8	4.7
	$\hat{C}_3$	0.0	3.2	4.5
LR_1	$\hat{C}_1$	34.0	74.7	167.5
	$\hat{C}_2$	–	–	–
	$\hat{C}_3$	–	–	–
LR_2	$\hat{C}_1$	29.0	22.0	26.0
	$\hat{C}_2$	0.0	44.0	35.0
	$\hat{C}_3$	5.0	1.0	65.0

Note: $\hat{C}_3$ denotes a union of clusters other than $\hat{C}_1$ and $\hat{C}_2$, and $\overline{C} = \Omega - C_1 - C_2$ denotes regions that are not in the true clusters. The simulation result is based on 200 replicates.

SaTScan, say LR_2, the radius of each candidate cluster is restricted to a maximum value of 2.5 km, which is twice the range of the estimated spherical correlation function shown in model (13). For each scan statistic method, we use $\hat{C}_k$ to denote the estimated cluster that has the greatest number of overlapping regions with C_k, $k = 1, 2$. Additionally, we use $\hat{C}_3$ to denote a union of estimated clusters other than $\hat{C}_1$ and $\hat{C}_2$. For convenience, we also use $\hat{C}_k^r$, $k = 1, 2, 3$, to denote the corresponding $\hat{C}_k$ in the rth simulation, $r = 1, \ldots, R$.

In the first simulation scenario, the TH scan statistic has very close identification results in all the simulation settings ($\sigma = 0.01, 0.02, 0.05$, and temporal correlation $\kappa = 0, 0.1, 0.3$). Also, the LR_1 and LR_2 scan statistics also present similar identification results in all the simulation settings. The simulation results may thus indicate that the proposed method is quite robust when the sample size and correlation satisfy certain conditions, as we discuss further in Section 6. Therefore, in the main context, we present only the simulation result for $\sigma = 0.01$ and $\kappa = 0$, while the other identification results are shown in the Supplementary Material. Table 1 lists average numbers of villages in C_k, $k = 1, 2$, which are classified into $\hat{C}_{k'}$, $k' = 1, 2, 3$, by each scan statistic. That is, in Table 1, we compute $\sum_{r=1}^{200} |\hat{C}_{k'}^r \cap C_k|/200$ for each $k = 1, 2$, and $k' = 1, 2, 3$. We also use Figure 3 to show the location of $\hat{C}_k$ identified by each scan statistic in part of the Kaohsiung map. (The identified clusters are very similar in all simulation runs.) For each $\hat{C}_k$, let $\overline{\xi}_k = (\overline{\xi}_{k,1}, \ldots, \overline{\xi}_{k,52})'$ denote an average temporal pattern over the simulation runs, where $\overline{\xi}_{k,t} = \sum_{r=1}^{200} \sum_{\boldsymbol{s}_i \in \hat{C}_k^r} Y_{i,t}^r / (200|\hat{C}_k|)$, $k = 1, 2, 3$. We use $\overline{\xi}_k$ as an estimated temporal pattern for cluster $\hat{C}_k$. Figure 4 depicts the estimated temporal pattern for each identified cluster by each scan statistic.

To ensure no ambiguity, in the following discussion, we also use $\hat{C}_k^{TH}$, $\hat{C}_k^{LR_1}$, and $\hat{C}_k^{LR_2}$ to denote the identified clusters $\hat{C}_k$, $k = 1, 2, 3$, by the TH, LR_1, and LR_2 scan statistics, respectively. First, for the LR_1 scan statistic, Table 1 shows that it identifies only one big cluster, $\hat{C}_1^{LR_1}$, which includes all the villages of C_1, 97% of the villages in C_2, and many villages not in the true clusters. Note that the number of villages in $\hat{C}_1^{LR_1}$ that do not belong to any true clusters is about 168. Figure 4B maps the identified cluster by the LR_1 scan statistic, indicating that about 61% of the villages in $\hat{C}_1^{LR_1}$ are not in any true clusters. Also, we find from Figure 3B that only one peak in the estimated temporal pattern has been identified by the LR_1 scan statistic, while the true simulation setting has two waves of outbreaks (Figure 2B). Since $\hat{C}_1^{LR_1}$ is mixed with almost all villages of C_1 and C_2, it is hard to compare the LR_1 scan statistic with the other scan statistics. So, only the TH and LR_2 scan statistics will be evaluated next for identification performance.

We use two criteria, the true positive rate (TPR) and precision, to compare the identification performance of $\hat{C}_k$ for C_k, $k = 1, 2$, between the TH and LR_2 scan statistics. Specifically, for $\hat{C}_k$, we define the TPR ϕ_{c_k} and precision

