



## RESEARCH ARTICLE

# Reclaiming independence in spatial-clustering datasets: A series of data-driven spatial weights matrices

Wei Wang<sup>1</sup> | Xiong Xiao<sup>1</sup> | Jian Qian<sup>1</sup> | Shiqi Chen<sup>2</sup> | Fang Liao<sup>3,4</sup> | Fei Yin<sup>1</sup> |  
Tao Zhang<sup>1</sup>  | Xiaosong Li<sup>1</sup> | Yue Ma<sup>1</sup> 

<sup>1</sup>West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

<sup>2</sup>Women and Children's Health Management Department, Sichuan Provincial Hospital for Women and Children, Chengdu, China

<sup>3</sup>Sichuan Provincial Center for Mental Health, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China

<sup>4</sup>Key Laboratory of Psychosomatic Medicine, Chinese Academy of Medical Sciences, Chengdu, China

## Correspondence

Yue Ma, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China.  
Email: gordonrozen@qq.com

## Funding information

The authors are grateful for the support of the National Natural Science Foundation of China (grant numbers 81803332 and 81872713) and Sichuan Science & Technology Program (grant number 2021YFS0181).

## Abstract

Most spatial models include a spatial weights matrix ( $\mathbf{W}$ ) derived from the first law of geography to adjust the spatial dependence to fulfill the independence assumption. In various fields such as epidemiological and environmental studies, the spatial dependence often shows clustering (or geographic discontinuity) due to natural or social factors. In such cases, adjustment using the first-law-of-geography-based  $\mathbf{W}$  might be inappropriate and leads to inaccuracy estimations and loss of statistical power. In this work, we propose a series of data-driven  $\mathbf{W}$ s (DDWs) built following the spatial pattern identified by the scan statistic, which can be easily carried out using existing tools such as SaTScan software. The DDWs take both the clustering (or discontinuous) and the intuitive first-law-of-geographic-based spatial dependence into consideration. Aiming at two common purposes in epidemiology studies (ie, estimating the effect value of explanatory variable  $X$  and estimating the risk of each spatial unit in disease mapping), the common spatial autoregressive models and the Leroux-prior-based conditional autoregressive (CAR) models were selected to evaluate performance of DDWs, respectively. Both simulation and case studies show that our DDWs achieve considerably better performance than the classic  $\mathbf{W}$  in datasets with clustering (or discontinuous) spatial dependence. Furthermore, the latest published density-based spatial clustering models, aiming at dealing with such clustering (or discontinuity) spatial dependence in disease mapping, were also compared as references. The DDWs, incorporated into the CAR models, still show considerable advantage, especially in the datasets for common diseases.

## KEYWORDS

clustering spatial dependence, conditional autoregressive model, disease mapping, spatial autoregressive model, spatial weights matrix

Wei Wang and Xiong Xiao contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

With the development of remote sensing, geographic information systems (GIS), global positioning systems, and computational science, a large number of spatial datasets have been collected.<sup>1,2</sup> Various spatial models have been developed to address spatial dependence in these datasets,<sup>2-5</sup> most of which employ a spatial weights matrix (**W**) to adjust the spatial dependence among observations to reclaim their independence between spatial units.<sup>6</sup>

**Ws** were first used in Moran's I and Geary's C statistics to explore whether spatial autocorrelation exists and then developed by a large number of geographers,<sup>7-15</sup> especially Cliff and Ord,<sup>10,11</sup> into a relatively mature tool in geographical analysis. In the 1980s, "Spatial Process: Methods and Application"<sup>3</sup> and "Spatial Econometrics: Methods and Models"<sup>4</sup> introduced spatial autocorrelation into general regression models in the form of **W**. Then, many spatial models involving **Ws**,<sup>16</sup> for example, spatial interpolations, spatial autoregressive (SAR) models, conditional autoregressive (CAR) models, were developed and employed in various fields,<sup>17-21</sup> for example, economical, epidemiological, environmental, and biological studies. As the fundamental factor in such spatial analyses, **W** is a  $n \times n$  non-negative matrix, and the element  $w_{ij}$  reflects the intensity of the spatial dependence between spatial units  $i$  and  $j$ .<sup>22</sup> Although many methods have been developed to construct **W**, with the lack of prior knowledge, adjacency-based **Ws** and distance-based **Ws** are commonly used in applications according to the first law of geography<sup>5,6,23,24</sup>: two units that are closer to each other have a stronger dependence than two units that are farther apart. Adjacency-based **Ws** specify that  $w_{ij}$  is equal to 1 if unit  $i$  and unit  $j$  are adjacent and equal to 0 otherwise, while the distance-based **Ws** define each  $w_{ij}$  as a function of distance (eg, the Euclidean distance).<sup>4</sup> However, spatial dependence in the entire study area is not always distributed only following the first law of geography, but also shows spatial clustering (or discontinuity) due to complex practical environment. Therefore, the classic **W** cannot adjust the spatial dependence sufficiently in such datasets containing clusters and leads to biased and inaccuracy parameter estimations.<sup>5,25</sup> For example, in Bhattacharjee and Jensen-Butler's simulation study,<sup>26</sup> different **Ws** yielded considerably different results in terms of the bias and root mean square error (RMSE). In Getis's study,<sup>15</sup> a more appropriate **W** can reduce the Akaike information criterion (AIC) by over half without increasing the bias.

Several methods have been developed to address this problem, such as subgroup analysis according to the spatial heterogeneity, adapting revised **Ws** based on some prior evidence and constructing **Ws** based on spatial pattern,<sup>27-29</sup> such as a multidirectional optimum ecotope-based algorithm (AMOEBa),<sup>14,15</sup> which adopts the local  $G_i^*$  statistic to identify clusters and further constructs **W**. However, subgroup analyses are often based on a subjective grouping and also reduce the statistical power. The revised **W** requires prior evidence that is not always available. A **W** constructed considering only clusters, as in AMOEBa, ignores the intuitive first-law-of-geography-based spatial dependence and also has no convenient implementation approach. Recently, a two-stage method based on agglomerative hierarchical clustering (AHC) algorithm or density-based spatial clustering (DBSC) algorithm was proposed to address the spatial clustering against the classic **W** in CAR models.<sup>30,31</sup> This method achieves smaller bias than the classic CAR model, but leads to a larger RMSE in disease mapping. A large RMSE may not be unsatisfactory when we focus more on estimating the exact risk value for each spatial unit than comparing the risks between spatial units. Therefore, an objective and effective **W**, which has a wide range of application, was necessary to deal with spatial dependence with clustering (or discontinuity).

In this work, we proposed a novel series of data-driven **Ws** (DDWs) which consider not only the intuitive first-law-of-geography-based spatial dependence but also the heterogeneity and homogeneity from clusters. DDWs do not depend on any prior evidence and can be used for various models including a spatial weights matrix. The two-stage procedure of constructing DDWs was detailed in Section 2. In Section 3, we introduced the DDWs into the commonly used SAR and CAR models. Then their performance in two common purposes in epidemiology studies (ie, estimating the effect of explanatory variable  $X$  and estimating the relative risk [RR] of each spatial unit in disease mapping), respectively, were evaluated in simulation datasets, with the references of the classic **W** and the latest DBSC model considering the spatial discontinuity. In Section 4, these novel DDWs were applied to a case study.

## 2 | METHODS: TWO-STAGE PROCEDURE TO CONSTRUCT DDWS

To sufficiently adjust spatial dependence in spatial datasets with clusters (or spatial discontinuity), the spatial dependence not following the first law of geography in clustering regions must be considered. To maintain the generalizability of the DDWs, we choose to build DDWs based on the classic **W**. A two-stage procedure is employed to identify the clusters first and then to construct DDWs according to the identified clusters.

In the first stage, to accurately identify the spatial pattern (ie, the locations of clusters), scan statistic is employed. According to previous studies,<sup>32</sup> as the accuracy of the detected clusters is strongly related to the scan parameters, the maximum clustering set-proportion (MCS-P) statistic is used for optimal parameter selection.<sup>33</sup> The parameter with the highest value of the MCS-P is selected, and the referred clusters represent the most likely accurate spatial pattern of the current dataset. The spatial units out of the clusters are defined as the baseline, in which the occurrence of events follows the intuitive first-law-of-geography-based spatial stochastic process. Since the spatial units in clusters are heterogeneous compared to those in the baseline region, a spatial stochastic process in clusters differing from the baseline region will be built to characterize the event occurrence. Therefore, the detected clusters reflect the spatial dependence which is presented as the heterogeneity between two spatial units in different clusters (or one in baseline and the other in a cluster) and the homogeneity in baseline (or the same one cluster).

