Statistica Neerlandica (2011) Vol. 65, nr. 1, pp. 84–100
doi:10.1111/j.1467-9574.2010.00475.x

Hierarchical Bayesian modeling of the space-time diffusion patterns of cholera epidemic in Kumasi, Ghana

Frank B. Osei* and Alfred A. Duker

Department of Geomatic Engineering, Kwame Nkrumah University of Science and Technology, Kumasi-Ghana

Alfred Stein

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

This study analyses the joint effects of the two transmission routes of cholera on the space-time diffusion dynamics. Statistical models are developed and presented to investigate the transmission network routes of cholera diffusion. A hierarchical Bayesian modelling approach is employed for a joint analysis of nonlinear effects of continuous covariates, spatially structured variation, and unstructured heterogeneity. Proximity to primary case locations and population density serve as continuous covariates. Reference to communities is modelled as a spatial effect. The study applied to the Kumasi area in Ghana shows that communities proximal to primary case locations are infected relatively early during the epidemics, with more remote communities infected at later dates. Similarly, more populous communities are infected relatively early and less populous communities at later dates. The rate of infection increases almost linearly with population density. A non systematic relation occurs between the rate of infection and proximity to primary case locations. It is discussed how these findings could serve as significant information to help health planners and policy makers in making effective decisions to limit cholera epidemics.

Keywords and Phrases: Cholera, *Vibrio cholera*, Geographic Information Systems, spatial statistics, Hierarchical, Bayesian.

1 Introduction

Mapping of disease transmission routes in human population and knowledge of its spatial and temporal transmission dynamics are essential for epidemiologist to understand better the population's interaction with its environment. Understanding the spatial distribution of diseases and transmission dynamics is facilitated by advancements in Geographic Information Systems (GIS) and spatial statistics. These

*osei23782@itc.nl; oseibadu2004@yahoo.co.uk

provide opportunities for epidemiologist to analyse disease distribution in space and interactions with the environment. Most of these approaches, however, ignore methodological difficulties that arise from the nature of the data, especially when the population distribution and environment is particularly variable and spatially structured.

Classical linear regression approaches, where the response variable is assumed to be Gaussian distributed with the covariates acting linearly on the response, have been used to model the diffusion dynamics of infectious diseases (Kuo and Fukui, 2007; Trevelyan, Smallman-Raynor and Cliff, 2005). Such diffusion models assume a strictly linear relationship between the dependent variable and the predictor variables, thereby ignoring the possible nonlinear and spatial effects of the predictor variables. Moreover, these diffusion models ignore the possibility and role of multiple index cases in the diffusion dynamics of the disease.

Cholera is a water-borne disease caused by *Vibrio cholera* (hereafter *V. cholera*). Comprehensive discussions about cholera are presented by Carpenter (1970), Colwell and Huq (1994), Finkelstein (1996), Prestero, Height and Hwang (2001), Sack *et al*. (2004), Huq *et al*. (2005). The disease has been scrutinized since the beginning of epidemiology (Snow, Frost and Richardson, 1936), yet it remains an important public health problem, especially in developing countries. Without treatment, case-fatality rate or death can be as high as 50% of severe cases (WHO, 1993; Sack *et al*., 2004). Cholera diffuses rapidly in environments that lack basic infrastructure with regard to access to safe water and proper sanitation. Provision of good sanitary conditions, sewage treatment, and provision of clean water have long been known as important critical measures for prevention and eradication. These measures have eliminated cholera from industrialized and developed countries. Chronic poverty in developing countries makes implementation of these measures almost unfeasible. A better understanding of the dynamics of cholera spread amongst communities could help to develop alternative and timely public health interventions to limit or prevent cholera epidemics.

Two routes of cholera transmission have been described. The primary route or *environment-to-human* transmission is the exposure of a human being to an aquatic reservoir of *V. cholera*. The secondary route or *human-to-human* transmission is through faecal-oral contacts induced by a previously infected person (Miller, Feachem and Drasar, 1985; Glass *et al*., 1991). Primary transmission is responsible for sparking initial outbreaks. Primary cases are therefore hypothesized to be scattered is space and time, occurring almost simultaneously in distant areas with no apparent connection. In contrast, once the outbreak has reached a threshold level, faecal-oral transmissions dominate and the disease becomes highly contagious. Consequently, geographic factors such as proximity to a primary case location and population density should spatially dominate the disease propagation. To examine these hypotheses, this study analyses the joint effects of primary and secondary transmission in the space-time diffusion dynamics of cholera. Specifically, the study seeks to (i) define and map the transmission routes of cholera diffusion from possible

multiple primary cases and (ii) model the joint effects of population density and proximity to primary cases on the space-time dynamics of cholera diffusion.

This paper is organized as follows. First, a variogram model is used to characterize the spatial auto-covariance structure of incidence rates to determine the threshold/extent of contagiousness of cholera. Thus, the variogram model is used to characterize the dominant scale at which cholera transmission occurs. Secondly, the threshold value and the times of cholera entrance in communities are applied to define the transmission routes and all probable primary cases. Third, a hierarchical Bayesian model is built, where the time ordered sequence of cholera entrance in each community is modelled as nonlinear functions of proximity to respective primary cases and the urban level. In such a modelling approach, the unknown parameters are treated as random variables arranged in a hierarchy such that the distributions at each level are determined by the random variables in the previous levels. Next, we present the results and conclude the paper with discussion on the results.

2 Methods and data

2.1 Study area and data

The area studied is the Kumasi Metropolis, an urban and the most populous city in Ashanti Region, at approximately 250 km (by road) northwest of Accra. It is centred at the intersection of latitude 6.04°N and longitude 1.28°W, covering an area of about 220 km^2 (See Figure 1). Kumasi has a population of approximately 1.2 million which accounts for just under a third (i.e. 32.4%) of the region's population. After cholera introduction in Ghana in the 1970s, the country has experienced a series of epidemic outbreaks. Surveillance and reporting of the disease before 2005 has been ineffective, and hence the existing data before 2005 have little or no

