Stochastic Environmental Research and Risk Assessment (2019) 33:1815–1833
https://doi.org/10.1007/s00477-019-01728-5

ORIGINAL PAPER

Disease relative risk downscaling model to localize spatial epidemiologic indicators for mapping hand, foot, and mouth disease over China

Chao Song[1,2,3] · Yaqian He[4] · Yanchen Bo[1] · Jinfeng Wang[2,5] · Zhoupeng Ren[2] · Jiangang Guo[2] · Huibin Yang[1]

Published online: 12 September 2019

Abstract

Given the limitations of current approaches for disease relative risk mapping, it is necessary to develop a comprehensive mapping method not only to simultaneously downscale various epidemiologic indicators, but also to be suitable for different disease outcomes. We proposed a three-step progressive statistical method, named disease relative risk downscaling (DRRD) model, to localize different spatial epidemiologic relative risk indicators for disease mapping, and applied it to the real world hand, foot, and mouth disease (HFMD) occurrence data over Mainland China. First, to generate a spatially complete crude risk map for disease binary variable, we employed ordinary and spatial logistic regression models under Bayesian hierarchical modeling framework to estimate county-level HFMD occurrence probabilities. Cross-validation showed that spatial logistic regression (average prediction accuracy: 80.68%) outperformed ordinary logistic regression (69.75%), indicating the effectiveness of incorporating spatial autocorrelation effect in modeling. Second, for the sake of designing a suitable spatial case–control study, we took spatial stratified heterogeneity impact expressed as Chinese seven geographical divisions into consideration. Third, for generating different types of disease relative risk maps, we proposed local-scale formulas for calculating three spatial epidemiologic indicators, i.e., spatial odds ratio, spatial risk ratio, and spatial attributable risk. The immediate achievement of this study is constructing a series of national disease relative risk maps for China's county-level HFMD interventions. The new DRRD model provides a more convenient and easily extended way for assessing local-scale relative risks in spatial and environmental epidemiology, as well as broader risk assessment sciences.

Keywords Disease mapping modeling · Local relative risk assessment · Spatial odds ratio · Spatial risk ratio · Spatial attributable risk · Hand, foot, and mouth disease

Abbreviations

AR	Attributable risk
BHM	Bayesian hierarchical modeling
CAR	Conditional autoregressive
CI	Confidence interval
CISDCP	System for Disease Control and Prevention
CMDC	China Meteorological Data Service Center
DIC	Deviance information criterion

Chao Song and Yaqian He have contributed equally to this article.

✉ Yanchen Bo
boyc@bnu.edu.cn

Chao Song
chaosong.gis@gmail.com

Yaqian He
yaqian.he@dartmouth.edu

Jinfeng Wang
wangjf@lreis.ac.cn

Zhoupeng Ren
renzp@lreis.ac.cn

Jiangang Guo
guojg@lreis.ac.cn

Huibin Yang
yanghb@mail.bnu.edu.cn

Extended author information available on the last page of the article

DRRD	Disease relative risk downscaling
EM	Expectation maximum
GDP	Gross domestic product
GIS	Geographic information science
HFMD	Hand, foot, and mouth disease
INLA	Integrated nested Laplace approximation
kNN	k-nearest neighbors
LS	Logarithmic score
OR	Odds ratio
PA	Prediction accuracy
PST	Progressive spatiotemporal
RF	Random forest
RR	Risk ratio or relative risk
SAR	Spatial attributable risk
SD	Standard deviation
SMR	Standardized mortality ratio
SOR	Spatial odds ratio
SRR	Spatial risk ratio
SSH	Spatial stratified heterogeneity
STVC	Spatiotemporally varying coefficients
SVD	Singular value decomposition
WAIC	Watanabe Akaike information criterion

1 Background

Disease mapping is widely used by spatial epidemiologists, medical geographers, and biostatisticians to highlight areas with elevated or lowered risk, understand geographical patterns and variations of disease, and further obtain disease etiological clues (Waller and Carlin 2010). In traditional epidemiology, ratio and difference are common ways for comparing the relative magnitude of disease risk between two groups in case–control or cohort studies and address the additive and multiplicative impact of exposures (Cummings 2009). The crude risk indicators like incidence rate or mortality, are maybe important in themselves, but the utility of these indices increases multiple-fold when their ratio is obtained relative to a comparison group. Therefore, the crude risk indicators are always converted to relative risk indicators to give more accurate results in epidemiological assessment (Indrayan and Malhotra 2017). The most widely applied relative risk indicators include odds ratio (OR), risk ratio (RR), and attributable risk (AR) (Schechtman 2002), among which OR and RR quantify the strength of the association between exposures and disease outcomes, which is more significant for etiology (Li et al. 2005; Schmidt and Kohlmann 2008), and AR reflects information about reducing the risk of exposures, playing a more critical role in disease prevention (Whittemore 1983). In work presented here, we focus on disease mapping methods especially employing those geospatial-based epidemiologic relative risk indicators to generate the so-called disease relative risk maps.

There have been several statistical ways on mapping disease relative risks (Berke 2005; Bithell 2000; Richardson et al. 2004; Ugarte et al. 2006), although they all adopted the same word “relative risk” to name the produced disease maps, the actual connotation of “relative risk” estimated by these methods were not exactly the same. Herein, we introduce three common ways for disease relative risk mapping for now.

A classical approach in mapping disease mortality or incidence rate data assumes that the number of disease outcomes follows a Poisson distribution, and estimates the relative risk in each spatial unit, which is known as the standardized mortality ratio (SMR) (Goicoa et al. 2018; Ugarte et al. 2006) or standardized incidence ratio (SIR) (Martínez-Bello et al. 2018). Unlike classical epidemiology, spatial SMR is obtained over the entire study area by grouping various geographical regions, instead of different age or gender (Adin et al. 2018; Roquette et al. 2018). Furthermore, as for rare diseases and small areas, a direct SMR map could be very unstable (Mollié 1996). A variety of statistical methods have been developed for producing smoothed estimates of the SMR map (Meza 2003; Ugarte et al. 2006). The smoothed SMR map may be a reliable measure of relative risk, nevertheless, it still has limitations. On the one hand, the spatial SMR estimator is only suitable for specific types, i.e., spatially disease rate or cases variables. However, it could do nothing for spatially disease binary data such as presence/absence or death/alive. On the other hand, SMR is a ratio indicator of which the epidemiological meaning is similar to effect indicator RR (Cummings 2009; Schechtman 2002), but the control group for calculating spatial SMR is the entire study area, which may hide sub-level regional heterogeneity, especially for a large scale geospatial study.

At present, the advances in spatial statistics and geographic information science (GIS) have provided more ways to produce different types of disease maps to offer rich local-scale information for policy-making (Lai et al. 2008; Lawson 2013). The spatial regression model is another category of disease relative risk mapping method, which was widely applied for areal data using Bayesian ideas, especially under the Bayesian hierarchical modeling (BHM) framework (Blangiardo et al. 2013; Rue et al. 2017). BHM is a powerful analytical technique to naturally represent the spatial information provided by neighboring regions as prior knowledge and to give robust posterior estimates of the local parameter in each spatial unit (Ugarte et al. 2014). The exponential form of the estimated local parameter is the so-called relative risk indicator, named as RR or OR according to different disease data distribution. For instance, as for hand, foot, and mouth disease (HFMD) concerned in this study, Zhang et al. (2018) fitted a spatial relative risk (RR) map for HFMD incidence in Henan, China using a Bayesian spatiotemporal hierarchical model. Song et al. (2018a) further concerned the mapping issue of excessive zero to produce the spatial RR map of HFMD incidence in Mainland China with spatiotemporal zero-inflated Bayesian hierarchical models. Moreover, Song et al. firstly generated a series of spatial odds ratio (OR) maps for HFMD occurrence and HFMD-climate associations in Sichuan, China by proposing a new Bayesian local regression method named Spatiotemporally Varying Coefficients (STVC) model (Song et al. 2019). Unfortunately, there are still several defects of implementing BHM-based spatial regression models for disease relative risk mapping. On the one hand, the estimated spatial local parameters are belonging to only one part of the regression process, which are not enough to represent the total disease risk effects that also include fixed effects such as covariates and spatial intercepts, thus the spatial-model-based smoothed RR/OR maps are not real maps representing disease total relative risks. On the other hand, the exponential calculation form is not based on the original epidemiologic theory, thus Bayesian spatial regression disease mapping method could only obtain two types of spatial indicators, i.e., RR and OR, not for the others, such as AR.

Considering the above issues, it is more reasonable to calculate local-scale spatial relative risks by directly drawing lessons from the original epidemiologic theories and formulas. The difficulty arising from this idea is how to define a suitable spatial case–control study framework. In particular, case–control studies start by identifying cases from those who have had the disease of interest as the case or exposed group, and identifying controls from those who have not as the control or unexposed group. Today, scan statistics and density estimation methods are applied to calculate the local-scale relative risk indicator (herein is risk ratio, RR), under the case–control studies. For instance, Bithell defined a relative risk function for point data (Bithell 2000), giving the risk of being affected by the disease incurred at a location (case group), which is typically estimated by kernel method (Bowman and Azzalini 1997), relative to the average risk in the region as a whole (control group) (Bithell 1990). Berke separated the population at risk into exposed and unexposed (or lesser exposed) groups by using spatial scan statistic to estimate RR for disease mapping for both spatial point and regional count data (Berke 2005). However, no matter for scan statistics or density estimation, there is no standard way to select the scan radius or search radius, which profoundly affects the case and control groups, as well as significantly leads to mapping uncertainties (Li et al. 2019b). Thus, to decrease uncertainties, a better way to select cases as the spatial control group is according to the real original observed data, not by estimation approaches. Furthermore, these estimation methods lack considering the spatial stratified heterogeneity (SSH) impact (Wang et al. 2016) to define various spatial control groups. Generally, at small extents and fine-scale spatial resolutions, as well as at large extents and coarse resolutions, spatial patterns may appear to be homogeneous, whereas, at intermediate spatial extents and resolutions (e.g., administrative division, landscape, land use types, and climate zones), spatial heterogeneity emerges (Fortin et al. 2012). This intermediate difference is defined as the SSH impact (Wang et al. 2016). Ignoring SSH impact within spatial control groups may hide the true disease risk distribution, and overlook outliers for disease control and prevention at the local scale, especially for large geospatial studies (Huang et al. 2014; Xu et al. 2019).

Under these circumstances, it is crucial to develop a general disease mapping approach not only for obtaining various epidemiologic relative risk indicators at the geospatial scale, e.g., spatial OR, spatial RR, and spatial AR, but also for different spatial disease outcomes, e.g., rate, binary, or cases variables.

To fill this gap in current disease mapping studies, we proposed a three-step statistical method in this work, named disease relative risk downscaling (DRRD) model, aiming to localize various spatial epidemiologic relative risk indicators simultaneously, as well as for different kinds of disease outcome data, such as binary, rate, or cases variables, jointly with the consideration of the SSH impact and covariate factors. In this study, we illustrated our newly proposed disease mapping method by using a real-world example, i.e., China's hand, foot, and mouth disease (HFMD) occurrence data to demonstrate its effectiveness. For one reason is that so far, there are no real relative risk maps of HFMD that have ever been published over entire Mainland China at the county level (Bo et al. 2014; Song et al. 2018a). The other reason is that disease occurrence data, which is a two-value binary variable, is a more difficult case for downscaling relative risks compared with rate or cases data, which is also unsolved well by the three disease mapping methods aforementioned.