Then, in the second stage, the spatial dependence between different spatial units can be specified. The relationship between any two spatial units can be divided into four types: both units are in the same cluster, both units are in the baseline region, two units are in different clusters, one is in the cluster and the other is in the baseline region. For the two former types, as the spatial units are located in the region in which events occur following the same spatial stochastic process, the spatial dependences mainly represent spatial homogeneity. For the two latter types, the spatial dependences mainly represent spatial heterogeneity, as spatial units in clusters follow significantly different stochastic processes from those in the baseline region, the spatial stochastic processes of different clusters might also be different. With the association between spatial units defined, the elements of  $\mathbf{W}$ ,  $w_{ij}$ , which indicate the spatial dependence in each pair, can be built as follows:

1. A cluster mainly represents the spatial heterogeneity between the cluster itself and other parts of the spatial region, so, for spatial units  $i$  and  $j$  from different clusters (or  $i$  from clusters and  $j$  from the baseline),  $w_{ij}$  was set to 0 to represent the heterogeneity between them, regardless of the positions. With  $w_{ij}$  set to 0, the heterogeneous regions (ie, clusters) are defined to be isolated from the regions in which events occur following different spatial stochastic processes.
2. For the rest of the elements  $w_{ij}$  indicating the spatial dependence between spatial units from homogeneous regions (ie, both are from the same cluster or both are from the baseline region), we defined three different weighting methods referring to three different spatial dependence distributions in homogenous regions, that is, geographic continuity weighting (GW), suggesting spatial dependence is distributed following the first law of geography; observed dependent variable weighting (RW), suggesting spatial dependence is distributed according to the dependent variable; and null weighting (NW), referring to uniformly distributed spatial dependence (shown in Table 1). However, considering that the baseline usually refers to a relatively large number of units with random variation, RW is excluded for baseline regions to avoid overfitting. As the spatial dependence in clusters and baseline might be different,  $w_{ij}$  for clusters and the baseline can be set different to build six DDWs (Table 2). Figure 1 shows the process of constructing DDWs.

Finally, each row of  $\mathbf{W}$  is proportionally standardized such that the sum equals 1 for better interpretability.

### 3 | SIMULATION STUDY

To validate the performance of DDWs over the classic  $\mathbf{W}$ , we used two batches of simulation datasets referring to two common purposes in epidemiology studies, respectively. One is to estimate the effect parameter of the risk factor  $X$  on

**TABLE 1** The method of assigning weights within a cluster or the baseline region

Method of assigning weights	Notation
Geographic continuity weighting (GW)	$w_{ij} = 1$ , when unit $i$ and $j$ are adjacent; otherwise, $w_{ij} = 0$
Null weighting (NW)	$w_{ij} = 1$ , regardless of the positional relationship
Risk weighting (RW)	$w_{ij} = \frac{1}{ y_i - y_j }$

Note:  $y_i$  and  $y_j$  are the observed values of units  $i$  and  $j$ , respectively.

TABLE 2 Six categories of DDWs

Assigning weights for the baseline	Assigning weights for clusters	Label of the DDW
GW	GW	GG
GW	NW	GN
GW	RW	GR
NW	GW	NG
NW	NW	NN
NW	RW	NR

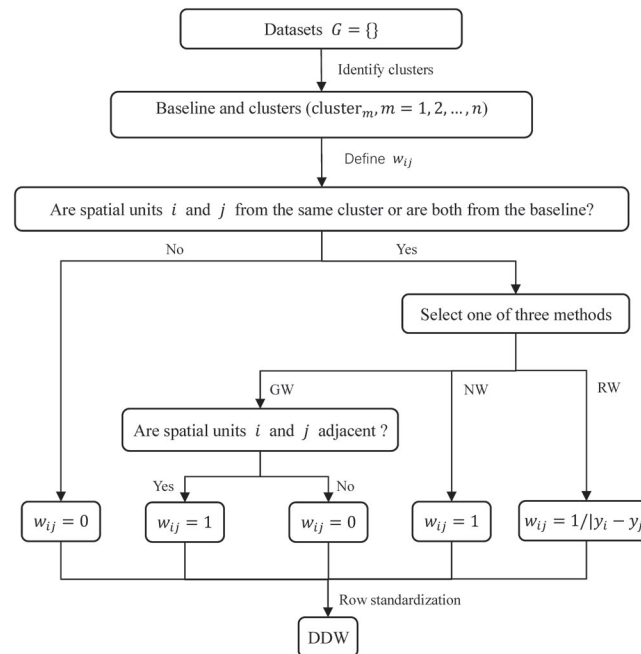


FIGURE 1 The process of constructing the DDWs

certain disease and the other is to estimate (or compare) the relative risk of each spatial unit in disease mapping. Though both the commonly used SAR and CAR models can achieve the two purposes, SAR is employed more often for the former purpose and CAR is employed more often for the latter purpose. Therefore, in our study, the SAR models was selected to compare the performance of DDWs over the classic  $\mathbf{W}$  in term of estimating the effective parameter of one risk factor  $X$ , and the CAR models was selected in term of estimating (or comparing) the relative risk of each spatial unit in disease mapping.

### 3.1 | SAR models for effect parameter estimating

The spatial error model (SEM) and spatial lag model (SLM) are two commonly used SAR models.<sup>2,34</sup> The SEM assumes that spatial dependence comes from unobserved explanatory variables and the SLM assumes that spatial dependence comes from the observed dependent variable.

SEM:

$$\mathbf{y} = \alpha_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \rho\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

where  $\mathbf{y}$ , the dependent variable, represents an  $n \times 1$  vector ( $n$  is the number of spatial units),  $\mathbf{1}_n$  is an  $n \times 1$  vector of ones associated with the constant term parameter  $\alpha$ .  $\mathbf{X}$ , the observed explanatory variables, represents an  $n \times p$  matrix associated with the  $p \times 1$  parameter vector  $\beta$ , and  $p$  is the number of explanatory variables.  $\mathbf{W}$  is the  $n \times n$  spatial weights matrix with zero diagonal elements.  $\mathbf{u}$  and  $\epsilon$  is an  $n \times 1$  vector.  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. The scalar  $\rho$  measures the strength of spatial dependence.

SLM:

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \alpha \mathbf{1}_n + \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where these variables and parameters are defined as in the SEM and  $\mathbf{W}\mathbf{y}$  denotes the endogenous interaction effects among the dependent variables.

The classic adjacency-based  $\mathbf{W}$  (hereafter AG) was constructed derived from a symmetric matrix  $\mathbf{C}$  with zero diagonal elements, in which the element  $c_{ij}$  is 1 when spatial unit  $i$  and  $j$  are neighbors, that is, share a boundary, and  $c_{ij} = 0$  otherwise. Finally, AG could be obtained by proportionally standardizing each row of  $\mathbf{C}$  such that the sum equals 1.

### 3.1.1 | Data generation process in SAR models

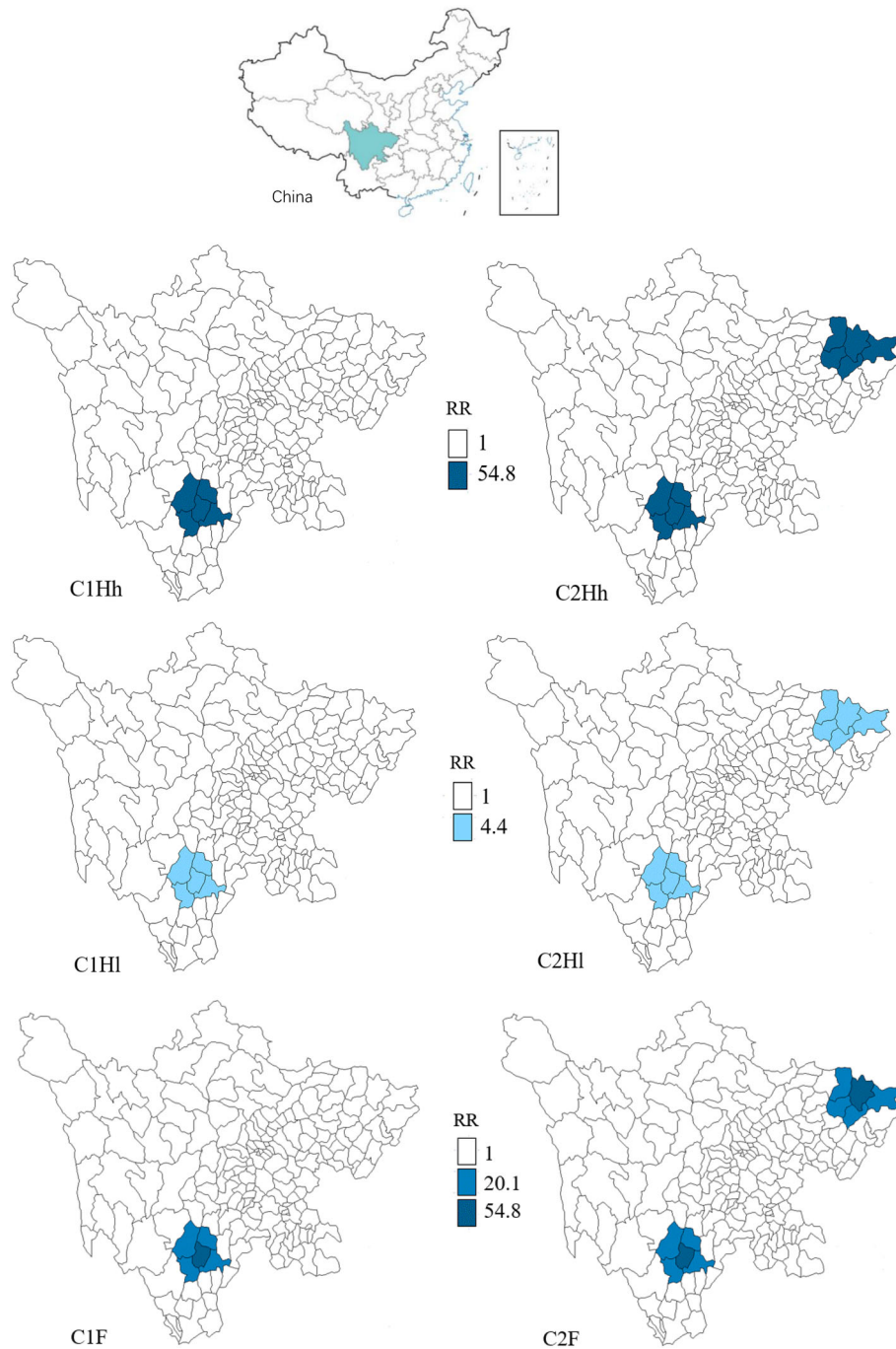
In this simulation study, the purpose is to estimate the effect of a particular factor  $X$  on a common infectious disease, which is a main problem to be addressed in real life. Sichuan Province in China, consisting of 180 counties, was selected as the study region. First, two batches of 10 000 base simulation datasets were generated based on the SEM and SLM with AG, respectively. Without loss of generality, both SEM and SLM include a response variable and an explanatory variable. Then one or two artificial clusters were added to the base simulation datasets, which seems to certain unobserved risk factors to emerge in corresponding spatial units in practical situations.

