Spatial Statistics 23 (2018) 36–51

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

A hierarchically adaptable spatial regression model to link aggregated health data and environmental data

Phuong N. Truong [*,1], Alfred Stein

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 16 January 2017
Accepted 6 November 2017
Available online 15 November 2017

Keywords:
Health geography
Aggregated health data
CoS
Ecological fallacy
Spatially varying coefficient
HFMD

ABSTRACT

Health data and environmental data are commonly collected at different levels of aggregation. A persistent challenge of using a spatial regression model to link these data is that their associations can vary as a function of aggregation. This results into ecological fallacy if association at one aggregation level is used for inferencing at another level. We address this challenge by presenting a hierarchically adaptable spatial regression model. In essence, the model extends the spatially varying coefficient model to allow the response to be count data at larger aggregation levels than that of the covariates. A Bayesian hierarchical approach is used for inferencing the model parameters. Robust inference and optimal prediction over geographical space and at different spatial aggregation levels are studied by simulated data sets. The spatial associations at different spatial supports are largely different, but can be efficiently inferred when prior knowledge of the associations is available. The model is applied to study hand, foot and mouth disease (HFMD) in Da Nang city, Viet Nam. Decrease in vegetated areas corresponds with elevated HFMD risks. A study to the identifiability of the parameters shows a strong need for a highly informative prior distribution. We conclude that the model is robust to the underlying aggregation levels of the calibrating data for association inference and it is ready for application in health geography.

* Corresponding author.
E-mail address: p.truongngocphuong@utwente.nl (P.N. Truong).
[1] Postal address: PO Box 217, 7500 AE Enschede, The Netherlands.
Abbreviations: SLM: spatial multilevel statistical model

https://doi.org/10.1016/j.spasta.2017.11.002
2211-6753/

1. Introduction

Important scientific research questions in health geography to study the effects of environmental exposure on human health are "What is the association between human health and the environment?" and "Which of the associations is statistically significant?" (Steenland and Dedden, 1997; Pekkanen and Pearce, 2001). In many cases, for the reasons of confidentiality and cost, extensive geo-referenced health data are only accessible as aggregated data at administrative levels. Meanwhile, environmental data are collected from stations monitoring air, soil or water at a relatively low number of locations or over smaller areas in order to capture their usually small scale variations. Remote sensing imagery such as MODIS, LANDSAT, etc. provides geographically extensive data of the environment but at various spatial resolutions. Unsurprisingly, in many circumstances, these data do not immediately have comparable spatial measurement units in terms of resolution or spatial support to health data. For example, provincial medical statistics in Viet Nam only report hand, food and mouth disease (HFMD) cases at the district level with area sizes ranging between 10 and 10^3 km^2; whereas the environmental risk factors for this disease such as daily air temperature and humidity are regularly recorded at only one or two meteorological monitoring stations per province with an average area of about 5×10^3 km^2. Persistent challenges to link these data are that their associations can vary as a function of aggregation, the well-known modifiable areal unit problem (MAUP) (Openshaw, 1983; Cressie, 1996). This results in ecological fallacy if association at one aggregation level is carelessly used for inferencing at another aggregation level (King, 1997).

Popular models for association analyses in health geography are the regression-based models (Keppel, 2005; Bender, 2009; Auchincloss et al., 2012). A well-known example of such models in health research is the spatial multilevel statistical model (SLM) (Langford et al., 1999; Goldstein, 2010; Arcaya et al., 2012). Here, the "spatial" prefix distinguishes the hierarchy of the geographical space from the hierarchy of the feature space of the data. Spatial hierarchy is defined by the differences of the spatial supports. For example, morbidity reporting of individual person contracted HFMD in Viet Nam (level 1) is aggregated to district level (level 2) and to regional level (level 3). SLM is widely applied to health data with the spatial hierarchy structuring from individual health outcomes to larger environmental surroundings. Such a desired spatial hierarchy, however, does not always hold in practice due to lack of exclusive sampling designs (Duncan and Jones, 2000).

The associations between health data and environmental risk factors can locally vary due to various facts such as averaging effects of aggregated health data, measurement units of environmental risk factors, spatially changing socio-economic and individual characteristics of the understudied population. Also, the spatial autocorrelation of health data, environmental risk factors or both have an effect. Previous research (e.g. Fotheringham et al., 1998, 2001; Saib et al., 2014; Hamm et al., 2015, amongst others) has demonstrated the superiority of geographically adaptable regression coefficient models in various real cases if spatial non-stationarity of the associations is present. The two best known models in current literature are the geographically weighted regression model (GWR) (Fotheringham et al., 1998, 2009) and the spatial varying coefficient model (SVC) (Gelfand et al., 2003). In GWR, the parameter surface is allowed to vary as a deterministic spatial surface; and its value at an individual location is estimated by geographical distance weighted least square of proximate locations (Fotheringham et al., 2001). In SVC, it is generated by a second-order stationary spatial random process that enables the probabilistic uncertainty quantification about the parameter estimators (Gelfand et al., 2003).

The underlying principle of both GWR and SVC is that the regression coefficient surfaces follow Tobler's first law of geography, i.e. that the regression coefficients are spatially correlated. GWR, however, is not robust to the MAUP (Fotheringham et al., 2001). Change of support (CoS) modelling (Cressie, 2015a) is important in spatial statistics and SVC has been shown to be efficient in modelling spatial non-stationarity of the regression coefficients (Wheeler and Calder, 2007; Finley, 2011). Therefore, we argue that SVC provides an efficient modelling framework to investigate the effect of CoS on the parameters of the regression-based model for geo-referenced data with different spatial supports. Nevertheless, the increasing number of the parameters of the SVC might pose other challenges, e.g. to inference the hyper-parameters of the unobservable stochastic regression coefficient surfaces. This identifiability issue has been experienced by many complex (spatial-temporal) statistical models (Brun et al., 2001; Bujosa et al., 2007; Lavielle and Aarons, 2016; Ugarte et al., 2017).

Emphasizing the advantages of the SVC over others, we aim at extending this model to enable it to link misaligned health and environmental data when the health data are available at larger aggregation levels than the environmental data. In this paper, we combine a bivariate Poisson log-normal model with SVC. We demonstrate its use with simulated data sets at regular spatial supports to investigate the effect of CoS on the regression coefficient surfaces. We also study identifiability of the parameters. The applications of the model to geographical health data at irregular spatial supports are illustrated with HFMD data at the district level from Da Nang city, Viet Nam.

2. Materials and methods

We assumed that count data of disease cases are available at spatial support level $\mathbf{v}$, called $\ell_{\mathbf{v}}$, where $\ell_{\mathbf{v}}$ is a geographical area of arbitrary size and shape. The count data are modelled as realizations of a spatially inhomogeneous Poisson process (IPP). Health data at two different aggregation levels are assumed to be realizations from two spatially dependent IPPs. Hence, the intensity of the IPP is a function of both the geographical location and the aggregation level. The intensity at every investigated location is modelled as a realization from a log-normal random process. Environmental or exposure data are available at smaller spatial support level $\mathbf{u}$, called $\ell_{\mathbf{u}}$, also of arbitrary size and shape. Other essential assumptions are specified in the following sections.

2.1. Models for health data

The count data of disease cases at $\ell_{\mathbf{u}}$ partitioning the study area D are in the form of a set of positive integers $\{z(\mathbf{u}_s),\ s = 1 : S\}$. The $z(\mathbf{u}_s)$ are assumed to be realizations of random count Z at location $\mathbf{u}_s$ in D. A marginal probability distribution of $Z(\mathbf{u}_s)$ has the form of a Poisson probability distribution with the intensity function $\lambda(\mathbf{u}_s)$. The $\lambda(\mathbf{u}_s)$ is defined as in Waller and Carlin (2010): $\lambda(\mathbf{u}_s) = \delta_s(\mathbf{u}_s) V(\mathbf{u}_s)$, where $\lambda(\mathbf{u}_s)$ is inhomogeneous, $\delta_s(\mathbf{u}_s)$ is the expected number of disease cases per person at $\ell_{\mathbf{u}}$. $V(\mathbf{u}_s)$ is the expected number of disease cases per spatial unit at $\ell_{\mathbf{u}}$, derived by: $V(\mathbf{u}_s) = \varphi \times po_s(\mathbf{u}_s)$, where φ is the global observed disease rate and $po_s(\mathbf{u}_s)$ is the population at risk at $\ell_{\mathbf{u}}$.