Pediatric hand, foot, and mouth disease is an emerging worldwide infectious disease occurring mainly in children under 5 years old and can lead to death (Koh et al. 2016). In China, the leading infectious HFMD has posed a severe threat to public health security since 2008 (Wang et al. 2011; Xing et al. 2014). We collected the county-level HFMD occurrence data, in conjunction with various climate and socioeconomic factors for our experiments, aiming at generating a series of China's national HFMD risk maps by using three DRRD-based spatial epidemiologic relative risk indicators, i.e., spatial OR (SOR), spatial RR (SRR), and spatial AR (SAR).

2 Methods

2.1 Data and preprocessing

For the study area of whole Mainland China, we acquired geospatial county-level HFMD occurrence data (1 and 0), and its related climate and socioeconomic variables in April 2009. April was the most serious month with the maximum number of disease cases in the year of 2009. The HFMD occurrence data in children aged between 0 and 9 years were provided by the China Information System for Disease Control and Prevention (CISDCP) (Huang et al. 2014). The monthly climate data was based on the raw data collected from 727 climate stations throughout Mainland China from the China Meteorological Data Service Center (CMDC) (http://data.cma.cn/en) (Bo et al. 2014). Data of yearly socioeconomic variables was integrated from the China County Statistical Yearbook, China Statistical Yearbook for Regional Economy, and China City Statistical Yearbook (Song et al. 2018b). We collected six climate and fourteen socioeconomic variables as candidate environmental-related covariates for China's HFMD occurrence (Song et al. 2018a). We performed z-score standardization for the twenty variables to make them dimensionless. Then we screened the most influencing covariates into modeling from the twenty environmental-related variables through multicollinearity assessment (Vatcheva et al. 2016) and forward stepwise regression (Wilkinson 1979). Expressly, we first calculated the variance inflation factor to exclude those variables with higher multicollinearity by setting 10 as the threshold value, and then we employed the forward stepwise regression to exclude those variables that were not statistically significant, in which we set 0.05 and 0.1 as the alpha cut (Bo et al. 2014; Song et al. 2018a). Hence, a total of six factors were selected as our final covariates for modeling in this study, as shown in Fig. 1.

Figure 1 illustrates the original spatial maps of HFMD occurrence condition and six environmental covariates, including ambient temperature, air pressure, population density, per capita household savings, per capita social consumption, and per capita industrial output values, across Mainland China at the county level in April 2009. In addition, according to the published national standards of Chinese geographical division, Mainland China is generally divided into seven geographical divisions, as shown in Fig. 1a, which includes North, East, South, Central, Northeast, Southwest, and Northwest China. The Chinese seven geographical divisions were applied as a second spatial level to represent the spatial stratified heterogeneity impact in this work.

2.2 Disease relative risk downscaling (DRRD) model

First of all, we introduce basic applicable conditions of the DRRD model, involving unit for basic data collection, geographical study area, and disease outcome variables, as discussed below.

Spatial data collection unit: In DRRD modeling framework, as well as in spatial epidemiology, the basic disease data unit is usually collected within a geospatial areal unit representing the overall population risk in a space area, such as a county, city, or regular grid. This is unlike in traditional epidemiology, in which a basic collection unit is an individual person within a specific population group (Lawson et al. 2016).

SSH-based geographical study area: Besides the basic spatial unit as the minimum spatial scale, the geographical study area should be able to be divided into several regions at an upper coarse level, such as provinces, ecological zones, or climate divisions, in order to consider the spatial stratified heterogeneity (SSH) impact to design more reasonable various spatial control groups under a spatial case–control study (Wang et al. 2016).

Disease outcome variable types: Generally, DRRD model is capable of taking three types of disease outcomes into consideration, i.e., rate variable (e.g., incidence rate, mortality, or morbidity), binary variable (e.g., yes/no, presence/absence, or death/alive), and cases variable. If it exists missing values in the original disease data, estimation methods (Song et al. 2018b) are requested to generate a complete disease crude risk map. We fully discuss how missing data issue is related to our DRRD model, and the choices of methods on estimating missing data in Sect. 4.

Furthermore, among three types of disease outcome variables, the disease binary variable is the most difficult case for downscaling spatial relative risk indicators, due to the fact that no local variations exist in the original two-value data. Using binary data from a real-world example, i.e., China's HFMD occurrence case, our method successfully address this issue. Figure 2 illustrates the overall three-step framework of the DRRD model, and how it worked for China's HFMD occurrence case. In general, the DRRD model contains three consecutive core steps, and within each step, clear tasks are required: (1) In the first step, estimating a complete disease crude risk map without missing data areas is a must. In this case, since the original HFMD occurrence data was a two-value binary variable, we employed the logistic regression method combined with various climate and socioeconomic covariates, to fit the local disease occurrence probability as a crude risk in each geographic unit, also to fill areas with missing values. (2) In the second step, as discussed above, defining spatial

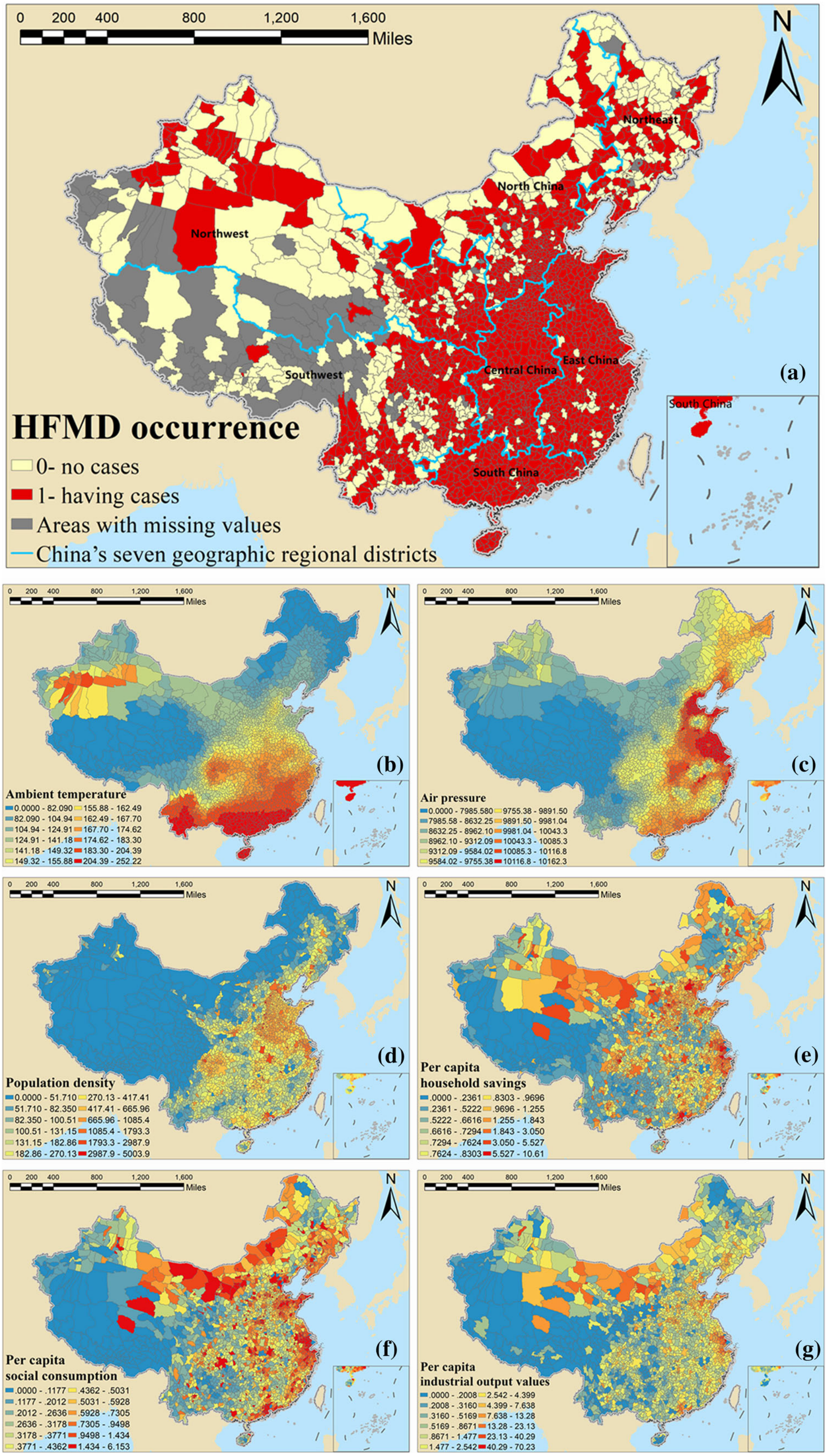

Fig. 1 **a** County-level hand, foot, and mouth disease (HFMD) occurrence data over Mainland China in April 2009, and its related climate and socioeconomic covariates: **b** ambient temperature, **c** air pressure, **d** population density, **e** per capita household savings, **f** per capita social consumption, and **g** per capita industrial output values

