A zero-inflated mixture spatially varying coefficient modeling of cholera incidences

Frank Badu Osei a,∗,1, Alfred Stein a, Veronica Andreo b,c

a Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Netherlands
b Instituto de Altos Estudios Espaciales “Mario Gulich” (IG). Comisión Nacional de Actividades Espaciales (CONAE) – Universidad Nacional de Córdoba (UNC), Argentina
c Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

A R T I C L E   I N F O

Article history:
Received 7 May 2021
Received in revised form 18 December 2021
Accepted 2 February 2022
Available online 9 February 2022

Keywords:
Cholera
Zero-Inflated Poisson
Zero-Inflated Negative Binomial
Bayesian
Spatially varying coefficients
Poisson

A B S T R A C T

Spatial disease modeling remains an important public health tool. For cholera, the presence of zero counts is common. The Poisson model is inadequate to (1) capture over-dispersion, and (2) distinguish between excess zeros arising from non-susceptible and susceptible populations. In this study, we develop zero-inflated (ZI) mixture spatially varying coefficient (SVC) models to (1) distinguish between the sources of the excess zeros and (2) uncover the spatially varying effects of precipitation and temperature (LST) on cholera. We demonstrate the potential of the models using cholera data from Ghana. A striking observation is that the Poisson model outperformed the ZI mixture models in terms of fit. The ZI Negative Binomial (ZINB) outperformed the ZI Poisson (ZIP) model. Subject to our objectives, we make inferences using the ZINB model. The proportion of zeros estimated with the ZINB model is 0.41 and exceeded what would have been estimated using a Poisson model which is 0.35. We observed the spatial trends of the effects of precipitation and LST to have both increasing and decreasing gradients; an observation implying that the use of only the global coefficients would lead to wrong inferences. We conclude that (1) the use of ZI mixture models has epidemiological significance. Therefore, its choice over the Poisson model should be based on an epidemiological concept rather than model fit and, (2) the extension of ZI mixture models to accommodate spatially

∗ Corresponding author.
E-mail addresses: f.b.osei@utwente.nl, oseibadu2004@yahoo.co.uk (F.B. Osei), a.stein@utwente.nl (A. Stein), veronica.andreo@ig.edu.ar (V. Andreo).

1 PhD.

https://doi.org/10.1016/j.spasta.2022.100635
2211-6753/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Introduction

Spatial modeling of diseases remains a vital public health surveillance tool for monitoring and evaluation of health challenges. After Snow’s seminal work on cholera in the 1850s, the use of spatial information to explore epidemics has been widely adopted. In a cholera endemic area of Bangladesh, spatial epidemiologic tools have severely been prominent for risk mapping and risk factor evaluation (Ali et al., 2006, 2002). Similarly in Africa, the same tools have been adopted (Lessler et al., 2018; Osei and Duker, 2008; Sasaki et al., 2008).

Cholera has been a major threat to the health of vulnerable populations, especially in developing countries where access to safe water and sanitation is limited. Major outbreaks are currently nearly limited to its new homeland in Africa. While concerted efforts are being made by various countries to increase access to safe water and sanitation, the achievement of this is most unlikely in the next few years since 2.1 billion and 4.5 billion of the world’s population still lack access to safe water and sanitation, respectively (WHO, 2017). Knowledge of the spatial epidemiologic dynamics, however, could lead to focused intervention strategies to manage scarce resources.

A challenge that arises with mapping area-level cholera incidences is the presence of excess zeros arising out of the absence of risk factors or nonreporting. In such circumstances, the data is said to be zero-inflated (ZI). In the presence of excess zeros, the Poisson spatial model could have limited ability to account for the over-dispersion. Statistical literature provides two strategies for modeling counts with excess zeros. These are the ZI mixture (Lambert, 1992) and ZI hurdles models (Heilbron, 1994; Mullahy, 1986). Each of them has had wide applications in epidemiology, including modeling cancer (Baetschmann and Winkelmann, 2013; Musio et al., 2010) and other infectious diseases (Musal and Aktekin, 2013; Neyens et al., 2017; Nyandwi et al., 2020). ZI modeling has a long history (Greene, 1994; Lambert, 1992); and its application to spatial counts has been documented in Agarwal et al. (2002). Disease mapping literature has seen many applications of ZI models (e.g. Neyens et al., 2017; Torabi, 2017). Some cholera studies have also utilized ZI models (Ali et al., 2017; Bwire et al., 2017). For instance, Carrel et al. (2011) and Root et al. (2013) applied a ZI mixture model to cholera data but ignored the spatial random effects.

Both ZI mixture and hurdle models have two components concerning disease modeling: there is a binary component that models the probability of onset (probability of observing at least one disease case) in a non-susceptible sub-population, and a count component that models the number of incidences in a susceptible population. Hurdle models consist of a binary component and a truncated count component, implying that only non-zero counts are expected once the hurdle is crossed. ZI mixture models on the other hand consist of a binary component and a non-truncated count component. This implies that after transitioning from the binary component, a subset of the population, being the susceptible subpopulation, still has zero occurrences. In contrast to hurdle models, ZI mixture models distinguish between two kinds of zeros. Statistical literature refers to these as structural zeros that occur because the disease is absent and sampling zeros that occur by chance. In epidemiological terms, sampling zeros can be interpreted as those incidences that never got captured or reported due to weak disease surveillance. In this manuscript, we refer to the structural zeros as those zeros that occur in the non-susceptible population. Sampling zeros on the other hand are referred to as those zeros that occur in the susceptible population. For cholera, ZI mixture inference is reasonable because asymptomatic or mild infections may lead to nonreporting to the health facilities even amongst the susceptible populations (Root et al., 2013). Thus, an observed zero can also imply a missing observation.
Since cholera remains a public health threat and therefore needs urgent evaluation of its spatial and environmental determinants, some unique challenges require consideration during modeling. First is the abundance of zeros. Transitioning from the binary to the count component, there is the need to select an appropriate base distribution to adequately capture dispersion in the counts. Therefore, we consider two base distributions: the Poisson and Negative Binomial distributions. Second, one of our aims is to distinguish between the zeros in the non-susceptible population (those that occurred due to the absence of cholera) and those in the susceptible population (those that might have occurred due to non-reporting). There is an epidemiological interest in this. Self-treatment with oral rehydration or other methods, or seeking medical attention outside the region of residence can lead to zero cases. In this sense, it is reasonable to consider the cholera data as being generated from the ZI mixture distribution rather than the ZI hurdle distribution. Third, given the complex nature of the relationship between cholera and the environment, modeling the risk factors as fixed is inadequate. A fixed-effect model averages the relationship into a single estimate by overlooking possible spatially varying effects across the various spatial entities under study. Geographically weighted regression (GWR) (Fotheringham et al., 2002) and spatially varying coefficient (SVC) models are common methods to capture possible spatially varying effects (Assunção, 2003; Finley, 2011). Finley (2011) has highlighted the strengths and weaknesses of GWR and SVC in a simulation study within a continuous spatial domain. Dambon et al. (2021) have recently shown how a Gaussian process-based SVC model can be estimated using maximum likelihood estimation (MLE). In this study, we seek also to extend ZI mixture models to accommodate non-stationary covariates effects within a discrete spatial domain. Thus, a ZI mixture SVC model within a Bayesian framework is proposed. In summary, this article describes a ZI mixture SVC cholera incidences models that address the unique challenges described above. We dwell on the 2014 cholera outbreak in Ghana as this fits the description of cholera incidences with excess zero counts (see Fig. 1). Using the probability of infection and incidence rate maps, we evaluate and discuss how geographical targeting could lead to efficient interventions. In doing so, a spatial risk map of cholera could serve as an important proxy for monitoring and evaluation. Hereafter, we describe the statistical methodology and its application to cholera data. We further present the results and discuss their implications for public health interventions.

