Bivariate geostatistical modelling of the relationship between *Loa loa* prevalence and intensity of infection

Emanuele Giorgi, Daniela K. Schlüter and Peter J. Diggle
(CHICAS, Lancaster Medical School, Lancaster University)

May 31, 2017

Abstract

Loiasis is a neglected tropical disease (NTD) caused by the parasitic roundworm *Loa loa*. A challenge faced by current multi-national programmes to control two other diseases, Lymphatic filariases and Onchocerciasis, by mass administration of prophylactic medication to at-risk communities is that individuals highly co-infected with *Loa loa* are at risk of developing serious adverse reactions to the medication. For this reason, understanding the geographical distribution of *Loa loa* prevalence and the distribution of microfilarial loads in communities have become of crucial importance. In this paper we extend the methodology developed by Schlüter et al. (2016) to analyse data on microfilariae counts per millilitre of blood. One feature of the data is the excess of zero counts which makes the use of standard geostatistical methods for prevalance data inappropriate. This phenomenon, also known as zero-inflation, is typical of count data from NTDs, whose endemic boundaries are often unknown, thus leading to the inclusion of disease-free communities in the sampling frame. We introduce a bivariate geostatistical model in order to study the relationship between the distributions of prevalence and intensity of *Loa loa* infections at community level. We show through a simulation study that the spatial model leads to more precise spatial predictions than the non-spatial approach used by Schlüter et al. (2016), and accordingly provide a geostatistical re-analysis of the *Loa loa* data.

Keywords: disease mapping; geostatistics; Loa loa; neglected tropical diseases; spatial correlation; zero-inflation.

1 Introduction

Loiasis, also known as the African eye worm, is a parasitic helminth disease caused by the nematode *Loa loa*. Although endemic in several countries across central and western Africa,

most individuals infected with *Loa loa* present at most mild symptoms such as itchy Calabar swellings and the movement of adult worms across the eye (CDC, 2015). Other filarial diseases that are endemic in the same regions, however, constitute serious public health problems, namely Lymphatic Filariasis (LF) and Onchocerciasis. LF adversely affects the immune system and can lead to lymphoendema and thickening and hardening of the skin due to an increase in bacterial infections, a condition called elephantiasis (WHO, 2016a). Onchocerciasis, commonly known as River Blindness, includes symptoms of disfiguring skin conditions and eye lesions leading to visual impairment or even permanent blindness (WHO, 2016b).

Over 120 million people in Africa, South America and Asia are affected by LF and more than 123 million are at risk of being infected with Onchocerciasis (CDC, 2013a,b). The World Health Organisation (WHO) has targeted both diseases for elimination and has launched mass drug administration (MDA) programs to interrupt transmission (Taylor et al., 2010; Keating et al., 20014). The LF elimination program is based on annual mass administration of a single dose of diethylcarbamazine or ivermectin combined with albendazole, while the onchocerciasis elimination program is based on mass administration of ivermectin only (Taylor et al., 2010). Although these drugs are generally considered safe, it is now known that individuals who are highly co-infected with *Loa loa* parasites are at risk of developing seriours adverse events, such as encephalopathy, which can lead to permanent brain damage or even death (Carme et al., 1991; Gardon et al., 1997; Boussinesq et al., 1999, 2001). Therefore, the MDA programme has been inhibited in regions where *Loa loa* is endemic (Mackenzie et al., 2012; Geary, 2012). Highly infected individuals are most likely to be found in communities with high *Loa loa* prevalence. Thus, efforts to date have focussed on mapping *Loa loa* prevalence to distinguish areas in which LF and Onchocerciasis can be safely treated from those in which caution is needed (Thomson et al., 2004; Diggle et al., 2007; WHO, 2012). Work has also been done to understand the distribution of *Loa loa* microfilarial loads in communities (Pion et al., 2006; Schlüter et al., 2016).