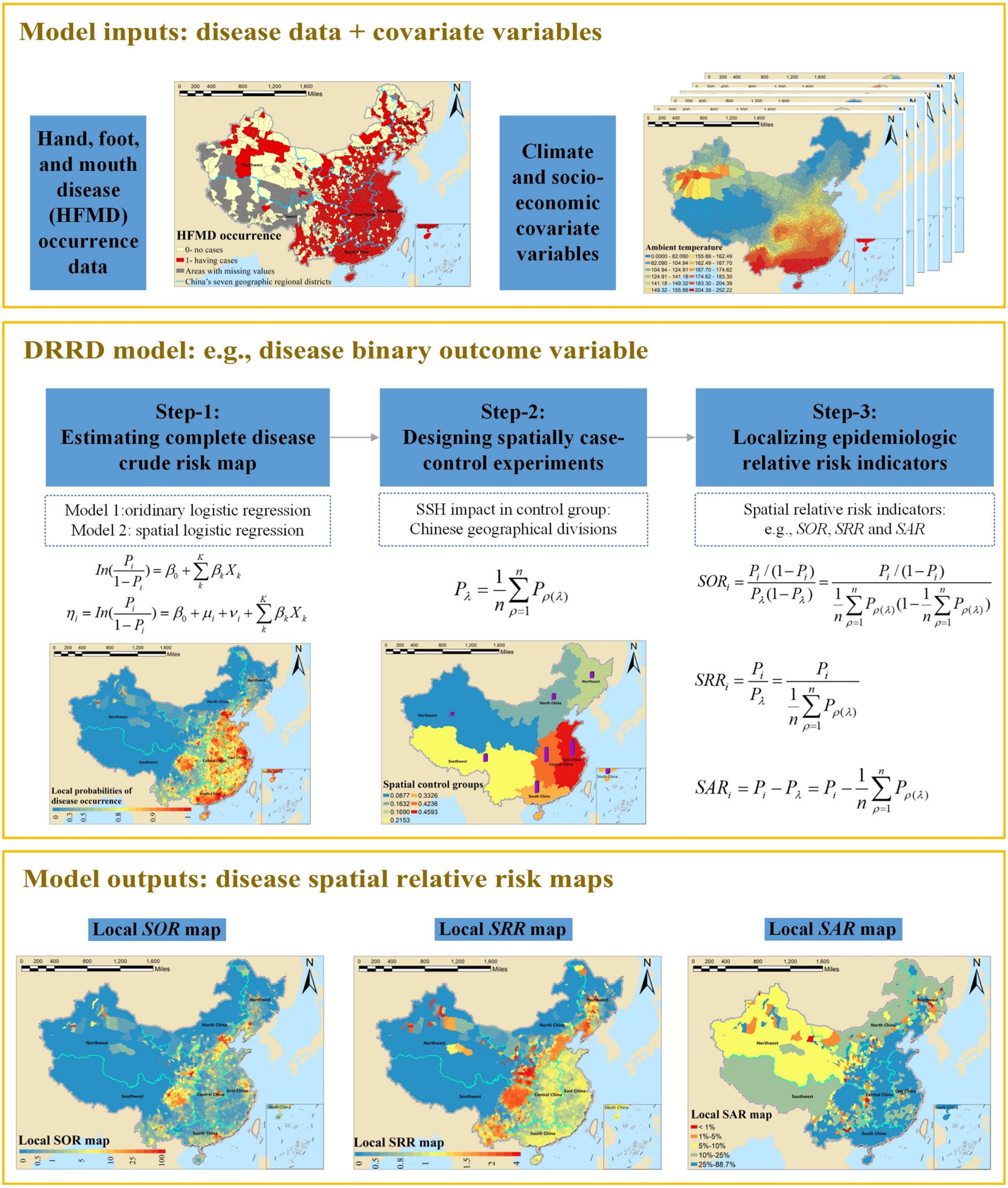
Model inputs: disease data + covariate variables
Hand, foot, and mouth disease (HFMD) occurrence data
HFMD occurrence
0- no cases
1- having cases
Areas with missing values
China's seven geographic regional districts
Climate and socio-economic covariate variables
Ambient temperature
0.0000 - 82.090
82.090 - 104.94
104.94 - 124.91
124.91 - 141.18
141.18 - 149.32
149.32 - 155.88
155.88 - 162.49
162.49 - 167.70
167.70 - 174.62
174.62 - 183.30
183.30 - 204.39
204.39 - 252.22
DRRD model: e.g., disease binary outcome variable
Step-1: Estimating complete disease crude risk map
Step-2: Designing spatially case-control experiments
Step-3: Localizing epidemiologic relative risk indicators
Model 1:oridinary logistic regression
Model 2: spatial logistic regression
SSH impact in control group: Chinese geographical divisions
Spatial relative risk indicators: e.g., SOR, SRR and SAR
$In(\frac{P_i}{1-P_i}) = \beta_0 + \sum_k^K \beta_k X_k$
$\eta_i = In(\frac{P_i}{1-P_i}) = \beta_0 + \mu_i + \nu_i + \sum_k^K \beta_k X_k$
$P_\lambda = \frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}$
$SOR_i = \frac{P_i/(1-P_i)}{P_\lambda(1-P_\lambda)} = \frac{P_i/(1-P_i)}{\frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}(1-\frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)})}$
$SRR_i = \frac{P_i}{P_\lambda} = \frac{P_i}{\frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}}$
$SAR_i = P_i - P_\lambda = P_i - \frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}$
Local probabilities of disease occurrence
0 0.3 0.5 0.8 0.9 1
Spatial control groups
0.0877
0.1632
0.1690
0.2153
0.3326
0.4236
0.4593
Model outputs: disease spatial relative risk maps
Local SOR map
0 0.5 1 5 10 25 100
Local SRR map
0 0.5 0.8 1 1.5 2 4
Local SAR map
< 1%
1%-5%
5%-10%
10%-25%
25%-88.7%
Northwest
North China
Northeast
Southwest
Central China
East China
South China
0 200 400 800 1,200 1,600 Miles
N

Fig. 2 Three steps of disease relative risk downscaling (DRRD) modeling flow-chart for disease binary outcome variable using China's county-level hand, foot, and mouth disease (HFMD) occurrence data as a case study. *SOR* spatial odds ratio, *SRR* spatial risk ratio, *SAR* spatial attributable risk, and *SSH* spatial stratified heterogeneity

control groups accounting for the SSH impact is recommended. For the study area of China, we chose Chinese seven geographical divisions as the second upper spatial level to represent the SSH impact and to further obtain their baseline risk values based on the first-step outputs. (3) In the third step, based on the former two steps' outputs, calculating local values of various spatial relative risk indicators in each geographic unit according to the newly developed local-scale calculation formulas. In this work, we were successful in downscaling three spatial epidemiologic relative risk indicators, i.e., spatial OR (SOR), spatial RR (SRR), and spatial AR (SAR), for China's national HFMD county-level risk mapping. The statistical modeling, formulas derivation, and other detailed settings of the three-step DRRD model for China's HFMD case are introduced one by one in the next three subsections.

2.2.1 Estimating complete disease crude risk map

The main task with the first step modeling is twofold. Firstly, to obtain/estimate area-specific disease crude risks at the local scale in order to collect different risk levels in both geographical case and control groups, so as to further retrieve spatial relative risks. Secondly, the disease crude risk map should be complete without any missing area, which is essential to make sure that the final outputs, i.e., disease relative risk maps, are also complete. Under these conditions, the ecological regression modeling method accounting for various covariate variables could be a priority approach to deal with both of the two essential requirements aforementioned.

For HFMD occurrence data of this case, it is a more difficult situation compared with other types of disease data, due to the fact that the original disease data is a two-value binary variable without local spatial variations. It is reasonable to assume that the risks in different spatial counties are in different-level occurrence risk, even they have the same value (1 or 0). The logistic regression model is the primary method to estimate such local-scale risk variations, i.e., occurrence probabilities, beyond two-value binary variables (Peng et al. 2002). Herein, disease occurrence probability is considered as a surrogate indicator of crude disease rate. We employed two alternative logistic regression models using both climate and socioeconomic covariates as influencing variables to fit the local-scale disease probability risk in each spatial unit as follows.

Firstly, the form of an ordinary logistic regression model (herein referred to as model 1) is given as follows (Harrell 2015):

$$In\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \sum_{k}^{K} \beta_k X_k, \quad P_i(y_i = 1|\beta_0, \beta_k) \tag{1}$$

where the basic spatial area is expressed as i, the estimated probability of HFMD occurrence is expressed as P_i, the potential explanatory variables are expressed as X_k, and the fitting regression coefficients are expressed as β_0 and β_k. The epidemiologic OR indicator is able to be directly calculated in logistic regression by $OR = e^{\beta}$. Specifically, $OR > 1$ indicates increased risk between the exposure variable and disease, $OR < 1$ indicates decreased risk between the exposure variable and disease, and $OR = 1$ indicates unrelated risk between the exposure variable and disease (Bland and Altman 2000; Szumilas 2010).

Secondly, a spatial logistic regression model is employed as an alternative approach (herein referred to as model 2) to estimate complete disease crude risk map with the consideration of spatial autocorrelation, which is frequently encountered in spatial disease data, and the neglect of such spatial property could result in a biased and underperforming model in health risk assessment. In this study, the predictor in a spatial logistic regression model is decomposed additively into components regarding both covariates fixed effects and spatial random effects (Yang et al. 2019a), as follows:

$$In\left(\frac{P_i}{1-P}\right) = \beta_0 + \mu_i + v_i + \sum_{k}^{K} \beta_k X_k, \quad P_i(y_i = 1|\beta_0, \beta_k, \mu_i, v_i) \tag{2}$$

where the estimated probability of HFMD occurrence in the spatial area i is expressed as P_i, the fixed effects' coefficients are expressed as β_0 and β_k, and the spatial random effects' coefficients are expressed as μ_i and v_i. Regarding spatial terms, μ_i is the unstructured spatial component with a Gaussian prior assumption $v_i \sim N(0, \delta_v^2)$, and v_i is the structured spatial component with the conditional autoregressive (CAR) prior assumption, formulated as (Lee 2011):

$$\mu_i|\mu_{j\neq i} \sim N\left(\frac{1}{m_i}\sum_{i\sim j}\mu_i, \frac{\sigma^2}{m_i}\right) \tag{3}$$

where two spatial units with adjacency relations are expressed as $i \sim j$, the count of spatial units j adjoining the spatial unit i is expressed as m_i and σ^2 is the variance term. This structured prior assumption is widely achieved by using the Besag model (Besag 1974).

Taking advantage of spatial regression modeling, the smoothed relative risk indicator could be obtained by using the exponential form of spatially structured coefficients μ_i, i.e., $OR_i = \exp(\mu_i)$. As in logistic regression, the relative risk is called as odds ratio (OR), which is especially for disease binary response variable. Many researchers have utilized such model-smoothed epidemiologic indicator for HFMD risk mapping (Song et al. 2018a, 2019; Zhang et al.

Table 1 Summary table of research data in a traditional epidemiologic cohort or case–control study

	Yes (1)	No (0)	Total	Rate
Exposed group	a	b	$a + b$	$P1 = a/(a + b)$
Unexposed group	c	d	$c + d$	$P0 = c/(c + d)$
Total	$a + c$	$b + d$	$a + b + c + d$	Ratio: *RR*, *AR*, *OR*

2018), whereas its main limitation is that the spatially structured coefficients μ_i are only in one part of the regression process, which is not enough to represent the total disease risk impacts that also include other components such as fixed effects of covariates or random effects of spatial heterogeneous intercepts.

2.2.2 Designing spatially case–control experiments

For the second step, the primary goal here is to define a spatially epidemiologic case–control experiment, of which the core task is how to specify a suitable spatial control group. In each geographical unit, we usually treat the first-step local-scale disease crude risk as the area-specific spatial case group (Richardson et al. 2004; Ugarte et al. 2006). With regards to design a more reasonable spatial control group, herein, we suggest accounting for the vital SSH impact by introducing an additional coarse-scale geographical level to obtain various spatial control groups (Wang et al. 2016), which is often ignored by previous disease mapping studies (Adin et al. 2018; Roquette et al. 2018), so as to replace the classical one and overall control group of the entire study area.

The original HFMD occurrence data were collected from the historical disease reported system (CISDCP), thus this work was under a spatially case–control experiment design. Follow the above guidelines, for HFMD spatial case groups, we specified the first-step fitted disease occurrence probability risk P_i as spatial case groups. Herein, disease occurrence probability P_i is considered as a surrogate indicator of crude rate for disease binary variable. Then, we specified Chinese geographical division as the SSH-based spatial control groups for the study area of Mainland China. For each regional division λ as a spatial control group, we selected n samples satisfying a double-screening requirement, i.e., for each spatial sample ρ, there should be no disease occurrence in the originally historic reported data, and it should also be predicted correctly as a non-occurrence area with the first-step logistic regression model. At last, the reference risk P_λ in each spatial control group is formulated as below:

$$P_\lambda = \frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)} \tag{4}$$

Various spatial control groups within different districts, rather than a single spatial control group of the entire study area in traditional disease mapping studies, is more reasonable and suitable, especially for large scale geospatial data.

2.2.3 Localizing relative risk indicators

For the last step, the main task herein is to calculate local-scale spatial epidemiologic relative risk indicators in each geographical unit for advanced disease mapping, under a spatially case–control study. The difficulty lies in how to define the new downscaling local-scale formulas that must base on the classical epidemiological theories and formulas.

In this case, we choose three indicators as references for HFMD mapping, i.e., risk ratio (*RR*), attributable risk (*AR*), and odds ratio (*OR*) that are widely used to measure the risk of a disease exposed to exposures in traditional epidemiology. Traditional global scale relative risk indices *RR*, *AR*, and *OR* are calculated based on research data in Table 1 (Schechtman 2002).