Because the intensity of heterogeneity between clusters and the baseline as well as inner heterogeneity within clusters may affect the efficiency of the DDWs, three types of clusters were generated labeled as Hh, Hl, and F. Hh type clusters represent homogeneous clusters with a high RR over the baseline, Hl type clusters represent homogeneous clusters with a second high RR, and F type clusters represent clusters with inner heterogeneity, namely, including both a second high and a high RR. In addition, the number of clusters in the datasets, also affecting the intensity of spatial heterogeneity in the whole dataset, was considered to vary from one to two. As a result, six simulation scenarios were set for the SEM and SLM, respectively (shown in Table 3). The location of the clusters was shown in Figure 2, and the other parameter settings and the detailed data generation process are shown in Figure S1 in the support information. In each dataset, the overall spatial dependence is composed of the intuitive adjacency-based spatial dependence and the local spatial-clustering dependence. The former is always distributed according to the first law of geography, while the latter is distributed according to the clustering intensity.

For the SLM, to add clustering spatial dependence to the dependent variable, we enlarged the values of the observed risk factor,  $X$ , to synchronize the spatial dependence from the observed explanatory variable and those from unobserved variable in the form of an SLM. As such, the derivation of spatial dependence in simulation datasets still satisfies the assumption of the focused SLM. Therefore, different scenarios in the SLM had different observed values for the risk factor,

**TABLE 3** Six scenarios built in the simulation study

Scenario	The number of clusters	The type of clusters
SEM(SLM)_C1Hh	1	Hh
SEM(SLM)_C1Hl	1	Hl
SEM(SLM)_C1F	1	F
SEM(SLM)_C2Hh	2	Hh
SEM(SLM)_C2Hl	2	Hl
SEM(SLM)_C2F	2	F



**FIGURE 2** The distribution of clusters in the simulation scenarios for the SAR models. The top map shows the location of Sichuan Province in China. The other six maps represent the position of the artificial clusters in Sichuan Province, where different colors correspond to different relative risks with respect to the white

shown in Table 4. To explore the stability of the DDWs in the SLM, we also tried other observed values of the risk factors, which will be mentioned in the discussion.

### 3.1.2 | Performance evaluation in SAR models

With different DDWs and AG, SEM and SLM were fitted to estimate the effect parameter of  $X$  for the two batches of simulation datasets, respectively. The absolute bias (ABias), mean squared error (MSE), and adjusted coefficient of determination



**TABLE 4** Generating the simulated  $X$  for each spatial unit in the SLM

<b>X</b>	<b>The location of X</b>	<b>Distribution</b>	<b>Available scenarios</b>
X0	In the baseline	$N(0.8, 0.25)$	All scenarios
X1	In a high H cluster	$N(1.5, 0.25)$	SLM_C1Hh and SLM_C2Hh
X2	In a second high H cluster	$N(1.0625, 0.25)$	SLM_C1Hl and SLM_C2Hl
X3	In a high-risk unit in a F cluster	$N(1.5, 0.25)$	SLM_C1F and SLM_C2F
X4	In a second high-risk unit in a F cluster	$N(1.325, 0.25)$	SLM_C1F and SLM_C2F

Note: The other parameters of the SLM are shown in Figure S1.

$(R^2_{\text{adj}})$  were used to evaluate the performance of models with different  $\mathbf{W}$ .

$$\text{MSE} = \sum_{i=1}^N \frac{(\hat{\beta}_i - \beta)^2}{N}, \quad \text{ABias} = \left| \sum_{i=1}^N \frac{(\hat{\beta}_i - \beta)}{N} \right|,$$

where  $\hat{\beta}$  is the estimated value obtained from the fitted model;  $\beta$  is the real value; and  $N$  is the number of replicas in the scenario.

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/\text{df}}{\text{SYY}/(n-1)},$$

where RSS, SYY, df represent the residual sum of squares, the total sum of squares, and the degrees of freedom, respectively, and  $n$  is the number of spatial units in this simulation study.

### 3.1.3 | Results in SAR models

Generally, the results show that the existence of clusters leads to considerably larger ABias, MSE and a low  $R^2_{\text{adj}}$  for the SAR models including the classic  $\mathbf{W}$  (hereafter AGSAR models consisting of AGSEM and AGSLM). Stronger clusters affect the AGSAR models more in terms of the bias and MSE. For the SAR models including DDWs (hereafter DDWSAR models consisting of DDWSEM and DDWSLM), most of them adjust the clustering spatial dependence well, leading to much smaller ABias, MSE and greater  $R^2_{\text{adj}}$  than the AGSAR models. Furthermore, stronger clustering leads to larger advantage of the DDWSAR models over the AGSAR models. The two DDWs (GG and GN) always keep a stable performance across various scenarios and SAR models. The details of the comparisons are presented below.

For the SEMs, shown in Table 5, all the SEMs including DDWs (hereafter DDWSEMs) outperform those including AG (hereafter AGSEM) in MSE and  $R^2_{\text{adj}}$ . The average improvements were a 77.93% reduction in the MSE and a 30.93% increase in the  $R^2_{\text{adj}}$ . For the ABias, the convergence plots, Figure S2, shows that it is difficult for the AGSEM to obtain a stable estimation of  $\beta$  in the scenarios containing strong clustering, while all the DDWSEMs achieve a stable estimation. Four kinds of DDWSEMs (ie, GGSEM, GNSEM, NGSEM, and NNSEM) achieve similar or smaller ABias values than the AGSEM, while the GRSEM and NRSEM obtain much larger ABias values, which will be discussed in Section 5. Stronger clustering makes larger advantage of the DDWSEMs over the AGSEM. Further comparison between DDWSEMs shows that those models with DDWs whose baseline weights are based on a geographic contiguity relationship (ie, GG, GN, and GR) outperform those with DDWs whose baseline weights are based on an equal-weighted relationship (ie, NG, NN, and NR) in almost all the scenarios.

For the SLMs, shown in Table 6, all the SLMs including DDWs (hereafter DDWSLMs) outperform those including AG (hereafter AGSLM) in ABias, MSE, and  $R^2_{\text{adj}}$ . Stronger clustering leads to larger advantage of the DDWSLMs over the AGSLM. The average improvements are 87.44% and 92.91% reductions in the ABias and MSE, respectively, as well as a 22.28% increase in the  $R^2_{\text{adj}}$ . Same as in the SEMs, the SLMs with GG, GN, and GR outperform those with NG, NN, and NR, respectively, in all the scenarios. Different from the result in the SEMs, the AGSLM obtains a stable estimation, shown in

TABLE 5 The simulation results in the SEM over 10 000 replicas

Scenario		AG	GG	GN	NG	NN	GR	NR
C1Hh	ABias	0.747	0.207	0.175	0.008	0.038	2.152	3.014
	MSE	12.414	0.956	0.949	1.110	1.102	0.964	1.125
	$R^2_{adj}$	0.639	0.978	0.978	0.973	0.973	0.978	0.973
C1Hl	ABias	0.030	0.003	0.017	0.030	0.016	2.643	3.467
	MSE	2.545	0.923	0.914	1.110	1.100	0.936	1.127
	$R^2_{adj}$	0.630	0.881	0.881	0.861	0.859	0.882	0.859
C1F	ABias	0.925	0.243	0.221	0.577	0.565	3.329	3.760
	RMSE	5.576	1.864	1.581	2.578	2.041	1.453	1.806
	$R^2_{adj}$	0.706	0.938	0.943	0.932	0.937	0.947	0.941
C2Hh	ABias	1.460	0.337	0.278	0.185	0.259	5.273	7.032
	MSE	25.213	0.967	0.961	1.115	1.103	1.020	1.191
	$R^2_{adj}$	0.579	0.988	0.988	0.986	0.986	0.988	0.986
C2Hl	ABias	0.526	0.120	0.146	0.130	0.159	5.831	7.285
	MSE	4.469	0.937	0.936	1.113	1.112	1.001	1.198
	$R^2_{adj}$	0.584	0.927	0.927	0.915	0.914	0.928	0.915
C2F	ABias	0.231	0.114	0.086	0.101	0.072	7.515	9.022
	MSE	11.674	2.666	2.226	3.613	2.846	1.983	2.457
	$R^2_{adj}$	0.654	0.952	0.956	0.949	0.954	0.961	0.958

Note: For the ABias and MSE, the values have been multiplied 1000 to clearly present the comparison.