The global mean values of the regression coefficients' surfaces at $\ell_{\mathbf{u}}$ are equal to those at $\ell_{\mathbf{v}}$ because they are not influenced by CoS (Cressie, 1996). Notably, $\ell_{\mathbf{v}}$ necessarily contains $\ell_{\mathbf{u}}$. The generalized linear mixed model (Stein, 1999) was used together with SVC (Gelfand et al., 2003) to describe the $\lambda(\mathbf{u}_s)$. In other words, a log-link function at $\ell_{\mathbf{u}}$ with known $P \times 1$ covariate vector $\mathbf{X} = \left(X_p; p = 1 \ldots P\right)$ has spatial varying regression coefficients that are assumed to be spatial Gaussian processes (MVN) with the Matérn spatial dependence structure (Stein, 1999).

$$\log\left(\delta_s(\mathbf{u}_s)\right) = \beta_{0u} + \sum_{p=1}^{P} \beta_{pu} X_p(\mathbf{u}_s) + \sum_{p=1}^{P} \varepsilon_{pu}(\mathbf{u}_s) X_p(\mathbf{u}_s) + \varepsilon_{0u}(\mathbf{u}_s), \tag{1}$$

$$\varepsilon_{(0,p)u} \sim MVN\left(0,\ \Sigma_{(0,p)u}\left(\theta_{(0,p)u}\right)\right), \tag{2}$$

where $\beta_{0u} = \beta_{0v}$ and $\beta_{pu} = \beta_{pv}$ are the global means of the intercept and the slope at $\ell_{\mathbf{u}}$, respectively, β_{0v} and β_{pv} are the global means of the intercept and the slope at $\ell_{\mathbf{v}}$, respectively, ε_{0u} and ε_{pu} are the autocorrelated variances of the intercept and the slope at $\ell_{\mathbf{u}}$, respectively, $\Sigma_{(0,p)u}$ is the S×S variance–covariance matrix of $\varepsilon_{(0,p)u}$, $\theta_{(0,p)u}$ is the vector of the variogram parameters.

Here ε_{0u} and ε_{pu} are modelled by an isotropic, second-order stationary Matérn random process (Pardo-Iguzquiza and Chica-Olmo, 2008) with the general form of the variogram model:

$$\gamma(\mathbf{h}) = \tau + c\left\{1 - \left[\left(\frac{2^{1-\nu}}{\Gamma(\nu)}\right)\left(\frac{\mathbf{h}}{\alpha}\right)^{\nu} K_{\nu}\left(\frac{\mathbf{h}}{\alpha}\right)\right]\right\}, \tag{3}$$

where $\mathbf{h}$ is the Euclidean distance, Γ is the gamma function, K_{ν} is the modified Bessel function of the second kind of order ν, τ is the nugget variance, c is the partial sill variance, α is the range parameter and ν is the smoothness parameter.

The log-normal distribution of $\delta_s(\mathbf{u}_s)$ has the following form (Izsák, 2008; Gelfand et al., 2003):

$$\log\mathrm{N}(\delta_s(\mathbf{u}_s)) = (2\pi)^{-\frac{S}{2}}\left(\prod_{s=1}^{S}\frac{\lambda(\mathbf{u}_s)}{V(\mathbf{u}_s)}\right)^{-1}\left|\sum_{p=1}^{P}D_{pu}^{T}\Sigma_{pu}D_{pu}+\Sigma_{0u}\right|^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(E_u)^{T}\left(\sum_{p=1}^{P}D_{pu}^{T}\Sigma_{pu}D_{pu}+\Sigma_{0u}\right)^{-1}E_u\right), \tag{4}$$

$$E_u = \log(\delta_s) - \beta_{0u} - \sum_{p=1}^{P}\beta_{pu}X_p(\mathbf{u}_s), \tag{5}$$

$$D_{pu} = \mathrm{Diag}\left(X_p(\mathbf{u}_s)\right). \tag{6}$$

Similarly, count data of disease cases at $\ell_{\mathbf{v}}$, also fully covering D are $\{z(\mathbf{v}_k),\ k=1:K\}$, where K is the total number of spatial units $\mathbf{v}$. The marginal probability distribution of $Z(\mathbf{v}_k)$ is also defined as an IPP with intensity function $\lambda(\mathbf{v}_k)$, varying in geographical space. The relationship between $\lambda(\mathbf{v}_k)$ and $\lambda(\mathbf{u}_s)$ that defines the linkage of the intercept and of the slopes between the two aggregation levels equals:

$$\lambda(\mathbf{v}_k) = \sum_{s=1}^{S_k}\lambda(\mathbf{u}_s),\ \mathbf{u}_s\in\mathbf{v}_k,\ S=\sum_{k=1}^{K}S_k, \text{ or} \tag{7}$$

$$\lambda(\mathbf{v}_k) = \sum_{s=1}^{S_k}V(\mathbf{u}_s)\,e^{\beta_{0v}+\sum_{p=1}^{P}\beta_{pv}X_p(\mathbf{u}_s)+\sum_{p=1}^{P}\varepsilon_{pu}(\mathbf{u}_s)X_p(\mathbf{u}_s)+\varepsilon_{0u}(\mathbf{u}_s)},\ \mathbf{u}_s\in\mathbf{v}_k,\ S=\sum_{k=1}^{K}S_k. \tag{8}$$

The log-link function of $\delta_k(\mathbf{v}_k)$, i.e. the expected number of disease cases per person at $\ell_{\mathbf{v}}$ is also specified as a linear function of the covariate vector $\mathbf{X}$ at the same aggregation level. It has spatially varying regression coefficients, but with different variation over geographical space and aggregation level as compared to those at $\ell_{\mathbf{u}}$:

$$\log(\delta_k(\mathbf{v}_k)) = \beta_{0v} + \sum_{p=1}^{P}\beta_{pv}X_p(\mathbf{v}_k) + \sum_{p=1}^{P}\varepsilon_{pv}(\mathbf{v}_k)X_p(\mathbf{v}_k) + \varepsilon_{0v}(\mathbf{v}_k), \tag{9}$$

$$\varepsilon_{(0,p)v} \sim MVN\left(0,\ \Sigma_{(0,p)v}\left(\theta_{(0,p)v}\right)\right) \tag{10}$$

where $\Sigma_{(0,p)\mathbf{v}}$ is the $K\times K$ variance–covariance matrix of $\varepsilon_{(0,p)v}$; ε_{0v} and ε_{pv} are the autocorrelated variances of the intercept and the slope at $\ell_{\mathbf{v}}$, respectively.

As we are interested in studying how the aggregation levels affect the association between health data and environmental data, both θ_{0u}, θ_{pu} and θ_{0v}, θ_{pv} were investigated using (1)–(10). The assumptions are that the intercept and the slope at $\ell_{\mathbf{v}}$ and $\ell_{\mathbf{u}}$ are dependent, but are independent from each other. This independence assumption is reasonable, but its validity may be questioned, in particular if a negative relation exists between the two (Gelfand et al., 2003).

2.2. A Bayesian hierarchical model

We follow the terminology of Cressie and Wikle (2011) to specify the spatial Bayesian hierarchical model.