In a classical epidemiologic cohort study, we could calculate *RR* and *AR* directly using Eqs. (5) and (6) (Li et al. 2005; Whittemore 1983). While, in a classical epidemiologic case–control study, the real occurring rate is unavailable as there are no observed numbers in both exposure and control groups, thus, we calculate *OR* instead of *RR* to quantitatively characterize relative risk levels, by using Eq. (7) (Bland and Altman 2000; Cummings 2009).

Epidemiologic *RR* (known as risk ratio, relative risk):

$$RR = \frac{P_1}{P_0} = \frac{a(c+d)}{c(a+b)} \tag{5}$$

Epidemiologic *AR* (known as attributable risk, rate difference):

$$AR = P_1 - P_0 = \frac{a}{(a+b)} - \frac{c}{(c+d)} \tag{6}$$

Epidemiologic *OR* (known as odds ratio, occurring ratio):

$$OR = \left(\frac{a}{a+c}\Big/\frac{c}{a+c}\right)\Big/\left(\frac{b}{b+d}\Big/\frac{d}{b+d}\right) = \frac{ad}{bc} \tag{7}$$

In traditional epidemiology, the basic study unit is an individual within a population group, thus, these classical population-based relative risk indicators are calculated as a single value for a population group, which are the so-called global scale indicator that cannot be downscaled for mapping directly.

However, in spatial epidemiology, as well as in our DRRD modeling framework, the basic study unit is a geospatial unit representing the overall population risk in a specific spatial area, and more importantly, for a same geographical study region, the spatial samples within a region are constant and stationary. For instance, in our case of China as the study area, 2310 spatial counties at most were collected for designing a spatial case–control study in the year 2009 (Song et al. 2018a). Thus, with this underlying difference from a traditional population-based case–control study, we are able to collect disease occurring rate P in each geospatial unit. Especially for OR, from the perspective of spatial epidemiology, we can build its connection with disease occurring rate P, as shown in Eq. (8).

$$OR = \frac{ad}{bc} = \frac{a}{b}\Big/\frac{c}{d} = \frac{P_1/(1-P_1)}{P_0/(1-P_0)} \tag{8}$$

where P_1 is the disease crude risk in the spatial case group, i.e., disease occurrence probability, and P_0 is the disease crude risk in the corresponding spatial control group. The calculation of spatial RR and spatial AR is the same as traditional ones.

Based on formulas (4), (5), (6), and (8), and further by taking advantage of the SSH-based spatially case–control experimental design, we infer three new kinds of local-scale formulas for each given geographical unit i, to calculate spatially epidemiologic relative risk indicators, namely, spatial OR, spatial RR, and spatial AR, as shown in Eqs. (9), (10), and (11), respectively.

Spatial odds ratio (SOR):

$$SOR_i = \frac{P_i/(1-P_i)}{P_\lambda(1-P_\lambda)} = \frac{P_i/(1-P_i)}{\frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}\left(1-\frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}\right)} \tag{9}$$

Spatial risk ratio (SRR):

$$SRR_i = \frac{P_i}{P_\lambda} = \frac{P_i}{\frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)}} \tag{10}$$

Spatial attributable risk (SAR):

$$SAR_i = P_i - P_\lambda = P_i - \frac{1}{n}\sum_{\rho=1}^{n} P_{\rho(\lambda)} \tag{11}$$

where local SOR_i, SRR_i, and SAR_i are the downscaled spatial relative risk indicators, unit i is the target geographic unit, unit ρ is the sample in the spatial control group λ, term P_i is the disease occurrence probability in geographical unit i, and term P_λ is the average probability of geographical units in the spatial control group of the corresponding geographical division.

For local SOR indicator, a SOR_i value greater than 1 indicates that the geographical unit i is a risky area, and a SOR_i value less than 1 indicates that geographic unit i is a relatively safe area. The larger the value of SOR_i, the riskier of the area. For instance, it is common to describe a spatial area with a SOR value of 2 in terms of a twice risk of disease occurrence compared with that with a SOR value of 1. The local SRR has the same explanation of local SOR map, but with different overall range and geographically local variations. Local SOR and SRR maps may detect different additional risk distribution for the same disease crude risk map.

For local SAR indicator, SAR_i represents the sensitivity of the selected covariates in each geographic unit i. The higher the SAR_i value, the more contribution that those selected covariates have made in this area, which indicates that those selected covariates are useful to fit the additional disease risk in this area. A SAR_i value of zero indicates that those selected covariates are not reprehensive enough to fit the local disease risk in unit i.

2.3 Model inference and evaluation

The logistic regression model 1 and model 2 were built under the Bayesian hierarchical modeling (BHM) framework in R software (Bakka et al. 2018). The integrated nested Laplace approximation (INLA) (Rue et al. 2009) was adopted as the Bayesian inference method to estimate the posterior disease occurrence probabilities in this study (Schrödle and Held 2011), due to its advantage of relatively short computation time (Rue et al. 2017). We assessed model 1 and model 2 under a cross-validation design by randomly removing 10%, 20%, and 30% samples. For each model, we calculated various indices to qualify model performance, which is summarized as follows (Song et al. 2019). First, the deviance information criterion (DIC) (Spiegelhalter et al. 2002) and Watanabe Akaike information criterion (WAIC) (Watanabe 2010) are two widely utilized indices to describe Bayesian model fitness, which both are the smaller, the better. Second, a logarithmic score (LS) extracted from leave-one-out cross-validation is employed to depict Bayesian model predictive ability (Held et al. 2010), which is also the smaller, the better. Last, for the logistic regression, the confusion matrix is earned to measure the actual prediction accuracy (PA), including PA(1) for disease-presence counties, PA(0) for disease-absence counties, and PA(1,0) for all counties within the study area (Yang et al. 2019b).

3 Results

3.1 Model evaluation and disease crude risk map

Table 2 summaries the model performance of both ordinary (model 1) and spatial (model 2) logistic regression for mapping HFMD occurrence probabilities over China. Through six selection criteria statistics listed in Table 2, model 2 surpassed model 1 in all three cross-validation experiments, indicating the effectiveness of concerning spatial autocorrelation impact in estimating disease occurrence probabilities. Notably, regarding actual prediction accuracy for all counties in China, model 2 (80.68%) was able to improve 10.93% on average compared with model 1 (69.75%). Therefore, we chose the optimal spatial logistic regression model for the first-step modeling in DRRD to generate the complete HFMD occurrence probability risk map in China.

An immediate outcome of spatial logistic regression (model 2) is a summary of posterior estimated regression parameters for those selected environmental covariates, including mean, standard deviation (SD), 2.5% and 97.5% confidence intervals (CI), and their overall OR values, as shown in Table 3. We found that both climate and socioeconomic variables had positive influences on increasing HFMD occurrence risk over China in April 2009. HFMD occurrence risk increased with increasing ambient temperature (OR = 4.95), air pressure (OR = 1.29), population density (OR = 2.23), per capita household savings (OR = 1.21), per capita social consumption (OR = 1.47), and per capita industrial output values (OR = 1.15). Among the six environmental covariates, ambient temperature is the most important explanatory variable with OR value close to five, followed by population density with OR value greater than two, while OR values of all the other four variables are less than two.

The final output of model 2 is a new estimated crude risk map of HFMD revealing the local variations of disease occurrence probabilities across Mainland China, as shown in Fig. 3a. More importantly, model 2 also estimated local disease occurrence risk for those missing data areas that were presented in Fig. 1, which was quite essential to generate complete disease risk maps. We further obtained the clustered hot spot map based on Fig. 3a to detect which regions were with significant clusters of the high-risk hot spot and low-risk cold spot, as further shown in Fig. 3b. Note that “not significant” regions do not necessarily indicate absence or presence of risk, just that the risks in these regions were not significant enough to form a cluster.

The probability risk map of Fig. 3a shows prominent spatially clustered characteristics, which suggested that the spatial autocorrelation component was reasonable and necessary to be accounted for disease mapping modeling. The estimated probability of HFMD occurrence at the county level reveals not only the overall spatial trends but also the local details of epidemic risk, even in those regions where there were no HFMD case records. For instance, from Fig. 1, we found that East China was hit hardest by HFMD generally, but we further found that there are more local variations of the HFMD occurrence risk in East China in Fig. 3a. Such local HFMD risk variations can also be found in the other six divisions. Further in Fig. 3b, we detected that high-risk hot spots were mostly concentrated in three divisions, which are Central, East, and South China, and primarily distributed in provinces including Beijing, Tianjin, Shandong, Henan, Guangdong, and Sichuan, among which officials need to pay more attention in practice. Correspondingly, low-risk cold spots were all distributed in the other four divisions, i.e., North, Northeast, Norwest, and Southwest China.

As a complete comparison, in Fig. 3c, d, we further generate the model-smoothed relative risk (OR) map, which represents the spatial structured autocorrelation random effects estimated by the selected spatial logistic regression (model 2), and its hot spot map, respectively. Overall, the model-smoothed OR map of Fig. 3c had a similar but much more smoothed geographical pattern compared with the final estimated disease occurrence

Table 2 Summary of model performance for estimating China’s HFMD occurrence probabilities (model 1: ordinary logistic regression; model 2: spatial logistic regression; *DIC* deviance information criterion, *LS* logarithmic score, *WAIC* Watanabe Akaike information criterion, *PA* prediction accuracy)

Cross-validation	Model	DIC	LS	WAIC	PA(1)	PA(0)	PA (1, 0)
10% test dataset	Model 1	2743.51	0.6928	2743.51	0.6519	0.8065	0.6955
	Model 2	1623.43	0.4100	1623.79	0.8354	0.7903	0.8227
20% test dataset	Model 1	2438.64	0.6928	2438.63	0.6512	0.7292	0.6682
	Model 2	1420.58	0.4037	1420.99	0.8285	0.6458	0.7886
30% test dataset	Model 1	2134.18	0.6929	2134.18	0.7173	0.7687	0.7288
	Model 2	1298.96	0.4219	1299.30	0.8187	0.7755	0.8091

Table 3 Odds ratio (OR) values and estimated posterior regression parameters of the climate and socioeconomic factors on HFMD occurrence

Covariates	OR	Mean	SD	0.025 CI	0.975 CI
Ambient temperature	4.95	1.5988	0.1678	1.2728	1.9311
Air pressure	1.29	0.2564	0.0868	0.0861	0.4267
Population density	2.33	0.8446	0.0940	0.6620	1.0309
Per capita household savings	1.21	0.1872	0.1061	− 0.0213	0.3952
Per capita social consumption	1.47	0.3829	0.1114	0.1645	0.6016
Per capita industrial output values	1.15	0.1386	0.1006	− 0.0585	0.3362