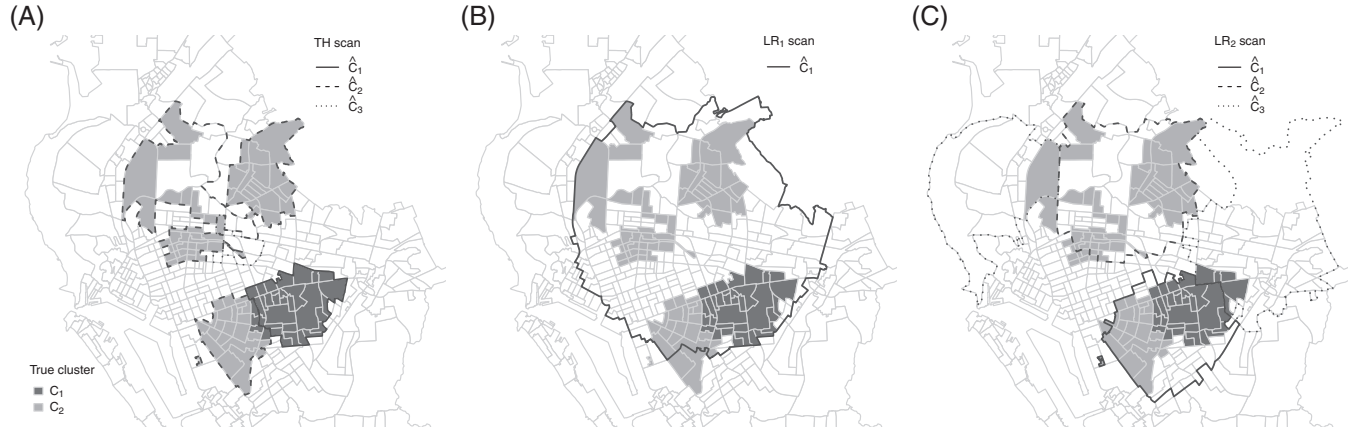

FIGURE 3 In the simulation, locations of the estimated clusters identified by the (A) TH scan statistic, (B) LR_1 scan statistic, and (C) LR_2 scan statistic

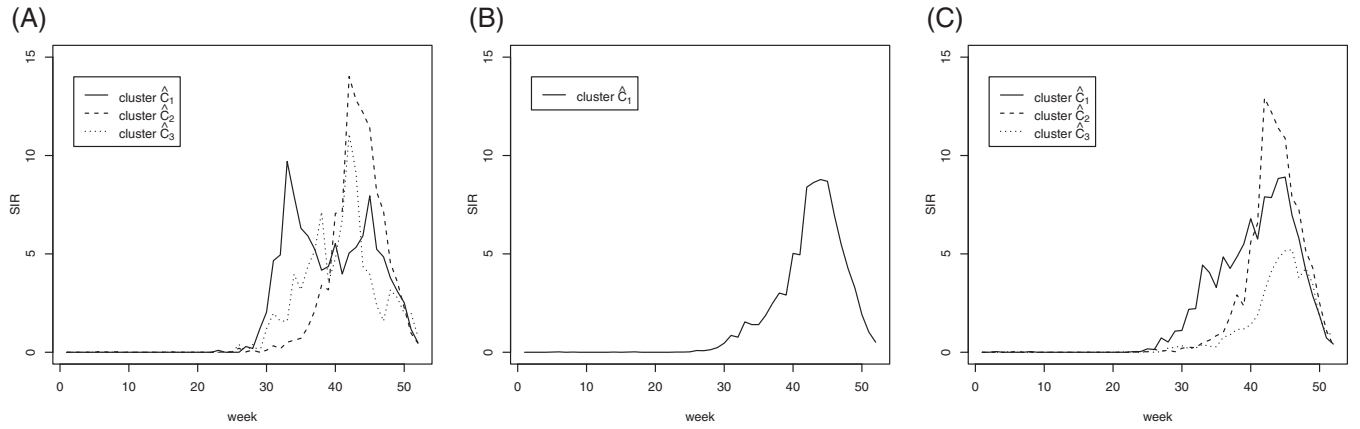

FIGURE 4 In the simulation, average temporal patterns within the identified clusters by the (A) TH scan statistic, (B) LR_1 scan statistic, and (C) LR_2 scan statistic

ψ_{c_k} by

$$\phi_{c_k} = |\hat{C}_k \cap C_k|/|C_k| \quad \text{and} \quad \psi_{c_k} = |\hat{C}_k \cap C_k|/|\hat{C}_k|, \tag{12}$$

$k = 1, 2$, respectively. We also use $\phi_{c_k}^T$ and $\phi_{c_k}^L$ to denote the corresponding value ϕ_{c_k} by the TH and LR_2 scan statistics, respectively, and likewise $\psi_{c_k}^T$ and $\psi_{c_k}^L$ for ψ_{c_k}. For the identification results of C_1, Table 1 shows that the TPRs of $\hat{C}_1$ by the TH and LR_2 scan statistics are $\phi_{c_1}^T = 0.88$ and $\phi_{c_1}^L = 0.85$, respectively. A t-test then shows no significant difference in the TPR of $\hat{C}_1$ between both scan statistics. Nevertheless, for the precision of $\hat{C}_1$, Table 1 gives $\psi_{c_1}^T = 1.0$ and $\psi_{c_1}^L = 0.38$ for the TH and LR_2 scan statistics, respectively. By using the t-test to evaluate the difference between the precisions of $\hat{C}_1$, we find that the TH scan statistic performs significantly better than the LR_2 scan statistic (P-value $\doteq 0$) in the simulation. As can be seen from Figure 2C, the LR_2 scan statistic is also unable to identify two outbreaks of epidemics in C_1, while Figure 3A indicates that the TH scan statistic can accurately predict the true temporal pattern in C_1.