In this paper we focus on the work reported in Schlüter et al. (2016), who analysed data on microfilarial loads of individuals in villages across Cameroon, the Democratic Republic of Congo and the Republic of the Congo to investigate the relationship between community-level prevalence and the proportion of highly infected individuals. The data in Schlüter et al. (2016) were collected in two field studies conducted in the West and East provinces of Cameroon (Takougang et al., 2002), and in the Republic of the Congo and the Bas-Congo and Orientale regions of the Democratic Republic of Congo (Wanji et al., 2012), respectively. In their analysis, the authors modelled the microfilariae (MF) counts per millilitre (ml) of blood from 19,128 individuals sampled across 222 villages. One of their objectives was to develop a statistical model to be used as an operational tool by public health workers, in order to predict the proportion of people in a village with an MF load exceeding a policy-relevant threshold of counts per ml of blood. An important feature of the data was the excess of zero counts, which invalidate the use of standard statistical models for count data.

This phenomenon, also known as zero-inflation, is a common feature of count data on neglected tropical diseases (NTDs). See, for example, Oluwole et al. (2015) for a case-study on soil-transmitted helminths. Since the natural boundaries of NTDs are usually unknown, zero-inflation can arise from the sampling of disease-free communities. This issue has been

addressed by Diggle & Giorgi (2016), who introduce a geostatistical framework in order to distinguish between zero cases as a result of this phenomenon from those that are a chance finding in endemic areas.

In the analysis by Schlüter et al. (2016), zero-inflation was taken into account by using a non-spatial, mixed effects Weibull-mixture model, with two distinct linear predictors for MF prevalence and intensity. Schlüter et al. (2016) assumed independent and identically distributed village-level random effects whilst allowing for cross-correlation between intensity and prevalence. They found a positive correlation between prevalence and intensity of infection which they then exploited to predict the proportion of individuals with infection levels above policy-relevant thresholds.

Our objective in this study is to extend the approach developed by Schlüter et al. (2016) within a geostatistical framework so as to model the relationship between prevalence and intensity. We use a simulation study to compare the predictive performance of the spatial and non-spatial models, and re-analyse the *Loa loa* data from Schlüter et al. (2016). In summary, our results strongly suggest that the use of this novel geostatistical approach leads to more precise prediction of community-level distributions of *Loa loa* infection.

The structure of the paper is the following. In Section 2, we review available methods and recent advances in the modelling of count data that exhibit zero-inflation. In Section 3, we introduce a bivariate geostatistical model to study the relationship between prevalence and intensity of *Loa loa* infections. In Section 4, we outline a Monte Carlo maximum likelihood procedure for likelihood-based inference. In Section 5, we illustrate an application of the model to the re-analysis of the *Loa loa* data. We then carry out a simulation study in Section 6 in order to validate the inferential properties of the model and to quantify the effects on spatial prediction that result from ignoring the residual spatial correlation in the data. Section 7 is a concluding discussion, where we also outline the wider applicability of the developed methodology.

2 Statistical models for zero-inflation

The problem known as zero-inflation arises whenever the distribution of a variable of interest, say Y, includes an additional probability mass at zero, hence $\mathrm{P}(Y \leq y) = (1-\pi) + \pi G(y)$, where $G(\cdot)$ is a distribution function with support confined to $\mathbb{R}^+$ or a sub-set thereof, for example the set of non-negative integers. In this paper, we use the equivalent specification of zero-inflation as

$$Y = Y_1 Y_2, \tag{1}$$

where Y_1 is binary variable such that $P(Y_1 = 1) = \pi$ and Y_2 is a random variable with probability function, if discrete, or density function, if continuous, $g(\cdot)$, and assume that Y_1 and Y_2 are independent. The resulting distribution of Y is a mixture given by

$$f(y) = \mathbf{1}(y = 0)(1-\pi) + \mathbf{1}(y > 0)\pi g(y), \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function. A common modelling choice for non-negative integer-valued Y_2 is a Binomial distribution if the counts are finite, or Poisson if open-ended (Lambert,

1992). In cases where it is desirable to model the zero observations separately from the positive counts, "hurdle" models (Mullahy, 1986) then define the distribution of Y_2 as a truncated model that modifies an ordinary distribution by conditioning on a positive outcome.

In disease mapping applications, Y typically corresponds to the number of diagnosed positive cases for a disease under investigation, in a community living at a geographical location x. Zero-inflation might then arise through the inclusion in the sampling frame of areas where environmental conditions do not allow disease transmission.