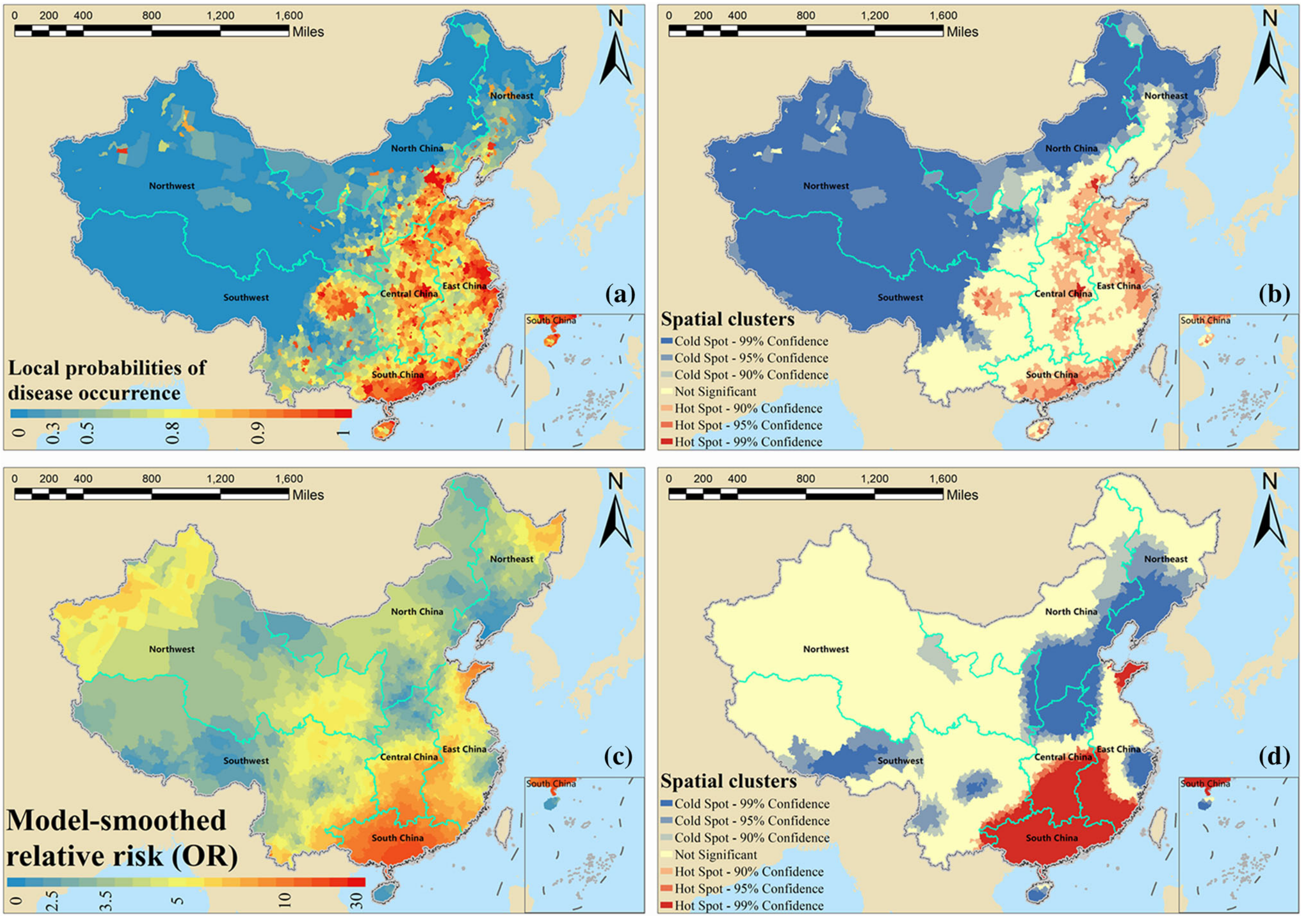
0 200 400 800 1,200 1,600 Miles
N
Northeast
North China
Northwest
Southwest
Central China
East China
South China
(a)
Local probabilities of disease occurrence
0 0.3 0.5 0.8 0.9 1
(b)
Spatial clusters
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
(c)
Model-smoothed relative risk (OR)
0 2.5 3.5 5 10 30
(d)

Fig. 3 **a** Disease occurrence probability (*P*) map and **b** its clustered hotspot map, **c** model- smoothed relative risk (OR) map and **d** its clustered hotspot map, for county-level hand, foot, and mouth disease (HFMD) occurrence across Mainland China in April, 2009

probability map in Fig. 3a. Besides the over-smoothing phenomenon, the regional risk levels were different in some divisions, e.g., North, Northwest, and Northeast districts were detected with a higher risk in Fig. 3c. This is due to the fact that the model-smoothed OR map is produced using only one part of a conventional spatial regression process, i.e., the spatially structured CAR component. Even though such spatial-model-based smoothed relative risk map has been widely applied, without considering some other important components, e.g., fixed effects of observed covariates, this type of relative risk map cannot be used to represent the real and total risk of disease outcome.

3.2 Spatial control groups mapping

With the first-step disease probability risk map and the new-designed spatially case–control experiment, we further obtained the SSH-based spatial control group map of HFMD in China, as shown in Fig. 4. We found that East and South China were with higher average disease occurrence risk, followed by Central China and Southwest

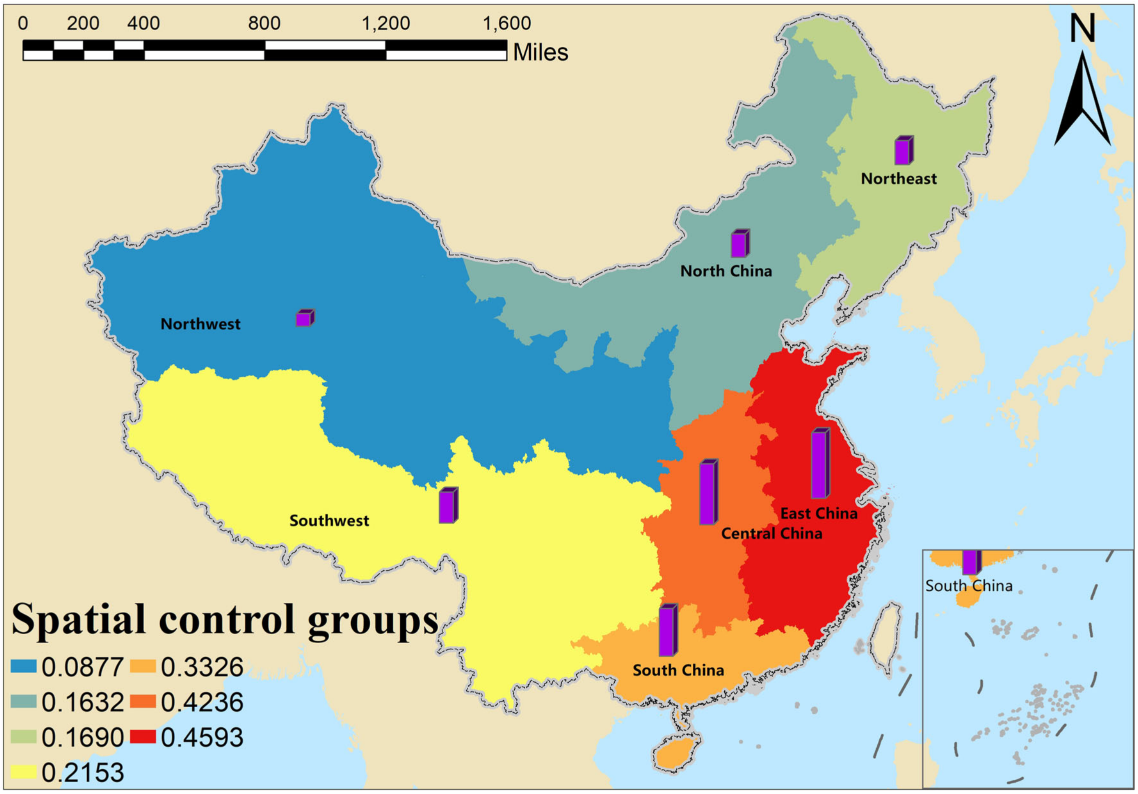

Fig. 4 Spatial control group map of HFMD considering the spatial stratified heterogeneity (SSH) impact over Mainland China

China, while North, Northeast, and Northwest China were at relatively lower risk levels. Utilizing the entire study area as the control group with a single value may underestimate high-risk areas in the lower-risk zone, e.g., Northwest China division, and hide super-high-risk areas in the higher risk zone, e.g., East China division. For instance, when in comparison with nearby counties within the same division, some counties might not be indeed at a low-risk level in Fig. 3a, especially in those relatively lower risk divisions in Fig. 4, so were in those higher risk divisions in Fig. 4, counties with super higher risk need to be further detected in Fig. 3a. Thus, it is considerable to utilize various divisions by taking consideration of the SSH impact, other than a single group of the whole study area as the spatial control group, to further improve mapping accuracy and detect hidden risk areas, especially for a sizeable geospatial study area such as China.

3.3 Disease relative risk maps

At the last step of the DRRD modeling framework, we obtained the three types of relative risk maps for HFMD across Mainland China with new downscaled indicators SOR, SRR, and SAR, as shown in Figs. 5a, c and 6, respectively. We also obtained corresponding hot spot maps of local SOR and local SRR maps to show significant spatial clusters, which were illustrated in Fig. 5b, d, respectively. The local SOR, SRR, and SAR relative risk maps revealed not only spatially local risks but also strong spatially clusters. More importantly, compared with the crude probability risk map of Fig. 3, these new relative risk maps were capable of detecting new risk clusters and outliers, which may offer new insights to develop policies for China’s HFMD control and prevention.

Regarding the SOR map and its hot pot map in Fig. 5a, b, we further detected new high-risk clustering regions compared with Fig. 3, mainly distributing around the so-called Hu Line (Heihe-Tengchong Line). Hu Line is a widely accepted geographic line to generally divide China into two big zones with remarkable differences in population density, natural environment, and socioeconomic conditions. However, we were unable to precisely detect these new HFMD high-risk regions around the Hu Line by merely using the crude risk map of Fig. 3. Moreover, with Fig. 5a, we can identify considerably super-high-risk counties in those higher risk divisions including East, Central, and South China, for instance, Shanghai, Hubei, and Guangdong provinces. However, these super-high-risk counties cannot be clearly identified in Fig. 3, due to the probability map did not consider SSH impact, thus hiding local variations in high-risk divisions.