2. Statistical modeling

We consider the observed cholera incidences \( y_i \) in \( m \) districts, each with population \( n_i \). Let \( p_i \) be the Bernoulli probability of reporting at least one case or the probability of non-zero response for district \( i \). We also refer to the probability of non-zero response \( p_i = \Pr^ZI(y_i > 0) \), as the onset probability in epidemiological terms. We assume that each observation is generated from a random variable whose probability density function is a mixture of the count and Bernoulli distributions

\[
\Pr^ZI(y_i | r_i, p_i) = \begin{cases} 1 - p_i + p_i \Pr^C(0 | r_i), & y_i = 0 \\ p_i \Pr^C(y_i | r_i), & y_i > 0 \end{cases}
\]

where \( \Pr^C(y_i | r_i, p_i) \) is a count base probability distribution and \( \Pr^C(0 | r_i, p_i) \) is the probability of observing zero from the count distribution. The likelihood function is thus the product of the individual likelihoods

\[
L(y_i | r_i, p_i) = \prod_{i=1}^{n} \left[ \left( 1 - p_i + p_i \Pr^C(0 | r_i) \right) I_{(y_i = 0)} + p_i \Pr^C(y_i | r_i) I_{(y_i > 0)} \right]
\]

The common base distributions for event counts are the Poisson and Negative Binomial functions, resulting in ZI Poisson (ZIP) or ZI Negative Binomial (ZINB) models. Suppose the data are generated by either the ZIP or ZINB processes, then their likelihood functions are given respectively as

\[
L(y_i | r_i, p_i) = \prod_{i=1}^{n} \left[ \left( 1 - p_i + p_i e^{-\lambda_i} \right) I_{(y_i = 0)} + p_i e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} I_{(y_i > 0)} \right]
\]
Fig. 1. The distribution of 2014 cholera incidences in Ghana. Top-left: Spatial distribution of raw incidences. Top-right: Spatial distribution of standardized incidence ratio (SIR). Bottom-right: Frequency distribution of the raw incidences.

\[
L(y_i|r_i, p_i, \omega) = \prod_{i=1}^{n} \left\{ 1 - p_i + p_i \left( \frac{\omega}{\omega + \lambda_i} \right)^\omega \right\} I_{y_i=0} \\
+ \left\{ p_i \frac{\Gamma(\omega + y_i)}{\Gamma(\omega) \Gamma(y_i + 1)} \left( \frac{\omega}{\omega + \lambda_i} \right)^\omega \left( \frac{\lambda_i}{\omega + \lambda_i} \right)^{y_i} \right\} I_{y_i>0}
\]

where \( \lambda_i = E_i \tilde{r} \) is the latent mean of the count component and \( r_i \) is the unknown relative risk. The expected number of cases is \( E_i = n_i \tilde{r} \) after accounting for district-level population differentials \( n_i \), where \( \tilde{r} = \sum_i y_i / \sum_i n_i \) is the overall risk within the study area. These choices of data generating processes relax the restrictive assumption of equivalent mean and variance such that the marginal expectation \( E(y_i|r_i, p_i) = p_i \lambda_i \) and the variance \( var(y_i|r_i, p_i) = p_i \lambda_i + p_i (1-p_i) \lambda_i^2 \) for the ZIP model. Thus, there is extra Poisson variation or over-dispersion when \( 1-p_i > 0 \). The ZIP model reduces to the standard Poisson model when \( p_i = 1 \). For the ZINB model, the expectation \( E(y_i|r_i, p_i, \omega) = p_i \lambda_i \) and variance \( var(y_i|r_i, p_i, \omega) = p_i \lambda_i + (1-p_i + 1/\omega) p_i \lambda_i^2 \). The difference is that the ZINB density has an extra parameter \( \omega \) to account for extra Poisson variation. Thus, the ZINB model is particularly suited if the count data part of the response distribution is over-dispersed. For brevity, we write the ZIP and ZINB models as \( y_i \sim ZIP(\lambda_i, p_i) \) and \( y_i \sim ZINB(\lambda_i, p_i, \omega) \), respectively. A special case of the ZINB model is the Poisson–Gamma model \( y_i \sim ZIP(\lambda_i, v_i, p_i) \) where the random effects are generated from a Gamma density of equal shape and scale parameters \( v_i | \omega \sim Ga(\omega, \omega) \) such that \( E(v_i|\omega) = 1 \) and \( var(v_i|\omega) = 1/\omega \). Thus, if we integrate the ZIP density \( y_i \sim ZIP(\lambda_i, v_i, p_i) \) with respect to \( v_i \) we get the ZINB density \( y_i \sim ZINB(\lambda_i, p_i, \omega) \).
To allow maximum flexibility, a model-based framework with suitable link functions is used to associate the onset probability $p_i$ and the latent mean $\lambda_i$ with covariates and random effects. First, we impose a Gaussian process with mean $\eta_{ri}$ (also termed as linear predictor) and variance $\sigma^2$ on the log of the relative risk. The variance $\sigma^2$ can be interpreted as representing extra-Poisson variation or over-dispersion in the counts (Tzala and Best, 2008). We further decompose the linear predictor into spatially varying covariate effects and spatially structured random effects. Thus

$$
\log \lambda_i = \log E_i + \log r_i
$$

$$
\log r_i \sim N(\eta_{ri}, \sigma^2)
$$

$$
\eta_{ri} = \alpha_r + \sum_q \beta_{qi}x_{qi} + u_{ri}
$$

$$
\logit p_i = \log E_i + \alpha_p + \sum_l \gamma_{li}z_{li} + u_{pi}
$$

Here, the expected number of cases $E_i$ is an offset term, $\alpha_r$ and $\alpha_p$ are the overall risk and probability of onset on the log and logit scales, respectively. The effects of parameters $\beta_{qi}$ are the spatially varying effects of the $q = 1, \ldots, Q$ covariates $x_{qi}$ for the risk, $\gamma_{li}$ are the spatially varying effects of the $l = 1, \ldots, L$ covariates $z_{li}$ for the probability of onset, and $u_{ri}$ and $u_{pi}$ are spatial random terms to account for spatial clustering. We call this log-Gaussian parameterization. An alternative common parameterization for the count component is:

$$
\log r_i = \alpha_r + \sum_q \beta_{qi}x_{qi} + u_{ri} + v_{ri}
$$

where the unstructured spatial heterogeneity $v_{ri}$ accounts for over-dispersion or random extra Poisson variation in the counts. We call this parameterization log-linear model. The log-Gaussian and log-linear models are inferentially the same. However, the MCMC implementation of the log-linear model resulted in convergence issues.