For the identification result of C_2, Table 1 shows that for the TH and LR_2 scan statistics, the respective TPRs are $\phi_{c_2}^T = 0.96$ and $\phi_{c_2}^L = 0.57$, and the respective precisions are $\psi_{c_2}^T = 0.89$ and $\psi_{c_2}^L = 0.56$. These values indicate that, in the simulation for C_2, the TH scan statistic has significantly better performance than the LR_2 scan statistic in the TPR (P-value $\doteq 0$) and precision (P-value $\doteq 0$). To see the reason, we note from Figure 4C that the geographic cluster G_5, which

belongs to C_2 but is next to C_1, is misclassified into the circular cluster $\hat{C}_1^{LR_2}$ by the LR_2 scan statistic, while Figure 4A depicts that the TH scan statistic can identify almost all villages in C_1 and C_2 into $\hat{C}_1^{TH}$ and $\hat{C}_2^{TH}$, respectively. Additionally, in the simulation result for C_2, Figure 4C also shows the LR_2 scan statistic groups the geographic clusters G_2 to G_5 and some villages not in the true cluster into a "big" circular cluster $\hat{C}_2^{LR_2}$. These results may reveal that the LR scan statistic has a tendency to overestimate cluster regions in circular shapes, while the TH scan statistic could identify localized clusters more precisely. So, by the simulation study, we learn that, under suitable conditions for cluster sizes, the TH scan statistic could be more flexible than the LR scan statistics in detecting TH clusters with arbitrary shapes.

In our study of how cluster size could affect scan statistics, from the second simulation scenario, we find that when $a = 1$, simulation results are close to those in the first simulation scenario. That is, the TH scan statistic can still identify $C_1(\equiv G_1)$ and $C_2(\equiv G_5)$ very well when $a = 1$. However, when $a = 0.5$, the TH scan statistic could fail to identify $0.5G_1$ and $0.5G_2$ as two different TH clusters. Specifically, the TH scan statistic would combine regions in $0.5G_1$ and $0.5G_5$ together as an estimated cluster. This simulation result may thus indicate that, for two different TH clusters in proximity, each cluster should contain at least a number of $n^{1/2}$ geographic units so that the TH scan statistic can work well. Note that $n_1 = |G_1| = 34$ and $n_5 = |G_5| = 23$, which are close to the value of $n^{1/2}$. For the *SatScan* method, both the LR_1 and LR_2 scan statistics fail to classify C_1 and C_2 separately in the second simulation scenario. That is, the LR_1 and LR_2 scan statistics identify only one big cluster in all the settings of the second simulation scenario. Finally, for the TH scan statistic, we also conduct another simulation study by using a collection of candidate clusters that include all villages as centroids. Corresponding simulation results are very similar to those shown in Table 1.

5 | ANALYSIS FOR DENGUE DATA

5.1 | Background

To illustrate the TH scan statistic, we use the 2014 Kaohsiung data collected by the Taiwan Centers for Disease Control (https://od.cdc.gov.tw/eic/DengueDailyEN.csv) to analyze the propagation pattern of the dengue infection. The Kaohsiung city consists of 891 villages with a total of 2,778,992 persons. In 2014, Kaohsiung experienced one of its largest dengue outbreaks with 15,043 confirmed cases and 20 deaths. Let N_i denote the number of population at risk in village $\boldsymbol{s}_i$, $i = 1, \ldots, 891$. Under the null hypothesis H, the expected number of dengue cases in village $\boldsymbol{s}_i$ at week t, $t = 1, \ldots, 52$, can be expressed by $E_{i,t} = \tau N_i$, where $\tau = 0.0001$ denotes the average weekly infection rate. Let $O_{i,t}$ denote the (weekly) number of cases in village $\boldsymbol{s}_i$ at week t. The spatial-temporal incidence rate (or, weekly incidence rate in $\boldsymbol{s}_i$) for the dengue infection in village $\boldsymbol{s}_i$ at week t is thus given by $Y_{i,t} = O_{i,t}/E_{i,t}$. Also, let $O_{i,+} = \sum_{t=1}^{52} O_{i,t}$ denote the (yearly) number of dengue cases in village $\boldsymbol{s}_i$. We then define the spatial incidence rate (or, yearly incidence rate at $\boldsymbol{s}_i$) of the dengue infection in village $\boldsymbol{s}_i$ by $Y_{i,+} = O_{i,+}/(52E_{i,t})$.

As have been explained in Section 4, we use the nearest neighborhood structure to define candidate clusters because of the geographic surface in Kaohsiung. To construct candidate clusters for the QL scan statistic, we first make two remarks for characteristics of the dengue data. First, in the Kaohsiung data, most dengue cases (87%) happened in downtown Kaohsiung (with 528 villages), while other dengue cases seemed to randomly scatter in remote villages. Figure 5 shows an image map for spatial incidence rates of the dengue infection in downtown Kaohsiung. So, candidate clusters are set only in downtown Kaohsiung. Second, in the dengue data, we find that some villages $\boldsymbol{s}_i$ had yearly incidence rates $Y_{i,+}$ lower than the expected value (ie, $Y_{i,+} < 1$), but $Y_{i,t} > 1$ for some $t \in T$. This phenomenon is probably due to dengue-control intervention that was immediately undertaken after dengue cases were found in these villages. To explore how the epidemics "naturally" propagated, the village with $Y_{i,+}$ less than the expected value would not be considered as a "center" of candidate clusters. (Nevertheless, these villages with $Y_{i,+} < 1$ could still be included in some candidate clusters if their neighbors had yearly incidence rates greater than one.) Finally, recall that in Assumption 2, the number of villages in each candidate cluster should be restricted. A candidate cluster with a center at $\boldsymbol{s}_i$ is thus limited to include villages up to $B_i^{(0)} \cup B_i^{(1)} \cup B_i^{(2)}$. Let $\Omega_+ = \{\boldsymbol{s}_{i'} : Y_{i',+} > 1\}$ denote a set of villages with $Y_{i',+}$ greater than one. The collection of candidate clusters is thus given by $\boldsymbol{\Lambda} = \{B_{i'}^{(0)}, B_{i'}^{(0)} \cup B_{i'}^{(1)}, B_{i'}^{(0)} \cup B_{i'}^{(1)} \cup B_{i'}^{(2)} : \boldsymbol{s}_{i'} \in \Omega_+\}$. The number of villages in the largest cluster among $\boldsymbol{\Lambda}$ is 41, which is just little higher than $n^{1/2} \approx 30$. In total, 855 candidate clusters are used in the data analysis. For convenience, we use $\Lambda_m \in \boldsymbol{\Lambda}$, $m = 1, \ldots, 855$, to denote a given candidate cluster.