Figure S4, but the ABias considerably inflates, especially in the scenarios with strong clustering. In addition, the GR and NR, which gives large ABias values in the SEM, achieve small ABias.

### 3.2 | CAR models for disease mapping

Since the work by Besag et al,<sup>35</sup> CAR models in Bayesian framework have become the most common tool to smooth the risks in disease mapping. Let  $O_i$  and  $E_i$  indicate the observed and expected cases, in spatial unit  $i$ , respectively.  $r_i$  is the RR of spatial unit  $i$  over the expected incidence.

$$O_i | r_i \sim \text{Poisson}(E_i r_i), \log(r_i) = \eta + \xi_i, \quad (1)$$

where  $\eta$ , an intercept, represents an overall level of risk.  $\xi_i$  is the spatially random effect. Let  $\xi = (\xi_1, \xi_2, \dots, \xi_n)'$ . The commonly used Leroux prior (hereafter LCAR model)<sup>36</sup> is given by

$$\xi \sim N\left(\mathbf{0}, \sigma_\xi^2 [\rho_\xi \mathbf{R}_\xi + (1 - \rho_\xi) \mathbf{I}_n]^{-1}\right), \quad (2)$$

where  $\sigma_\xi^2$  is the variance parameter,  $\rho_\xi$  represents strength of spatial dependence taking values between 0 and 1.  $\mathbf{R}_\xi$  is a  $n \times n$  symmetric matrix with the diagonal elements equal to the number of neighbors around the  $i$ th unit. For the off-diagonal element,  $(R_\xi)_{ij} = -1$  if spatial unit  $i$  and  $j$  are neighbors,  $(R_\xi)_{ij} = 0$  otherwise.

Based on LCAR model, AHC-based CAR model (hereafter AHC model) was proposed to deal with the spatial clustering (or discontinuity).<sup>30,37</sup> As the AHC model is computation-intensive especially with large number of spatial units, DBSC-based CAR model (hereafter DBSC model),<sup>31</sup> which has similar performance but with much lower computation cost than AHC model, was subsequently proposed as an alternative. DBSC model first identifies the clusters using the



**TABLE 6** The simulation results in the SLM over 10 000 replicas

Scenario		AG	GG	GN	NG	NN	GR	NR
C1Hh	ABias	436.661	3.7674	4.062	34.011	36.932	0.706	33.648
	MSE	198.881	1.087	1.092	2.485	2.706	1.067	2.469
	$R^2_{adj}$	0.728	0.976	0.976	0.971	0.970	0.976	0.970
C1Hl	ABias	63.584	0.514	0.241	31.956	34.805	3.966	30.992
	RMSE	6.554	1.062	1.064	2.354	2.556	1.080	2.305
	$R^2_{adj}$	0.685	0.879	0.879	0.851	0.849	0.880	0.850
C1F	ABias	248.806	30.916	28.407	56.611	55.975	24.731	52.213
	RMSE	66.986	2.690	2.397	5.248	5.010	2.044	4.388
	$R^2_{adj}$	0.771	0.942	0.946	0.933	0.937	0.951	0.944
C2Hh	ABias	865.224	7.087	7.5550	35.081	38.235	0.094	30.890
	RMSE	760.985	1.143	1.149	2.563	2.806	1.087	2.297
	$R^2_{adj}$	0.712	0.987	0.987	0.984	0.984	0.987	0.984
C2Hl	ABias	144.303	2.790	3.254	33.653	36.765	4.387	29.169
	RMSE	24.933	1.097	1.100	2.474	2.705	1.110	2.206
	$R^2_{adj}$	0.659	0.924	0.924	0.908	0.907	0.925	0.908
C2F	ABias	524.901	62.686	58.240	79.726	77.186	48.519	67.418
	RMSE	283.911	6.244	5.472	8.965	8.314	4.083	6.483
	$R^2_{adj}$	0.749	0.956	0.960	0.953	0.956	0.966	0.963

Note: For the ABias and MSE, the values have been multiplied 1000 to clearly present the comparison.

DBSC algorithm and then incorporates the identified clusters into LCAR model by adding cluster-level spatial structure to the model. When the number of identified clusters is not large, a fixed cluster-level spatial effect is used, as such, Equation (1) is modified to

$$\log(r_i) = \eta + \xi_i + \sum_{j=1}^k I[U_i \in G_j] \beta_j, \quad (3)$$

where  $k$  is the total number of clusters,  $I[\cdot]$  is an indicator function so that  $I[U_i \in G_j] = 1$  if spatial unit  $i$  lies in cluster  $G_j$ ,  $I[U_i \in G_j] = 0$  otherwise.  $\beta_j$  is the fixed parameter to estimate. When the number of identified clusters is large, a random cluster-level spatial structure is added, as such, Equation (3) becomes

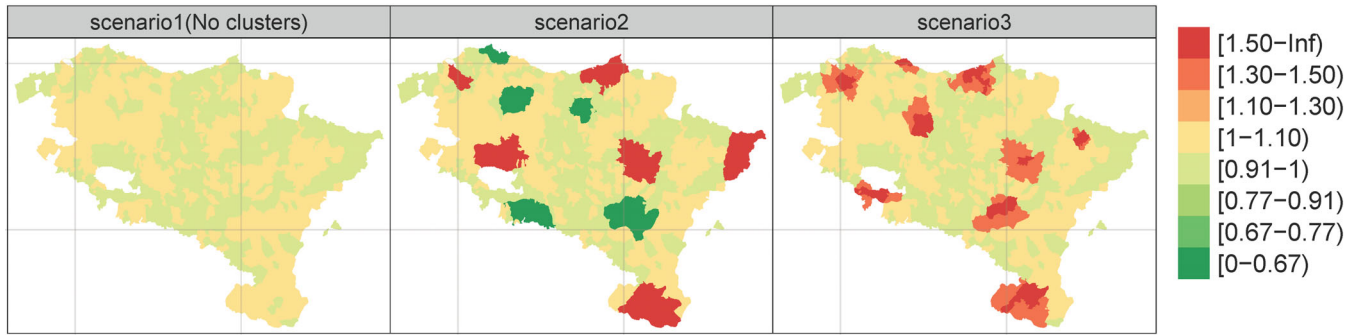
$$\log(r_i) = \eta + \xi_i + \varphi_{j(i)}, \quad (4)$$

where  $j(i)$  indicates that spatial unit  $i$  lies in cluster  $G_j$ . Let  $\boldsymbol{\varphi} = c(\varphi_1, \dots, \varphi_k)$ , like the random spatial structure for  $\xi$ , the cluster-level spatial structure can be represented as

$$\boldsymbol{\varphi} \sim N(\mathbf{0}, \sigma_\varphi^2 [\rho_\varphi \mathbf{R}_\varphi + (1 - \rho_\varphi) \mathbf{I}_k]^{-1}),$$

where  $\sigma_\varphi^2$  is the cluster-level variance parameter,  $\rho_\varphi$  represents strength of spatial dependence between clusters.  $\mathbf{R}_\varphi$  is a  $k \times k$  symmetric matrix with the diagonal elements equal to the number of neighbors around the  $j$ th cluster. For the off-diagonal element,  $(R_\varphi)_{ij} = -1$  when cluster  $i$  and  $j$  are neighbors,  $(R_\varphi)_{ij} = 0$  otherwise. Another noteworthy thing is that two hyperparameters (background parameter and neighbor-level parameter) need to be prespecified or selected by certain standard, such as the logarithm score (LS).<sup>38</sup>

To clearly understand how to employ DDWs in the CAR model, we get insight into the correlated spatial structure above that



**FIGURE 3** The distribution of clusters in the simulation scenarios for the CAR models. No clusters exist in scenario 1, 11 high/low-risk clusters in scenario 2, and 9 intra-heterogeneous clusters in scenario 3

$$\mathbf{R}_\xi = \mathbf{D}_\mathcal{N} (\mathbf{I}_n - \mathbf{W}),$$

where  $\mathbf{D}_\mathcal{N}$  is a diagonal matrix with the diagonal vector equal to  $\mathcal{N}$  whose  $i$ th element is the number of neighbors around spatial unit  $i$ , and  $\mathbf{W}$  is just the classic row-standardized adjacence-based  $\mathbf{W}$  as in the SAR models. When  $\rho = 1$  in the LCAR,  $\xi \sim N(\mathbf{0}, \sigma_\xi^2 \mathbf{R}_\xi^{-1})$ , we obtain the intrinsic CAR (ICAR) model, in which an intuitive explanation about  $\mathbf{W}$  is given as.

$$\xi_i | \xi_{-i} \sim N\left(\sum_{j=1}^n w_{ij} \xi_j, \frac{\sigma_\xi^2}{\mathcal{N}_i}\right).$$

Therefore, the LCAR and DBSC model can be written as the spatial models including the classic  $\mathbf{W}$ , and then, it is reasonable to replace the classic  $\mathbf{W}$  with a DDW to construct a new CAR model. In this work, based on the DBSC model dealing with the clustering (or discontinuous) spatial dependence, we use DDWs to substitute the classic  $\mathbf{W}$  in  $\mathbf{R}_\xi$  to construct DDWCAR models, in which the random correlated spatial prior could be defined as.

$$\xi \sim N\left(\mathbf{0}, \sigma_\xi^2 [\rho_\xi \mathbf{D}_\mathcal{N} (\mathbf{I}_n - \mathbf{W}_{\text{DDW}}) + (1 - \rho_\xi) \mathbf{I}_n]^{-1}\right).$$

And like that in the DBSC model, the cluster-level effect is introduced into the DDWCAR model using either fixed effect or random effect according to the number of identified clusters. As  $\mathbf{W}_{\text{NR}}$  and  $\mathbf{W}_{\text{GR}}$  fail to achieve the symmetric attribute of  $\mathbf{D}_\mathcal{N} (\mathbf{I}_n - \mathbf{W}_{\text{DDW}})$  as in  $\mathbf{R}_\xi$ , only the other four DDWs could be used to build a total of four DDWCAR models (ie, NNCAR, NGCAR, GNCAR, GGCAR).