2.1. Process and prior layers

We now look at the random effects and covariates effects processes. We call these process layers. For our areal-level data (based on districts), assigning a Conditional Autoregressive (CAR) prior process (Besag and Kooperberg, 1995) to the structured spatial processes is arguably the best practice. An alternative is a Gaussian process which has received much attention for situations where the exact observation locations are known (Banerjee and Gelfand, 2002). In this study, we assign a CAR prior to the random processes $u_{ki}$, $k = \{r, p\}$, where the joint density is proportional to their pairwise differences

$$
\Pr(u_{k1}, \ldots, u_{kn}) \propto \frac{1}{\sigma_{ku}^2} \exp \left( -\frac{1}{2\sigma_{ku}^2} \sum_{i=1}^n \sum_{j<i} w_{ij} (u_{ki} - u_{kj})^2 \right).
$$

The $w_{ij}$ are the elements of the spatial adjacency matrix $W$, typically defined as $w_{ii} = 0$, $w_{ij} = 1$ if districts $i$ and $j$ are neighbors ($i \sim j$), and 0 otherwise. The joint density $p(u_k) = p(u_{k1}, \ldots, u_{kn})$ is a multivariate normal $u_k \sim MVN(0, \Sigma_{ku}^{-1})$ with variance $\sigma_{ku}^2$ dependent precision matrix $\Sigma_{ku}^{-1} = 1/\sigma_{ku}^2(D - W)$, where the diagonal matrix $D = \text{Diag}(1/\sum_{j<i} w_{ij})$. Despite the limitations in terms of its impropriety, this CAR structure is simple and easy to fit (Jin et al., 2005). For brevity, we write the univariate CAR model as $u_{ki} \sim \text{CAR}(\sigma_{ku}^2)$.

For the effects of the covariates, we assume non-stationary processes rather than the common fixed effects. Consider for instance those covariates of the risk $x_{qi}$, we write the SVC as $\beta_{qi} = \bar{\beta}_q + \hat{\beta}_{qi}$, where the fixed effect $\bar{\beta}_q$ is the overall mean effect and $\hat{\beta}_{qi}$ are spatially varying processes. Similar
to the spatial random intercepts, it is simpler to assign $\tilde{\beta}_{qi} \sim \text{CAR}(\sigma_{\tilde{\beta}}^2)$. The CAR model for $\tilde{\beta}_{qi}$ captures the residual spatial correlation between areas but ignores potential correlations between the random effects within the same area when there is more than one covariate. Alternatively, assigning a multivariate CAR (MCAR) to $\tilde{\beta}_{qi}$ can improve model fitting by accommodating correlation between the random effects. The univariate CAR extends directly to the MCAR by the inclusion of a symmetric positive definite correlation matrix $\Lambda$ of dimension equal to the number of covariates. We can write $\tilde{\beta}_{qi} \sim \text{MVN}(0, \Sigma_u^{-1})$. Here, the precision matrix becomes $\Sigma_u^{-1} = (D - W) \otimes \Lambda$, where $\otimes$ is a Kronecker product. For brevity, we write the MCAR as $\tilde{\beta}_{qi} \sim \text{MCAR}(\Lambda)$. Lastly, a prior layer consisting of prior distributions for the fixed effects and the variance parameters is needed to complete the model. The common prior for the fixed effects is a zero-centered normal density with a large variance to reflect a lack of information. The priors for the variance parameters are commonly assigned to their reciprocals via a Gamma density (Wakefield et al., 2001). The reciprocals are also termed precision parameters; thus, for instance, $\tau_{ku} = \sigma_{ku}^{-2} \sim \text{Ga}(a, b)$, where $a$ and $b$ are the shape and rate parameters, respectively. An alternative, suggested by Gelman (2006), is to assign a strictly positive half-normal (e.g. $\sigma_{ku} \sim N_{+\infty}(0, 10)$) or uniform (e.g. $\sigma_{ku} \sim U(0, 10)$) densities to their standard deviations.

3. Application to cholera data in Ghana

Ghana is centrally located on the west coast of Africa and bordered by Cote d’Ivoire to the west, Togo to the east, Burkina Faso to the north, and the Atlantic Ocean to the south. Cholera outbreaks have occurred in Ghana since the first appearance in the 1970s. In the year 2014, a massive cholera epidemic hit the West Africa sub-region. Ghana and Nigeria were amongst the five countries that together reported 84% of the worldwide cases. Although Ghana is amongst the 8 countries that reported a case fatality rate of <1%, it is sobering to report over 23000 cases of cholera in Ghana. In this study, we relied on the incidences reported for the 216 districts in Ghana. The data were acquired from the Centre for Health Information and Management (CHIM) of the Ghana Health Services (GHS). The dataset contained district aggregated counts of cholera cases for the 261 districts.

To proceed to apply the models described to cholera incidences from Ghana, we need to justify two issues. First, is the ZI model necessary? Out of the 216 districts in Ghana as of 2014, 93 of them reported no cholera incidence; thus, creating a high proportion of zeros. The number of non-zero incidences ranged from 10 to 4239 (see histogram in Fig. 1). The spatial distribution of the standardized incidence ratios (SIR), the ratio of the observed incidences to the expected incidences $SIR_i = y_i / E_i$, show clustering of elevated risks within the north-east, central, south-west, and southern parts of Ghana (Fig. 1). A similar pattern is observed for the incidences. The mean number of cases under the Poisson assumption is $\lambda = 108.05$. Therefore, the observed proportion of zeros, 0.43, far exceeds the proportion expected under the Poisson assumption. Second is the choice of the ZI mixture model over the ZI hurdle model. For the cholera incidence data, two kinds of zeros could arise as explained above. Those that occurred due to the absence of the disease (zeros in non-susceptible population) and those that occurred but never got reported (zeros in susceptible population). The latter can be due to individuals not seeking treatment at the hospitals, either because of self-treatment with oral rehydration or other methods. Since one of the objectives is to differentiate and estimate the occurrences of these two kinds of zeros, a ZI mixture model is conceptually appropriate.

To examine the performances of the ZI mixture models in the determination of significant environmental factors on the occurrences of cholera, we used precipitation and temperature variables. We averaged remote sensing derived products, precipitation ($x_{\text{prec}}$), and land surface temperature ($x_{\text{LST}}$), over each district as continuous covariates. We used MOD11A2 (ca. 1 Km spatial resolution, 8-day composites) product to obtain the LST estimates. The precipitation data were obtained from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) version 2.0 database, a product that combines satellite and meteorological stations’ rainfall data (Funk et al., 2014). Monthly (from June to December) time series of the variables were temporally averaged and spatially aggregated to match Ghana’s 216 districts (see Fig. 2).
We fitted Poisson and Negative Binomial versions of the ZI mixture model; thus, ZIP and ZINB model respectively. For the choice of the base distributions, we considered two sub-models: those with fixed probability of onset and those with spatially varying probabilities. For all the models, the log link function is used to associate the count component with the risk factors. Except for models with a fixed probability of onset, the logit link function is used for the binary component. To avoid identifiability issues, it is advisable to avoid the use of the same covariates in the count and the binary components (Lambert, 1992). Due to the physical relationship between cholera and the covariates, no covariates were included in the binary component. Exposure to predisposing conditions like precipitation and temperature variations can rarely trigger cholera onset. To trigger cholera onset in a subpopulation, one must be exposed to an aquatic reservoir of the cholera bacteria (primary infection) or exposed to an already infected person (secondary infection). Precipitation and temperature variations can increase cholera occurrences and progression rather than triggering an onset. Therefore, we do not include the precipitation and temperature variables, which are the only two covariates, in the binary components. The models, therefore, follow as:

Model 1: ZIP models

\[
y_i \sim ZIP(\lambda_i, p_i)\\
\lambda_i = E_i r_i \\
\log r_i \sim N(\eta_{ri}, \sigma_r^2)\\
\eta_{ri} = \alpha_r + \beta_{\text{prec},i} x_{\text{prec},i} + \beta_{\text{LST},i} x_{\text{LST},i} + u_{ri}\\
p_i \sim U(0, 1) \text{ or } p_i \sim \text{Beta}(1, 1) : \text{Only for Model 1a}\\
\text{logit } p_i = \alpha_p + \log E_i + u_{pi} : \text{Only for Model 1b}\\
\beta_{\text{prec},i} = \bar{\beta}_{\text{prec},i} = \bar{\beta}_{\text{prec},i}\\
\beta_{\text{LST},i} = \bar{\beta}_{\text{LST},i} = \bar{\beta}_{\text{LST},i}\\
\bar{\beta}_{\text{prec},\text{LST},i} \sim \text{MCAR}(\Lambda_{2 \times 2})\\
u_{ri} \sim \text{CAR}(\sigma_{ru}^2) ; u_{pi} \sim \text{CAR}(\sigma_{pu}^2)\\
\]

Model 2: ZINB models

\[
y_i \sim ZIP(\lambda_i v_i, p_i)\\
\]
\[
\lambda_i = n_i r_i; \quad v_i \sim \text{Gd}(\omega, \omega)
\]

\[
\log r_i \sim N(\eta_{ni}, \sigma_i^2)
\]

\[
\eta_{ni} = \alpha_r + \beta_{\text{prec},i} x_{\text{prec},i} + \beta_{\text{LST},i} x_{\text{LST},i} + u_{ni}
\]

\[
p_i \sim U(0, 1) \text{ or } p_i \sim \text{Beta}(1, 1): \text{ Only for Model 2a}
\]

\[
\logit p_i = \alpha_p + \log E_i + u_{pi}: \text{ Only for Model 2b}
\]

\[
\beta_{\text{prec},i} = \bar{p}_{\text{prec},i} + \tilde{u}_{\text{prec},i}
\]

\[
\beta_{\text{LST},i} = \bar{p}_{\text{LST},i} + \tilde{u}_{\text{LST},i}
\]

\[
\tilde{u}_{\text{prec},\text{LST},i} \sim \text{MCAR}(A_{2 \times 2})
\]

\[
u_{ni} \sim \text{CAR}(\sigma^2_{nu}); \quad u_{pi} \sim \text{CAR}(\sigma^2_{pu})
\]

### 3.1. Bayesian inference

We estimated the model parameters via a fully Bayesian framework. Under a Bayesian inference, we use evidence from the data to update our prior beliefs about the processes and priors. If we were to assume our prior processes \( \Theta_1 | \Theta_2 = \{\alpha_r, \alpha_p, \bar{p}_q, \bar{\gamma}_q, \bar{\rho}_q, \bar{\gamma}_q, \bar{u}_k\} \) and beliefs \( \Theta_2 = \{\sigma_{i}^{2}\ldots, \sigma_{2}^{2}\} \), to be true, then \( \text{Pr}(y|\Theta_1, \Theta_2) \) equals either the ZIP or ZINB likelihoods defined in the previous section. This describes the conditional probability that our data would be observed. Using the Bayes theorem, our posterior density, also called updated beliefs about the processes and parameters. Let the current state be \( \Theta^t = \{\theta_i^1, \ldots, \theta_j^f\} \) and \( \theta_j^{t-j} = \{\theta_j^t, \ldots, \theta_j^{t-1}, \theta_j^{t+1}, \ldots, \theta_j^{f}\} \) a vector containing all parameters except \( \theta_j^t \). For each parameter at each step of the iteration \( \theta_{j+1}^{t} \), \( t = 0, \ldots, T + M \), with the starting value \( \theta_{0}^{t} \), we simulate from the full conditional \( \theta_{j+1}^{t} \sim \text{Pr}(\theta_j | \theta_{-j}^t) \). Notice that \( \text{Pr}(\theta_j | \theta_{-j}^t) = \text{Pr}(\theta_j | \theta_{-j}^t) / \text{Pr}(\theta_{-j}^t) \). In brief, at each step of the iteration, we sample a new value for each parameter \( \theta_{j+1}^{t} \) according to its distribution based on the values of all the other variables; and the new parameters are used in the next sampling as soon as they are obtained. When the samples have reached stationarity after sufficiently large samples, the posterior marginals are computed as \( \mathbb{E}[\theta_t | y] = \hat{\theta}_t = T-1 \sum \theta_t^j \), where \( T \) is the number of remaining samples after \( M \) initial samples have been discarded.

As required, we assigned prior distributions to all the unknown parameters. For the fixed effects, we assigned diffuse normal priors of zero mean with low precision: \( \{\alpha_r, \alpha_p, \bar{p}_q, \bar{\gamma}_q\} \sim N(0, 0.001) \). We assigned the Gamma prior \( 1/\sigma_{i}^{2} \sim \text{Ga}(0.5, 0.05) \) to all the variance parameters and a Wishart prior for the covariance \( A_{2 \times 2} \sim \text{W}(\Omega, 2) \). Following simulation studies on the multivariate CAR modeling, we set \( \Omega \) as a scaled identity matrix with diagonal entries equal to one and off-diagonal entries equal to zero (Moraga and Lawson, 2012).

To make objective inferences, we generated 200,000 iterations each of three chains, discarded the first 100,000 iterations retained every 50th iteration of each chain for posterior summaries. A visual examination of the trace plots of the parameters (see Fig. 3 as an example of some of the parameters, \( \alpha_r, \alpha_p, \bar{p}_q, \bar{\gamma}_q \)) indicates convergence and efficient mixing of the chains. We formally assessed convergence using the Brooks and Gelman plots and shrink factor (Brooks and Gelman, 1998). Here, convergence is said to be attained when the shrink factor approaches 1. In Fig. 4, we show the Brooks and Gelman plots for some fixed parameters \( (\alpha_r, \alpha_p, \bar{p}_q, \bar{\gamma}_q) \) where the shrink factor approaches 1 as the number of iterations increases. We fitted the models using the R2WinBUGS package (Spiegelhalter et al., 2008) together with the R software (R Core Team, 2016).
Fig. 3. Trace plots of the posterior estimates of $\alpha_r$ (A), $\alpha_r$ (B), $\beta_{\text{prec}}$ (C), $\beta_{\text{prec}}$ (D).

Fig. 4. Books and Gelman plots for $\alpha_r$ (A), $\alpha_r$ (B), $\beta_{\text{prec}}$ (C), $\beta_{\text{prec}}$ (D).
Table 1
PsML estimates for each model.

| Parameters       | Model 1a  | Model 1b  | Model 2a  | Model 2b  | Poisson Model |
|------------------|-----------|-----------|-----------|-----------|---------------|
| PsML             | −728.25   | −657.22   | −688.45   | −616.20   | −615.4        |

4. Model evaluation and comparison

To evaluate the predictive performance of the models, we used conditional predictive ordinates (CPO). The CPO expresses the posterior probability of observing the outcome $y_i$ when the model is fitted to all data except $y_i$. Using the Monte Carlo estimate of the CPO, $\hat{CPO}_i = \left( T^{-1} \sum_T (f(y_i|\theta^T))^{-1} \right)^{-1}$, the inverse of the posterior predictive ordinates (PPO) for individual observations can be calculated as $PPO_i = (f(y_i|\theta^i))^{-1}$ at each step of the MCMC iteration in WinBUGS. An estimate of the CPO is directly obtained by the inverse of the PPO (Ntzoufras, 2009). A larger value implies a better fit of the model to $y_i$, and very low CPO values suggest that $y_i$ is an outlier concerning the model being fitted. For the comparison of the models, we used the pseudo marginal likelihood (PsML) measure, also a pseudo-Bayes factor. This is estimated as the product of the CPOs or the sum of their logged values (Gelfand, 1996). A higher value of the PsML implies a better fit of a model to the observations.