FIGURE 5 A heat map for the spatial incidence rates of the dengue infection in downtown Kaohsiung in 2014

5.2 | Cluster identification for spatial data

For the candidate clusters in $\mathbf{\Lambda}$, we first use the independent scan statistic to find localized clusters. In the first step of the independent scan statistic, we test significance for each single cluster $\Lambda_m \in \mathbf{\Lambda}$. Given Λ_m, we fit a scan model by $Y_{i,+} = \exp\{\mu_s + \xi_m \delta_{\Lambda_m}(\boldsymbol{s}_i) + \epsilon_i^s\}$, and then estimate parameters μ_s and ξ_m by the estimating Equation (4) under the independence assumption. We thus have 855 estimated cluster coefficients $\hat{\xi}_1, \ldots, \hat{\xi}_{855}$. To address the multiple-testing issue, we use the BH procedure to control the FDR at $\alpha = 0.05$. Specifically, let $\hat{p}_m$ denote a P-value associated with $\hat{\xi}_m$, and let $\hat{p}_{(m)}$ denote the mth order statistic for $\{\hat{p}_1, \ldots, \hat{p}_{855}\}$. (That is, $\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(855)}$.) Also, let $\Lambda_{(m)}$ denote the cluster associated with $\hat{p}_{(m)}$. The BH procedure will select a candidate cluster $\Lambda_{(m)}$ as an estimated cluster if $\hat{p}_{(m)} \leq 0.05m/855$. In the data analysis, 22 single clusters, $\Lambda_{(1)}, \ldots, \Lambda_{(22)}$, are selected as (initial) estimated clusters after the BH procedure. The total number of villages in $\{\Lambda_{(1)}, \ldots, \Lambda_{(22)}\}$ is 137.

However, some of $\Lambda_{(1)}, \ldots, \Lambda_{(22)}$ had overlapping regions, which would cause a collinearity problem in the multiple regression model. To avoid collinearity, we use a forward selection procedure to partition $\Lambda_{(1)}, \ldots, \Lambda_{(22)}$ into disjoint clusters. Specifically, for $J = 1, \ldots, 22$, we sequentially updated each initial estimated cluster by the following criteria:

Algorithm 1. Partition procedure

(i) If $\left(\bigcup_{j=1}^{J} \Lambda_{(j)}\right)^c \cap \Lambda_{(J+1)} \neq \emptyset$, then $\Lambda_{(J+1)}$ is replaced by $\left(\bigcup_{j=1}^{J} \Lambda_{(j)}\right)^c \cap \Lambda_{(J+1)}$ and $\{\Lambda_{(1)}, \ldots, \Lambda_{(J)}\}$ remains the same, or

(ii) if $\left(\bigcup_{j=1}^{J} \Lambda_{(j)}\right)^c \cap \Lambda_{(J+1)} = \emptyset$ (that is, $\Lambda_{(J+1)} \subseteq \bigcup_{j=1}^{J} \Lambda_{(j)}$), then $\Lambda_{(J+1)}$ is removed.

After the partition procedure, we have 13 disjoint clusters, $I_1, \ldots, I_{13}$. Note that the total number of villages in $I_1, \ldots, I_{15}$ is still 137.

To investigate whether the independence assumption is suitable for the cluster model associated with $I_1, \ldots, I_{13}$, we fit a multiple regression model by $\log(Y_{i,+}) = \mu_s + \sum_{k=1}^{13} \xi_{k,+} \delta_{I_k}(\boldsymbol{s}_i) + \epsilon_i^s$, and obtained residuals $\hat{\epsilon}_i^s$ under the independence assumption. By applying the variogram method (9) for the residuals $\hat{\epsilon}_i$, we get an estimate for the spatial correlation function by

$$\hat{\rho}_{i,j}^S(h) = 0.55 - 0.83(h/1.23) + 0.28(h/1.23)^3, \tag{13}$$

where the value of 1.23 (km) denotes the range for the spherical correlation. Also, an estimate for the spatial variance is given by $\hat{\sigma}_s^2 = 1.03$.

A permutation test for the variogram estimates $\hat{\sigma}_s$ and $\hat{\rho}_{i,j}^S \equiv \hat{\rho}_{i,j}^S(h)$ suggests the existence of spatial correlation for the cluster model. We hence implement $\hat{\sigma}_s$ and $\hat{\rho}_{i,j}^S$ into the spatial estimating Equation (4) to reestimate parameters in the multiple cluster model. The limiting distribution (7) is used to evaluate P-values for estimated parameters. We find that only five geographic clusters are significant with associated P-values less than .05. Let $G_1, \ldots, G_5$ denote the resulting estimated (geographic) clusters by the spatial scan statistic. The total number of villages in $G_1, \ldots, G_5$ is 111 with $|G_1| = 34$, $|G_2| = 21$, $|G_3| = 10$, $|G_4| = 23$, and $|G_5| = 23$. Figure 1A shows the locations of the five geographic clusters, which include most historic hot spots of the dengue infection.