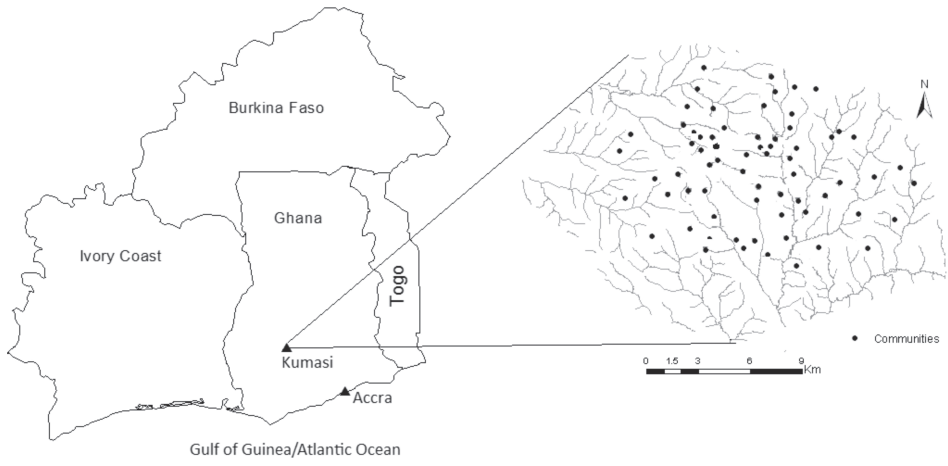

Fig. 1. District map of Ghana (left), and Kumasi (right). Dots indicate the centroids of communities.

spatial and temporal information. With intensified surveillance and reporting systems during an outbreak in 2005, disease cases in Kumasi are being recorded daily at community level spatial units. Kumasi is therefore suitable for studying the dynamics of cholera in space and time.

The topographic map of the metropolis and the $n = 68$ communities where cholera records are available was digitized. Cholera data for each community was extracted from case records obtained from the Kumasi Metropolitan Disease Control Unit (DCU). This data contains information about the index case records for each community, thus the time of cholera onset for each community. Residential addresses were not recorded during the period of diagnoses; therefore, the centroids of the communities were used as the spatial references of the case locations. Also, index case records for each community were extracted and assigned with unique identification codes. These were cross-linked to the communities by unique identification codes to facilitate easy geo-referencing and further analysis.

2.2 Defining the extent of contagiousness: variogram modelling

Let the n communities belonging to a domain D be given by the spatial locations $S = (s_1, \ldots, s_n)$, $\mathrm{Chol_C}(s_\alpha)$ denote the number of cholera cases, and $n(s_\alpha)$ denote the size of the population at risk. Then the incidence rate at s_α is expressed as $\mathrm{Chol_R}(s_\alpha) = \mathrm{Chol_C}(s_\alpha)/n(s_\alpha)$. To make a statistical inference, $\mathrm{Chol_C}(s_\alpha)$ is interpreted as the realization of a random variable that follows a one parameter Poisson distribution with intensity $n(s_\alpha) \cdot \mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha)$. That is:

$$\mathrm{Chol_C}(s_\alpha) \,|\, \mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha) \sim P(n(s_\alpha) \cdot \mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha)), \quad \alpha = 1, \ldots, n. \tag{1}$$

Such modelling is based on the assumption that the spatial correlation among cholera cases is caused by spatial trends in either the population sizes or in the local individual risks given the risk value $\mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha)$. Therefore, the count variables $\mathrm{Chol_C}(s_\alpha)$ are assumed to be conditionally independent. The risk variable $\mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha)$ is modelled as a stationary random field with mean m, variance $\sigma^2_{\mathrm{Chol}_{\hat{\mathrm{R}}}}$, and covariance function $C_{\mathrm{Chol}_{\hat{\mathrm{R}}}}(h) = \mathrm{Cov}[\mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha), \mathrm{Chol}_{\hat{\mathrm{R}}}(s_\beta)]$ which depends only on the distance h between observation pairs s_α and s_β. Following MATHERON's (1963, 1965) intrinsic hypothesis on expected mean differences and variances, the equivalent traditional variogram is:

$$\gamma_{\mathrm{Chol}_{\hat{\mathrm{R}}}}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [(\mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha) - \mathrm{Chol}_{\hat{\mathrm{R}}}(s_\beta))^2] \mathrm{I}_{d_{\alpha\beta} \sim h}, \quad \forall \mathrm{Chol_R} \geq 0, \tag{2}$$

where $\mathrm{I}_{d_{\alpha\beta} \sim h}$ is the indicator function for observations pairs (s_α, s_β) separated by the distance h and $N(h)$ is the number of observation pairs separated by h. Strictly speaking, the above model is a semi-variogram but the prefix 'semi' shall be omitted and this convention is followed in our subsequent references to it. The above variogram model, however, is not suited for the analysis of disease incidences since

it does not account for the heterogeneous population distributions. Following the approach developed by Monestiez *et al.* (2005, 2006), the experimental variogram is estimated as:

$$\hat{\gamma}_{\mathrm{Chol}_{\hat{\mathrm{R}}}}(h) = \frac{1}{2N(h)} \times \sum_{\alpha,\beta=1}^{N(h)} \left(\frac{n(s_\alpha)\cdot n(s_\beta)}{n(s_\alpha)+n(s_\beta)} [\mathrm{Chol}_{\mathrm{R}}(s_\alpha) - \mathrm{Chol}_{\mathrm{R}}(s_\beta)]^2 - m^* \right) \mathrm{I}_{d_{s_\alpha\beta}\sim h}, \tag{3}$$

where now $N(h) = \sum_{\alpha,\beta} \frac{n(s_\alpha)\cdot n(s_\beta)}{n(s_\alpha)+n(s_\beta)} \mathrm{I}_{d_{s_\alpha\beta}\sim h}$ is a normalizing constant and is an estimate of the mean of $\mathrm{Chol}_{\hat{\mathrm{R}}}(s_\alpha)$ expressed as the population weighted mean of the rates. Thus:

$$m^* = \frac{\sum_{\alpha=1}^{N} n(s_\alpha)\cdot \mathrm{Chol}_{\mathrm{R}}(s_\alpha)}{\sum_{\alpha=1}^{N} n(s_\alpha)}. \tag{4}$$

In Equation (3) the different pairs $[\mathrm{Chol}_{\mathrm{R}}(s_\alpha) - \mathrm{Chol}_{\mathrm{R}}(s_\beta)]$ are weighted by the corresponding population sizes $\frac{n(s_\alpha)\cdot n(s_\beta)}{n(s_\alpha)+n(s_\beta)}$ to homogenize their variance terms by dividing them by a weight proportional to the standard deviation $\sqrt{m\frac{n(s_\alpha)+n(s_\beta)}{n(s_\alpha)\cdot n(s_\beta)}}$. Monestiez *et al.* (2005, 2006) developed the above variogram to account for the spatially heterogeneous observation efforts and sparse animal sightings for mapping the relative abundance of species (*Balenoptera physalus*). In their approach, the heterogeneous distribution of the observation efforts was modelled as Poisson distribution. Simulation studies indicated that this approach outperforms simple population-weighted approaches and Bayesian smoothers (Goovaerts, 2005). Generalization of this approach for disease mapping can be seen in Goovaerts (2005). The approach, however, is similar to Oliver *et al.* (1998), except that Poisson distribution in Monestiez *et al.* approach replaces the Binomial distribution. In this study, the approach developed by Monestiez *et al.* is employed to model the spatial autocovariance structure of cholera variability in Kumasi. Next, a permissible variogram model by means of least squares $\gamma_{\mathrm{Chol}_{\hat{\mathrm{R}}}}(h)$ is fitted to the experimental variogram. From the fitted model, the maximum distance at which no spatial autocorrelation occurs (i.e. the range) is noted as d^{Th}.