5. Results and analysis

The results of the four models, i.e. the two ZIP and the two ZINB SVC models, are shown in Table 1. In this study, we make a decisive choice of one model over the other if the difference in their PsML is greater than 2 (Kass and Raftery, 1995). This has also been used to compare ZIP models in Musal and Aktekin (2013). The table clearly shows that models with spatially varying probabilities of onset performed better than those without by comparing their PsML. Thus, the ZIP Model 1a performed better than that of its counterpart, ZIP Model 1b (PsML = −728.25 for Model 1a and PsML = −657.22 for Model 1b). Likewise, the ZINB Model 2a performed better than that of its counterpart, ZINB Model 2b (PsML = −688.45 for Model 1a and PsML = −616.20 for Model 1b). The ZINB model with varying probabilities of onset performed better than all the models; the closest competitor, the ZIP model with varying probabilities of onset, has a difference of approximately 40 in their PsML estimates. Thus, we can conclude that the addition of Gamma random effects to the ZIP SVC model to obtain the ZINB SVC model significantly improves the model fit according to the PsML measure.

Neyens et al. (2017) concluded that models that consider zero inflation do not always provide better fits to data with excess zeros than less complex models. We, therefore, compared our ZI mixture models with a Poisson SVC model. While the traditional SVC Poisson model does not consider zero inflation, it was unexpectedly observed to be superior in terms of fit. That is the Poisson SVC has a PsML = −615.4, higher than all the ZI mixture models except. Yet, we draw further inferences based on the ZI mixture models, specifically the ZINB Model 2b. We will consider this observation further in our discussion section.

To determine whether fitting the ZI mixture models was worthwhile, we tested if the proportion of zeros in the non-susceptible population estimated from the ZI mixture models is higher than those estimated from a Poisson SVC model. Thus, if $1 - p_i|ZI > \exp(-E_r_i)|Pois$, then the data are indeed ZI, and therefore ZI mixture SVC models are worthwhile relative to the Poisson SVC. Considering Model 2b, i.e. the ZINB SVC, we observed the average probability of onset to be $p_i = 0.59$; thus, the average predicted number of zeros in the non-susceptible population equals $1 - p_i = 0.41$. We observed the average proportion of zeros relative to the Poisson SVC model to be $\exp(-E_r_i) = 0.35$, a clear indication that the proportion of zeros estimated from the ZINB SVC model for the non-susceptible population exceeds the proportion of zeros estimated from a Poisson SVC. Fig. 5 shows the spatial distribution of the predicted proportions of zeros in the susceptible
and non-susceptible populations. The predicted zeros in the susceptible population ranged from 0 to 11.5%. Those in the non-susceptible population ranged from approximately 0 to 1.

We now analyze the estimates of the posterior estimates of the relative risks, onset probabilities, and the spatially varying effects of precipitation and temperature. We consider these estimates from the ZINB SVC model, i.e. Model 2b. The overall estimated relative risk is $\exp(\alpha_i) = 1.12$. This is the relative risk after accounting for the spatially varying effects of precipitation and LST, and the structured and unstructured spatial random effects. Fig. 6 shows the posterior estimates of the relative risks (Fig. 6 A) and exceedance probabilities (Fig. 6B). There are 138 out of the 216 districts with higher than expected relative risks. Clustering of the relative risks is assessed using their exceedance probabilities, i.e. $\Pr(r_i|\theta > 1)$. Clustering of high relative risks dominates within the north-east, south-west, and southern parts (Fig. 6B). The spatial trends of the onset probabilities...
are also shown in Fig. 6C. Large clusters of high probabilities of cholera onset ($0.8 < p_i < 1$) are found within the southern and southwestern areas. These are also the areas where high relative risks cluster. Here, we are also interested in the probability that the probability of cholera onset at a given district will exceed the average probability of onset (Fig. 6D). Out of the 112 districts identified to have higher probabilities of cholera onset, i.e. $(0.8 < p_i < 1)$, 89 of them have high exceedance probabilities.

Estimates of the global linear effects of precipitation $x_{\text{prec}}$ and LST $x_{\text{LST}}$ are $\beta_{r,\text{prec}} = 0.311$ and $\beta_{r,\text{LST}} = 0.137$ (Table 2). This implies that the multiplicative effects of precipitation and LST are $\exp(\beta_{r,\text{prec}}) = 1.36$ and $\exp(\beta_{r,\text{LST}}) = 1.15$, respectively. These estimates, however, fall outside the 95%, but within the 50% Bayesian credible intervals. The local effects varied considerably with large clusters of decreasing and increasing trends trend (Fig. 7). The increasing trends dominated over the decreasing trends for both precipitation and LST; over 80% of districts have multiplicative effects exceeding 1. These spatially varying effects for precipitation and LST are moderately correlated as indicated by the estimate of the within area correlation ($\rho_{\text{prec, LST}} = 0.25$). The interval, however, falls outside the 95% Bayesian credible interval although within the 50%.

6. Discussion

The structure of Model 2 (both 2a and 2b), except for the inclusion of zero-inflation and spatially varying effects, is similar to the so-called combined model presented in Neyens et al. (2012). The
combined model is a Poisson–Gamma model where the spatial process is a CAR random effect, to account for the structured heterogeneity (spatial clustering), plus a random Gamma term for the
unstructured heterogeneity. With the strong conjugacy of the Poisson and the Gamma distributions and the performance as observed in our study, we also allude to the fact that Poisson–Gamma model is a valuable alternative to the convolution specification for ZI mixture models. Similar to the BYM convolution model (Besag et al., 1991), our ZINB may also suffer from identification issues due to the addition of the Gamma random effects to capture unstructured heterogeneities (Eberly and Carlin, 2000). For this reason, the shape and rate parameters of the Gamma random effects density were set to be equal (Chebon et al., 2017). The results of the convergence diagnostics, however, suggest that identifiability has not been problematic in this study (Fig. 3). That said, we focus our inferences on the interpretation of the relative risks, probabilities of onset and zeros in the non-susceptible and susceptible populations, and the spatially varying effects. We do not focus on structured and unstructured heterogeneities.

In terms of model fit, it would be safe to conclude that ZI SVC mixture models do not always perform better than their traditional Poisson SVC counterparts. Although this is not our objective, its observation deserves some discussion. This finding can be attributed to the log-Gaussian parameterization of the relative risk, whose precision parameter seems to capture a greater proportion of the variations in the counts. Out of curiosity, we fitted the log-linear counterparts for the Poisson SVC, ZIP SVC, and ZINB SVC models. We observed severe convergence issues for the fixed effects and random effects at the nodes with non-zero counts. Fast convergence, however, was observed for all parameters for the log-Gaussian parameterization. A similar parameterization has been used by Tzala and Best (2008) who studied the multivariate spatio-temporal variation in cancer mortality. They argued in favor of this parameterization concerning MCMC convergence issues, while the reason for their argument was the avoidance of intercept terms in the linear predictor. Best and Hansell (2009) have also used the log-Gaussian parameterization to study the geographical variation in disease risk, but did not discuss its influences on MCMC convergence and random-effects estimates. We plan to examine these influences in a future study.