The count data $\{z(\mathbf{v}_k),\ k=1:K\}$ are known. Because the Poisson probability distribution ensures that the first and second moments are identical, we assume that the data $z(\mathbf{v}_k)$ are mutually

independent and that any dependence is taken care of by the intensity function. Hence, the marginal log-likelihood function of $z(\mathbf{v}_k)$ is:

$$\log L\left(z\left(\mathbf{v}_k\right)|\lambda\left(\mathbf{v}_k\right)\right) = \sum_{k=1}^{K} z\left(\mathbf{v}_k\right)\log\left(\lambda\left(\mathbf{v}_k\right)\right) + \sum_{k=1}^{K}\lambda\left(\mathbf{v}_k\right) - \sum_{k=1}^{K}\log\left(z\left(\mathbf{v}_k\right)!\right). \tag{11}$$

The intensity (process) models at support levels $\mathbf{v}$ and $\mathbf{u}$ are defined by (7, 8) and (12, 13):

$$\left[\lambda\left(\mathbf{u}_s\right)|\beta_{0v}, \beta_{pv}, \theta_{0u}, \theta_{pu}\right] \sim \log N\left(\delta_s\left(\mathbf{u}_s\right)\right), \tag{12}$$

$$\left[\lambda\left(\mathbf{v}_k\right)|\beta_{0v}, \beta_{pv}, \theta_{0v}, \theta_{pv}\right] \sim \log N\left(\delta_k\left(\mathbf{v}_k\right)\right). \tag{13}$$

2.3. Hybrid MCMC estimators

Hybrid Metropolis–Hastings and Gibbs estimators based upon Markov chains (Jackman, 2009) were used for estimating $\beta_{0v}, \beta_{pv}, \theta_{0v}, \theta_{pv}$ and θ_{0u}, θ_{pu}. Posterior probability distributions for all parameters are specified using (14)–(16).

— For β_{0v} and $\beta_{(1\cdots P)v}$:

$$\begin{aligned}&\left[\beta_{0v}, \beta_{p=1\cdots Pv}|z\left(\mathbf{v}_k\right), X_{p=1\cdots P}\left(\mathbf{u}_s\right)\right] \sim \left[z\left(\mathbf{v}_k\right)|\lambda\left(\mathbf{v}_k\right)\right]\\ &\left[\lambda\left(\mathbf{u}_s\right)|X_{1\cdots P}\left(\mathbf{u}_s\right), \beta_{0v}, \beta_{(1\cdots P)v}, \theta_{0u}, \theta_{(1\cdots P)u}\right]\left[\beta_{0v}, \beta_{(1\cdots P)v}\right]\end{aligned} \tag{14}$$

— For θ_{0u} and $\theta_{(1\cdots P)u}$:

$$\begin{aligned}\left[\theta_{0u}, \theta_{(1\cdots P)u}|z\left(\mathbf{v}_k\right), X_{(1\cdots P)}\left(\mathbf{u}_s\right)\right] \sim &\left[z\left(\mathbf{v}_k\right)|\lambda\left(\mathbf{v}_k\right)\right]\left[\lambda\left(\mathbf{u}_s\right)|X\left(\mathbf{u}_s\right), \beta_{0v}, \beta_{1v}, \theta_{0u}, \theta_{1u}\right]\\ &\times\left[\theta_{0u}, \theta_{pu}, \right]\end{aligned} \tag{15}$$

— For θ_{0v} and $\theta_{(1\cdots P)v}$:

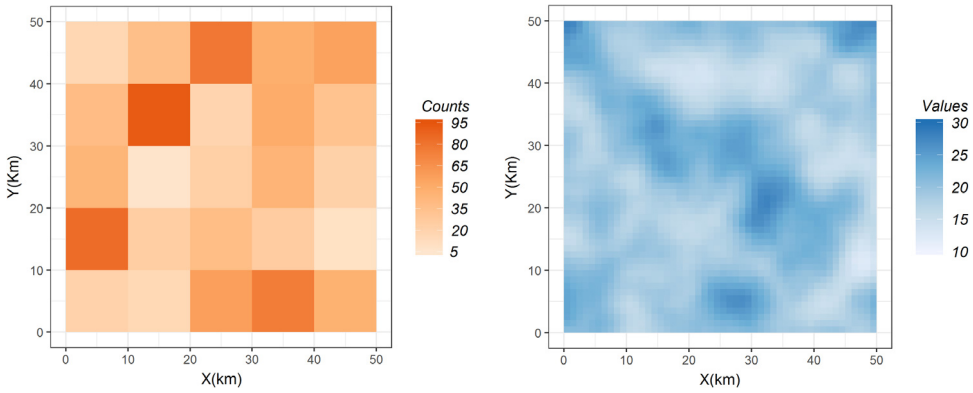
$$\left[\theta_{0v}, \theta_{(1\cdots P)v}|\lambda\left(\mathbf{v}_k\right), X_{1\cdots P}\left(\mathbf{v}_k\right)\right] \sim \left[\lambda\left(\mathbf{v}_k\right)|\theta_{0v}, \theta_{(1\cdots P)v}, X_{1\cdots P}\left(\mathbf{v}_k\right)\right]\left[\theta_{0v}, \theta_{(1\cdots P)v}.\right] \tag{16}$$

All models and MCMC estimators were implemented in R (R Core Team, 2017). We partly reused the R scripts from Truong et al. (2014) and mainly used the gstat-package (Pebesma, 2004) for implementing the models and the coda-package (Plummer et al., 2006) for testing the convergence of the MCMC runs.

2.4. Description of simulated data sets

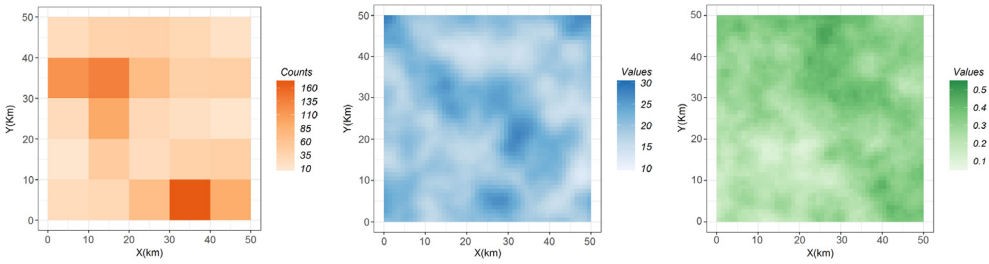
We simulated hypothetical cases of health and environmental data at regular spatial support with arbitrary sets of parameters that produce non-negative and finite intensity values. Simulated health data at $\ell_{\mathbf{u}}\left(z\left(\mathbf{u}_s\right)\right)$ were generated at 1×1 km grid cells based upon (1)–(3) with $p = 1$ and $p = 2$. The global observed disease rate was assumed to be equal to $\varphi = 0.3 \times 10^{-3}$. Simulated data of $z(\mathbf{u}_s)$ were then aggregated to 10×10 km grid cells to create the $z(\mathbf{v}_k)$ data sets (Figs. 1(A) and 2(A)). Hence, $K = 25$ and $S = 2,500$. Gaussian unconditional simulation generated the unobservable slope and intercept surfaces.

Covariates $X_p(\mathbf{u}_s)$ were simulated at 1×1 km grid cell by Gaussian random simulation. Table 1 presents the specifications in detail. The nugget effects are not relevant to regression coefficients and also cannot be identified from aggregated data (Truong et al., 2014). It is immaterial to specify the unit of all parameters, except for defining the Euclidean distance. Root mean square error (RMSE) was used to measure the degree of model fitting.



(A) Count of disease cases. (B) Covariate 1.

Fig. 1. Simulated data of disease cases at $\ell_{\mathbf{v}}$ (A) and one environmental covariate at $\ell_{\mathbf{u}}$ (B).



(A) Count of disease cases. (B) Covariate 1. (C) Covariate 2.

Fig. 2. Simulated data of disease cases at $\ell_{\mathbf{v}}$ (A) and two environmental covariates at $\ell_{\mathbf{u}}$ (B, C).

Table 1
Parameter specifications for simulating gridded data sets.