5.3 | Clusters for temporal heterogeneity

We next propose a combination procedure based on the test given in Section 3.2 to regroup the estimated clusters $G_1, \ldots, G_5$ by their temporal heterogeneity. For each G_k, we fit a spatial-temporal model by

$$\log(Y_{i,t}) = -0.077 + \xi_{k,t} \delta_{G_k}(\boldsymbol{s}_i) + \epsilon_{i,t}, \tag{14}$$

where $\epsilon_{i,t}$ follows a Gaussian process with mean 0, variance $\sigma^2 = 0.02(\equiv\hat{\sigma}_s^2/T)$, and the covariance structure given by (5) with $\rho_{i,j}^S$ being implemented by $\hat{\rho}_{i,j}^S$ of (13). The value of −0.07 in (14) comes from a logarithm of an average of all spatial-temporal responses. Since the dengue infection was seasonal in Taiwan in 2014 (Figure 2A), to reduce estimation variation caused by the log-transformation of small disease rates, we directly set $\xi_{k,t} = 0$ if $\sum_{i\in G_k} Y_{i,t} < |G_k|$. That is, we set $\xi_{k,t} = 0$ if $\log(\overline{Y}_{k,t}) < 0$, where $\overline{Y}_{k,t} = \sum_{i\in G_k} Y_{i,t}/|G_k|$.

Let $\boldsymbol{\xi}_k = (\xi_{k,1}, \ldots, \xi_{k,52})'$ denote a vector of the log-spatial-temporal risks. We use the spatial-temporal estimating Equation (6) to estimate $\boldsymbol{\xi}_k$ for each cluster G_k. The test statistic $U_{k,k'}$ in (11) is then computed to measure heterogeneity between $\hat{\boldsymbol{\xi}}_k$ and $\hat{\boldsymbol{\xi}}_{k'}$. Let $\chi^2_{52,0.95}$ denote a 95th percentile of the chi-squared distribution with 52 degrees of freedom. If some values of $U_{k,k'}$ are less than $\chi^2_{52,0.95}$, then the pair of geographic clusters G_k and $G_{k'}$ with the smallest value of $U_{k,k'}$ will be combined. When G_k and $G_{k'}$ are evaluated to be combined as $G_{k''}$, we then update the corresponding risk to $\xi_{k'',t} = (\xi_{k,t}|G_k| + \xi_{k',t}|G_{k'}|)/(|G_k| + |G_{k'}|)$. Table 2 shows the regrouping process for temporal heterogeneity between the geographic clusters. After the regrouping process for the dengue data, the previous five geographic clusters, $G_1, \ldots, G_5$, are combined into two TH clusters, say C_1 and C_2, with $C_1 \equiv G_1$ and $C_2 \equiv (G_2 \cup \cdots \cup G_5)$. The total number of villages in $C_1 \cup C_2$ is still 111 with $|C_1| = 34$ and $|C_2| = 77$. Figure 1B maps the locations of the TH clusters C_1 and C_2.

Based on the TH clusters C_1 and C_2, a final model associated with $Y_{i,t}$ becomes

$$\log(Y_{i,t}) = -0.077 + \sum_{k=1}^{2} \xi_{k,t}\delta_{C_k}(\boldsymbol{s}_i) + \epsilon_{i,t}. \tag{15}$$

TABLE 2 A flow chart for the regrouping process based on $U_{k,k'}$ for geographic clusters $G_1, \ldots, G_5$

	G1	G2	G3	G4	G5
G1	—	—	—	—	—
G2	258.7	—	—	—	—
G3	189.2	**16.4**	—	—	—
G4	161.7	25.8	22.4	—	—
G5	89.5	31.6	16.5	19.4	—
⇓G2+G3→G2					
	G1	**G2**	**G4**	**G5**	
G1	—	—	—	—	
G2	239.8	—	—	—	
G4	161.7	24.9	—	—	
G5	89.5	25.0	**19.4**	—	
⇓G4+G5→G4					
	G1	**G2**	**G4**		
G1	—	—	—		
G2	239.8	—	—		
G4	121.0	**31.3**	—		
⇓G2+G4→G2					
	G1	**G2**			
G1	—	—			
G2	171.9	—			

Note: The bold number indicates the smallest value of $U_{k,k'}$ in each step.

Again, we set $\xi_{k,t} = 0$ if $\sum_{i \in C_k} Y_{i,t} \leq |C_k|$. Applying the estimating Equation (6) for (15) gives an estimate for the expected disease rate by

$$\hat{\theta}_{i,t} = \exp\{-0.077 + \hat{\xi}_{1,t}\delta_{C_1}(\boldsymbol{s}_i) + \hat{\xi}_{2,t}\delta_{C_2}(\boldsymbol{s}_i)\}, \; i = 1, \ldots, 891, \quad t = 1, \ldots, 52. \tag{16}$$

Figure 2A shows temporal patterns for averaged incidence rates in C_1, C_2, and Ω, while Figure 2B depicts estimated incidence rates for the TH clusters C_1 and C_2. Details for the estimated log-risks $\hat{\xi}_{1,t}$ and $\hat{\xi}_{2,t}$ can be seen in the Supplementary Material.