### 3.2.1 | Data generation process in the CAR

To obtain comparable results, the simulation scenarios used to validate the efficiency of the latest DBSC model are employed to evaluate the DDWCAR performance in disease mapping.<sup>31</sup> They derive from the cancer mortality data for 2011 to 2015 and are composed of three simulation scenarios in 508 Spanish municipalities, shown in Figure 3, the first scenario with a spatial smooth surface based on the intuitive adjacence-based spatial structure without clusters (scenario 1), the second scenario with 11 high/low-risk homogeneous clusters without overlap (scenario 2), and the last scenario with 9 high-risk but intra-heterogeneous clusters without overlap (scenario 3). Each scenario consists of three sub-scenarios (labeled as A, B, and C): the expected cases derive from the real cancer mortality data (A), the expected cases of A are multiplied by 0.1 and 1/30 to get ones of B and C, respectively. The three subscenarios imitate the occurrence of common diseases, rare diseases, and much rare diseases, respectively. According to the expected cases and risk level of each municipality, the Poisson-distributed observed cases were generated with 100 replicas. As such, the average proportions of spatial units with observed cases equal to zero are 3.3%, 36%, and 57% for A, B, and C, respectively. The detailed data generation process could be found in the work by Santafé et al.<sup>31</sup>

### 3.2.2 | Performance evaluation in the CAR

Four DDWCAR models were fitted along with the DBSC model and the classic LCAR model as references. For DBSC models, three neighbor-level parameters ( $\ell = 1, 2, 3$ ) with or without background cluster were selected as those in Santafé et al.<sup>31</sup> So, one LCAR, six DBSC, and four DDWCAR models were fitted to estimate the relative risks of each spatial units for each simulation datasets. The performances of these models were compared in terms of mean absolute relative bias (MARB) and mean relative root mean square error (MRRMSE) which are defined as

$$\text{MARB} = \frac{1}{n} \sum_{i=1}^n \frac{1}{100} \left| \sum_{s=1}^{100} \frac{\hat{r}_i^s - r_i}{r_i} \right|, \quad \text{MRRMSE} = \frac{1}{n} \sum_{i=1}^n \sqrt{\frac{1}{100} \sum_{s=1}^{100} \left( \frac{\hat{r}_i^s - r_i}{r_i} \right)^2},$$

where  $r_i$  is the artificial real relative risk of spatial unit  $i$ ,  $\hat{r}_i^s$  is the estimated relative risk in  $s$ th simulation dataset and  $n$  is the number of spatial units. MARB reflects the predictive bias of models, which is of more interest in risk comparison between spatial units, while MRRMSE reflect the predictive accuracy of models, which is more important for predicting the exact risk values of spatial units. In addition, the LS is also given as a secondary measure of the model predictive ability as a cross-validation-based measure.<sup>38</sup> With no relying on the real values, the LS can be used to select a preferred model.

### 3.2.3 | Results in the CAR

The comparison results are shown in Table 7. Generally, when no clusters exist, all the models obtain similar MARB values. The LCAR and the DDWCAR models obtain competitive MRRMSE values but considerably lower than the DBSC models. When clusters exist, in term of MRRMSE, the DDWCAR models achieve the best performance, while the DBSC models achieve the poorest performance, even get considerably larger MRRMSE values than the classic model, LCAR. In term of MARB, the DDWCAR models achieve similar or better performance than the DBSC models in the scenarios for the common diseases. Although the DDWCAR models get larger MARB values than the DBSC models in most of the scenarios for rare (or much rare) diseases, they still get considerably lower MARB values than the LCAR model. In all the scenarios, the LS scores of the DDWCAR models are lower than the other models and capable of selecting an acceptably optimal model. The details of the comparisons are presented below.

In the scenarios without clusters (scenario 1), the LCAR model achieve the best performance in term of both MARB and MRRMSE as expected. Compared with the LCAR model, the DDWCAR models obtains the same MARB and competitive MRRMSE (ie, the largest difference is 0.052 vs 0.032 seen in scenario 1C). For the DBSC models, they obtained considerably larger MRRMSE values than the LCAR and DDWCAR models, even in the optimal DBSC model, the inflations of MRRMSE over the LCAR models reach 107.1% (0.015/0.014), 91.3% (0.021/0.023), and 215.6% (0.069/0.032) in scenario 1A, scenario 1B, and scenario 1C, respectively.

In the scenarios with clusters (ie, scenario 2 and scenario 3), in term of MARB, the LCAR models show the poorest performance, and both the DBSC and the DDWCAR models considerably improve the MARB. Comparing the DBSC models with the DDWCAR models, in the scenarios for common diseases (ie, scenario 2A and scenario 3A), the DDWCAR models obtain the smaller (0.047 vs 0.054 seen in scenario 2A) or similar (0.04 vs 0.04 seen in scenario 3A) MARB values. In the scenarios for rare (or much rare) diseases (ie, scenario 2B, 2C, 3B, and 3C), the DDWCAR models obtain larger MARB values in most of scenarios, but also get smaller MARB in scenario 2C. In term of MRRMSE, the DBSC models still inflates the MRRMSE over the LCAR model as that in scenario 1, while the DDWCAR models get considerable improvement in almost all the scenarios. Another interesting result is that GNCAR and GGCAR has similar or better performance than NNCAR and NGCAR in terms of both MARB and MRRMSE.

## 4 | CASE STUDIES OF INCORPORATING DDWS INTO THE SAR AND CAR MODELS

Hand, foot, and mouth disease (HFMD) is an acute contagious disease transmitted via the oral-fecal route and by contact with respiratory droplets and contaminated food.<sup>39</sup> It primarily affects children less than 5 years old and has

TABLE 7 Average values of MARB, MRRMSE, and LS in the CAR models

		No background cluster				With background cluster						
		LCAR	DBSC $\ell=1$	DBSC $\ell=2$	DBSC $\ell=3$	DBSC $\ell=1$	DBSC $\ell=2$	DBSC $\ell=3$	NNCAR	NGCAR	GNCAR	GGCAR
Scenario 1A	MARB	0.012	0.013	0.014	0.013	0.012	0.012	0.012	0.012	0.012	0.012	0.012
	MRRMSE	0.014	0.034	0.038	0.036	0.029	0.033	0.032	0.016	0.016	0.016	0.016
	LS	1394.7	1420.4	1540.7	1417.7	1414.1	1411.5	1405.9	1394.5	1394.5	1394.4	1394.4
Scenario 1B	MARB	0.012	0.017	0.034	0.086	0.014	0.017	0.031	0.012	0.012	0.012	0.012
	MRRMSE	0.023	0.05	0.093	0.169	0.044	0.056	0.1	0.026	0.026	0.026	0.026
	LS	776.5	775.6	774.9	783.4	776.2	776.0	776.6	776.6	776.6	776.5	776.5
Scenario 1C	MARB	0.012	0.047	0.057	0.078	0.033	0.033	0.032	0.013	0.013	0.013	0.013
	MRRMSE	0.032	0.112	0.129	0.165	0.101	0.101	0.112	0.053	0.052	0.054	0.054
	LS	521.4	520.6	521.1	524.5	520.8	521.1	522.1	522.6	523.6	5214	521.4
Scenario 2A	MARB	0.058	0.054	0.054	0.057	0.054	0.055	0.057	0.053	0.052	0.052	0.047
	MRRMSE	0.125	0.129	0.131	0.133	0.125	0.13	0.132	0.138	0.138	0.123	0.124
	LS	1533.9	1530.0	1532.7	1547.1	1514.4	1522.2	1539.8	1539.5	1539.5	1509.5	1516.6
Scenario 2B	MARB	0.096	0.099	0.113	0.127	0.099	0.102	0.104	0.117	0.117	0.11	0.112
	MRRMSE	0.17	0.213	0.243	0.268	0.207	0.213	0.219	0.143	0.143	0.145	0.143
	LS	826.5	817.0	820.8	833.8	817.0	819.6	822.7	807.6	807.6	803.3	804.9
Scenario 2C	MARB	0.132	0.13	0.134	0.125	0.121	0.119	0.115	0.111	0.111	0.108	0.108
	MRRMSE	0.202	0.266	0.276	0.285	0.268	0.267	0.266	0.187	0.186	0.187	0.183
	LS	555.3	547.8	551.9	558.9	547.5	550.3	552.4	545.5	545.3	541.8	542.9
Scenario 3A	MARB	0.049	0.04	0.045	0.046	0.041	0.045	0.047	0.051	0.051	0.041	0.04
	MRRMSE	0.097	0.1	0.104	0.104	0.098	0.1	0.103	0.116	0.116	0.095	0.100
	LS	1522.6	1522.9	1536.4	1531.8	1508.6	1519.5	1525.4	1520.9	1520.9	1493.1	1498.6
Scenario 3B	MARB	0.105	0.075	0.075	0.098	0.079	0.08	0.078	0.106	0.105	0.087	0.090
	MRRMSE	0.139	0.158	0.196	0.233	0.162	0.171	0.184	0.129	0.128	0.118	0.118
	LS	833.6	825.1	827.5	834.5	825.1	828.4	829.7	830.9	830.5	824.0	824.9
Scenario 3C	MARB	0.152	0.081	0.087	0.096	0.081	0.089	0.088	0.139	0.138	0.124	0.125
	MRRMSE	0.174	0.207	0.226	0.243	0.206	0.216	0.224	0.166	0.165	0.156	0.156
	LS	558.8	552.0	553.9	561.1	552.6	553.9	556.3	556.3	556.3	552.2	552.4