Parameter	Specification for simulating data
β_{0v}	Constant (0.5)
β_{1v}	Constant (0.08)
β_{2v}	Constant (0.005)
θ_{0u}	MVN with the Matérn variogram model: $c_{0u} = 0.005$, $\alpha_{0u} = 5$ km, $\tau_{0u} = 0$, $\nu_{0u} = 2$
θ_{1u}	MVN with the Matérn variogram model: $c_{1u} = 0.0025$, $\alpha_{1u} = 2.5$ km, $\tau_{1u} = 0$, $\nu_{1u} = 1.5$
θ_{2u}	MVN with the Matérn variogram model: $c_{2u} = 0.007$, $\alpha_{2u} = 5$ km, $\tau_{2u} = 0$, $\nu_{2u} = 1.5$
$x_1(u_s)$	MVN with mean 20, the Matérn variogram model: $c_{x1} = 10$, $\alpha_{x1} = 3$ km, $\tau_{x1} = 0$, $\nu_{x1} = 2$
$x_2(u_s)$	MVN with mean 0.3, the Matérn variogram model: $c_{x2} = 0.006$, $\alpha_{x2} = 6$ km, $\tau_{x2} = 0$, $\nu_{x2} = 1$

2.5. HFMD in Da Nang city, Viet Nam

Outbreak of HFMD was first reported in Viet Nam in 2003 (WPRO, 2011). Since 2011, HFMD has been classified as a severely infectious disease by Vietnamese Ministry of Health. The majority of

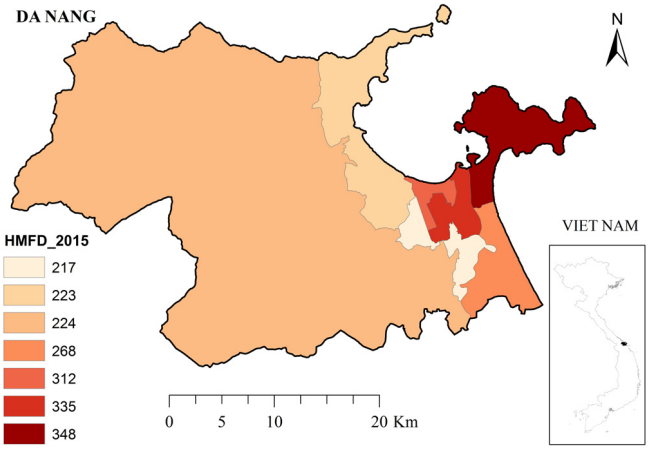


Fig. 3. HFMD cases at the district level in Da Nang city, Viet Nam, 2015.

reported and fatal cases in Viet Nam were children younger than five years old. In total, there were 170 fatal cases amongst 113,121 total cases in 2011 (Nguyen et al., 2014). HFMD caused by *Coxsackievirus A16* (CAV16) and *Enterovirus 71* (EV71) infections has the symptoms of fever, skin rashes on hands and feet, sore throat and mouth ulcers (WPRO, 2011). Currently, neither vaccine nor specific treatment for HFMD is available in Viet Nam. HFMD is contagious and transmitted both through the air by free-living viruses and through physical contact. Previous studies of HFMD in China and South East Asia show that specific local meteorological conditions strongly stimulate the transmission of this disease (Li et al., 2014; Dong et al., 2016; Wang et al., 2016).

In Viet Nam, 72.8% of the HFMD cases were reported from the southern part for the first quarter of 2015. The largest number of cases (36.8%) was reported from Da Nang city (WPRO, 2015), located in central Viet Nam with a population of more than 1 million people (GSO, 2015). The city has one urban district and six rural districts covering an area of approximately 1280 km^2. Due to the recent outbreak, the Da Nang Preventative Medicine Centre weekly reports to the public (via the website: http://www.yteduphongdanang.vn) the total number of HFMD cases for each district in the city. In order to prevent the small number problem in our data analysis (Pekkanen and Pearce, 2001), we used the yearly sum data of HFMD cases in 2015 for each district based upon those reports (Fig. 3). The total number of reported HFMD cases per district in Da Nang city in 2015 varies from approximately 200 to 300 cases. Environmental conditions for each district are indicated by the percentage of vegetated land. The reason is that the areas of vegetated lands are approximately ten times lower in the urban and semi-urban districts than in the rural districts. This can lead to different environmental conditions for the spread of HFMD. The vegetation data set was obtained from the Terra MODIS Vegetation Continuous Fields 2015 product (Fig. 4). Gridded population data were obtained from the WorldPop Viet Nam 2015 product (http://www.worldpop.org.uk).

3. Results and discussion

3.1. Simulation studies

Following the simulation strategy recommended by Gelman and Shirley (2011), three chains of 10 thousands MCMC runs at $\ell_{\mathbf{u}}$ and 100 thousands MCMC runs at $\ell_{\mathbf{v}}$ were executed for all parameters. The prior distributions of all parameters were defined as in Table 5 of Appendix. Acceptance rates vary from 0.25 to 0.35. Geweke's convergence tests (Geweke, 1992) show well convergence for the second-half of the chains with all Z-scores satisfactorily falling within the two standard deviations. The first halves of the chains were discarded as burn-in. The second-halves of all convergent chains were mixed together and thinned out at each of 15^{th} step at $\ell_{\mathbf{u}}$ and 150^{th} step at $\ell_{\mathbf{v}}$.

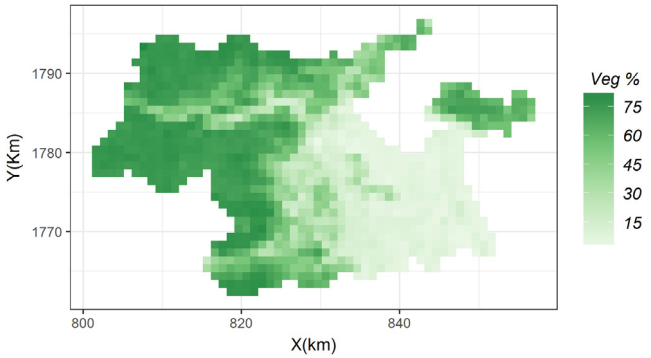


Fig. 4. Percentage of vegetated areas at a 1 km resolution in Da Nang city, Viet Nam, 2015.

Table 2
Summary statistics of the posteriors for simulation with one covariate.

Parameter	Posterior mode ± standard deviation			RMSE
$\beta_{0v} = \beta_{0u}$		0.47 ± 0.47		0.47
$\beta_{1v} = \beta_{1u}$		0.08 ± 0.03		0.03
	$c\ (\times 10^{-3})$	α (km)	ν	$(c \times 10^{-3}, \alpha, \nu)$
θ_{0u}	3.5 ± 0.1	$4.8 \pm 2.7 \times 10^{-3}$	1.9 ± 0.34	(1.5, 0.2, 0.35)
θ_{1u}	2.1 ± 0.1	$2.11 \pm 8.5 \times 10^{-3}$	1.5 ± 0.34	(0.4, 0.39, 0.34)
θ_{0v}	0.6 ± 0.1	$5 \pm 5 \times 10^{-3}$	2.05 ± 0.05	–
θ_{1v}	1.3 ± 0.6	2.05 ± 0.02	1.97 ± 0.17	–

3.1.1. Simulation with one covariate

Summary statistics of the posterior distributions of the parameters are given in Table 2.

Table 2 shows that the total variance of all regression coefficients' surfaces at $\ell_{\mathbf{v}}$ ($c_{0v} = 0.6 \times 10^{-3}$ and $c_{1v} = 1.3 \times 10^{-3}$) deviates largely from those at $\ell_{\mathbf{u}}$ ($c_{0u} = 3.5 \times 10^{-3}$ and $c_{1u} = 2.1 \times 10^{-3}$). The averaging effects of aggregating to a larger spatial support can be seen from Figs. 5 and 6. The maps in these figures are the realizations of the Gaussian unconditional simulations of the regression coefficient surfaces at both spatial supports based upon the posterior modes of the variogram parameters in Table 2. Their geographic extents are 36 times larger than those in Fig. 1 for the purpose of visualization. The large reduction in the total sill values of the slope and the intercept of approximately 38% and 83%, respectively (see Fig. 9 in Appendix) leads to smoother regression coefficient surfaces at $\ell_{\mathbf{v}}$.