2.3 Defining transmission network routes of cholera diffusion

Let $T = (t_1, \ldots, t_n)$ be a vector of serially ordered observed times of cholera onset at each community and ds_{ij} be the distance between pairs of communities s_i and s_j. The elements in the vector T are ordered such that $t_i \leq t_j \forall i < j$. Using the date of index case in each community, a pairwise $n \times n$ directional transmission matrix

$\overleftarrow{\mathrm{V}}=(\overleftarrow{\mathrm{v}}_{i,j})$ is constructed based on neighbourhood with previously infected community. The elements in the matrix $\overleftarrow{\mathrm{V}}=(\overleftarrow{\mathrm{v}}_{ij})$ represent the probability of transmission from spatial unit s_j to s_i with respect to time and distance.

First, a binary neighbourhood weight matrix $\omega=(\omega_{i,j})$ is defined, with elements representing the probability of transmission between pairs of communities with respect to distance from each other or threshold distance at which cholera is considered contagious, thus $\omega_{i,j}\in[0,1]$. Formally:

$$\omega_{i,j}=\begin{cases}1 & \text{if } ds_{i,j}\leq d^{\mathrm{Th}} \quad \forall i\neq j\\ 0 & \text{otherwise}\end{cases}. \tag{5}$$

Since no knowledge of the extent of contagiousness of cholera exists in the study area, the threshold distance d^{Th} is used as the threshold distance at which cholera is considered contagious.

Next, a directional matrix $\overleftarrow{\mathrm{Z}}=(\overleftarrow{\mathrm{z}}_{i,j})$ is defined, with elements representing the probability of a transmission from s_j to s_i with respect to time. More precisely, the elements in the matrix represent the probability that a spatial unit s_i with time of onset $t_{i\geq 2}$ can be infected by another spatial unit s_j with time of onset t; such that:

$$\overleftarrow{\mathrm{z}}_{i,j}=\begin{cases}1 & \text{iff } t_i>t_j \quad \forall t_{i\geq 2}\ \&\ i\neq j\\ 0 & \text{otherwise}\end{cases}. \tag{6}$$

Here the assumption is that $t_{i=1,}\sum_{j=1}^{n}\overleftarrow{\mathrm{Z}}_{i,j}=0$ that for the index case, i.e. for $t=1$.

The final transmission matrix $\overleftarrow{\mathrm{V}}=(\overleftarrow{\mathrm{v}}_{i,j})$ is defined based on an element-wise multiplication of the spatial neighbourhood matrix $\omega=(\omega_{i,j})$ and the directional matrix $\overleftarrow{\mathrm{Z}}=(\overleftarrow{\mathrm{z}}_{i,j})$. Thus, $\mathrm{V}=W\odot\overleftarrow{\mathrm{Z}}$. Formally:

$$v_{i,j}=\begin{cases}1 & \text{iff}\begin{cases}z_{i,j}=1,\forall i\neq j\\ \omega_{i,j}=1,\forall i\neq j\end{cases}\\ 0 & \text{otherwise}\end{cases}. \tag{7}$$

The matrix $\overleftarrow{\mathrm{V}}=(\overleftarrow{\mathrm{v}}_{i,j})$ is made up of several transmission trees with $1-1$ (transmission from one community to one community), $M-1$ (transmission from many communities to one community) and $1-M$ (transmission from one community to many communities) relationships. Theoretically, however, only $1-M$ and $1-1$ transmissions are feasible; therefore, nearest neighbor approach based on direct distance is chosen to extract all $1-1$ and $1-M$ transmissions. These are subsequently mapped with GIS to identify the various possible index cases and their locations.

2.4 Time-ordered diffusion modelling

This study hypothesizes that the time-ordered sequence of appearance of cholera patterns has a dynamic relationship with the urban level and proximity to the primary cases location of the diffusion system. As such, the urban population

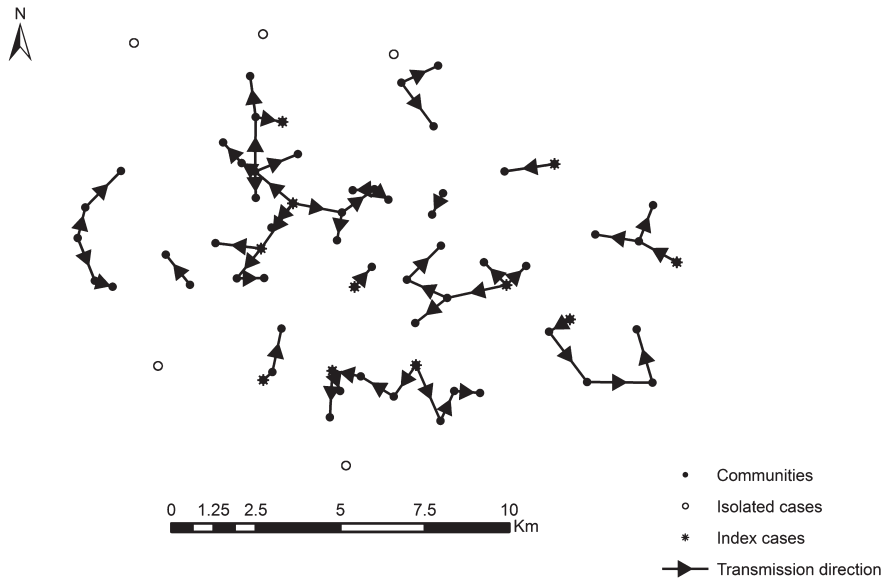

Fig. 2. The diffusion network patterns showing 1–M and 1–1 transmissions network routes.

represents the hierarchical component in the spread process, whereas the geographic distance from a *respective index case* represents the contagious component in the diffusion. The term *respective index case* is used because there are multiple index cases and each infected community corresponds to a particular index case. Here, the study adopts a nonlinear non-parametric Bayesian modelling approach for the effect of population and proximity on the diffusion of cholera.