The choice of the ZI mixture models over the traditional Poisson model should not necessarily be based on only model fit criteria. While our Poisson SVC model can capture the over-dispersion and has relatively better performance in terms of fit, it cannot distinguish between zeros in the non-susceptible and susceptible populations which is an important aspect of this study. As argued by Gilthorpe et al. (2009) and also pointed out by Long et al. (2015), prior knowledge of the data generating mechanism should drive the choice of models from which to choose. In a similar study, Neyens et al. (2017) observed that the traditional Poisson models that do not take into account the excess zeros but contain at least one random effects term that models the extra variance in the counts have better fits compared to their ZI counterparts. Their purpose was to investigate the distribution of the disease using Bayesian models that account for both excess zeros and over-dispersion, without recourse to distinguish between the different kinds of zeros. Putting these together, we argue that, if the purpose of a ZI model is to distinguish between the different sources of zero incidences, then there is no point to undertake a model fit comparison with the traditional Poisson model. In our study area, and perhaps other sub-Saharan African countries of similar settings, limitations in cholera surveillance may influence the ability to capture all incidences. It is, therefore, reasonable to distinguish between zero incidences arising from the non-susceptible and susceptible populations.
We have avoided the use of the same covariates in the count and binary components to avoid identification issues. Lambert (1992) proposed a variant of the ZIP model to allow the inclusion of the same covariates in the count and binary components by thinking of the onset probability as a function of the latent mean of the count component. This is the focus of another research. However, the relationship between cholera and these determinants, i.e., precipitation and LST, conceptually precludes their inclusion in the binary component. Cholera has been described by two forms of transmission: primary or environment-to-human transmission induced by exposure to the aquatic environments where cholera bacterial are viable, and secondary or human-to-human transmission induced by contaminated water or food by an infected person (Miller et al., 1985). Cholera onset in a population, as implied in the binary component of our model, can be characterized as the probability of initiating transmission. This is modulated by biotic and abiotic factors such as sea surface temperature, sea surface topography, and organic nutrients, that affect the aquatic environment. A contemporary review has discussed these factors (Lutz et al., 2013). Cholera onset, arguably, cannot be induced by exposure variations in precipitation and LST. In contrast, variations in precipitation and LST rather influence secondary transmissions and the progression of incidences once a cholera outbreak has already been initiated. Cholera onset via importation can be also characterized as primary transmission if it is the index case in the sub-population. Therefore, factors that can modulate the importation of cholera may rather also be included in the binary component instead of those that enhance the progression.

Our results indicated that the proportions of zeros in the non-susceptible population exceeded what would have been observed in a traditional Poisson model, and therefore fitting the ZIP model was necessary. We observed the average probability of onset to be \( p_i = 0.59 \). The implication is that 59% of the population was exposed to cholera during the outbreak. The proportion of zeros attributable to the susceptible population ranged from 0 to 11%. This can be ascribed as the proportion of asymptomatic cases or cases that could not be captured via the surveillance system. A proportion of cases going unnoticed is conceivable due to the nature of cholera reporting in developing countries. In Ghana, for instance, case detection is by clinical manifestation after an individual falling ill has visited the hospital, hence mild cases can go unnoticed. The WHO recognizes this (Ali et al., 2012), but estimating the percentages seems to have gained no attention. Even in Bangladesh where a ZINB non-spatial model for cholera has been used (Carrel et al., 2011), the results and discussion so far ignored estimates and distribution of zeros attributable to the susceptible and non-susceptible populations.

The risk and probability of onset maps show clustering of high rates along the coast. Such patterns are as expected since primary transmission of cholera is triggered from coastal environments which also normally serve as the natural habitat of the cholera vibrios. This conforms with several studies which have asserted that areas closer to the coast are always at a high risk of cholera infection (Jutla et al., 2010; Sack et al., 2003).

Our study has combined the SVC modeling idea of Assunção (2003) with the ZI mixture data generating process to uncover non-monotonic spatial trends of the effects of precipitation and LST on cholera incidences. Thus, we have allowed the intrinsic interested parameters, the coefficients, to vary across space to capture the complex relationship between cholera and its determinants. Is this of scientific interest for cholera epidemiology? The non-monotonic spatial trends of the effects of precipitation and LST on cholera incidences are a striking observation that suggests that they have both negative and positive effects on cholera incidences. The use of only the global coefficients for inferences would rather lead to wrong conclusions. The plausibility of this observation can be explained by how cholera occurrences jointly relate with precipitation and LST, and other physical conditions. Similarly in Bangladesh, cholera incidences have been observed to increase with both high and low precipitation (Hashizume et al., 2008). What are the possible mechanisms? A suggestive hypothesis is that runoff from higher rainfall can serve as the mechanical pathway for bacterial dissemination in the environment or contamination of surface water. On the other hand, higher rainfall can also wash away or dilute the bacterial in the environment (Pascual et al., 2002). Intermediate factors may play a role in these contrasting mechanisms. We may implicate the topographical patterns which have a direct influence on surface water runoff during heavy and light rains. Some studies on cholera during the 1850s outbreak drew a related conclusion (Bingham et al.,
Also, the importance of considering topographical elevation as a geographical and environmental risk factor has been demonstrated elsewhere (Luque Fernandez et al., 2012). The limitation here is that, within the ZI mixture framework, the effects of the covariates do not have the same direct interpretation as the traditional Poisson Model. The marginal inference of covariates on the sampled population is a challenge in the ZI mixture structure. The covariates are associated with the mean of the susceptible population $\lambda_i$ rather than with that of the entire population $\mathbb{E}(y_i|\mathbf{r}_i, p_i) = p_i\lambda_i$. Thus, the modeling structure fails to discriminate between inference for the susceptible as defined by the model and inference for the entire population. This can be seen as a limitation and grounds for developing extensions for the ZI mixture models as discussed elsewhere (Preisser et al., 2012). The prominent suggestions have been marginalized ZI mixture modeling (Long et al., 2015, 2014) but this has not been extended to accommodate for spatial random effects. Also, while the ZI mixture modeling is conceptually appropriate for modeling cholera incidences, the presence of very large counts of incidences in the data can create extreme over-dispersion. In the future, we propose to attach a second mixture to the non-zero component for extremely large counts.

7. Summary of findings and conclusions

In this study, we have proposed a ZI mixture SVC modeling framework to investigate the spatial patterns of cholera counts with excess zeros. The models have been applied to the 2014 cholera incidences in Ghana. We have combined the ideas of ZI mixture and SVC modeling to (1) develop a ZINB SVC model for cholera incidences with excess zeros, (2) estimate the zeros arising from the non-susceptible and susceptible populations, and (3) examine the performance of ZINB SVC in the determination of the spatially varying effects of precipitation and LST on cholera. A consideration of log-Gaussian parameterization over the log-normal normal parameterization was necessary due to MCMC convergence issues. The traditional Poisson SVC model was comparable to the ZI mixture models in terms of fit. Yet inferences based on the ZI mixture are necessary because of its ability to distinguish between zeros within the non-susceptible and susceptible populations. The proportion of zeros in the non-susceptible population estimated from the ZINB SVC model is higher than those estimated from a Poisson SVC model; a piece of evidence that suggests the importance of ZI mixture modeling. That said, if the objective is only to capture over-dispersion due to excess zero counts, then the Poisson SVC model with log-Gaussian parameterization is adequate. Otherwise, if the objective is to distinguish between the sources of the zero counts, then it may not be important to use the Poisson counterpart as a benchmark for model comparison. On this, we conclude that the choice of a ZI mixture model over a Poisson model should be based on the epidemiological concept rather than the model fit.