Aggregation from 1 km^2 to 100 km^2 significantly influences the association between health data and environmental data by increasing the smallest association and decreasing the largest association (Figs. 5 and 6). In other words, the slope and intercept values at $\ell_{\mathbf{v}}$ vary less around their global mean values over geographical space than at $\ell_{\mathbf{u}}$. It is thus insensible to infer local association (at $\ell_{\mathbf{u}}$) between health data and environmental data based on global associate (β_{pv}). This also holds for the aggregated associate, i.e. the slope surface at $\ell_{\mathbf{v}}$.

The results in this section once more corroborate the MAUP in ecological inference. The results also show that not only the random part of the mixed effect models (the intercept) is affected by CoS but also its fixed part (the slope). This implies that correcting for only the inferred intercept as in the studies of Kerry et al. (2012) or Wang et al. (2015) is not sufficient to produce unbiased association inferences or to produce unbiased predictions. Prediction accuracy with one covariates at $\ell_{\mathbf{u}}$ and $\ell_{\mathbf{v}}$ has RMSE values equal to 0.023 and 28.5, respectively.

3.1.2. Simulation with two covariates

For two covariates, the number of parameters increases to 12. The estimates of β_0, β_1, θ_0 and θ_1 are mostly unchanged as compared to the case of one covariate (Table 3). However, their standard

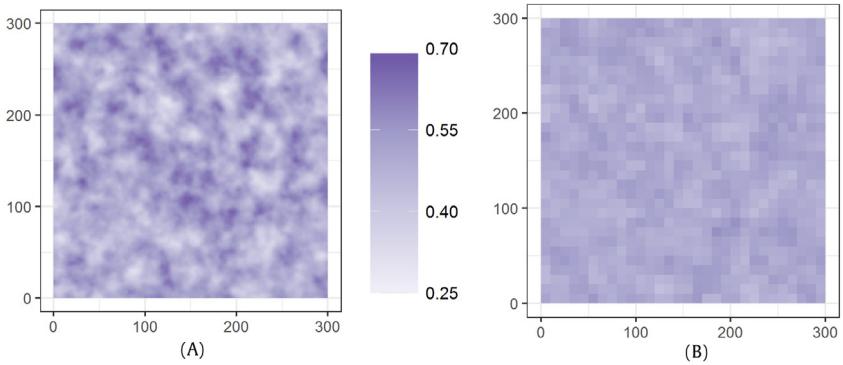


Fig. 5. Intercept surfaces at $\ell_{\mathbf{u}}$ (A) and at $\ell_{\mathbf{v}}$ (B). The intercept values at $\ell_{\mathbf{u}}$ vary from 0.2 to 0.7; at $\ell_{\mathbf{v}}$, they vary between 0.4 and 0.6.

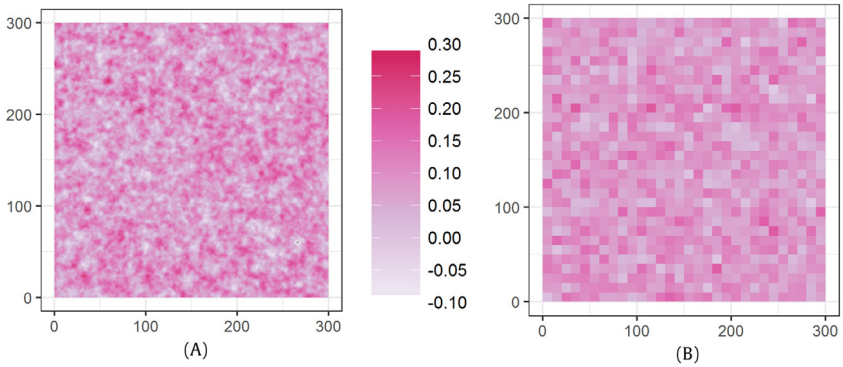


Fig. 6. Slope surfaces at $\ell_{\mathbf{u}}$ (A) and $\ell_{\mathbf{v}}$ (B). Similarly, the slope values at $\ell_{\mathbf{u}}$ have larger variance from -0.1 to 0.3, whereas those values at $\ell_{\mathbf{v}}$ vary between -0.05 and 0.25.

Table 3
Summary statistics of the posteriors for simulation with two covariates.

Parameter	Posterior mode ± standard deviation			RMSE
$\beta_{0v} = \beta_{0u}$		0.6 ± 0.6		0.6
$\beta_{1v} = \beta_{1u}$		0.05 ± 0.03		0.04
$\beta_{2v} = \beta_{2u}$		0.007 ± 0.003		0.004
	c ($\times 10^{-3}$)	α (km)	ν	$(c \times 10^{-3}, \alpha, \nu)$
θ_{0u}	3 ± 0.3	4.8 ± 0.01	1.8 ± 1.5	(2, 0.2, 1.5)
θ_{1u}	2.1 ± 0.16	$2.1 \pm 8.4 \times 10^{-3}$	1.47 ± 0.3	(0.4, 0.4, 0.3)
θ_{2u}	8 ± 2.3	$4.8 \pm 23 \times 10^{-3}$	2.1 ± 0.9	(2.3, 0.2, 1.1)
θ_{0v}	1 ± 0.2	5 ± 0.02	4.5 ± 0.8	
θ_{1v}	0.1 ± 0.0	$2 \pm 6 \times 10^{-3}$	2 ± 0.6	
θ_{2v}	0.5 ± 0.01	2 ± 0.03	3 ± 1.4	

deviations and thus their RMSE values increase. The RMSEs in Table 3 are almost equal to their standard deviations. Thus, estimates β_2, θ_2 have high accuracy. Convergence of the MCMC chains of the fixed parts of the regression coefficients, i.e. the β_0, β_1 and β_2 is approximately three times slower than that of the other parameters.

We based our model on the BYM modelling framework of disease mapping (Besag et al., 1991) but our model allows to model data at various spatial aggregation levels and allows the impacts of

Table 4
Summary statistics of the posteriors in the HFMD study.

Parameters	Mode ± standard deviation		
$\beta_{0v} = \beta_{0u}$		2.5 ± 0.2	
$\beta_{1v} = \beta_{1u}$		$-0.77 \pm 0.78 \times 10^{-3}$	
	c ($\times 10^{-3}$)	α (km)	ν
θ_{0u}	0.36 ± 1.3	3.6 ± 13	0.24 ± 0.21
θ_{1u}	0.08 ± 0.00002	0.24 ± 0.21	$0.98 \pm 0.64 \times 10^{-3}$
θ_{0v}	0.03 ± 0.0005	$4.8 \pm 4 \times 10^{-3}$	0.8 ± 0.02
θ_{1v}	0.005 ± 0	$5.0 \pm 1 \times 10^{-3}$	1.3 ± 0.01

environmental covariates on disease occurrences. Note that the CAR mixture multiscale models of Lawson and Clark (2002) and Aregay et al. (2015, 2016) do not take the impacts of environmental covariates on diseases into account. In terms of estimating the spatial associations between diseases and environmental risk factors, our model allows to estimate the non-stationary spatial associations over multiple aggregation levels. The estimates are achieved simultaneously instead of the separate modelling at each aggregation level of Saib et al. (2014) or Louie and Kolaczyk (2006).

The log-linear model with a Poison distribution is widely applied in mapping both disease rate and disease count. It is easily extended to model data at different aggregation levels. When applied to the log-link function with non-stationary covariates, the model of Gotway and Young (2007) would produce biased estimates (Cressie et al., 2004). The binomial beta model of King et al. (1999) or the Bayesian wavelet-based model of Kolaczyk and Huang (2001) also deals with CoS issues, but extending their models to include covariates at different aggregation levels is much more complicated. Our model assumes that the spatial processes at all aggregation levels share the same underlying spatial point latent process. Hence, modelling the spatial structures with the variogram or the covariance function is straightforward as compared to CAR modelling or wavelet modelling.