5.4 | Analysis result

By comparing Figure 2A,B, we find that the estimated disease rates by the TH scan statistic are quite close to the observed values, although for cluster C_1, the TH scan statistic would slightly underestimate the average disease rate in week 34. Also, as can be seen from Figure 2A, the identified clusters, C_1 and C_2, had very different temporal patterns for dengue epidemics. On the other hand, an exploratory analysis for weekly disease rates across the study area shows a relatively flat AR(1) model for dengue infection. Figure 2A also indicates three waves of dengue outbreaks in the year 2014: the first was in cluster C_1, which was transmitted to cluster C_2 and finally went back to cluster C_1. Note that part of cluster C_2 was next to cluster C_1, as can be seen from Figure 1B.

When exploring why the first strike of the dengue epidemics happened in cluster C_1, we find that cluster C_1 included several industrial parks, where guest workers from Southeast Asia were often employed. In Taiwan, some scientists believed that no local dengue virus existed, due to cold currents in winter.[10] Since dengue epidemics occur very frequently in Southeast Asian countries, in cluster C_1, the imported cases could play a crucial role to initiate the dengue outbreaks. On the other hand, cluster C_2, which included two major areas of Kaohsiung, had more local people and merchant activities, which could speed up dengue transmission. Specifically, the geographic cluster G_3 includes a station for the high-speed train system, and G_4 has the main train station for Kaohsiung. Also, as shown in Figure 1A, the geographic cluster G_2 is close to a lake, which could provide a suitable environmental situation for the growth of mosquitos. This could be why cluster C_2 appears to have more severe and widespread dengue outbreaks than cluster C_1, as shown in Figure 2A.

Another interesting finding from Figure 2A is that the dengue epidemic in cluster C_2 started in week 38, about 6 weeks later than that in cluster C_1. A cross-correlation map (not shown here) between the number of cases and temperature also indicates that, for the 2014 Kaohsiung dengue infection, higher temperatures were associated with a larger number of cases in a 6-week lag. Since an extrinsic incubation period of dengue is normally considered to be positively correlated with temperature, the 6-week lag between the outbreaks in clusters C_1 and C_2 may show that the proposed method can reflect the extrinsic incubation period.

We also use the *SaTScan* method with two different settings, the LR_1 and LR_2 statistics, to analyze the dengue data. However, the LR_1 scan statistic identifies only one big circular cluster, which is very similar to the one identified by the LR_1 scan statistic in the simulation. The big circular cluster can be seen in Figure 4B. So, in the data analysis, we compare the identification performance between only the TH and LR_2 scan statistics. The LR_2 scan statistic identifies nine geographic clusters, say, $D_1, \ldots, D_9$, in various time intervals. Information about the spatial-temporal clusters identified by the LR_2 scan statistic can be seen in Table A4 of the Supplementary Material. Let T_k denote the corresponding time interval of elevated disease rates associated with D_k, $k = 1, \ldots, 9$. We compute a sum of squared errors (SSE) for each scan statistic to compare the performance. In the computation of SSEs, only responses in the identified spatial-temporal clusters with $Y_{i,t} > 0$ are considered. For the LR_2 scan statistic, the sum of squared errors is given by $\text{SSE}_1 = \sum_{k=1}^{9} \sum_{i \in D_k} \sum_{t \in T_k} \{\log(Y_{i,t}) - \log(\overline{Y}_{k,t})\}^2 = 4418$, where $\overline{Y}_{k,t} = \sum_{\boldsymbol{s}_i \in D_k} Y_{i,t}/|D_k|$. The total number of spatial-temporal units identified by the LR_2 scan statistic is 3778. For the TH scan statistic, let $\hat{\epsilon}_{i,t} = \log(Y_{i,t}) - \log(\hat{\theta}_{i,t})$ denote the residual, and let $\hat{\boldsymbol{\epsilon}}$ denote a vector formed by $\epsilon_{i,t}$. The sum of squared errors for the TH scan statistic is $\text{SSE}_2 = \hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{V}}_{\epsilon}^{-1} \hat{\boldsymbol{\epsilon}} = 1100$, where $\hat{\boldsymbol{V}}_{\epsilon}$ is an estimated covariance matrix by the variogram estimation. The total number of spatial-temporal units identified by the TH scan statistic is 1196. So, the mean squared errors (MSEs) are $\text{MSE}_1 = 4418/3778 \doteq 1.17$ for the LR_2 scan statistic, and $\text{MSE}_2 = 1100/1196 \doteq 0.92$ for the TH scan statistic. A comparison between the value of 1.27 ($=\text{MSE}_1/\text{MSE}_2$) and an F-distribution with degrees of freedom 3378 and 1196 gives a P-value around 3×10^{-7}, which may provide evidence that the TH scan statistic could have better identification results than the LR_2 scan statistic in this data analysis.

6 | DISCUSSION

Identifying transmission of hot spots and characterizing the temporal trends are important for understanding infectious disease propagation. In this article, we propose a novel scan statistic by combining the spatial scan statistic for geographic clusters and chi-squared test for temporal heterogeneity to identify clusters whose temporal patterns are similar within clusters but different between clusters. The proposed scan statistic could be more flexible than traditional methods in the sense that contiguous and nonproximate regions with similar temporal patterns can be identified simultaneously. Although some scan statistics, such as those produced by *SaTScan*, can also be used to search for geographic clusters with similar temporal trends, they are mainly designed to group contiguous hot-spot regions as clusters in certain types of shapes. The simulation indicates that such scan statistics may thus have difficulty in distinguishing proximate clusters that have different temporal patterns. The proposed approach, on the other hand, can be used to explore different routes of epidemic propagation from a common source of an infectious disease, as we illustrated in the data analysis for the dengue infection. Although Lin and Zhu[11] developed a method that also can connect spatial clustering and classification approaches, their work focuses on dealing with spatial heterogeneity.