*Note:* In the DDWCAR models, the fixed cluster-level spatial effects are considered for a relatively small number of identified clusters. We also tried the random cluster-level spatial effects, which performs a similar result. Only the high-risk clusters were considered in the scan statistic for the rare (or much rare) diseases.

become a major public health issue in the Asia-Pacific region.<sup>40-42</sup> As studies have demonstrated that HFMD incidence is affected by natural and social environments, such as meteorological factors and economic and medical care levels,<sup>43-45</sup> discontinuous (or clustering) spatial dependence of HFMD risk can easily be found due to specific terrain or social development levels. Therefore, in this section, we aim to investigate the effects of meteorological variables on HFMD and estimate the excess risk of HFMD for each county to provide a practical application of DDWs in SAR and CAR models, respectively.

For each county in Sichuan Province, the incidences under the age of 5 years in April 2014 were collected from the Chinese Center for Disease Control and Prevention. With respect to previous studies, we collected climatic factors from the China Meteorological Bureau, including the monthly average wind speed, sunshine time, temperature, and humidity.<sup>46,47</sup>

**TABLE 8** The results of parameter estimation in the SAR models for the case study

X	AG	NG	NR	NN	GG	GR	GN
Temperature	0.0169*** (0.0036)	0.0207*** (0.0031)	0.0176*** (0.0027)	0.0220*** (0.0031)	0.0160*** (0.0031)	0.0110*** (0.0025)	0.0174*** (0.0032)
Sunshine	−0.0230 (0.0141)	−0.0265** (0.0131)	−0.0198* (0.0117)	−0.0241* (0.0132)	−0.0244** (0.0127)	−0.0156 (0.0105)	−0.0224* (0.0127)
Wind	0.0066 (0.0378)	0.0263 (0.0355)	0.0650** (0.0314)	0.0567 (0.0355)	0.0204 (0.0341)	0.0582** (0.0282)	0.0494 (0.0341)
Humidity	−0.0164 (0.0197)	−0.0239 (0.0188)	−0.0072 (0.0166)	−0.0203 (0.0190)	−0.0199 (0.0178)	−0.0019 (0.0148)	−0.0165 (0.0179)
$R^2_{adj}$	0.58	0.63	0.71	0.63	0.66	0.77	0.66
AIC	713.08	691.48	650.05	685.75	685.42	628.71	681.22

Note: The values in parentheses are the standard errors corresponding to the estimated parameters.

\* $P < .1$ .

\*\* $P < .05$ .

\*\*\* $P < .001$ .

## 4.1 | SAR models in application

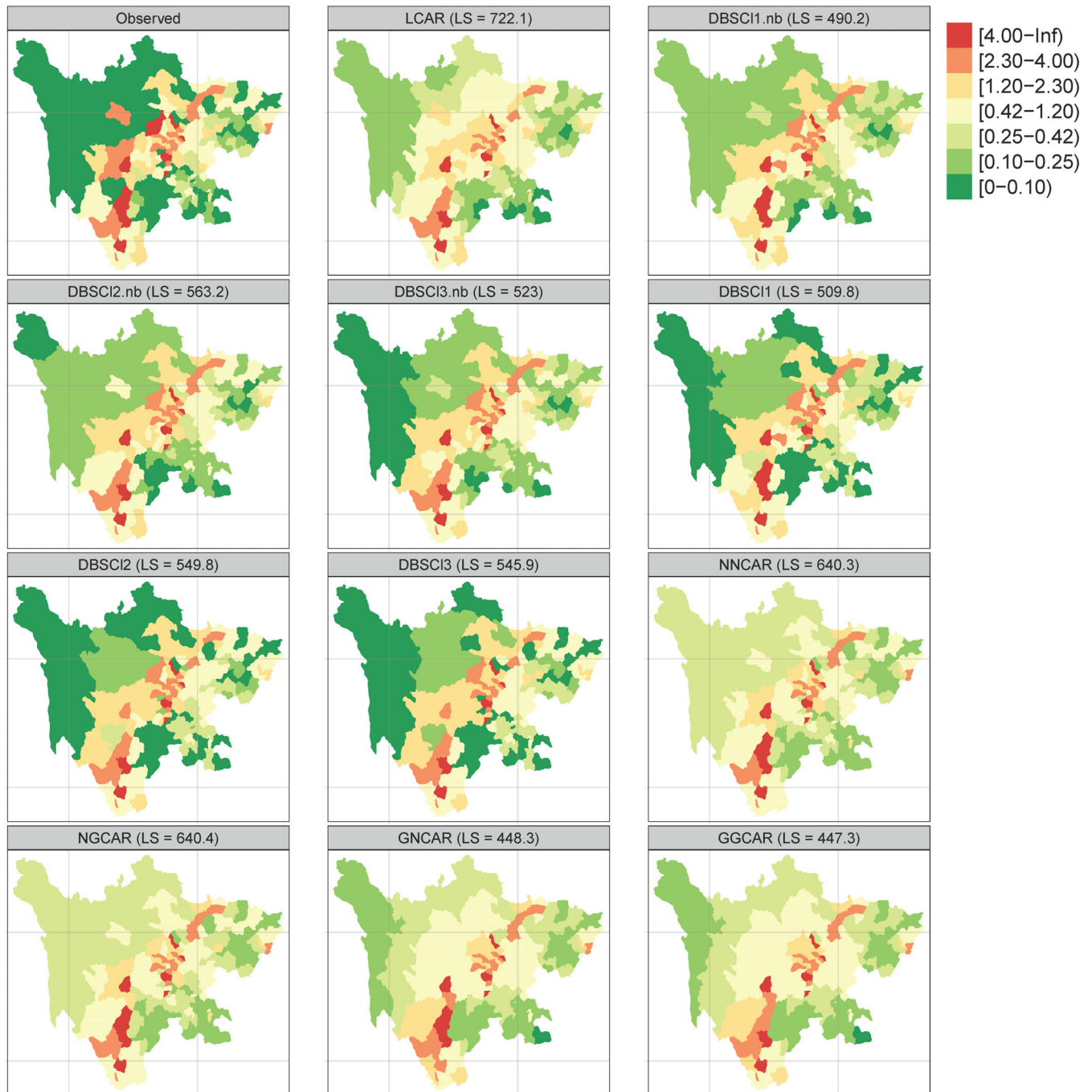
First, the scan statistic and the MCS-P were employed to find clusters in this dataset, and the selected maximum scan window was 14%. Seven clusters that comprise 95 counties are identified, seen in Figure S6. Then, based on the clusters, six kinds of DDWs were constructed. Finally, based on the Lagrange multiplier (LM) test, the SLM was selected, and the six DDWs and AG were included in the model. The logarithm for  $y$  is used in the model under the assumption of a lognormal distribution. As no true  $\beta$  is known in practice, AIC and  $R^2_{adj}$  were used to evaluate and compare the performance of different DDWs and AG.

The results in Table 8 show that all the models with DDWs obtain a higher  $R^2_{adj}$ , a lower AIC and smaller standard errors than that with the classic **W** (ie, AG), suggesting that the DDWs achieve more stable parameter estimations and more statistical power than the classic **W**. This advantage in the statistical power leads to a stronger capacity to find existing associations. For example, in the models with NG and GG, sunshine time is identified as a protective factor for HFMD, as found in the previous study.<sup>47</sup> However, the SLM with AG fails to identify such an association because of the smaller statistical power. The other DDWs have similar performances. In addition, the models with GR or NR obtains considerably higher  $R^2_{adj}$  and smaller AIC than the models with other DDWs. The reason for this finding may be that more than half of the counties are found as clustering regions. Therefore, the assigned weights based on the observed values in clusters utilize too much information from random errors and leads to overfitting. Overall, the models with DDWs outperform those with the classic **W**, AG.

## 4.2 | CAR models in application

The same method as in SAR was used to identify the clusters. The expected cases of each county were calculated based on the population. Like in the simulation study for CAR models, four DDWCAR models (ie, NN CAR, NG CAR, GN CAR, and GG CAR) were fitted, along with the references, the classic LCAR and six DBSC models. Because of the real risk values unknown, the LS was used as the main performance measure. A smaller LS means a better model predictive ability.