Consider the observations $(y_i, x_i), i = 1, \ldots, n,$ with response $y_i = \ln(t_i)$, and t_i the time of cholera onset in communities $s_i \in 1, \ldots, S$. The time of cholera onset t_i is relative to a respective index case. The vector $x_i = (d(s_i), n(s_i))'$ contains two metrical covariates; the population size $n(s_i)$, and the direct distance from an infected community to a respective index case location $d(s_i)$ (See Figure 2).

The study assumes that the response variable follows Gaussian distribution, i.e. $y_i \mid \eta_i, \sigma^2 \sim N(\eta_i, \sigma^2/c_i)$, with unknown mean η_i of a nonparametric *geo-additive* model of the form:

$$\eta_i = f_n(n(s_i)) + f_d(d(s_i)) + f_{\text{str}}(s_i) + f_{\text{unstr}}(s_i). \tag{8}$$

Here, $f_n(n(s))$ and $f_d(d(s))$ are nonlinear smooth functions of the metrical covariates $n(s), d(s)$, respectively and $f_{\text{str}}(s)$ and $f_{\text{unstr}}(s)$ are the structured and unstructured spatial effects of the spatial covariates $s_i \in \{1, \ldots, S\}$.

To explore the dependence of the rate of cholera infection r on $d(s)$ and $n(s)$, we used the higher order moments of the frequency distribution of cholera infection against time (Cliff, Haggett and Ord 1986). Thus, the response variable $y = \ln(t)$

is replaced with $y = r = m_3^2/m_2^3$, where m_i is the ith central moment about the mean (average) time to infection $\bar{t}$. For each community, $\bar{t}$ is defined as:

$$\bar{t} = \frac{1}{n}\sum_{t=1}^{t_n} t\mathrm{Chol}_t,$$

where Chol_t is the number of cholera cases at time t, t_n is the number of days of cholera existence, and $n = \sum \mathrm{Chol}_t$ for all t. The ith central moment about $\bar{t}$ is expressed as:

$$m_i = \frac{1}{n}\sum_{t=1}^{t_n} (t-\bar{t})^i \mathrm{Chol}_t.$$

The rate of infection r was evaluated for all communities, excluding communities for which cases where recorded in only one day of the epidemic period. The model in equation (8) was fitted for the set of communities for which r was available.

2.4.1 Prior assumptions

The unknown model parameters are estimated by a fully Bayesian approach. Prior assumptions for the smooth functions $f_n(n(s))$ and $f_d(d(s))$ are specified. First, second order random walk prior is imposed on the function evolutions $f_n(n(s))$ and $f_d(d(s))$. Following Lang and Brezger (2004), we suppose that $x_{(1)} < \cdots < x_{(t)} < \cdots < x_{(m)}$ are m ordered distinct values with equally spaced observations $x_i, i = 1, \ldots, m$ with $m \leq n$ that are observed for the covariates x and define $\xi_t = f_j(x_{(t)})$. Then $f_j(x)$ can be written as $f_j(x) = \upsilon'\xi$, where υ is a 0/1 incidence vector taking the value of one if $x = x_{(t)}$ and zero otherwise, and $\xi = (\xi_1, \ldots, \xi_m)'$ is a vector of regression coefficients. The first order random walk prior for non-equidistance observations of adjacent values is defined as:

$$\xi_t = \xi_{t-1} + u_t, \quad t = 2, \ldots, m \tag{9}$$

with Gaussian distributed error terms $u_{(t)} \sim N(0, \delta_t\tau^2)$, where the variance depends on $\delta_t = x_{(t)} - x_{(t-1)}$. Random walks of the second order are defined by:

$$\xi_t = \left(1 + \frac{\delta_t}{\delta_{t-1}}\right)\xi_{t-1} - \frac{\delta_t}{\delta_{t-1}}\xi_{t-2} + u_t. \tag{10}$$

Here, $u_t \sim N(0, w_t\tau^2)$, where the weights w_t define the variances of the random walks. In this study we chose the simplest approach, where $w_t = \delta_t$.

A first order random walk penalizes abrupt jumps $\xi_t - \xi_{t-1}$ between successive states while a second order random walk penalizes deviations from the linear trend $2\xi_{t-1} - \xi_{t-2}$. The joint distribution of the regression parameters $\xi_j = (\xi_1, \ldots, \xi_m)'$ is computed as the product of conditional densities defined by Equation (11). Diffuse priors $\xi_1 \propto \mathrm{const}$, or ξ_1 and $\xi_2 \propto \mathrm{const}$, are chosen as initial values respectively. These specifications act as smoothness priors that penalize too rough functions. The

general form of the prior for ξ_j is a multivariate Gaussian distribution with density

$$p(\xi_j \mid \tau_j^2) \propto \exp\left(\frac{-\xi_j' K \xi_j}{2\tau_j^2}\right). \tag{11}$$

The penalty matrix of order k is of the form $K = D_k' D_k$ where D_k is a first or second order difference matrix for k = 1 or 2 respectively. Since the penalty matrix K is often not a full rank, it follows that $\xi_j \mid \tau_j^2$ is improper Gaussian prior, $\xi_j \mid \tau_j^2 \propto N(0; \tau_j^2 K^-)$, where K^- is a generalized inverse of K. The tradeoff between flexibility and smoothness is controlled by the variance parameter τ_j^2. Thus, a small (large) value of τ_j^2 correspond to an increase (decrease) of the penalty or shrinkage. Here, a weakly informative inverse Gamma prior $IG(a; b)$ with hyper-parameters for τ_j^2 is used.

For the structured spatial effects, the neighbourhood matrix $\omega = (\omega_{i,j})$ is modelled as a Gaussian random filed prior (Besag, York and Mollie, 1991; Rue and Held, 2005). This prior is defined by the conditional distribution of $\omega = (\omega_{i,j})$. Spatial units near the edges of the study area are likely to have fewer neighbours than those in the centre of the study area. Estimates of spatial units near the edges are less reliable than estimates of spatial units in the centre of the study area as fewer neighbours may distort any estimates for spatial units near the edges, the so called *edge effects*. To reduce edge effects, the conditional mean of $f_{\text{str}}(s_i)$ is chosen to be a weighted average of the function evaluations of neighbouring spatial units, with weighting scheme based on the proportion of the number of observed neighbour. Thus:

$$f_{\text{str}}(s_i) \mid f_{\text{str}}(s_j), s_j \neq s_i, \tau^2 \sim N\left(\sum_{s_j \in \partial s_i} \frac{w_{ij}}{\sum_{s_j \in \partial s_i} w_{ij}} f_{\text{str}}(s_i), \frac{\tau^2}{\sum_{s_j \in \partial s_i} w_{ij}}\right), \tag{12}$$

where $s_j \in \partial s_i$ denotes that spatial unit s_j is a neighbour of spatial unit s_i. Here, the weights $w_{ij} = \frac{|\partial s_i|}{N_s}$, where $|\partial s_i|$ is the number of neighbours of spatial unit s_i and N_s is the total number of spatial units. The design matrix ψ of the spatial effects is 0/1 incidence matrix where the number of columns is equal to the number of spatial units. The variance parameter τ^2 controls the amount of smoothing of the spatial covariates and the degree of similarity.