To emphasize the spatial context of the current paper, we write $Y(x) = Y_1(x)Y_2(x)$. We also assume that $Y_2(x)$ has expectation

$$E[Y_2(x)] = m\mu(x), m > 0,$$

where m is an off-set. Let $h_1\{\cdot\}$ and $h_2\{\cdot\}$ denote two link functions such that $h_1\{\pi(x)\}$ and $h_2\{\mu(x)\}$ can each take any real value. Models for $Y(x)$ that do not take account of spatial correlation use spatially referenced, explanatory variables, say $d(x)$, as terms in the linear predictors for $\pi(x)$ and $\mu(x)$, hence

$$h_1\{\pi(x)\} = d(x)^\top \beta_1 \tag{3}$$

and

$$h_2\{\mu(x)\} = d(x)^\top \beta_2, \tag{4}$$

where β_1 and β_2 are vectors of regression coefficients.

However, the data might also exhibit over-dispersion as a result of residual correlation induced by unmeasured explanatory variables. For example, Min & Agresti (2005) study the phenomenon of zero-inflation in the case of longitudinal data and propose zero-inlflated models with random effects so as to account for within subject-correlation.

In Schlüter et al. (2016), the outcome of interest $Y(x)$ represents the MF density of an individual living at a village location x. They assume that $Y(x)$ is dependent on a latent bivariate zero-mean Gaussian random variable $Z(x) = (Z_1(x), Z_2(x))$, and that the $Z(x)$ at different locations are stochastically independent. They then assume that conditionally on $Z_1(x)$ and $Z_2(x)$, $Y_1(x)$ is a Bernoulli variable with probability of success $\pi(x)$, such that

$$h_1\{\pi(x)\} = d(x)^\top \beta_1 + Z_1(x), \tag{5}$$

and $Y_2(x)$ has a Weibull distribution with shape parameter κ and expectation $E[Y_2(x)] = \Gamma(1 + 1/\kappa)\mu(x)$, where $\mu(x)$ is modelled as

$$h_2\{\mu(x)\} = d(x)^\top \beta_2 + Z_2(x). \tag{6}$$

Geostatistical zero-inflated models have previously been used in ecology (Agarwal et al., 2002) and in epidemiology (Amek et al., 2011; Giardina et al., 2012). All of these authors assume that $Y_2(x)$ is either Binomial or Poisson, conditionally on spatially structured random effects, $S(x)$, and unstructured random effects, $Z(x)$. The linear predictor for $\mu(x)$ is then defined as

$$h_2\{\mu(x)\} = d(x)^\top \beta_2 + S(x) + Z(x).$$

In contast, they assume that $\pi(x)$ depends on a limited set of explanatory variables, as defined by (3). They then model the spatial process $S(x)$ as a stationary Gaussian process, and the unstructured component $Z(x)$ as Gaussian white noise. If the conditional distribution of $Y_2(x)$ is Poisson, an alternative choice is to assume $\exp\{Z(x)\}$ to be Gamma noise, in which case integrating out $Z(x)$ leads to a negative-binomial distribution for $Y_2(x)$, conditionally on $S(x)$. However the differences between these two approaches are, in general, negligible (Firth, 1988).

Diggle & Giorgi (2016) introduce a more general framework by incorporating spatial random effects into both $\mu(x)$ and $\pi(x)$. They then model the resulting linear predictors $h_1\{\cdot\}$ and $h_2\{\cdot\}$ using a tri-variate spatial process $(S_1(x), S_2(x), T(x))$ to give

$$h_1\{\pi(x)\} = d(x)^\top \beta_1 + S_1(x) + T(x)$$

and

$$h_2\{\mu(x)\} = d(x)^\top \beta_2 + S_2(x) + T(x).$$

where $T(x)$ is used to model residual spatial variation that jointly affects $\pi(x)$ and $\mu(x)$. However, recovering $S_1(x)$, $S_2(x)$ and $T(x)$ from the data is a challenge, that requires a pragmatic response owing to the limited identifiability of the model parameters of all three latent processes. For this reason, in their application to river blindness mapping Diggle & Giorgi (2016) then set $T(x) = 0$ for all x. In the next Section, we adapt this framework to develop a parsimonious model for relationship between *Loa loa* prevalence and intensity while retaining all three of the latent spatial components.