By using our hierarchical modelling approach, spatial associations at both $\ell_{\mathbf{u}}$ and $\ell_{\mathbf{v}}$ can be obtained with high accuracy. Furthermore, Bayesian inferencing quantifies confident intervals for all estimates. A non-Bayesian approach would not be straightforward to deal with our non-linear modelling. Our model thus yields a robust approach to different aggregation levels for ecological inference. The computation cost, however can be high if the number of parameters increases (Finley, 2011). Section 3.2 presents the case of actual data at an irregular spatial support level.

3.2. HFMD study

In the HFMD study, the same MCMC simulation strategy and statistical convergence test as in Section 3.1 were adopted for sampling the posterior distributions. Unlike in the simulation study, normal prior distributions were used for all parameters. Table 4 presents summary statistics of all posteriors. Note that the irregular support level $\mathbf{v}$ ($\ell_{\mathbf{v}}$) is at the district size of Da Nang city, varying from 9 to 748 km^2 with various shapes, and that the spatial support level $\mathbf{u}$ ($\ell_{\mathbf{u}}$) is at the 1 km^2 square grid of the vegetation data.

The global mean slope value from Table 4 shows a negative association between HFMD and vegetated areas. At the two spatial support levels $\mathbf{u}$ and $\mathbf{v}$, every 1% decrease in vegetated areas elevates the risk of HFMD by 0.77%. This finding corroborates previous findings on the impacts of climatic conditions on HFMD incidence in South East Asia. It is well-known that decreasing the vegetated areas leads to increasing temperature (Lim et al., 2008). Studies of Ma et al. (2010) in Hong Kong, Hii et al. (2011) in Singapore, and recently, Dong et al. (2016) in Beijing show that increasing temperature significantly results in increasing HFMD risk. Sill values of the slope variograms at both support levels, i.e. 0.08×10^{-3} at $\ell_{\mathbf{u}}$ and 0.005×10^{-3} at $\ell_{\mathbf{v}}$, respectively, are low as compared to their mean values. Spatial variation of the association between HFMD and the percentage of vegetation is thus not so large that it can lead to changing the association from negative towards a positive impact of vegetation on the HFMD risk. The assumption of unchanged global mean value of the association between health data and environmental data in our models is therefore both mathematically convenient and biologically sensible.

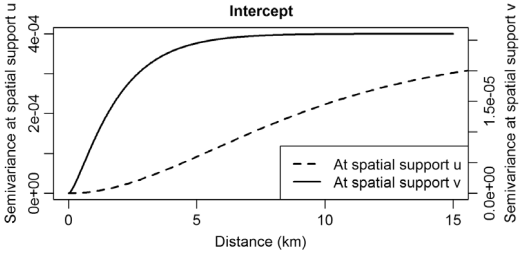


Fig. 7. Variogram models of the intercept surfaces at $\ell_{\mathbf{u}}$ and $\ell_{\mathbf{v}}$ for HFMD. The intercept has a stronger autocorrelation and larger values at $\ell_{\mathbf{u}}$ than at $\ell_{\mathbf{v}}$.

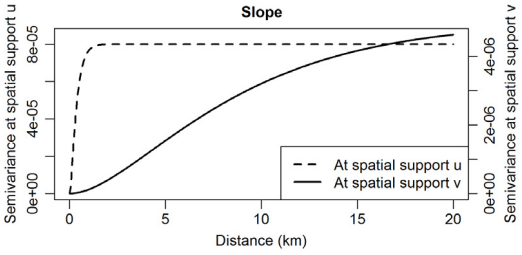


Fig. 8. Variogram models of the slope surfaces at $\ell_{\mathbf{u}}$ and $\ell_{\mathbf{v}}$ for HFMD. The slope values at $\ell_{\mathbf{u}}$ vary almost randomly with a larger sill value in contrast to those at $\ell_{\mathbf{v}}$.

In Figs. 7 and 8, the left y-axis presents the values of the semivariances corresponding to lag distance $\mathbf{h}$ at $\ell_{\mathbf{u}}$, whereas the right y-axis presents those at $\ell_{\mathbf{v}}$. In Fig. 7, the sill of the intercept variogram at $\ell_{\mathbf{v}}$ decreases considerably (92.8%) as compared to that at $\ell_{\mathbf{u}}$. Similarly, the sill of the slope variogram at $\ell_{\mathbf{v}}$ considerably decreases (93.8%) as compared to that at $\ell_{\mathbf{u}}$ (Fig. 8). When $\ell_{\mathbf{v}}$ is irregular and much larger than $\ell_{\mathbf{u}}$, the averaging effect increases. Figs. 11 and 12 in Appendix present Gaussian simulations of the slope and the intercept surface.

In Fig. 8, the variogram of the slope surface shows no spatial structure in the association between HFMD and vegetation at the local scale of 1 km^2. As a result, slope values change almost randomly amongst the grid cells (Fig. 12(A)). The association is stronger at Da Nang's district level than at 1 km^2. Similar findings are observed on the spatial structure of the association between HFMD and climatic factors at county-level of China with areas from 100 to 1,000 km^2 (Hu et al., 2012).

The spatial factor is recently recognized as an important factor in spatial epidemiology of HFMD in South East Asia (Hii et al., 2011; Dong et al., 2016). Our model thus takes account of the non-stationary spatial associations between HFMD and the environmental conditions corresponding to the aggregation levels. Aggregation health data are common in practice of health geography. Using aggregated health data for association inference should be done cautiously because this association is influenced largely by the aggregation level as shown in this study.

3.3. Identifiability

To access whether or not the estimated spatial associations are spurious, we studied the identifiability of model parameters using simulated data with known parameter values. We investigated Bayesian inferences for the parameters of core model (1) of Section 2.1, i.e. the spatial generalized linear mixed model with SVC. The model has a non-stationary mean spatial random process as an addition to the fixed and the random effects. The model parameters are identifiable if the inferencing methods can identify the parameters of the fixed effects and the parameters of the random effects

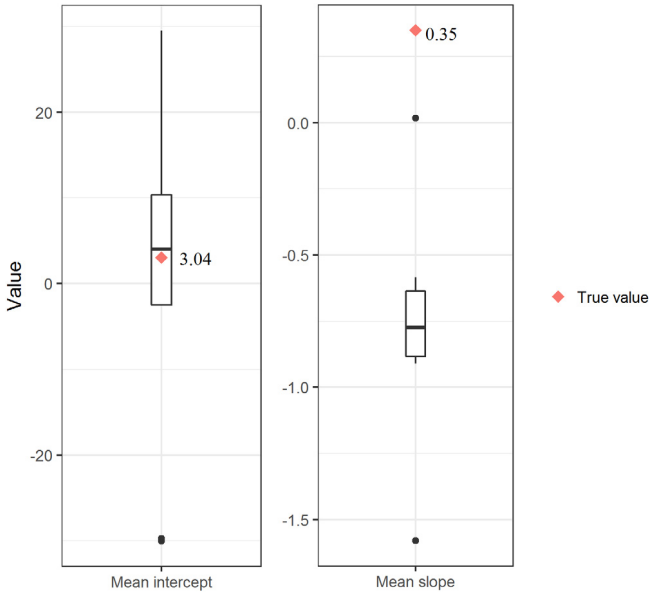


Fig. 9. Variation of the values of the modes of the posteriors of the global mean intercept and the global mean slope. The true value of the mean slope does not fall within the variation interval of the posterior modes.

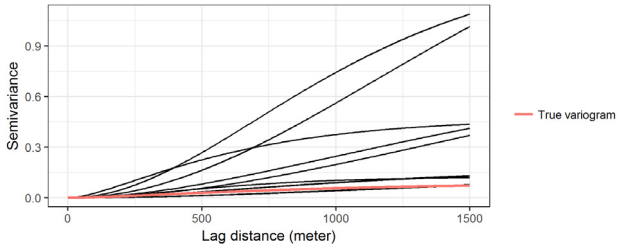


Fig. 10. The variogram models plotted using the values of posterior modes of all parameters of the random slopes. The true values of the Matérn variogram model of the random intercept are $c_0 = 0.09$, $\alpha_0 = 658$ m, $\nu_0 = 0.9$, and those of the Matérn variogram model of the random slope are $c_1 = 0.59$, $\alpha_1 = 700$ m, $\nu_1 = 2$.

given available observations. Identifiability is measured by the consistency of multiple simulation-based inferenced results. Simulating data was done as presented in Section 2.4. The simulated data have 100 observations at a regular spatial support of 100 m^2.