For the unstructured effects, the parameters $f_{\text{unstr}}(s)$ are assumed to be *i.i.d.* Gaussian. Thus:

$$f_{\text{unstr}}(s_i) \mid \tau_{\text{unstr}}^2 \sim N(0; \tau_{\text{unstr}}^2). \tag{13}$$

In a fully Bayesian approach, the variance parameters τ_j^2, $j = n, d, \text{str}, \text{unstr}$ are also considered as unknown and estimated simultaneously with the corresponding unknown functions $f_n(n(s))$, $f_d(d(s))$, $f_{\text{str}}(s)$, $f_{\text{unstr}}(s)$. Highly dispersed inverse gamma distribution IG$(a; b)$, with hyper-parameters are assigned to them in a second stage of the hierarchy.

2.4.2 Posterior estimation

Fully Bayesian inference is based on the posterior distribution of the unknown parameters. In this approach, samples are drawn from the full conditionals of the unknown parameters given the data through MCMC simulations. Let β represent the vector of all unknown functions to be evaluated and spatial effects (i.e., $\beta = (f_n(n(s)); f_d(d(s)); f_{\text{str}}(s); f_{\text{unstr}}(s))$ and τ represent a vector of all variance components; the posterior distribution then equals

$$p(\beta, \tau \mid y) \propto p\ (y \mid \beta)\ p\ (\beta \mid \tau) p(\tau), \tag{14}$$

where $p(y \mid \beta)$ is the likelihood function of the data given the parameters and $p(\cdot)$ represents the probability density function. Full conditionals for the unknown functions $f_n(n(s)), f_d(d(s)), f_{\text{str}}(s), f_{\text{unstr}}(s)$ are multivariate Gaussian and, as a consequence, a Gibbs sampler for MCMC simulation is employed. Cholesky decompositions for band matrices have been used to efficiently draw random samples from the full conditional (Rue and Held, 2005; Rue, 2000). The model has been implemented in public domain software for Bayesian analysis, BayesX ver 2.0 (Brezger, Kneib and Lang, 2005; Belitz *et al.*, 2009). We used a total number of 40,000 MCMC iterations and 10,000 number of burn-in samples. Since, in general, these random numbers are correlated, only every 20th sampled parameter of the Markov chain were stored.

3 Results

The population distribution in the study area is highly variable ranging from 587 to 56,417 people and standard deviation of approximately 13,506. Such spatially varying populations induced heteroscedasticity in the disease rates as well as non-stationarity in the variances. Consequently, the experimental variogram computed with the raw disease rates is uneven and exhibits less continuous patterns, depicting little or no spatial correlation and/or structure among communities (Figure 3b). This, however, necessitated an alternative to the Matheron's variogram estimator to characterize the spatial variability of the disease rates. Monestiez *et al.* variogram model can reveal structures that might be blurred by the random variability of extreme population values.

Experimental variograms were computed for the 68 community-level incidence rates using the traditional variogram (Equation 2) and the Monestiez *et al.* variogram models (Equation 3). The spatial variability is considered isotropic since no systematic differences are observed between the directional variograms; hence, only the omni-directional variograms are displayed in Figure 3. The traditional variogram model for cholera rates is exponential with a practical range of 1.36 km (Figure 3b); while the Monestiez *et al.* variogram model is spherical with a practical range of 2.3 km (Figure 3a). The relatively larger range of autocorrelation for the Monestiez *et al.* variogram model indicates a better spatial structure for

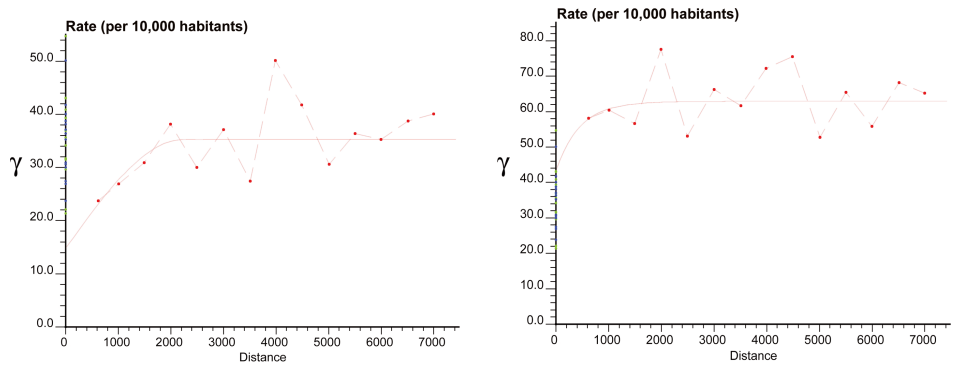

Fig. 3. Experimental variograms computed for the 68 community-level cholera incidence rates. (left) Monestiez *et al.* variogram model and (right) traditional variogram model.

cholera rates after heterogeneity in population distribution is accounted for. Therefore $d^{\mathrm{Th}} = 2.3$ km is used for the subsequent analysis.

Figure 2 shows the transmission networks routes of cholera diffusion in Kumasi. The red spots show the location of primary cases and starting points of different diffusion systems. The primary case locations are shown to be scattered, occurring simultaneously at distant locations. Twelve main primary cases are indentified, each corresponding to a different diffusion system. The largest diffusion system involves 19 communities and recorded approximately 37% of cholera cases during the outbreak period. From the transmission network routes, five isolated communities are observed. These are not included in any of the diffusion systems (Figure 2). The geographic locations of the primary cases are used for modelling the effect of $d(s)$ on $\ln(t)$ and r.

The nonlinear effects of the metrical covariates $n(s)$ and $d(s)$ on $\ln(t)$ and r are shown in Figure 4. The effect of $n(s)$ on $\ln(t)$ is nonlinear with decreasing posterior mean (Figure 4a). For $d(s)$, the posterior mean increases with increasing $\ln(t)$ (Figure 4b). The effect of $n(s)$ on r is almost linear with increasing posterior mean (Figure 4c). No systematic relationship is observed between r and $d(s)$ at $d(s) \leq 2.4$ km (Figure 4d). Thus, at $d(s) \leq 2.4$ km, the relationship between r and $d(s)$ is fixed with neither decreasing nor increasing effect. At $d(s) > 2.4$ km, however, a decreasing relationship is observed between r and $d(s)$.