3 A geostatistical zero-inflated Weibull-mixture model

Let $Y_j(x_i)$ denote MF density, measured as the number of MF per ml in a blood sample, for the j-th sampled individual at the village location x_i. Let $\mathcal{S}_h = \{S(x) : x \in A\}$, for $h = 1, 2$, and $\mathcal{T} = \{T(x) : x \in A\}$ denote a set of three independent stationary zero-mean Gaussian processes with unit variance. For each spatial process, we assume isotropic exponential covariance functions, hence

$$\text{corr}\{S_h(x), S_h(x')\} = \exp\{-u/\phi_{S_h}\}, h = 1, 2,$$

and

$$\text{corr}\{T(x), T(x')\} = \exp\{-u/\phi_T\},$$

where u is the Euclidean distance between x and x'.

We then assume that conditionally on $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{T}$, the cumulative distribution function (cdf) of MF density $Y(x)$ at a village location x is given by

$$F\{y(x)\} = 1 - \pi(x) + \pi(x)G\{y(x); \kappa\}, \text{ if } y(x) > 0, \tag{7}$$

where $G\{\cdot; \kappa\}$ is a continuous cdf indexed by the parameter κ, and $\pi(x) = 1 - F(0)$ is the disease prevalence, at location x. We model $G\{\cdot; \kappa\}$ as a Weibull distribution with cdf

$$G\{y(x); \kappa\} = 1 - \exp\left\{-\left[\frac{y(x)}{\mu(x)}\right]^\kappa\right\},$$

where $\mu(x)$ is a spatially varying scale parameter and κ is a shape parameter, assumed to be common to all locations. Finally, we model the spatial variation in $\pi(x)$ and $\mu(x)$ as

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \mu_1 + \sigma_1[S_1(x) + T(x)]$$

and

$$\log\{\mu(x)\} = \mu_2 + \sigma_2[S_2(x) + T(x)],$$

where σ_1^2 and σ_2^2 are the variances of the linear predictors for prevalence and intensity, respectively. This constrains the component spatial processes to have the same variance whilst allowing the two linear predictors to have different variances. In this formulation, $T(x)$ contributes to the spatial variation of both prevalence and intensity whereas $S_1(x)$ and $S_2(x)$ account for independent residual spatial variation in prevalence and intensity, respectively.

The resulting standardized variogram for the linear predictors of prevalence and intensity is

$$\gamma_h(u) = 1 - \frac{1}{2}\left(\exp\{-u/\phi_{S_h}\} + \exp\{-u/\phi_T\}\right), \text{ for } h = 1, 2. \tag{8}$$

Finally, using the definition of Cressie (1993, page 66, equation 2.3.19), the standardized cross-variogram between the two linear predictors is

$$\begin{aligned}\gamma_{12}(u) &= \frac{1}{2}E[(S_1(x) + T(x) - S_2(x') - T(x'))^2] \\ &= 1 - \exp\{-u/\phi_T\}.\end{aligned} \tag{9}$$

4 Inference

Let $y^\top = (y_{ij}, i = 1, \ldots, n, j = 1, \ldots, m_i)$ denote the vector of the observed MF densities y_{ij}, for the j-th person at the i-th village. Also, let $W^\top = (W_1^\top, \ldots, W_n^\top)$, where $W_i^\top = (S_1(x_i)+T(x_i), S_2(x_i)+T(x_i))$ is the bivariate vector of random effects associated with location x_i for $i = 1, \ldots, n$.

The marginal distribution of W is multivariate Gaussian with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2\left[\Sigma_{S_1} + \Sigma_T\right] & \sigma_1\sigma_2\Sigma_T \\ \sigma_1\sigma_2\Sigma_T & \sigma_2^2\left[\Sigma_{S_2} + \Sigma_T\right] \end{pmatrix},$$

where Σ_{S_h} and Σ_T are spatial covariance matrices associated with the processes $\mathcal{S}_h$, for $h = 1, 2$, and $\mathcal{T}$, respectively. Using $f(a)$ as a shorthand notation for "the density function f, of a random variable A," the likelihood function for the vector parameter θ is

$$L(\theta) = \int_{\mathbb{R}^{2n}} f(w)f(y|w)\, dw \tag{10}$$

where

$$f(y|w) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} f(y_{ij}|w_i)$$

and $f(y|w)$ is the density function corresponding to the cdf given by (7). We now rewrite the intractable integral in (10) as