Note that the priors of all parameters in (1) are independent and have uniform distributions on a wide but finite interval. Ten repetitions of 50,000 MCMC runs were conducted with arbitrary initial values. MCMC chains of all parameters passed the Geweke's convergence test with proper acceptance rates. Fig. 9 shows the modes of the posteriors of the global mean intercept and those of the global mean slope. The modes of the global mean intercept are consistent, in contrast to those of the global mean slope. We interpreted this that the first is better identifiable than the latter. In passing we noted that the global mean intercept and the global mean slope have a consistently negative correlation as previously observed by Gelfand et al. (2003). Inconsistency also extends to the posterior variograms of the spatial random slope based upon the modes of the posteriors of all parameters (Fig. 10).

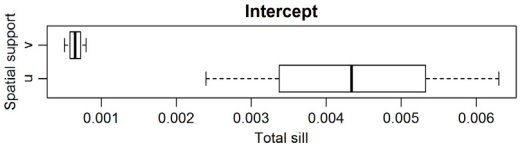


Fig. 11. Decrease of the sill intercept at $\ell_{\mathbf{v}}$ as compared to that at $\ell_{\mathbf{u}}$.

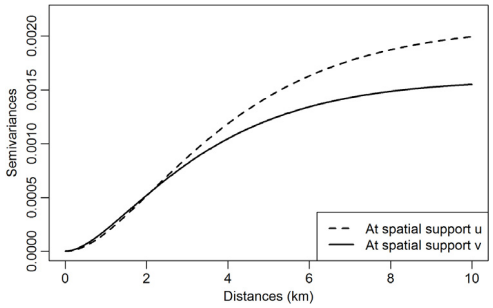


Fig. 12. Variogram models of the slope surfaces at $\ell_{\mathbf{v}}$ and $\ell_{\mathbf{u}}$ in the simulation with one covariate. Reduction in the total sill at $\ell_{\mathbf{v}}$ starts from the lag distance of about 2 km onwards.

Those results as compared to those in Section 3.1 where the priors are informative and the initial values of all parameters were dedicated guesses, indicate that Bayesian inference needs to be guided with informative priors to result in consistent inference. A complete lacking of such information leads to inconsistency and requires that Bayesian inference for the spatial generalized mixed midel with SVC needs to be replaced, e.g. by empirical Bayesian inference (Cressie, 2015a, b Section 3.4.4).

4. Conclusions

In this study, we have presented a Bayesian hierarchical approach to associate health data and environmental data at different aggregation levels. The modelling approach extends SVC to take account of the spatially non-stationary association. Its application on two simulation data sets showed that the model is robust to the underlying aggregation levels of the calibrating data for spatial association inference between health and environmental conditions. In a practical study on the HFMD in Da Nang city, Viet Nam, the model showed that a decrease in vegetation elevates the HFMD risk.

We have thus introduced a modelling approach that is widely applicable for studying with aggregated health data in environmental epidemiology or disease mapping. More delicate modelling approaches can be extended from our models to allow for incorporating dependences in the model parameters. Rigorous study of identifiability taking account of other effects on the consistency of the Bayesian inferences for the parameters of the spatial GLMM with SVC is also recommended for future research.

Acknowledgements

The authors would like to thank the associate editor and anonymous reviewers for their constructive comments.

Appendix

See Figs. 11–14 and Table 5.

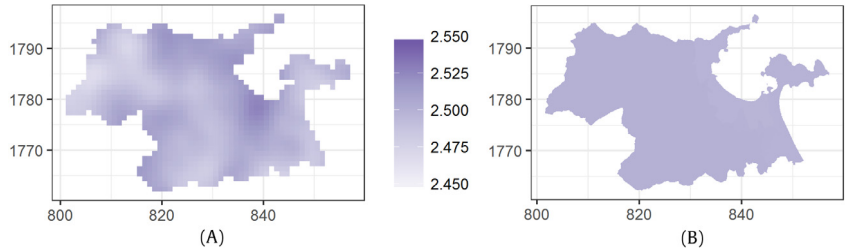


Fig. 13. Intercept surfaces at $\ell_{\mathbf{u}}$ (A) and $\ell_{\mathbf{v}}$ (B) for Da Nang city. Intercept values at the 1 km^2 grid vary between 2.45 and 2.55, whereas those at the Da Nang districts vary between 2.49 and 2.50.

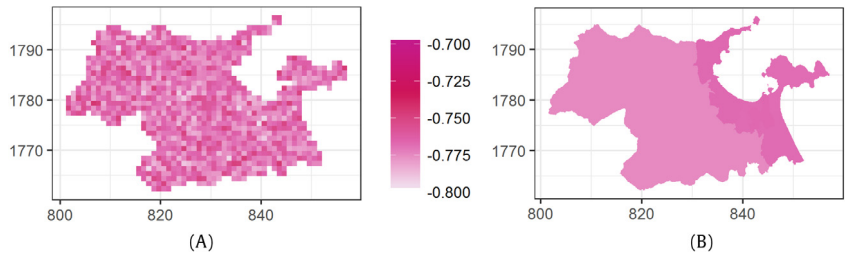


Fig. 14. Slope surfaces at $\ell_{\mathbf{u}}$ (A) and $\ell_{\mathbf{v}}$ (B) for Da Nang city. The slope value at the 1 km^2 grid varies between −0.8 and −0.75, whereas, at the Da Nang districts it varies within a smaller interval, i.e. between −0.769 and −0.766.

Table 5
Prior distributions of the parameters for the simulation study with one and two covariates.

Parameter	Priors		
$\beta_{0v} = \beta_{0u}$		U(0, 1)	
$\beta_{1v} = \beta_{1u}$		U(0, 0.16)	
$\beta_{2v} = \beta_{2u}$		U(0, 0.01)	
	c	α (km)	ν
θ_{0u}	N(0.005, 0.5),	N(5,000, 10,000),	N(2, 0.5)
θ_{1u}	N(0.0025, 0.5),	N(2,500, 10,000).	N(1.5, 0.5)
θ_{2u}	N(0.007, 0.1),	N(5,000, 10,000).	N(1.5, 0.5)
θ_{0v}	U(0.0005,10),	U(1,000, 10,000),	U(0,5)
θ_{1v}	U(0.00005,10),	U(500, 10,000),	U(0,5)
θ_{2v}	U(0.00005,10),	U(500, 10,000),	U(0,5)

References

Arcaya, M., Brewster, M., Zigler, C.M., Subramanian, S.V., 2012. Area variations in health: A spatial multilevel modeling approach. Health Place. 18, 824–831.
Aregay, M., Lawson, A.B., Faes, C., Kirby, R.S., 2015. Bayesian multi-scale modeling for aggregated disease mapping data. Stat. Methods Med. Res. 0, 1–20.
Aregay, M., Lawson, A.B., Faes, C., Kirby, R.S., Carroll, R., Watjou, K., 2016. Spatial mixture multiscale modeling for aggregated health data. Biom. J. 58, 1091–1112.
Auchincloss, A.H., Gebreab, S.Y., Mair, C., Roux, A.V.D., 2012. A review of spatial methods in epidemiology, 2000–2010. Annu. Rev. Public Health 33, 107–122.
Bender, R., 2009. Introduction to the use of regression models in epidemiology. In: Verma, M. (Ed.), Cancer Epidemiology. Humana Press, Totowa, N.J., pp. 179–195.
Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43, 1–20.