The proportion of zeros attributable to the susceptible population varied spatially, and therefore improving cholera surveillance should be focused within districts with higher proportions. The spatial patterns of the effects of precipitation and LST on cholera incidences are non-monotonic, with large clusters of increasing and decreasing trends. We have suggested the possibility of an intermediate mechanism, such as the undulating nature of the topography, to impart the observed trends. Investigating this is beyond the scope of this study and therefore reserved for the future. We must state that the aggregation of the covariates, precipitation and LST, over the districts may lead to errors whose propagation and effects in the models have not been considered in this study. This will be considered in our future study. From an epidemiological standpoint, this study forms a clearer picture of the spatially varying effects of precipitation and LST on cholera occurrences. The ZI mixture modeling structure in this study has a latent class formulation that overlooks marginal inferences of the expectation of the mixture distribution. This limitation is an avenue for future study into marginalized ZI mixture modeling. Indeed, it will be interesting to consider the space–time trends if surveillance data are organized in space–time data frames. Extending these models for spatial forecasting of the disease at fine-scale resolution will also be valuable to enhance monitoring and evaluation of interventions. This can be achieved via the conceptualization of spatial correlation as Gaussian random fields. In Rathbun and Fei (2006), unlike our study, the proportion of excess zeros is assumed to be generated by a spatial probit model based on the realizations of a Gaussian
random field falling below a given threshold. A future comparative study between the spatial logistic and spatial probit parametrizations of the excess zeros could further complement the literature on ZI models.

References

Agarwal, D.K., Gelfand, A.E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environ. Ecol. Stat. 9, 341–355. http://dx.doi.org/10.1023/A:1020910605990.

Ali, M., Emch, M., Donnay, J.P., Yunus, M., Sack, D.A., 2002. Identifying environmental risk factors for endemic cholera: a raster GIS approach. Health Place 8, 201–210. http://dx.doi.org/10.1016/S1353-8292(01)00043-0.

Ali, M., Goovaerts, P., Nazia, N., Haq, M.Z., Yunus, M., Emch, M., 2006. Application of Poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh. Int. J. Health Geogr. 5, 45. http://dx.doi.org/10.1186/1476-072X-5-45.

Ali, M., Gupta, S.S., Arora, N., Khasnobis, P., Venkatesh, S., Sur, D., Nair, G.B., Sack, D.A., Ganguly, N.K., 2017. Identification of burden hotspots and risk factors for cholera in India: An observational study. PLoS One 12, e0183100. http://dx.doi.org/10.1371/journal.pone.0183100.

Ali, M., Lopez, A.L., You, Y.A., Kim, Y.E., Sah, B., Maskery, B., Clemens, J., 2012. The global burden of cholera. Bull. World Health Organ. 90, 209–218A. http://dx.doi.org/10.2471/BLT.11.093427.

Assunção, R.M., 2003. Space varying coefficient models for small area data. Environmetrics 14, 453–473. http://dx.doi.org/10.1002/env.599.

Baetschmann, G., Winkelmann, R., 2013. Modeling zero-inflated count data when exposure varies: With an application to tumor counts. Biom. J. Biom. Z 55, 679–686. http://dx.doi.org/10.1002/bimj.201200021.

Banerjee, S., Gelfand, A.E., 2002. Prediction, interpolation and regression for spatially misaligned data. Sankhyā Indian J. Stat. Ser. A 64 (1961–2002), 227–245.

Besag, J., Kooijenberg, C., 1995. On conditional and intrinsic autoregressions俵 pg. 733–746.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 43, 1–20. http://dx.doi.org/10.1007/BF00116466.

Best, N., Hansell, A.L., 2009. Geographic variations in risk: Adjusting for unmeasured confounders through joint modeling of multiple diseases. Epidemiology 20, 400–410. http://dx.doi.org/10.1097/EDE.0b013e31819d90f9.

Bingham, P., Verlander, N.Q., Che, M.J., 2004. John snow, william farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. Public Health 118, 387–394. http://dx.doi.org/10.1016/j.jpuhe.2004.05.007.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat 7, 434–455. http://dx.doi.org/10.1080/10618609.1998.1047487.

Bwire, G., Ali, M., Sack, D.A., Nakinsige, A., Naigaga, M., Debes, A.K., Ngwa, M.C., Brooks, W.A., Orach, C., Garimoi., 2017. Identifying cholera hotspots in Uganda: An analysis of cholera surveillance data from 2011 to 2016. PLoS Negl. Trop. Dis 11, e0006118. http://dx.doi.org/10.1371/journal.pntd.0006118.

Carrel, M., Escamilla, V., Messina, J., Giebultowicz, S., Winston, J., Yunus, M., Pk, Streatfield., Emch, M., 2011. Diarrheal disease risk in rural Bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis. Int. J. Health Geogr. 10, 41. http://dx.doi.org/10.1186/1476-072X-10-41.

Chebon, S., Faes, C., Cools, F., Geys, H., 2017. Models for zero-inflated, correlated count data with extra heterogeneity: when is it too complex? Stat. Med. 36, 254–361. http://dx.doi.org/10.1002/sim.7142.

Dambon, J.A., Sigrist, E., Furrer, R., 2021. Maximum likelihood estimation of spatially varying coefficient models for large data with an application to real estate price prediction. Stat. Med. 41, 100470. http://dx.doi.org/10.1002/sim.8320. 100470.

Eberly, L.E., Carlin, B.P., 2000. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Stat. Med. 19, 2279–2294. http://dx.doi.org/10.1002/1097-0258(20000915)30:19<17::AID-SIM569>3.0.CO;2-r.

Farr, W., 1852. Influence of elevation on the fatality of Cholera. J. Stat. Soc. Lond. 15, 155–183. http://dx.doi.org/10.1177/0308875001010002.

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. http://dx.doi.org/10.1111/j.2041-210X.2010.00060.x.

Fotheringham, A.S., Brunsdon, C., Charlton, M.E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons, Chichester.

Funk, C.C., Peterson, P.J., Landsfeld, M.F., Pedreros, D.H., Verdin, J.P., Rowland, J.D., Romero, B.E., Husak, G.J., Michaelson, J.C., Verdin, A.P., 2014. A Quasi-Global Precipitation Time Series for Drought Monitoring (USGS Numbered Series No. 832). In: A quasi-global precipitation time series for drought monitoring, Data Series, U.S. Geological Survey, Reston, VA, http://dx.doi.org/10.3133/ds832.

Gelfand, A., 1996. Model determination using sampling-based methods. http://dx.doi.org/10.1201/b14835-14.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). Bayesian Anal. 1, 515–534. http://dx.doi.org/10.1214/06-BA117A.

Gilthorpe, M.S., Frydenberg, M., Cheng, Y., Baelum, V., 2009. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. Stat. Med. 28, 3539–3553. http://dx.doi.org/10.1002/sim.3699.
Greene, W.H., 1994. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models (SSRN Scholarly Paper No. ID 1293115). Social Science Research Network, Rochester, NY.

Hashizume, M., Armstrong, B., Hajat, S., Wagatsuma, Y., Faruque, A.S.G., Hayashi, T., Sack, D.A., 2008. The effect of rainfall on the incidence of cholera in Bangladesh. Epidemiology 19, 103–110. http://dx.doi.org/10.1097/EDE.0b013e31815c09ea.

Heilbron, D.C., 1994. Zero-altered and other regression models for count data with added zeros. Biom. J. 36, 531–547. http://dx.doi.org/10.1002/bimj.4710360505.