Similar spatial patterns is observed for both $\ln(t)$ and r, hence only the patterns exhibited by r is shown. Figure 5 shows the estimated total spatial effects (left) and the corresponding 80% (credible interval) posterior probability map (right) of r. Areas shaded black show strictly negative credible intervals, whereas white areas depict strictly positive credible intervals, and grey indicate areas of non-significant spatial effects. There is a considerable spatial variation in the rate of cholera infection. Major spatial effects are observed at central part of the study area. The structured and unstructured spatial effects are given by the caterpillar plots in Figure 6. The wider variations in the caterpillar plots of Figure 6a compared with

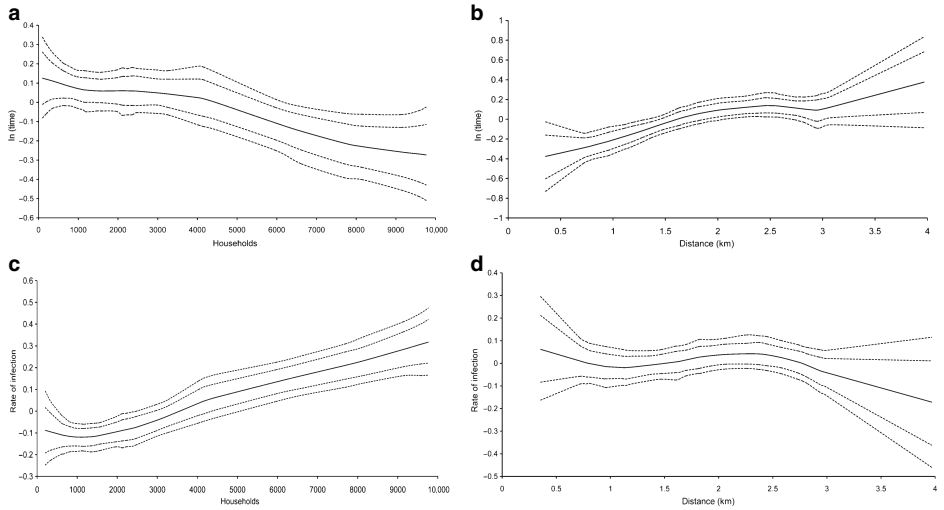

Fig. 4. The estimated nonlinear effects of the metrical covariates (a) $n(s)$ on $\ln(t)$, (b) $d(s)$ on $\ln(t)$, (c) $n(s)$ on r, (d) $d(s)$ on r. The posterior mean together with the 80% and 90% credible intervals are also shown.

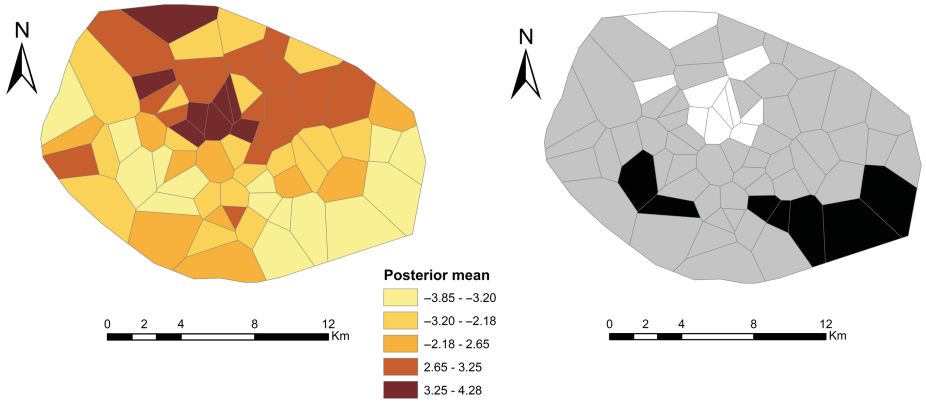

Fig. 5. Spatial distribution of the posterior means of the total spatial effects of modeling the effects $n(s)$ and $d(s)$ on $\ln(t)$ (left), and posterior probabilities at nominal level of 80% (right). (left) Black denotes areas with strictly negative credible intervals; white denotes areas with strictly positive credible intervals, whereas grey shows areas of no significant difference.

Figure 6b show that the spatially structured effects are dominant over the unstructured effects.

4 Discussion

This study utilizes statistical methods to explore the space-time diffusion dynamics of cholera incidences in Kumasi-Ghana. Variogram models are used to characterize

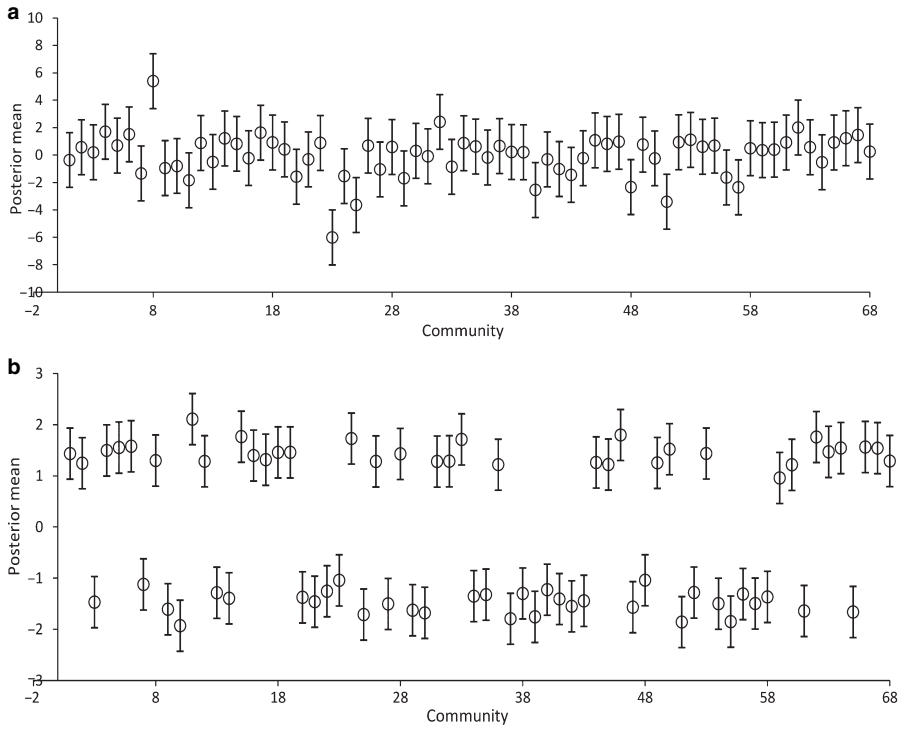

Fig. 6. Caterpillar plots of the posterior means of the structured (a) and unstructured (b) spatial effects in Figure 5, with 90% error bars.

the threshold of contagiousness of cholera. This threshold is subsequently used with the times of cholera entrance in each community to characterize all probable primary cases and diffusion systems during the outbreak. Finally, a hierarchical Bayesian modelling approach is used to explore the space-time diffusion dynamics of cholera in Kumasi.