The result shows that the LCAR model, as expected, obtains the largest LS values (722.1), that is, the poorest risk estimation, due to the strong clustering (or discontinuous) spatial dependence of HFMD risk in Sichuan province. As shown in Figure 4, low-risk clusters are found in the western plateau and high-risk clusters are found in the central plains. The DBSC and DDWCAR models considerably improve the risk estimation, that is, achieve smaller LS values. Furthermore, the smallest LS in the DDWCAR models is 447.3 found in the GG CAR model, which is considerably smaller than 490.2



**FIGURE 4** The estimated risk surface and logarithm score (LS) values using LCAR, DBSC, and DDWCAR models in the case study

achieved by the optimal DBSC model (ie, DBSC11.nb model with one level of neighbor and without background), suggesting the DDWCAR models achieve a better risk estimation than the DBSC models. In addition, we found the GNCAR models also has similar LS values (448.3 vs 447.3) with the GGCAR model, which conforms to simulated results.

**5 | DISCUSSION**

Clustering (or discontinuous) spatial dependence widely exists in spatial datasets, which will reduce the performance of the spatial models with the classic spatial weights matrix (AG). In this work, we proposed six novel spatial weights



matrices, called DDWs, to deal with this problem. The DDWs are constructed based on the identified clusters by the scan statistic and the natural adjacency relationship, which take both the clustering and the intuitive adjacency-based spatial dependence into consideration. Aiming at two common purposes in epidemiology studies (ie, estimating the effect of explanatory variable and estimating the RR of each spatial unit in disease mapping), the common SAR models, that is, SEM and SLM, and the CAR models were built to evaluate performance of DDWs, respectively.

In the SAR models, all the six DDWs are available for the SEM and SLM. In the simulation scenarios, the SAR models incorporating the four DDWs (GG, GN, NN, NG) achieve considerable improvement over those including the AG in terms of bias, MSE and  $R^2_{adj}$ , and the stronger clustering leads to larger advantages, suggesting the DDWs have a good ability to adjust the clustering (or discontinuous) spatial dependence. But the other two DDWs (GR and NR), which may achieve a better performance in the scenarios with inner-heterogeneous clusters, lead to unacceptable large bias in the SEM for the reason that the risk-weighting method employs too much information from random errors. Thus, it is risky to adopt the risk-weighting method in complex real-world datasets. Furthermore, as the baseline spatial dependence was set to contain only the first-law-of-geography-based spatial dependence, in the SAR models, the DDWs with baseline weights based on the geographic contiguity relationship (ie, GG, GN, and GR) outperform those DDWs with baseline weights based on the equal-weighted relationship (ie, NG, NN, and NR), as expected. This suggests that the spatial weights matrices ignoring the intuitive first-law-of-geography-based spatial dependence, such as AMOEBA,<sup>14</sup> NN, NG, and NR, might not be appropriate choices for real-world datasets in which such spatial dependence commonly exists. The case study shows similar results to those from simulation study, that is, the DDWs achieves stronger statistical power, higher  $R^2_{adj}$ , and lower SE than the AG.

Noteworthy, in the simulation datasets for the SLM, we increased the spatial dependence from the observed explanatory variable by appropriately enlarging this variable to satisfy the condition of the SLM. Otherwise, the strong clusters will make the spatial dependence from the observed explanatory variable ignorable, in which conditions, the SEM or spatial Durbin model should be used.<sup>5</sup> Several other values of the observed independent variable were simulated to validate the robustness of DDWs in the SLM and showed similar results. This may suggest a SLM including DDWs still has a stable performance even in situations that the spatial dependence does not exactly satisfy the assumption of the SLM.

In the CAR models, because the two DDWs (GR and NR) cannot achieve the symmetry of the random correlated spatial prior, seen in Section 3.2, only four DDWs (GG, GN, NN, NG) are available for the CAR models. Based on the latest proposed DBSC model which aims to deal with the clustering spatial dependence, four DDWCAR models were constructed. In simulation studies, the DDWCAR models achieve the considerably smaller MRRMSE values than the common LCAR and DBSC models in the datasets with clustering spatial dependence. Moreover, in the datasets without clusters, they also achieve similar performance with the LCAR models. In addition, similar to the pervious study,<sup>31</sup> DBSC models considerably inflate the MRRMSE over the LCAR model in all the simulation datasets. A smaller MRRMSE means a more accuracy risk prediction, which suggests that the DDWCAR models should be recommended for predicting the risk of each spatial unit. On the other hand, the bias is of more interest for the risk comparison between the spatial units. As the simulation study shows, in the clustering datasets for the common diseases, the DDWCAR models improve the MARB over the DBSC and LCAR models, while in the clustering datasets for rare (or much rare) diseases, the DBSC models achieves the best MARB values. This suggests that, focusing on the risk comparison, the DDWCAR models should be recommended for common diseases, while the DBSC models should be recommended for rare (or much rare) diseases. Same as that in SAR models, the GG and GN achieve better performance than NG and NN with the widely existed geographic adjacency-based relationship considered, which also supported by the case study where the GG and GN achieve the smallest LS values.

In addition, we must acknowledge that, due to the accessibility of data, an ancillary history dataset was not included in this case study, which may lead to overfitting of the DBSC models.<sup>31</sup> Taking a further insight into the overfitting for the two methods, the DBSC models do not test the identified clusters, which may find excessive number of clusters; while the DDWCAR models only identify statistically significantly clustering regions as clusters, which will substantially alleviate the overfitting. So, under the condition without ancillary datasets, the DDWCAR models may be further recommended.

Another issue worth noting, the proposed DDWSAR and DDWCAR models may suffer from the issues of identifiability between spatial random effects and the fixed effects (including the intercept).<sup>48-51</sup> Luckily, the previous methods, such as the reparameterization method,<sup>50,51</sup> restricted spatial regression,<sup>49-51</sup> and constraining the random effects to be orthogonal to the fixed effects,<sup>51</sup> are still available at least for the DDWSEM and DDWCAR models. The preliminary derivation could be seen in the supplementary material, while the specific performance in practical applications needs further research. For the DDWSLM model, because a part of spatial random effect comes from the observed response variable, which is

different from that in the SEM and CAR model, the previous method to deal with the identifiability issue may not be simply expanded to DDWSLM.

In this study, though CAR models can both smooth the risk and estimate the coefficients, they were only evaluated for risk smoothing to enhance the comparability between the DDWCAR and the previously developed clustering-dealing-with method, that is, DBSC model. The performance of DDWCAR models on the regression coefficient estimates were not considered, which would be one of our future works.

In conclusion, the proposed DDWs, especially the GG and GN, are capable of adjusting both the clustering (or discontinuous) and the intuitive first-law-of-geographic-based spatial dependence, and therefore, improving the performance of the commonly used SAR and CAR models. Although the advantage of DDWs over the classic **W** was only evaluated in the SAR and CAR models for effect estimating and disease mapping, respectively, the DDWs may also be used in other spatial models with the similar usage of **W**, such as spatial interpolation model, which would be part of our future work. On the other hand, the DDWs are based on the identified clusters by the scan statistic, which could be conveniently implemented by the SaTScan software. As many methods have been carried out to obtain more accuracy clusters according to specific studies,<sup>52-55</sup> such accuracy improvement may lead to more effective DDWs.

## ACKNOWLEDGEMENTS

The authors would like to thank the editor, two anonymous reviewers and Xingyu Zhang for their constructive comments, which improved the quality of this paper. In addition, we appreciate our colleagues, Sheng Li, Xuelin Li, Jingfei Huang, Siwei Zhai, Xinyue Tian, and Yi Zhang, for collecting the case dataset and running the large number of simulation codes. Specifically, we appreciate Zihao Niu (Bottle 152) for bringing us joy and inspiration during the hard-working days.

## CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

## DATA AVAILABILITY STATEMENT

The simulation datasets that support the findings of this study are available at <https://github.com/winkey1230/DDW-article>. And the datasets in case study are available from the corresponding author upon reasonable request.

## ORCID

Tao Zhang  <https://orcid.org/0000-0001-7279-8929>

Yue Ma  <https://orcid.org/0000-0002-1980-7520>

## REFERENCES

1. Beale CM, Lennon JJ, Yearsley JM, Brewer MJ, Elston DA. Regression analysis of spatial data. *Ecol Lett*. 2010;13(2):246-264. doi:10.1111/j.1461-0248.2009.01422.x
2. Auchincloss AH, Gebreab SY, Mair C, Diez Roux AV. A review of spatial methods in epidemiology, 2000-2010. *Annu Rev Public Health*. 2012;33:107-122. doi:10.1146/annurev-publhealth-031811-124655
3. Ripley BD. Review of spatial process: models and applications, by AD Cliff, JK Ord. *J Am Stat Assoc*. 1984;79(385):238. doi:10.2307/2288381
4. Griffith DA. Review of spatial econometrics: methods and models, by L Anselin. *Econ Geogr*. 1989;65(2):160-162. doi:10.2307/143780
5. Vega SH, Elhorst JP. The Slx model. *J Reg Sci*. 2015;55(3):339-363. doi:10.1111/jors.12188
6. Getis A. Spatial weights matrices. *Geogr Anal*. 2009;41(4):404-410. doi:10.1111/j.1538-4632.2009.00768.x
7. Moran PAP. The interpretation of statistical maps. *J R Stat Soc Ser B Stat Methodol*. 1948;10(2):243-251.
8. Moran PAP. Notes on continuous stochastic phenomena. *Biometrika*. 1950;37(1-2):17-23. doi:10.2307/2332142
9. Geary RC. The contiguity ratio and statistical mapping. *Incorporated Stat*. 1954;5(3):115-146.
10. Cliff AD, Ord K. Spatial autocorrelation: a review of existing and new measures with applications. *Econ Geogr*. 1970;46(sup1):269-292.
11. Cliff AD, Ord K. Testing for spatial autocorrelation among regression residuals. *Geogr Anal*. 1972;4(3):267-284.
12. Anselin L. Local indicators of spatial association—LISA. *Geogr Anal*. 1995;27(2):93-115. doi:10.1111/j.1538-4632.1995.tb00338.x
13. Getis A, Ord JK. The analysis of Spatial association by use of distance statistics. *Geogr Anal*. 1992;24(3):189-206. doi:10.1111/j.1538-4632.1992.tb00261.x
14. Aldstadt J, Getis A. Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geogr Anal*. 2006;38(4):327-343. doi:10.1111/j.1538-4632.2006.00689.x
15. Getis A, Aldstadt J. Constructing the spatial weights matrix using a local statistic. *Geogr Anal*. 2004;36(2):90-104. doi:10.1111/j.1538-4632.2004.tb01127.x
16. Dormann CF, McPherson JM, Araújo MB, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*. 2007;30(5):609-628. doi:10.1111/j.2007.0906-7590.05171.x