Brun, R., Reichert, P., Künsch, H.R., 2001. Practical identifiability analysis of large environmental simulation models. Water Resour. Res. 37, 1015–1030.
Bujosa, M., García-Ferrer, A., Young, P.C., 2007. Linear dynamic harmonic regression. Comput. Statist. Data Anal. 52, 999–1024.
Cressie, N., 1996. Change of support and the modifiable areal unit problem. J. Geogr. Syst. 3, 159–180.
Cressie, N.A.C., 2015a. 3.4.4. Bayesian kriging. In: Statistics for Spatial Data, revised ed. Wiley Interscience, New York, pp. 284–289.
Cressie, N.A.C., 2015b. Change of support. In: Statistics for Spatial Data, revised ed. Wiley Interscience, New York, pp. 284–289.
Cressie, N., Richardson, S., Jaussent, I., 2004. Ecological Bias: Use of maximum-entropy approximations. Aust. N. Z. J. Stat. 46, 233–255.
Cressie, N., Wikle, C.K., 2011. Statistics for Spatio - Temporal Data. Wiley & Sons, Hoboken.
Dong, W., Li, X.e., Yang, P., Liao, H., Wang, X., Wang, Q., 2016. The effects of weather factors on hand, foot and mouth disease in beijing. Sci. Rep. 6, 19247.
Duncan, C., Jones, K., 2000. Using multilevel models to model heterogeneity: Potential and pitfalls. Geogr. Anal. 32, 279–305.
Finley, A.O., 2011. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. Methods Ecol. Evol. 2, 143–154.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2001. Scale issues and geographically weighted regression. In: Tate, N.J., Atkinson, P.M. (Eds.), Modelling Scale in Geographical Information Science. Wiley & Sons, Chichester, pp. 123–140.
Fotheringham, A.S., Charlton, M.E., Brunsdon, C., 1998. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. Environ. Plann. A 30, 1905–1927.
Fotheringham, A.S., Charlton, M., Demšar, U., 2009. Looking for a relationship? Try GWR. In: Miller, H.J., Han, J. (Eds.), Geographic Data Mining and Knowledge Discovery, second ed.. CRC Press, pp. 227–254.
Gelfand, A.E., Kim, H.J., Sirmans, C.F., Banerjee, S., 2003. Spatial modeling with spatially varying coefficient processes. JASA 98, 387–396.
Gelman, A., Shirley, K., 2011. Inference from simulations and monitoring convergence. In: Brooks, S., Gelman, A., Jones, G.L., Meng, X.L. (Eds.), Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC, pp. 163–174.
Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. Oxford University Press, pp. 169–193.
Goldstein, H., 2010. An introduction to multilevel models. In: Multilevel Statistical Models. John Wiley & Sons, Ltd, pp. 1–14.
Gotway, C.A., Young, L.J., 2007. A geostatistical approach to linking geographically aggregated data from different sources. J. Comput. Graph. Statist. 16, 115–135.
GSO: Vietnam General Statistics Office, 2015. Statistic Yearbook of Vietnam. Statistical Publishing House, Viet Nam, pp. 85–86.
Hamm, N.A.S., Finley, A.O., Schaap, M., Stein, A., 2015. A spatially varying coefficient model for mapping PM10 air quality at the European scale. Atmos. Environ. 102, 393–405.
Hii, Y.L., Rocklöv, J., Ng, N., 2011. Short term effects of weather on hand, foot and mouth disease. PLoS One 6, e16796.
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.
Izsák, R., 2008. Maximum likelihood fitting of the Poisson lognormal distribution. Environ. Ecol. Stat. 15, 143–156.
Jackman, S., 2009. Markov chain Monte Carlo. In: Bayesian Analysis for the Social Sciences. John Wiley & Sons, Ltd, pp. 201–250.
Keppel, K., 2005. Methodological issues in measuring health disparities. Vital. Health Stat. 2, 1–16.
Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P., 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma 170, 347–358.
King, G., 1997. A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton University Press, Princeton.
King, G., Rosen, O., Tanner, M.A., 1999. Binomial-beta hierarchical models for ecological inference. Sociol. Methods Res. 28, 61–90.
Kolaczyk, E.D., Huang, H., 2001. Multiscale statistical models for hierarchical spatial aggregation. Geogr. Anal. 33, 95–118.
Langford, I.H., Leyl, A.H., Rasbash, J., Goldstein, H., 1999. Multilevel modelling of the geographical distributions of diseases. J. Roy. Statist. Soc. Ser. C 48, 253–268.
Lavielle, M., Aarons, L., 2016. What do we mean by identifiability in mixed effects models? J. Pharmacokinet. Pharmacodyn. 43, 111–122.
Lawson, A.B., Clark, A., 2002. Spatial mixture relative risk models applied to disease mapping. Stat. Med. 21, 359–370.
Li, T., Yang, Z., Liu, X., Kang, Y., Wang, M., 2014. Hand-foot-and-mouth disease epidemiological status and relationship with meteorological variables in Guangzhou, Southern China, 2008-2012. Rev. Inst. Med. Trop. São Paulo 56, 533–539.
Lim, Y.K., Cai, M., Kalnay, E., Zhou, L., 2008. Impact of vegetation types on surface temperature change. J. Appl. Meteorol. Climatol. 47, 411–424.
Louie, M.M., Kolaczyk, E.D., 2006. A multiscale method for disease mapping in spatial epidemiology. Statist. Medicine 25, 1287–1306.
Ma, E., Lam, T., Wong, C., Chuang, S.K., 2010. Is hand, foot and mouth disease associated with meteorological parameters? Epidemiol. Infect. 138.
Nguyen, N.T., Pham, H.V., Hoang, C.Q., Nguyen, T.M., Nguyen, L.T., Phan, H.C., Phan, L.T., Vu, L.N., Tran Minh, N.N., 2014. Epidemiological and clinical characteristics of children who died from hand, foot and mouth disease in Vietnam. BMC Infect. Dis. 14, 341.
Openshaw, S., 1983. The modifiable areal unit problem. CATMOG. 38.
Pardo-Iguzquiza, E., Chica-Olmo, M., 2008. Geostatistics with the Matern semivariogram model: A library of computer programs for inference, kriging and simulation. Comput. Geosci. 34, 1073–1079.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.
Pekkanen, J., Pearce, N., 2001. Environmental epidemiology: challenges and opportunities. Environ. Health Perspect. 109, 1–5.
Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6, 7–11.
R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Saib, M.S., Caudeville, J., Carre, F., Ganry, O., Trugeon, A., Cicolella, A., 2014. Spatial relationship quantification between environmental, socioeconomic and health data at different geographic levels. Int. J. Environ. Res. Public Health 11, 3765–3786.
Steenland, K., Dedden, J., 1997. Design and analysis of studies in environmental epidemiology. In: Steenland, K., Savitz, D.A. (Eds.), Topics in Environmental Epidemiology. Oxford UniversityPress, Oxford, pp. 9–27.
Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. In: Springer Series in Statistics.
Truong, P.N., Heuvelink, G.B.M., Pebesma, E., 2014. Bayesian area-to-point kriging using expert knowledge as informative priors. Int. J. Appl. Earth Obs. Geoinf. 30, 128–138.
Ugarte, M.D., Adin, A., Goicoa, T., 2017. One-dimensional, two-dimensional, and three dimensional B-splines to specify space –time interactions in Bayesian disease mapping: Model fitting and model identifiability. Spat. Stat. (in press).
Waller, L., Carlin, B., 2010. Disease mapping. In: Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (Eds.), Handbook of Spatial Statistics. CRC Press, pp. 217–243.
Wang, Q., Shi, W., Atkinson, P.M., Zhao, Y., 2015. Downscaling MODIS images with area-to-point regression kriging. Remote Sens. Environ. 166, 191–204.
Wang, J., Xiao, Y., Peng, Z., 2016. Modelling seasonal HFMD infections with the effects of contaminated environments in mainland China. Appl. Math. Comput. 274, 615–627.
Wheeler, D.C., Calder, C.A., 2007. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. J. Geogr. Syst. 9, 145–166.
WPRO, 2011. A Guide to Clinical Management and Public Health Response for Hand, Foot and Mouth Disease (HFMD), http://www.wpro.who.int/emerging_diseases/documents/HFMDGuidance/en/.
WPRO, 2015. Hand, Foot, and Mouth Disease Situation Update, http://www.wpro.who.int/emerging_diseases/hfmd_biweekly_20150210.pdf.