Several primary cases have been identified, each corresponding to a different diffusion system. This is an indication that the transmission of cholera during epidemic situations can start from several sources. This confirms the fact that primary transmission of cholera is responsible for sparking initial outbreaks. The primary case locations have been shown to be scattered, occurring almost simultaneously in distant areas with no apparent connection. This may be explained by the fact that *V. cholera* concentration in the environment is dominated by environmental drivers and the stochastic nature of these processes (Pascual, Koelle and Dobson, 2006).

The findings also imply that communities proximal to primary case locations are infected relatively early during the epidemics, with more distant communities infected at increasing later dates. Similarly, more populous communities are infected relatively early, with less populous communities infected at increasingly later dates. The plausibility of these implications could be explained by: (i) the existing

hypothesis about the propagation of cholera and (ii) the mode of cholera spread in a population and the demographic structure of the study area. Firstly, cholera diffuses contagiously between surrounding communities following order of social interactions and/or geographic proximity (Pyle, 1969; Smallman-Raynor & Cliff, 1998a, b; Trevelyan *et al.*, 2005; Stock, 1976). Thus, it is likely for the disease to propagate from its origin to proximal communities earlier than communities which are farther away.

Secondly, cholera is a disease of deficient sanitation (Ackers *et al.*, 1998) which is primarily driven by environmental (Huq *et al.*, 2005) and demographic factors (Borroto and Martinez-Piedra, 2000; Osei and Duker, 2008). High population density puts pressure on existing sanitation systems, thus increasing the risk of early cholera infection in populous communities than less populous communities during an epidemic period.

When the response variable $y = \ln(t)$ is replaced by $y = r$, an expected observation is that r increases with $n(s)$, with an almost linear relationship. Thus, the effect of $n(s)$ on r is almost fixed. Such a relationship is plausible because in highly populous communities, many people live close together which results in shorter disease transmission paths; and therefore, higher rate of cholera infection. The passage of *V. cholerae* through the human host transiently increases the infectivity of *V. cholera*. Therefore, the existence of short-lived hyper-infective stage of *V. cholerae* could provide a mechanism for exhibiting a strong feedback between the past and present levels of infection (Merrell *et al.*, 2002; Hartley, Morris and Smith, 2005), especially in a population where faecal contamination of water sources is high. The rate of exposure to short-lived hyper-infective *V. cholerae* could also be dominated by spatial interactions among infected communities, population density and/or urban level (Pyle, 1969; Smallman-Raynor and Cliff, 1998a, b; Stock, 1976). Codeço (2001) reports that in an epidemic situation the initial reproduction rate of secondary cases is positively affected by the degree of contamination of water supply as well as the frequency of contacts with these waters, which is in turn influenced by demographic factors such as population density. The non systematic relation between r and $d(s)$ at approximately $d(s) \leq 2.4$ km may be explained by the contagious nature of cholera. Cholera is communicable which spreads contagiously amongst inhabitants from one community to another. Since proximal communities tend to exhibit similar socioeconomic and environmental characteristics, similar rate of cholera infection may be observed. Thus, for $d(s) \leq 2.4$ km, no systematic relationship is observed between $r$and $d(s)$. The negative relationship, however, between r and $d(s)$ amongst distant communities, i.e. $d(s) > 2.4$ km m, is seemingly grounds for questioning the acceptance of this hypothesis.

This study cannot conclude that only the covariates $d(s)$ and $n(s)$ influence $\ln(t)$ and r due to the possibility of other unobserved influential covariates. Therefore, the inclusion of $f_{\text{str}}(s)$ and $f_{\text{unstr}}(s)$ in the model is meant to mimic the nature of unobserved influential covariates on $\ln(t)$ and r. From Figure 5, there is evidence of significant increased rate of cholera infection at the central part, and a significant

reduced rate of infection at the south-eastern part (the periphery) of Kumasi. The plausibility of these patterns may be explained by the fact that communities at the central part are highly populated with lots of slum settlers, while communities at the peripheries are moderately populated. As a consequence, shorter disease transmission path (higher rate of infection) is expected at the central part and longer disease transmission path is expected at the peripheries (lower rate of infection). These patterns also indicate the existence of possible unobserved covariates, some of which may be individual or household level, giving leads for further epidemiological research using purpose collected data.

Although several of these findings confirm existing hypothesis of cholera, this study resolves the methodological deficiencies of exploring the space-time diffusion patterns of infectious diseases. For instance in Trevelyan *et al*. (2005), a strictly linear model is imposed on the relationship between the period of observation of poliomyelitis and the joint effects population and distance from the epidemic centre. Also, in Kuo and Fukui (2007) the time-ordered cholera diffusion sequence is modelled as a linear logarithmic regression model, which is the functional relationship between the residents of infected counties and distances from epidemic origins. The strictly linear effects imposed in such models can obscure important nonlinear effects. Moreover, such models also underestimate important effects of spatial interactions amongst communities on the space-time diffusion patterns of infectious diseases.

5 Conclusion

This study applies statistical methods to explore the space-time diffusion patterns of cholera in Kumasi. We use hierarchical Bayesian modelling approaches which allow joint analysis of the nonlinear effects of population hierarchy and geographic proximity on cholera infection. Our study reveals that the time-ordered sequence of appearance of cholera in a community has a dynamic relationship with the population hierarchy and proximity to primary case locations. Likewise, the rate of cholera infection increases with high population density. The geographic proximity to a primary case location, however, does not influence the rate of cholera infection. These findings provide significant information to help health planners and policy makers about the dynamics of cholera spread amongst communities.

Acknowledgements

We extend our sincere appreciation to the Kumasi Metropolitan Health Directorate for providing all the necessary data and background information for this research. We also extend our appreciation to Ellen-Wien Augustijn for her fruitful comments and suggestions.

References

Ackers, M.-L., R. E. Quick, C. J. Drasbek, L. Hutwagner and R. V. Tauxe (1998), Are there national risk factors for epidemic cholera? The correlation between socioeconomic and demographic indices and cholera incidence in Latin America, *International Journal of Epidemiology* **27,** 330–334.

Belitz, C., A. Brezger, T. Kneib and S. Lang (2009), BayesX – Software for Bayesian inference in structured additive regression models, Version: 2.0 Available at: http://www.stat.uni-muenchen.de/∼bayesx. Last accessed: 3 November 2010.