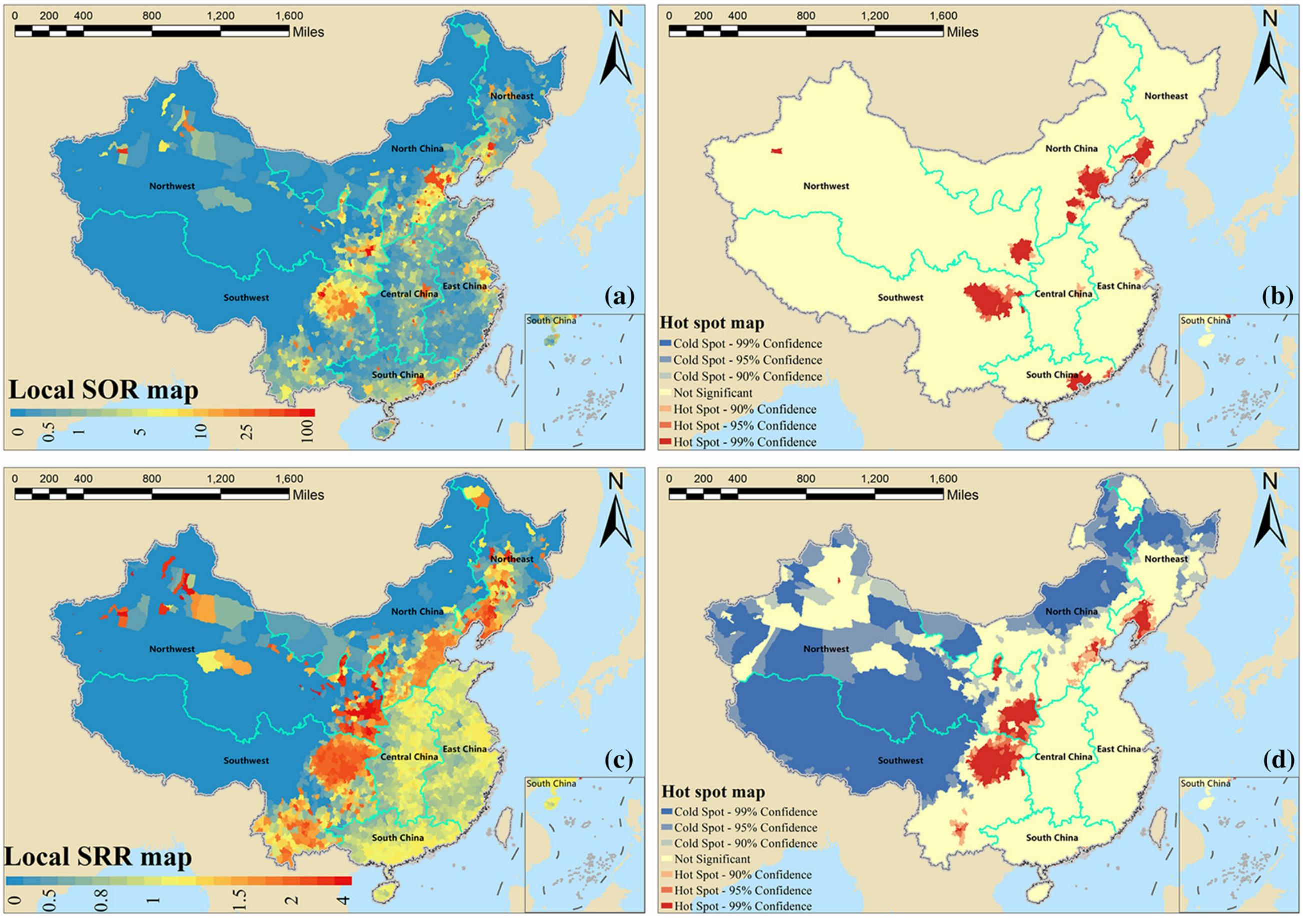

Fig. 5 Spatial odds ratio (SOR) and spatial risk ratio (SRR) maps of county-level HFMD occurrence across Mainland China: **a** local SOR map, **b** SOR hot spot map, **c** local SRR map, and **d** SRR hot spot map

The SRR map and its hot pot map of Fig. 5c, d had a similar spatial explanation of SOR-related maps. We also detected new high-risk clusters concentrated around Hu Line, but with a broader and more significant range in the SRR map. More importantly, the core difference between two relative risk maps was that the SOR map was useful to detect super-high-risk areas especially in those divisions with higher average risk. Whereas, unlike SOR map, SRR map was useful to detect high-risk areas in those divisions with lower average risk, such as Northwest and Northeast China, which couldn't be detected clearly by utilizing either the probability risk map of Fig. 3 or the SOR map of Fig. 5a. In other words, the SOR map reflects more details of the local variation in high-risk regions, while the SRR map reflects more details of the local variations in low-risk regions. For practical disease control and prevention, decision-makers may combine different kinds of disease risk maps to highlight noteworthy areas with additional or hidden risks.

At last, regarding the SAR map of Fig. 6, it represents disease risk contribution attributable to the selected risk factors (i.e., various climate and socioeconomic variables for HFMD). The higher the local SAR value, the better, indicating that the selected risk factors were appropriate and effective with more explanatory ability in disease mapping modeling. The SAR map is also a representation of the uncertainty map to show the reliability of the model prediction at the local scale. For instance, in those divisions with higher local SAR values, including South, East, and Central China, we confirmed that the selected risk variables were useful and significantly affect the disease occurrence risk (25–88.7% contribution) at the local scale. For parts with blue-green color in North, Northeast, and Northeast China where local SOR and SRR values were lower, local SAR values were still high, which means the risks of HFMD occurrence in these areas were lower compared with the others, the explanatory ability of the risk variables was well acceptable (10–25% contribution). However, geographical divisions with lowest local SAR values (< 10%) were mainly distributed in Northwest China, indicating that we may need to introduce more potential

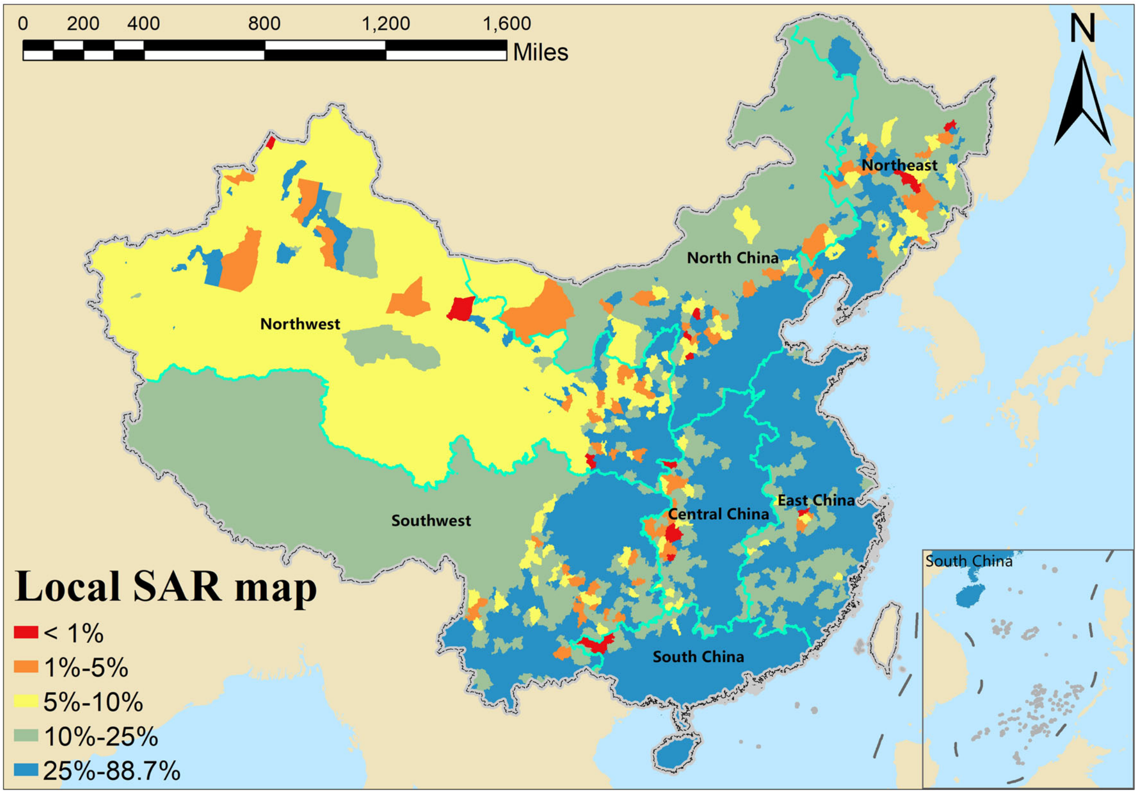

Fig. 6 Spatial attributable risk (SAR) map of HFMD occurrence attributable to climate and socioeconomic factors across Mainland China

risk factors in modeling to fit the noticeable local differences in this division further individually.

4 Discussion

Above all, with regards to HFMD environmental pathogenesis in China, both climate and socioeconomic variables were found to be related to rising disease occurrence probability. First, our study revealed that higher temperature could be an important influential factor for helping spread the virus of HFMD (Guo et al. 2016; Song et al. 2018a; Xiao et al. 2017; Zhao et al. 2018), besides, higher air pressure had an enhancing impact on disease occurrence. Second, we identified that three socioeconomic variables (i.e., per capita household savings, per capita social consumption, and per capita industrial output values) had positive influences on disease occurrence probability, indicating that the economic development was significantly associated with HFMD occurrence over China. Former literature also supported this finding by utilizing similar factors, e.g., per capita GDP (Huang et al. 2014; Li et al. 2018; Song et al. 2018a), the income of citizens (Xu et al. 2019), or a ratio of urban to rural population (Zhang et al. 2018). In addition, for demographic aspect, we found that the higher population density may lead to an ideal environment for easy spreading of virus related to HFMD, which is also consistent with former studies (Bo et al. 2014; Hu et al. 2012; Li et al. 2018).

Regarding disease mapping and to our best knowledge, this study is the first one to generate real and various kinds of disease relative risk maps for HFMD occurrence, not to mention within the whole study area of Mainland China, as well as at the most fine-scale administrative county level (Bo et al. 2014; Song et al. 2018a), which should be of great significance for locally disease control and prevention of HFMD in China. As for China's HFMD case, three new kinds of disease relative risk maps fitted by the DRRD model were capable of offering new information from different perspectives. First, local SOR and SRR maps were similar in explanation, and with both two kinds of maps, we detected new risk clusters around the Hu Line with prominent variations in nature and socioeconomic variations across Mainland China. More importantly, the local SOR map of HFMD was effective to identify locally super-high-risk areas in geographical divisions with higher average risk, e.g., East, South and Central China. While, unlike SOR mapping, the local SRR map of HFMD was effective to highlight locally hidden-high-risk areas in geographical divisions with lower average risk, e.g., North, Northeast, and Northwest China. At last, the local SAR map further showed that our selected climate and socioeconomic factors were effective to explain locally disease occurrence risk in all six geographical divisions, except for Northwest China, in which more environmental factors should be further introduced to extract local variations. Three kinds of national HFMD relative risk maps could help decision-makers to develop specific county-level strategies for control and surveillance, and to monitor progress achieved by ongoing efforts aimed at the elimination of HFMD in China.

Going beyond the findings associated with HFMD aforementioned, the most important contribution of this study is that we proposed a sophisticated but generally applied method, i.e., DRRD model, for advanced disease relative risk mapping. This study is a pilot study on applying this framework to a real-world and large-scale disease mapping study. The DRRD model downscales spatial epidemiology relative risk indicators at the local scale using a three-step progressive statistical strategy, which is fully discussed as follows.

In the first step of DRRD, obtaining a complete disease crude risk map without any missing area is a priority in the DRRD modeling framework. The challenges of the HFMD occurrence data we applied, in this case, were twofold.

Firstly, disease binary data should be the most challenging example among different disease data distribution due to that the original occurrence data is a two-value variable without any local variations (Harrell 2015; Peng et al. 2002). This first task was implemented by constructing logistic regression, in particular, a spatial logistic regression model that incorporates items of both spatial autocorrelation and various environmental covariates, under a Bayesian hierarchical modeling (BHM) framework (Yang et al. 2019b). The BHM method is effective in taking into account locally spatial associations as prior information and estimating posterior values to fill areas with missing data (Rue et al. 2017; Ugarte et al. 2014). The evidence on considering spatial autocorrelated random effects in a regression for estimating spatial disease occurrence probability is summarized as blow. First of all, theoretically speaking, spatial agglomeration is a common disease phenomenon, thus disease data usually have spatial structures, particularly spatial autocorrelation, and the neglect of such spatial autocorrelation could result in a biased and under-performing model in epidemiological and public health risk assessment (Miller 2004; Wu et al. 2004). Second, for HFMD case, the literature review shows that strong spatial autocorrelation effects, e.g., disease geographical clusters, existed across whole Mainland China at the county level (Bo et al. 2014; Song et al. 2018a). Moreover, in this study, the cross-validation evaluation results in Table 2 suggest that the spatial logistic regression surpassed the ordinary logistic regression by improving 10.93% prediction accuracy, indicating it is necessary and promising to account for spatial autocorrelation in regression modeling. Last but not least, the model-based coefficients map of spatial autocorrelated component, as shown in Fig. 3c, showed a similar geographical pattern compared with the disease occurrence probability map in Fig. 3a, further indicating the availability of taking into account spatial autocorrelation in regression modeling.