17. Messner SF, Anselin L, Baller RD, et al. The spatial patterning of county homicide rates: an application of exploratory spatial data analysis. *J Quant Criminol*. 1999;15(4):423-450. doi:10.1023/A:1007544208712
18. Kim CW, Phipps TT, Anselin L. Measuring the benefits of air quality improvement: a spatial hedonic approach. *J Environ Econ Manage*. 2003;45(1):24-39. doi:10.1016/S0095-0696(02)00013-X
19. Harouvi O, Ben-Elia E, Factor R, de Hoogh K, Kloog I. Noise estimation model development using high-resolution transportation and land use regression. *J Expo Sci Environ Epidemiol*. 2018;28(6):559-567. doi:10.1038/s41370-018-0035-z
20. Mahara G, Yang K, Chen S, Wang W, Guo X. Socio-economic predictors and distribution of tuberculosis incidence in Beijing, China: a study using a combination of spatial statistics and GIS technology. *Med Sci*. 2018;6(2):26. doi:10.3390/medsci6020026
21. Morley DW, Gulliver J. A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment. *Environ Model Softw*. 2018;105:17-23. doi:10.1016/j.envsoft.2018.03.030
22. Stakhovych S, Bijmolt THA. Specification of spatial models: a simulation study on weights matrices. *Pap Reg Sci*. 2009;88(2):389-408. doi:10.1111/j.1435-5957.2008.00213.x
23. Jerrett M, Gale S, Kontgis C. Spatial modeling in environmental and public health research. *Int J Environ Res Public Health*. 2010;7(4):1302-1329. doi:10.3390/ijerph7041302
24. Tobler WR. A computer movie simulating urban growth in the detroit region. *Econ Geogr*. 1970;46(sup1):234-240. doi:10.2307/143141
25. Harris R, Moffat J, Kravtsova V. In search of "W". *Spat Econ Anal*. 2011;6(3):249-270. doi:10.1080/17421772.2011.586721
26. Bhattacharjee A, Jensen-Butler C. Estimation of spatial weights matrix in a spatial error model, with an application to diffusion in housing demand; 2006; St. Andrews, Centre for Research into Industry, Enterprise, Finance and the Firm.
27. Xia M, Yang S, Ho SL. A new topology optimization methodology based on constraint maximum-weight connected graph theorem. *IEEE Trans Magn*. 2017;54(3):1-4. doi:10.1109/TMAG.2017.2757001
28. Parent O, LeSage JP. Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers. *J Appl Economet*. 2008;23(2):235-256. doi:10.1002/jae.981
29. Duncan EW, White NM, Mengersen K. Spatial smoothing in Bayesian models: a comparison of weights matrix specifications and their impact on inference. *Int J Health Geogr*. 2017;16(1):47. doi:10.1186/s12942-017-0120-x
30. Adin A, Lee D, Goicoa T, Ugarte MD. A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. *Stat Methods Med Res*. 2019;28(9):2595-2613. doi:10.1177/0962280218767975
31. Santafé G, Adin A, Lee D, Ugarte MAD. Dealing with risk discontinuities to estimate cancer mortality risks when the number of small areas is large. *Stat Methods Med Res*. 2021;30(1):6-21. doi:10.1177/0962280220946502
32. Ribeiro SH, Costa MA. Optimal selection of the spatial scan parameters for cluster detection: a simulation study. *Spat Spatiotemp Epidemiol*. 2012;3(2):107-120. doi:10.1016/j.sste.2012.04.004
33. Ma Y, Yin F, Zhang T, Zhou XA, Li X. Selection of the maximum spatial cluster size of the spatial scan statistic by using the maximum clustering set-proportion statistic. *PLoS One*. 2016;11(1):e0147918. doi:10.1371/journal.pone.0147918
34. Elhorst JP. Applied spatial econometrics: raising the bar. *Spat Econ Anal*. 2010;5(1):9-28. doi:10.1080/17421770903541772
35. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*. 1991;43(1):1-20. doi:10.1007/BF00116466
36. Leroux BG, Lei X, Breslow N. *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*. New York, NY: Springer; 2000.
37. Anderson C, Lee D, Dean N. Identifying clusters in Bayesian disease mapping. *Biostatistics*. 2014;15(3):457-469. doi:10.1093/biostatistics/kxu005
38. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*. 2007;102(477):359-378. doi:10.1198/016214506000001437
39. Xu C. Spatio-temporal pattern and risk Factor analysis of hand, foot and mouth disease associated with under-five morbidity in the Beijing-Tianjin-Hebei region of China. *Int J Environ Res Public Health*. 2017;14(4):416. doi:10.3390/ijerph14040416
40. Xing W, Liao Q, Viboud C, et al. Hand, foot, and mouth disease in China, 2008-12: an epidemiological study. *Lancet Infect Dis*. 2014;14(4):308-318. doi:10.1016/S1473-3099(13)70342-6
41. Upala P, Apidechkul T, Suttana W, Kullawong N, Tamornpark R, Inta C. Molecular epidemiology and clinical features of hand, foot and mouth disease in northern Thailand in 2016: a prospective cohort study. *BMC Infect Dis*. 2018;18(1):630. doi:10.1186/s12879-018-3560-4
42. Cobbin JCA, Britton PN, Burrell R, et al. A complex mosaic of enteroviruses shapes community-acquired hand, foot and mouth disease transmission and evolution within a single hospital. *Virus Evol*. 2018;4(2):vey020. doi:10.1093/ve/vey020
43. Yin F, Ma Y, Zhao X, et al. The association between diurnal temperature range and childhood hand, foot, and mouth disease: a distributed lag non-linear analysis. *Epidemiol Infect*. 2017;145(15):3264-3273. doi:10.1017/S0950268817002321
44. Tian L, Liang F, Xu M, Jia L, Pan X, Clements ACA. Spatio-temporal analysis of the relationship between meteorological factors and hand-foot-mouth disease in Beijing, China. *BMC Infect Dis*. 2018;18(1):158. doi:10.1186/s12879-018-3071-3
45. Nguyen THT, Everaert G, Boets P, et al. Modelling tools to analyze and assess the ecological impact of hydropower dams. *Water*. 2018;10(3):259. doi:10.3390/w10030259
46. Bo YC, Song C, Wang JF, Li XW. Using an autologistic regression model to identify spatial risk factors and spatial risk patterns of hand, foot and mouth disease (HFMD) in mainland China. *BMC Public Health*. 2014;14:358. doi:10.1186/1471-2458-14-358
47. Deng T, Huang Y, Yu S, et al. Spatial-temporal clusters and risk factors of hand, foot, and mouth disease at the district level in Guangdong Province, China. *PLoS One*. 2013;8(2):e56943. doi:10.1371/journal.pone.0056943

48. Zadnik V, Reich BJ. Analysis of the relationship between socioeconomic factors and stomach cancer incidence in Slovenia. *Neoplasma*. 2006;53(2):103-110.
49. Hodges JS, Reich BJ. Adding spatially-correlated errors can mess up the fixed effect you love. *Am Stat*. 2010;64(4):325-334. doi:10.1198/tast.2010.10052
50. Goicoa T, Adin A, Ugarte MD, Hodges JS. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stoch Env Res Risk A*. 2018;32(3):749-770. doi:10.1007/s00477-017-1405-0
51. Adin A, Goicoa T, Hodges JS, Schnell PM, Ugarte MD. Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India. *Stat Model*. 2021;1471082X2110154. doi:10.1177/1471082X211015452
52. Wang W, Zhang T, Yin F, et al. Using the maximum clustering heterogeneous set-proportion to select the maximum window size for the spatial scan statistic. *Sci Rep*. 2020;10(1):4900. doi:10.1038/s41598-020-61829-y
53. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Stat Med*. 2006;25(22):3929-3943. doi:10.1002/sim.2490
54. Tango T, Takahashi K. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Stat Med*. 2012;31(30):4207-4218. doi:10.1002/sim.5478
55. Lin PS, Kung YH, Clayton M. Spatial scan statistics for detection of multiple clusters with arbitrary shapes. *Biometrics*. 2016;72(4):1226-1234. doi:10.1111/biom.12509

## 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:** Wang W, Xiao X, Qian J, et al. Reclaiming independence in spatial-clustering datasets: A series of data-driven spatial weights matrices. *Statistics in Medicine*. 2022;1-18. doi: 10.1002/sim.9395