To see when the proposed method could work well, three remarks summarized from the asymptotic property and simulation study are listed below. We recall that in the limiting distribution (7), the error rate for the spatial scan statistic is an order of $n^{1/2}$. So, first, if one geographic cluster with a disease rate $p\%$ higher than the expected value does exist, then the true cluster should consist of at least $(\log(1+0.01p))^{-2}$ geographic units to make the statistical inference valid. For example, if we would like all geographic clusters that are not in proximity with disease rates at least 40% higher than the expected value to be identified, then each of the true clusters should contain at least nine geographic units. In the simulation and data analysis, we find that the TH scan statistic is able to identify geographic cluster G_3, which consists of 10 villages. Second, if two clusters are in proximity but have different temporal patterns, then by the simulation result in the second scenario, each of the true clusters should contain at least a number of $\max\{n^{1/2}, (\log(1+0.01p))^{-2}\}$ geographic units such that the TH scan statistic can work well. On the other hand, if a study area consists of n geographic regions, then only TH clusters with disease rates $(\exp(n^{1/4})-1)\%$ higher than the expected value could be identified by the proposed method. Third, in the identification process of the TH scan statistic, we characterize the temporal pattern of each cluster by estimating its risk coefficient at each time t. This approach could work well for data with separable spatial-temporal correlations, as can be seen from the simulation result. Nevertheless, the number of temporal coefficients in the corresponding model would then be an order of T, and this could make estimation less efficient, unless the study area consists of geographic units many enough. In the data analysis and simulation, the largest geographic cluster consists of $(n_1 =)34$ villages, while the whole Kaohsiung city has $891(\approx n_1^2)$ villages. We find that in such a situation, the temporal pattern in each TH cluster can still be estimated well.

While traditional scan statistics can generally allow larger sizes of candidate clusters, the TH scan statistic is required to restrict the size of candidate clusters up to $n^{1/2}$ for valid statistical inference. To evaluate whether the size limitation on candidate clusters would make the TH scan statistic infeasible to identify "big" clusters, in the simulation, the number of villages in the true clusters is set to be proportional to the number of sampling villages n. The simulation shows that, with a suitable regrouping procedure, the TH scan statistic can still identify the true clusters very accurately. Finally, because the dengue virus is transmitted mainly by the mosquito, whose range of activities and life period are usually short, the dengue infection may have different epidemic circles in different TH clusters. In this article, we consider that each cluster has its own risk coefficient at each time period. Nevertheless, an interesting extension to this article would be to combine the TH scan statistic and generalized linear dynamic model for temporal coefficients. We leave this topic for future research.

ACKNOWLEDGEMENT

The author would like to thank the editor, the associate editor, and two anonymous referees for their constructive comments toward improving the paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://od.cdc.gov.tw/eic/DengueDailyEN.csv.

ORCID

Pei-Sheng Lin https://orcid.org/0000-0002-3057-094X

REFERENCES

1. Waller L, Gotway C. *Applied Spatial Statistics for Public Health Data*. New York, NY: Wiley; 2004.
2. Yuan H, Wen T, Kung Y, et al. Prediction of annual dengue incidence by hydro-climatic extremes for southern Taiwan. *Int J Biometeorol*. 2019;63:259-268.
3. Cliff A, Haggett P, Ord J, Versey G. *Spatial Diffusion: A Historical Geography of Epidemic in an Island Community*. Cambridge, UK: Cambridge University Press; 1981.
4. Kuo F, Wen T, Sabel C. Characterizing diffusion dynamics of disease clustering: a modified space-time DBSCAN (MST-DBSCAN) algorithm. *Ann Am Assoc Geogr*. 2018;108:1168-1186.
5. Lin P, Kung Y, Clayton M. Spatial scan statistics for detection of multiple clusters with arbitrary shapes. *Biometrics*. 2016;72:1226-1234.
6. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Methods*. 1997;26:1487-1496.
7. McCullagh P, Nelder J. *Generalized Linear Models*. 2nd ed. Cambridge, UK: Chapman & Hall; 1989.
8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Soc Ser B Stat Methodol*. 1995;57:289-300.
9. Cressie N. *Statistics for Spatial Data*. 2nd ed. New York, NY: Wiley; 1993.
10. Wang S, Wang W, Chang K, et al. Severe dengue fever outbreaks in Taiwan. *Am J Trop Med Hyg*. 2016;94:193-197.
11. Lin P, Zhu J. A heterogeneity measure for cluster identification with application to disease mapping. *Biometrics*. 2020;76:403-413.

SUPPORTING INFORMATION

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

How to cite this article: Lin P-S. Identification of geographic clusters for temporal heterogeneity with application to dengue surveillance. *Statistics in Medicine*. 2022;41(1):146-162. doi: 10.1002/sim.9227

APPENDIX

Derivation of (2): Let $b_{i,+} = \exp\left(\sum_{k=1}^{K}\xi_{k,+}\delta_{C_k}(\boldsymbol{s}_i)\right)$ and let $b_{i,t} = \exp\left(\sum_{k=1}^{K}\xi_{k,t}\delta_{C_k}(\boldsymbol{s}_i)\right)$. Then, $\log(Y_{i,+}) = \log(b_{i,+}) + \mu_s + \log\left\{\sum_{t=1}^{T}\left(b_{i,t}\exp(\epsilon_{i,t})\right)/b_{i,+}\right\}$, where $\mu_s = \mu_0 + 1/T$. Let $\sum_{t=1}^{T}\left(b_{i,t}\exp(\epsilon_{i,t})\right)/b_{i,+} = B_{i,+}$. Since the logarithm for sum of log-normal variables can be approximated by another log-normal variable, $B_{i,+}$ approximately follows a log-normal distribution. So, under suitable conditions, we have $B_{i,+} \doteq \exp(\epsilon_i^s)$, where $\epsilon_i^s \sim N(0, T\sigma^2)$. This then leads to the desired result.