Besag, J., Y. York, and A. Mollie (1991), Bayesian image-restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Borroto, R. J. and R. Martinez-Piedra (2000), Geographical patterns of cholera in Mexico, 1991–1996, *International Journal of Epidemiology* **29**, 764–772.

Brezger, A., T. Kneib and S. Lang (2005), BayesX: Analyzing Bayesian structured additive regression models, *Journal of Statistical Software* **14**, 11.

Carpenter, C. C. J. (1970), Principles and practice of cholera control, *Public health papers* WHO No. 40, Geneva.

Cliff, A. D., P. Haggett and J. K. Ord (1986), *Spatial aspects of influenza epidemics*, Pion, London.

Codeço C. T. (2001) , Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir, *BMC Infectious Diseases* **1**, 1.

Colwell, R. R. and A. Huq (1994), Vibrios in the environment: viable but nonculturable vibrio cholerae, in: *Vibrio cholera and cholera: molecular to global perspectives* I. Wachsmoth, P. Blaker and O. Olsvik, (eds), American Society for Microbiology, Washington, DC, pp. 117–134.

Finkelstein, R. A. (1996), Cholera, Vibrio cholerae O1 and O139 and other pathogenic vibrios, in: S. Baron (ed.), *Medical Microbiology*. 4th ed. Churchill Livingstone, New York.

Glass, R., M. Claeson, P. Blake, R. Waldman and N. Pierce (1991), Cholera in Africa: Lessons on transmission and control for Latin America, *Lancet* **338**, 791–795.

Goovaerts, P. (2005), Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging, *International Journal of Health Geographics* **4**, 31.

Hartley, D. M., M. Morris, and D. L. Smith (2005), Hyperinfectivity: a critical element in the ability of V. cholerae to cause epidemics, *PLoS Medicine* **3**, e7.

Huq A., R. B. Sack, A. Nizam, I. M. Longini, G. B. Nair, A. Ali, J. G. Morris, M. N. H. Khan, A. K. Siddique, M. Yunus, M. J. Albert, D. A. Sack, and R. R. Colwell (2005), Critical factors influencing the occurrence of Vibrio cholerae in an environment of Bangladesh, *Applied and Environmental Microbiology* **71**, 4645–4654.

Kuo, C. L. and H. Fukui (2007), Geographical structures and the cholera epidemic in modern Japan: Fukushima prefecture in 1882 and 1895, *International Journal of Health Geographics* **6**, 25.

Lang, S. and A. Brezger (2004), Bayesian P-splines, *Journal of Computational and Graphical Statistics* **13**, 183–212.

Matheron, G. (1963), Principles of geostatistics, *Economic Geology* **58**, 1246–1266.

Matheron, G. (1965), *Les variables re'gionalise'es et leur estimation*, Masson, Paris.

Merrell, D. S., S. M. Butler, F. Qadri, N. A. Dolganov, A. Alam, M. B. Cohen, S. B. Calderwood, G. K. Schoolnik, and A. Camilli (2002), Host-induced epidemic spread of the cholera bacterium, *Nature* **417** (6889): 642–645.

Miller, C. J., R. G. Feachem, and B. S. Drasar (1985), Cholera epidemiology in developed and developing countries new thoughts on transmission, seasonality, and control, *Lancet* **1**, 261–263.

Monestiez, P., L. Dubroca, E. Bonnin, J. P. Durbec and C. Guinet (2005), Comparison of model based geostatistical methods in ecology: application to fin whale spatial distribution in northwestern Mediterranean sea, in: *Geostatistics Banff 2004*, Vol. **2**, O. Leuangthong,

and C. V. Deutsch (eds). Kluwer Academic Publishers, Dordrecht, The Netherlands, 777–786.

Monestiez, P., L. Dubroca, E. Bonnin, J. P. Durbec and C. Guinet (2006), Geostatistical modelling of spatial distribution of Balenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts, *Ecological Modeling* **193** (3–4): 615–628.

Oliver, M. A., R. Webster, C. Lajuanie, K. R. Muir, S. E. Parkes, A. H. Cameron, M. C. G. Stevens and J. R. Mann (1998), Binomial cokriging for estimating and mapping the risk of childhood cancer, *IMA Journal of Mathematics Applied in Medicine and Biology* **15**, 279–297.

Osei, F. B. and A. A. Duker (2008), Spatial and demographic patterns of cholera in Ashanti Region-Ghana, *International Journal Health Geographics* **7**, 44.

Pascual, M., K. Koelle and A. P. Dobson (2006), Hyperinfectivity in cholera: a new mechanism for an old epidemiological model, *PLoS Medicine* **3**, e280.

Prestero, T., M. Height, and R. Hwang (2001), *Rapid cholera treatment: exploring alternative IV treatment devices*, MIT Media Lab-Development by Design Workshop.

Pyle, G. F. (1969), The diffusion of cholera in the United States in the nineteenth century, *Geographical Analysis* **1**, 59–79.

Rue, H. (2000), Fast sampling of Gaussian Markov random fields with applicatins, *Journal of the Royal Statistical Society B* **63**, 325–338.

Rue, H. and L. Held (2005), Gaussian markov random fields: theory and applications, *Monographs on statistics and applied probability 104*, CRC/Chapman and Hall, Boca Raton.

Sack, D. A., R. B. Sack, G. B. Nair and A. K. Siddique (2004) Cholera, *The Lancet* **365**, 1. Avaliable at: http://www.thelancet.com, accessed 17 January 2004.

Smallman-Raynor, M. and A. Cliff (1998a), The Philippines insurrection and the 1902-4 cholera epidemic: part 1- epidemiological diffusion process in war, *Journal of Historical Geography* **24**, 69–89.

Smallman-Raynor, M. and A. Cliff (1998b), The Philippines insurrection and the 1902-4 cholera epidemic: part 2- diffusion patterns in war and peace, *Journal of Historical Geography* **24**, 188–210.

Snow, J., W. H. Frost, and B. W. Richardson (1936), *Snow on cholera*, Commonwealth Fund, New York.

Stock, R. F. (1976), *Cholera in Africa-Diffusion of the Disease 1970–1975 with particular emphasis on West Africa*, African Environment Special Report 3, International African Institute, Plymouth, 1–21.

Trevelyan, B., M. Smallman-Raynor, and A. D. Cliff (2005), The spatial structure of epidemic emergence: geographical aspects of poliomyelitis in north-eastern USA, July–October 1916, *Journal of the Royal Statistical Society Series A* **168**, 701–722.

WHO (1993), *Guidelines for cholera control-Geneva*, World Health Organization, Geneva, p.1.

Received: April 2010. Revised: July 2010.