Secondly, missing data is a common issue existing in large scale geospatial medical research (Lawson et al. 2016). The DRRD modeling framework considers disease missing data as an important issue, and the first-step modeling of DRRD is specifically applied to fill missing data to generate a complete crude risk map. Herein, we summarized some advanced and useful missing data estimation methods based on different situations. Generally, with sufficient explanatory covariates information available, we could employ multiple ecological regression to estimate disease missing data. For instance, we may utilize logistic ecological regression for disease binary variable (Bo et al. 2014; Song et al. 2019), or Poisson ecological regression for disease cases and rate variables (Adin et al. 2018; Song et al. 2018a), among which, spatial types of regression could help increasing model fitness and predict ability by further incorporating spatial structured random effects, e.g., spatial autocorrelation, as priors (Bakka et al. 2018).

Furthermore, when neither samples nor auxiliary data, e.g., explanatory covariates information, are available for missing data estimation, which is a more difficult but common situation, model-based imputation methods could be employed to estimate missing data, such as k-nearest neighbors (kNN) (Troyanskaya et al. 2001), expectation maximum (EM) (Xiong et al. 2015), singular value decomposition (SVD) (Yuan et al. 2019), random forest (RF) (Tang and Ishwaran 2017), and progressive spatiotemporal (PST) (Song et al. 2018b). It's worth mentioning that for a large-scale spatiotemporal official statistics dataset, the recently proposed PST method outperformed the other four imputation methods aforementioned, by taking advantage of sophisticatedly incorporating additional spatiotemporal information, as well as progressively utilizing covariates information (Song et al. 2018b).

In the second step of DRRD, defining spatial control groups considering the SSH impact is reasonable and highly suggested, especially for a large scale spatial case–control experiment. For the HFMD case, we chose Chinese seven geographical divisions as the second SSH-based spatial level to obtain different spatial control groups for every local county. SSH is especially useful for analyzing small area data that are grouped in larger regions (Wang et al. 2016), such as the county-level HFMD area data within the Chinese geographical division. We found that

East, South and Central China were with much higher risk baseline than North, Northeast, and Northwest China, which further suggested that the traditional assumption (a single control group) that the baseline is the same over the whole study area for disease relative risk mapping may be unreasonable (Berke 2005; Bithell 2000; Meza 2003). Ignoring SSH impact could lead to mapping issues such as smoothing super-high-risk areas in higher-baseline regions, or overlooking hidden-high-risk areas in lower-baseline regions. Thus, the second step of the DRRD model should be designed to solve these issues by accounting for the SSH-based spatial control groups.

In the third step of DRRD, following the classical epidemiological relative risk assessment theories and basic formulas, we further developed three new local-scale spatial epidemiological indicators, namely, SRR, SOR, and SAR, for advanced disease relative risk mapping in order to detect new regions with invisible and hidden crude risks. For China's HFMD case, local SOR and SRR maps showed similar macro-scale geographical patterns, but with different local spatial heterogeneity, e.g., the local SOR map could detect super-high-risk areas in the higher risk divisions, whereas the local SRR map could detect hidden-high-risk areas in lower risk divisions. Local SAR map made the disease presence risk assessment complete, which represents the sensitivity of selected variables in different geographical units. Compared with crude probability risk map, all three kinds of relative risk maps could offer extra local spatial information to detect additional risk clusters/outliers, as has been thoroughly discussed above, which could offer new insights into Chinese HFMD detection, control, and prevention at both regional and local scales.

In summary, our newly proposed DRRD model may provide an easy and general way to localize various spatial epidemiological indicators for mapping disease relative risks to detect additional clusters or outliers of disease outcomes, which has some notable advantages. First of all, the DRRD model is applicable for different types of disease outcomes by changing first-step regression models to obtain various types of disease crude risk maps, e.g., we may employ Poisson regression for disease cases variables. Second, to the best knowledge, the DRRD model is the first method indeed considering the SSH impact for disease relative risk mapping, by providing a feasible way to define various spatial control groups, especially useful and necessary for large-scale geospatial studies. Last but not least, on account of DRRD model is following classical epidemiology theories, thus, besides three spatial relative risk indicators that have been localized in this study, future improved DRRD model could localize more epidemiological indicators, such as spatial AR % or PAR %. Under these conditions, the DRRD model is a more general and convenient approach compared with previous disease relative risk mapping methods.

To end our discussion, we need to mention that there might be underreported HFMD cases in CISDCP reporting system in a few counties, which could lead to disease mapping uncertainties (Hu et al. 2012). Furthermore, introducing large-scale environmental remote sensing (He et al. 2016, 2017) and air pollution (Du et al. 2019; Yu et al. 2019) covariates data may improve model interpretation ability for producing disease risk maps with higher prediction accuracy. Finally, the DRRD model only works for spatial areal data at present, whereas not for spatial point data which needs to be further interpolated to a continuous surface (He et al. 2019; Shi et al. 2019). Our future work for this research is to test our proposed DRRD model with different types of disease data, various infectious diseases or public health events, and more meaningfully, with more epidemiological risk assessment indicators (Li et al. 2019a).

5 Conclusions

In this article, we propose a new disease relative risk mapping approach in a general way, named disease relative risk downscaling (DRRD) model, with the aim of providing a guideline for epidemiologists and public health researchers to localize various spatial epidemiological indicators into local scales for exploring additional risk areas of disease outcomes. We successfully applied this sophisticated model for China's HFMD case by generating three spatial relative risk indicators, i.e., SOR, SRR, and SAR. The immediate outcome of this work is a series of complete county-level disease relative risk maps of HFMD covering the whole Mainland China, which should be the first of its kind. These new maps expanded the limited knowledge of the complex local-scale risks of HFMD, revealing additional spatially clustered areas over China, such as the Hu-line region of China. Notably, the local SOR map could detect super-high-risk areas in the higher risk divisions of East, South, and Central China, and the local SRR map could detect hidden-high-risk areas in lower risk divisions of North, Northeast, and Northwest China. The local SAR map further revealed that locally HFMD risks were significantly attributable to climate and socioeconomic factors in all geographical divisions of China except Northwest China. More importantly, the proposed DRRD model is more natural expanded for various researches, not only due to it is based on the original epidemiological formulas, but also as it takes into account of the imperative SSH impact within a spatially case–control experimental design. The DRRD model could be a new alternative method for local-scale relative risk assessment and advanced disease mapping, as well as offers new insights into broader spatial and environmental epidemiology.

Acknowledgements The authors appreciate A-Xing Zhu (University of Wisconsin–Madison, US) and Xun Shi (Dartmouth College, US) for improving the quality of our article, Henry Chung (Michigan State University, US) for proofreading our paper carefully, María Dolores Ugarte (Public University of Navarre, Spain) for her help with coding. We would like to thank the editor and anonymous reviewers for their constructive comments and valuable suggestions on improving this manuscript. The work was jointly supported by the National Natural Science Foundation of China (No. 41701448), a grant from State Key Laboratory of Resources and Environmental Information System (No. 201811), the State Key Laboratory of Remote Sensing Science, the Young Scholars Development Fund of Southwest Petroleum University (No. 201699010064), the Technology Project of the Sichuan Bureau of Surveying, Mapping and Geoinformation (No. J2017ZC05), and the Science and Technology Strategy School Cooperation Projects of the Nanchong City Science and Technology Bureau (No. NC17SY4016, 18SXHZ0025).

References

Adin A, Martínez-Bello DA, López-Quílez A, Ugarte MD (2018) Two-level resolution of relative risk of dengue disease in a hyperendemic city of Colombia. PLoS ONE 13:e0203382. https://doi.org/10.1371/journal.pone.0203382

Bakka H, Rue H, Fuglstad GA, Riebler A, Bolin D, Illian J, Krainski E, Simpson D, Lindgren F (2018) Spatial modeling with R-INLA: A review. Wiley Interdiscip Rev Comput Stat 10:e1443. https://doi.org/10.1002/wics.1443

Berke O (2005) Exploratory spatial relative risk mapping. Prev Vet Med 71:173–182. https://doi.org/10.1016/j.prevetmed.2005.07.003

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B Methodol. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

Bithell JF (1990) An application of density estimation to geographical epidemiology. Stat Med 9:691–701. https://doi.org/10.1002/sim.4780090616

Bithell J (2000) A classification of disease mapping methods. Stat Med 19:2203–2215. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18%3c2203:AID-SIM564%3e3.0.CO;2-U

Bland JM, Altman DG (2000) The odds ratio. BMJ 320:1468. https://doi.org/10.1136/bmj.320.7247.1468

Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with R-INLA. Spatial Spatio-Temporal Epidemiol 4:33–49. https://doi.org/10.1016/j.sste.2012.12.001

Bo Y, Song C, Wang J, Li X (2014) 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 14:358. https://doi.org/10.1186/1471-2458-14-358

Bowman AW, Azzalini A (1997) Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations, vol 18. OUP, Oxford

Cummings P (2009) The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med 163:438–445. https://doi.org/10.1001/archpediatrics.2009.31

Du Z, Lawrence WR, Zhang W, Zhang D, Yu S, Hao Y (2019) Interactions between climate factors and air pollution on daily HFMD cases: A time series study in Guangdong, China. Sci Total Environ 656:1358–1364. https://doi.org/10.1016/j.scitotenv.2018.11.391

Fortin M, James P, MacKenzie A, Melles S, Rayfield B (2012) Spatial statistics, spatial regression, and graph theory in ecology. Spat Stat 1:100–109. https://doi.org/10.1016/j.spasta.2012.02.004

Goicoa T, Adin A, Ugarte M, Hodges J (2018) In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. Stoch Environ Res Risk Assess 32:749–770. https://doi.org/10.1007/s00477-017-1405-0

Guo C, Yang J, Guo Y, Ou Q, Shen S, Ou C, Liu Q (2016) Short-term effects of meteorological factors on pediatric hand, foot, and mouth disease in Guangdong, China: a multi-city time-series analysis. BMC Infect Dis 16:524. https://doi.org/10.1186/s12879-016-1846-y

Harrell FE (2015) Ordinal logistic regression. In: Harrell FE Jr (ed) Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, Berlin, pp 311–325

He Y, Bo Y, Chai L, Liu X, Li A (2016) Linking in situ LAI and fine resolution remote sensing data to map reference LAI over cropland and grassland using geostatistical regression method. Int J Appl Earth Obs Geoinf 50:26–38. https://doi.org/10.1016/j.jag.2016.02.010

He Y, Lee E, Warner TA (2017) A time series of annual land use and land cover maps of China from 1982 to 2013 generated using AVHRR GIMMS NDVI3 g data. Remote Sens Environ 199:201–217. https://doi.org/10.1016/j.rse.2017.07.010

He J, Christakos G, Wu J, Jankowski P, Langousis A, Wang Y, Yin W, Zhang W (2019) Probabilistic logic analysis of the highly heterogeneous spatiotemporal HFRS incidence distribution in Heilongjiang province (China) during 2005-2013. PLoS Negl Trop Dis 13:e0007091. https://doi.org/10.1371/journal.pntd.0007091