Jim, X., Carlin, B.P., Banerjee, S., 2005. Generalized hierarchical multivariate CAR models for areal data. Biometrics 61, 950–961. http://dx.doi.org/10.1111/j.1541-0420.2005.00359.x.

Jutla, A.S., Akanda, A.S., Islam, S., 2010. Tracking cholera in coastal regions using satellite observations. J. Am. Water Resour. Assoc. AWRA 46, 651–662. http://dx.doi.org/10.1111/j.1752-1688.2010.00448.x.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14. http://dx.doi.org/10.2307/1269547.

Lessler, J., Moore, S.M., Luquero, F.J., McKay, H.S., Grais, R., Henkens, M., Mengel, M., Dunoyer, J., M’bangombe, M., Lee, E.C., Djingarey, M.H., Sudre, B., Bompangue, D., Fraser, R.S.M., Abubakar, A., Pereia, W., Legros, D., Azman, A.S., 2018. Mapping the burden of cholera in sub-saharan africa and implications for control: an analysis of data across geographical scales. Lancet http://dx.doi.org/10.1016/S0140-6736(17)33050-7.

Long, D.L., Preisser, J.S., Herring, A.H., Golin, C.E., 2014. A marginalized zero-inflated Poisson regression model with overall exposure effects. Stat. Med. 33, 5151–5165. http://dx.doi.org/10.1002/sim.6293.

Long, D.L., Preisser, J.S., Herring, A.H., Golin, C.E., 2015. A marginalized zero-inflated Poisson regression model with random effects. J. R. Stat. Soc. Ser. C Appl. Stat. 64, 815–830. http://dx.doi.org/10.1111/rssc.12104.

Luque Fernandez, M.A., Schomaker, M., Mason, P.R., Fesselet, J.F., Baudot, Y., Boulle, A., Maes, P., 2012. Elevation and cholera: An Epidemiological Spatial Analysis of the Cholera Epidemic in Harare, Zimbabwe, 2008–2009. vol 12, BMC Public Health. p. 442. http://dx.doi.org/10.1186/1471-2458-12-442.

Lutz, C., Erken, M., Noorian, P., Sun, S., McDougal, D., 2013. Environmental reservoirs and mechanisms of persistence of vibrio cholerae. Front. Microbiol. 4 (375), http://dx.doi.org/10.3389/fmicb.2013.00357.

Miller, Christopher J., Feachem, Richard G., Drasar, Bohumil S, 1985. Cholera epidemiology in developed and developing countries: New thoughts on transmission, seasonality, and control. Lancet 1 (8423), 261–263. http://dx.doi.org/10.1016/S0140-6736(85)91036-0, 325.

Moraga, P., Lawson, A.B., 2012. Gaussian component mixtures and CAR models in Bayesian disease mapping. Comput. Stat. Data Anal. 56, 1417–1433. http://dx.doi.org/10.1016/j.csda.2011.11.011.

Mullahy, J., 1986. Specification and testing of some modified count data models. J. Econom. 33, 341–365. http://dx.doi.org/10.1016/0304-4076(86)90002-3.

Musial, M., Aktekin, T., 2013. Bayesian spatial modeling of HIV mortality via zero-inflated Poisson models. Stat. Med. 32, 267–281. http://dx.doi.org/10.1002/sim.5457.

Musio, M., Sauleau, E.A., Buemi, A., 2010. Bayesian semi-parametric ZIP models with space–time interactions: an application to cancer registry data. Math. Med. Biol. J. IMA 27, 181–194. http://dx.doi.org/10.1093/imambq/dqp025.

Neyens, T., Faes, C., Molenberghs, G., 2012. A generalized Poisson-gamma model for spatially overdispersed data. Spat. Temporal Epidemiol. 3, 185–194. http://dx.doi.org/10.1016/j.sste.2011.10.004.

Neyens, T., Lawson, A.B., Kirby, R.S., Nuyts, V., Watjou, K., Aregay, M., Carroll, R., Nawrot, T.S., Faes, C., 2017. Disease mapping of zero-excessive mesothelioma data in flanders. Ann. Epidemiol. 27, 59–66. http://dx.doi.org/10.1016/j.annepidem.2016.10.006.e3.

Ntzoufras, I., 2009. Bayesian Modeling using WinBUGS. Wiley.

Nyandwi, E., Osei, F.B., Amer, S., Veldkamp, A., 2020. Modeling schistosomiasis spatial risk dynamics over time in rwanda using zero-inflated Poisson regression. Sci. Rep. 10, 1–9. http://dx.doi.org/10.1038/s41598-020-76288-8.

Osei, F.B., Duker, A.A., 2008. Spatial and demographic patterns of cholera in ashatini region - ghana. Int. J. Health Geogr. 7, 44. http://dx.doi.org/10.1186/1476-072X-7-44.

Pascual, M., Bouma, M.J., Dobson, A.P., 2002. Cholera and climate: revisiting the quantitative evidence. Microbes Infect. 4, 237–245. http://dx.doi.org/10.1016/S1286-4579(01)01533-7.

Preisser, J.S., Stamm, J.W., Long, D.L., Kincade, M.E., 2012. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Res. 46, 413–423. http://dx.doi.org/10.1159/000338992.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rathbun, S.L., Fei, S., 2006. A spatial zero-inflated Poisson regression model for oak regeneration. Environ. Ecol. Stat. 13, 409–426. http://dx.doi.org/10.1007/s10651-006-0020-x.

Root, E.D., Rodd, J., Yunus, M., Emch, M., 2013. The role of socioeconomic status in longitudinal trends of Cholera in matlaf, Bangladesh, 1993–2007. PLoS Negl. Trop. Dis. 7, e1997. http://dx.doi.org/10.1371/journal.pntd.0001997.

Sack, R.B., Siddique, A.K., Longini, I.M., Nizam, A., Yunus, M., Islam, M.S., Morris, J.G., Ali, A., Huq, A., Nair, G.B., Qadri, F., Faruque, S.M., Sack, D.A., Colwell, R.R., 2003. A 4-year study of the epidemiology of vibrio cholerae in four rural areas of Bangladesh. J. Infect. Dis. 187, 96–101. http://dx.doi.org/10.1086/354865.

Sasaki, S., Suzuki, H., Igarashi, K., Tambatamba, B., Mulenga, P., 2008. Spatial analysis of risk factor of Cholera outbreak for 2003–2004 in a peri-urban area of Lusaka. Zambia J. Trop. Med. Hyg. 79, 414–421. Spiegelhalter, D.J., Thomas, A., Best, N.G., 2008. Winbugs version 1.4.3.

Torbabi, M., 2017. Zero-inflated spatio-temporal models for disease mapping. Biom. J. Biom. Z 59, 430–444. http://dx.doi.org/10.1002/bimj.201600120.
Tzala, E., Best, N., 2008. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. Stat. Methods Med. Res. 17, 97–118. http://dx.doi.org/10.1177/0962280207081243.

Wakefield, J.C., Best, N.G., Waller, L., 2001. Bayesian approaches to disease mapping. In: Spatial Epidemiology. Oxford University Press, Oxford, http://dx.doi.org/10.1093/acprof:oso/9780198515326.003.0007.

WHO, W., 2017. 2.1 Billion people lack safe drinking water at home, more than twice as many lack safe sanitation [www document]. https://www.who.int/news/item/12-07-2017-2-1-billion-people-lack-safe-drinking-water-at-home-more-than-twice-as-many-lack-safe-sanitation. (Accessed 22 January 2021).