Proof of (7). For convenience, let $\boldsymbol{\xi}_{s,+} = (\mu_s, \xi_{1,+}, \ldots, \xi_{K,+})'$. Let $\boldsymbol{Q}_s(\boldsymbol{\xi}_{s,+}) = \boldsymbol{D}_s'\boldsymbol{V}_s^{-1}(\boldsymbol{Y}_s - \boldsymbol{\theta}_s)$. Also, let $\dot{\boldsymbol{Q}}_s(\boldsymbol{\xi}_{s,+})$ denote a derivative function of $\boldsymbol{Q}_s(\boldsymbol{\xi}_{s,+})$ with respect to $\boldsymbol{\xi}_{s,+}$. Since $\boldsymbol{D}_s'\boldsymbol{V}_s^{-1} = \boldsymbol{O}(1)$ by Assumption 1(a), consistency of $\hat{\boldsymbol{\xi}}_{s,+}$ comes from a method similar to that by Lin et al.[5] By a first-order Taylor expansion, we have $\boldsymbol{Q}_s(\hat{\boldsymbol{\xi}}_{s,+}) = \boldsymbol{Q}_s(\boldsymbol{\xi}_{s,+}) + \dot{\boldsymbol{Q}}_s(\hat{\boldsymbol{\xi}}_{s,+})\cdot(\hat{\boldsymbol{\xi}}_{s,+} - \boldsymbol{\xi}_{s,+}) + \boldsymbol{o}_p(\|\hat{\boldsymbol{\xi}}_{s,+} - \boldsymbol{\xi}_{s,+}\|)$, and therefore $\hat{\boldsymbol{\xi}}_{s,+} - \boldsymbol{\xi}_{s,+} = -\{\dot{\boldsymbol{Q}}_s(\hat{\boldsymbol{\xi}}_{s,+})\}^{-1}\boldsymbol{Q}_s(\boldsymbol{\xi}_{s,+}) + o_p(\|\hat{\boldsymbol{\xi}}_{s,+} - \boldsymbol{\xi}_{s,+}\|)$ as $n \to \infty$. Furthermore, it can be shown that $-n^{-1}\dot{\boldsymbol{Q}}_s(\hat{\boldsymbol{\xi}}_{s,+}) = n^{-1}\hat{\boldsymbol{D}}_s^{-1}\hat{\boldsymbol{V}}_s^{-1}\hat{\boldsymbol{D}}_s + \boldsymbol{o}_p(n^{-1/2})$ as $n \to \infty$. A central limit theorem thus gives the desired result.

Consistency of the chi-squared test (11): Since $\boldsymbol{V}$ can be approximated by a spatially block-diagonal matrix by the discussion for Assumption 1, we have $\boldsymbol{D}_{\xi_k}'\boldsymbol{V}^{-1}\boldsymbol{D}_{\xi_k} = \boldsymbol{O}(n)$ by Assumption 1(a). It is thus reasonable to assume that $n^{-1}\boldsymbol{D}_{\xi_k}'\boldsymbol{V}^{-1}\boldsymbol{D}_{\xi_k}$ converges to a positive-definite matrix $\boldsymbol{\Upsilon}_{\xi_k}$ as $n \to \infty$. An argument similar to the proof of (7) then gives consistency of $\hat{\boldsymbol{\xi}}_k$ and $\hat{\boldsymbol{\xi}}_{k'}$. Also, we have $n^{1/2}\hat{\boldsymbol{\xi}}_k \to N(\boldsymbol{\xi}_k, \boldsymbol{\Upsilon}_{\xi_k}^{-1})$ and $n^{1/2}\hat{\boldsymbol{\xi}}_{k'} \to N(\boldsymbol{\xi}_{k'}, \boldsymbol{\Upsilon}_{\xi_{k'}}^{-1})$ as $n \to \infty$, where $\boldsymbol{\Upsilon}_{\xi_k}^{-1}$ denotes an inverse matrix of $\boldsymbol{\Upsilon}_{\xi_k}$. Under suitable conditions, it follows from consistency of $\hat{\boldsymbol{\xi}}_k$ and $\hat{\boldsymbol{\xi}}_{k'}$ that $\hat{\boldsymbol{\Upsilon}}_{\xi_k} \to \boldsymbol{\Upsilon}_{\xi_k}$ and $\hat{\boldsymbol{\Upsilon}}_{\xi_{k'}} \to \boldsymbol{\Upsilon}_{\xi_{k'}}$ as $N_T \to \infty$. Consequently, $U_{k,k'} = n(\hat{\boldsymbol{\xi}}_k - \hat{\boldsymbol{\xi}}_{k'})'(\hat{\boldsymbol{\Upsilon}}_{\xi_k}^{-1} + \hat{\boldsymbol{\Upsilon}}_{\xi_{k'}}^{-1})^{-1}(\hat{\boldsymbol{\xi}}_k - \hat{\boldsymbol{\xi}}_{k'})$ approximately follows a noncentral chi-squared distribution with degrees of freedom at T and mean $\boldsymbol{\xi}_k - \boldsymbol{\xi}_{k'}$ as $n \to \infty$, which validates the consistency of selection for $U_{k,k'}$.