Held L, Schrödle B, Rue H (2010) Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. In: Kneib T, Tutz G (eds) Statistical modelling and regression structures. Springer, Berlin, Heidelberg, pp 91–110

Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, Guo Y, Zhao D, Yang W (2012) Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. PLoS ONE 7:e38978. https://doi.org/10.1371/journal.pone.0038978

Huang J, Wang J, Bo Y, Xu C, Hu M, Huang D (2014) Identification of health risks of hand, foot and mouth disease in China using the geographical detector technique. Int J Environ Res Public Health 11:3407–3423. https://doi.org/10.3390/ijerph110303407

Indrayan A, Malhotra RK (2017) Medical biostatistics. Chapman and Hall/CRC, London

Koh WM, Bogich T, Siegel K, Jin J, Chong EY, Tan CY, Chen MI, Horby P, Cook AR (2016) The epidemiology of hand, foot and mouth disease in Asia: a systematic review and analysis. Pediatr Infect Dis J 35:e285. https://doi.org/10.1097/INF.0000000000001242

Lai P, So F, Chan K (2008) Spatial epidemiological approaches in disease mapping and analysis. CRC Press, Boca Raton. https://doi.org/10.1201/9781420045536

Lawson AB (2013) Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Chapman and Hall/CRC, London

Lawson AB, Banerjee S, Haining RP, Ugarte MD (2016) Handbook of spatial epidemiology. CRC Press, Boca Raton

Lee D (2011) A comparison of conditional autoregressive models used in Bayesian disease mapping. Spat Spatio-temporal Epidemiol 2:79–89. https://doi.org/10.1016/j.sste.2011.03.001

Li H, Li J, Wong L, Feng M, Tan Y (2005) Relative risk and odds ratio: a data mining perspective. In: Proceedings of the twenty-fourth ACM SIGMOD–SIGACT–SIGART symposium on

Principles of database systems. ACM, pp 368–377. https://doi.org/10.1145/1065167.1065215

Li L, Qiu W, Xu C, Wang J (2018) A spatiotemporal mixed model to assess the influence of environmental and socioeconomic factors on the incidence of hand, foot and mouth disease. BMC Public Health 18:274. https://doi.org/10.1186/s12889-018-5169-3

Li M, Shi X, Li X, Ma W, He J, Liu T (2019a) Epidemic Forest: A Spatiotemporal Model for Communicable Diseases. Ann Am Assoc Geogra 109:812–836. https://doi.org/10.1080/24694452.2018.1511413

Li M, Shi X, Li X, Ma W, He J, Liu T (2019b) Sensitivity of disease cluster detection to spatial scales: an analysis with the spatial scan statistic method. Int J Geogr Inf Sci. https://doi.org/10.1080/13658816.2019.1616741

Martínez-Bello D, López-Quílez A, Torres Prieto A (2018) Spatiotemporal modeling of Zika and dengue infections within Colombia. Int J Environ Res Public Health 15:1376. https://doi.org/10.3390/ijerph15071376

Meza JL (2003) Empirical Bayes estimation smoothing of relative risks in disease mapping. J Stat Plan Inference 112:43–62. https://doi.org/10.1016/S0378-3758(02)00322-1

Miller HJ (2004) Tobler's first law and spatial analysis. Ann Assoc Am Geogr 94:284–289. https://doi.org/10.1111/j.1467-8306.2004.09402005.x

Mollié A (1996) Bayesian mapping of disease. Markov Chain Monte Carlo Pract 1:359–379

Peng C-YJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96:3–14. https://doi.org/10.1080/00220670209598786

Richardson S, Thomson A, Best N, Elliott P (2004) Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect 112:1016. https://doi.org/10.1289/ehp.6740

Roquette R, Nunes B, Painho M (2018) The relevance of spatial aggregation level and of applied methods in the analysis of geographical distribution of cancer mortality in mainland Portugal (2009–2013). Popul Health Metr 16:6. https://doi.org/10.1186/s12963-018-0164-6

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc B Stat Methodol 71:319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK (2017) Bayesian computing with INLA: a review. Annu Rev Stat Its Appl 4:395–421. https://doi.org/10.1146/annurev-statistics-060116-054045

Schechtman E (2002) Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use? Value Health 5:431–436. https://doi.org/10.1046/j.1524-4733.2002.55150.x

Schmidt CO, Kohlmann T (2008) When to use the odds ratio or the relative risk? Int J Public Health 53:165–167. https://doi.org/10.1007/s00038-008-7068-3

Schrödle B, Held L (2011) Spatio-temporal disease mapping using INLA. Environmetrics 22:725–734. https://doi.org/10.1002/env.1065

Shi X, Li M, Hunter O, Guetti B, Andrew A, Stommel E, Bradley W, Karagas M (2019) Estimation of environmental exposure: interpolation, kernel density estimation or snapshotting. Ann GIS 25:1–8. https://doi.org/10.1080/19475683.2018.1555188

Song C, He Y, Bo Y, Wang J, Ren Z, Yang H (2018a) Risk Assessment and Mapping of Hand, Foot, and Mouth Disease at the County Level in Mainland China Using Spatiotemporal Zero-Inflated Bayesian Hierarchical Models. Int J Environ Res Public Health 15:1476. https://doi.org/10.3390/ijerph15071476

Song C, Yang X, Shi X, Bo Y, Wang J (2018b) Estimating missing values in China's official socioeconomic statistics using progressive spatiotemporal bayesian hierarchical modeling. Sci Rep 8:10055. https://doi.org/10.1038/s41598-018-28322-z

Song C, Shi X, Bo Y, Wang J, Wang Y, Huang D (2019) Exploring spatiotemporal nonstationary effects of climate factors on hand, foot, and mouth disease using Bayesian Spatiotemporally Varying Coefficients (STVC) model in Sichuan, China. Sci Total Environ 648:550–560. https://doi.org/10.1016/j.scitotenv.2018.08.114

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc B Stat Methodol 64:583–639. https://doi.org/10.1111/1467-9868.00353

Szumilas M (2010) Explaining odds ratios. J Can Acad Child Adoles Psychiatry 19:227–229

Tang F, Ishwaran H (2017) Random forest missing data algorithms. Stat Anal Data Min ASA Data Sci J 10:363–377. https://doi.org/10.1002/sam.11348

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Ugarte M, Ibáñez B, Militino A (2006) Modelling risks in disease mapping. Stat Methods Med Res 15:21–35. https://doi.org/10.1191/0962280206sm424oa

Ugarte M, Adin A, Goicoa T, Militino A (2014) On fitting spatio-temporal disease mapping models using approximate Bayesian inference. Stat Methods Med Res 23:507–530. https://doi.org/10.1177/0962280214527528

Vatcheva KP, Lee M, McCormick JB, Rahbar MH (2016) Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology. https://doi.org/10.4172/2161-1165.1000227

Waller LA, Carlin BP (2010) Disease mapping. Chapman & Hall/CRC handbooks of modern statistical methods, London, p 217

Wang J, Guo Y, Christakos G, Yang W, Liao Y, Li Z, Li X, Lai S, Chen H (2011) Hand, foot and mouth disease: spatiotemporal transmission and climate. Int J Health Geogr 10:25. https://doi.org/10.1186/1476-072X-10-25

Wang J, Zhang T, Fu B (2016) A measure of spatial stratified heterogeneity. Ecol Ind 67:250–256. https://doi.org/10.1016/j.ecolind.2016.02.052

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594

Whittemore AS (1983) Estimating attributable risk from case-control studies. Am J Epidemiol 117:76–85. https://doi.org/10.1093/oxfordjournals.aje.a113518

Wilkinson L (1979) Tests of significance in stepwise regression. Psychol Bull 86:168. https://doi.org/10.1037/0033-2909.86.1.168

Wu J, Wang J, Meng B, Chen G, Pang L, Song X, Zhang K, Zhang T, Zheng X (2004) Exploratory spatial data analysis for the identification of risk factors to birth defects. BMC Public Health 4:23. https://doi.org/10.1186/1471-2458-4-23

Xiao X, Gasparrini A, Huang J, Liao Q, Liu F, Yin F, Yu H, Li X (2017) The exposure-response relationship between temperature and childhood hand, foot and mouth disease: a multicity study from mainland China. Environ Int 100:102–109. https://doi.org/10.1016/j.envint.2016.11.021

Xing W, Liao Q, Viboud C, Zhang J, Sun J, Wu JT, Chang Z, Liu F, Fang VJ, Zheng Y (2014) Hand, foot, and mouth disease in China, 2008–12: an epidemiological study. Lancet Infect Dis 14:308–318. https://doi.org/10.1016/S1473-3099(13)70342-6

Xiong W, Yang X, Ke L, Xu B (2015) EM algorithm-based identification of a class of nonlinear Wiener systems with missing output data. Nonlinear Dyn 80:329–339. https://doi.org/10.1007/s11071-014-1871-6

Xu C, Zhang X, Xiao G (2019) Spatiotemporal decomposition and risk determinants of hand, foot and mouth disease in Henan, China. Sci Total Environ 657:509–516. https://doi.org/10.1016/j.scitotenv.2018.12.039

Yang J, Song C, Yang Y, Xu C, Guo F, Xie L (2019a) New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway Basin, Sichuan Province, China. Geomorphology 324:62–71. https://doi.org/10.1016/j.geomorph.2018.09.019

Yang Y, Yang J, Xu C, Xu C, Song C (2019b) Local-scale landslide susceptibility mapping using the B-GeoSVC model. Landslides 16:1301–1312. https://doi.org/10.1007/s10346-019-01174-y

Yu G, Li Y, Cai J, Yu D, Tang J, Zhai W, Wei Y, Chen S, Chen Q, Qin J (2019) Short-term effects of meteorological factors and air pollution on childhood hand-foot-mouth disease in Guilin, China. Sci Total Environ 646:460–470. https://doi.org/10.1016/j.scitotenv.2018.07.329

Yuan X, Han L, Qian S, Xu G, Yan H (2019) Singular value decomposition based recommendation using imputed data. Knowl Based Syst 163:485–494. https://doi.org/10.1016/j.knosys.2018.09.011

Zhang X, Xu C, Xiao G (2018) Space-time heterogeneity of hand, foot and mouth disease in children and its potential driving factors in Henan, China. BMC Infect Dis 18:638. https://doi.org/10.1186/s12879-018-3546-2

Zhao Q, Li S, Cao W, Liu DL, Qian Q, Ren H, Ding F, Williams G, Huxley R, Zhang W, Guo Y (2018) Modeling the Present and Future Incidence of Pediatric Hand, Foot, and Mouth Disease Associated with Ambient Temperature in Mainland China. Environ Health Perspect 126:047010. https://doi.org/10.1289/EHP3062

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Chao Song[1,2,3] · **Yaqian He**[4] · **Yanchen Bo**[1] · **Jinfeng Wang**[2,5] · **Zhoupeng Ren**[2] · **Jiangang Guo**[2] · **Huibin Yang**[1]

[1] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

[2] State Key Laboratory of Resources and Environmental Information System (LREIS), Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

[3] State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, School of Geoscience and Technology, Southwest Petroleum University, Chengdu 610500, Sichuan, China

[4] Department of Geography, Dartmouth College, Hanover, NH 03755, USA

[5] University of Chinese Academy of Sciences, Beijing 100049, China