$$\begin{aligned} L(\theta) &= \int_{\mathbb{R}^{2n}} \frac{f(w)f(y|w)}{f_0(w)f_0(y|w)} f_0(y,w)\, dw \\ &\propto \int_{\mathbb{R}^{2n}} \frac{f(w)f(y|w)}{f_0(w)f_0(y|w)} f_0(w|y)\, dw \\ &= E_{f_0(w|y)}\left[\frac{f(w)f(y|w)}{f_0(w)f_0(y|w)}\right], \end{aligned} \tag{11}$$

where $f_0(w)$ and $f_0(y|w)$ have the same distributions of $f(w)$ and $f(y|w)$ but with parameter vector θ_0 chosen as a "best guess" of the true value of θ. We then simulate B samples, say $w_{(k)}$, from $f_0(w|y)$ and approximate (11) as

$$L(\theta) \approx L_B(\theta) = \frac{1}{B}\sum_{k=1}^{B} \frac{f(w_{(k)})f(y|w_{(k)})}{f_0(w_{(k)})f_0(y|w_{(k)})}. \tag{12}$$

We then maximize (12), using numerical procedures, to obtain the MCML estimate $\hat{\theta}_B$ of θ. To improve the approximation of the likelihood function, we then repeat this procedure by setting $\theta_0 = \hat{\theta}_B$ and re-iterate until convergence.

To simulate from the distribution of $f_0(w|y)$, we adapt the MCMC algorithm proposed in Diggle & Giorgi (2016). Specifically, we use a Metropolis-adjusted Langevin MCMC algorithm to update the standardized vector of random effects $\tilde{W} = \hat{\Sigma}^{-1/2}(W - \hat{w})$, where $\hat{w}$ and $\hat{\Sigma}$ are the mode and the the inverse of the negative Hessian of $f_0(w|y)$ at $\hat{w}$, respectively.

Our predictive target, $R(x)$, is the probability that a randomly sampled individual at location x has an MF density above a predefined threshold c, hence

$$R(x_i) = \mathrm{P}\left\{Y(x_i) > c|W_i\right\} = \pi(x_i)\exp\left\{-\left[\frac{c}{\lambda(x_i)}\right]^{\kappa}\right\}, \text{ for } i = 1,\ldots,n. \tag{13}$$

5 Re-analysis of the Loa loa data

Table 1 reports MCML estimates for the model parameters. These were obtained by repeating the MCML algorithm five times and, each time, using 10,000 samples by retaining every tenth sample in 100,000 iterations after a burn-in of 10,000 iterations. The retained samples showed very good mixing and small autocorrelation up to lag five.

We also set $\phi_{S_1} = \phi_{S_2} = \phi_S$, as a pragmatic strategy to circumvent a rather flat likelihood surface for these parameters. A Wald test of the null hypothesis that $\log\{\phi_{S_1}/\phi_{S_2}\} = 0$ give a non-significant result ($p > 0.05$).

The estimates of μ_1, μ_2, σ_1^2, σ_2^2 and κ are comparable with those reported by Schlüter et al. (2016). Additionally, the processes $\mathcal{S}_1$ and $\mathcal{S}_2$ account for spatial variation in MF density up to about 35 km, beyond which their correlation function falls below 0.05, whilst the corresponding

spatial range of the $\mathcal{T}$ process is about 300 km. This spatial structure is clearly visible from the map of point predictions (conditional expectations) of $R(x)$ shown in Figure 1.

Figure 2 shows the standardized empirical variograms and cross-covariograms based on the point predictions of $Z_1(x)$ and $Z_2(x)$ delivered by the non-spatial model of Schlüter et al. (2016), as defined in equations (5) and (6). Each panel suggests the presence of residual spatial correlation, and does not show strong evidence against the fitted correlation structure.

The two panels of Figure 3 compare the point predictions and lengths of the 95% predictive intervals for the fraction of individuals with more than 8000 MF per ml of blood, from the fitted spatial and non-spatial models. Whilst the point estimates are in good agreement, the spatial model provides more accurate predictions, with lengths for the 95% predictive intervals almost always shorter than those from the non-spatial model. An intuitive explanation for this is that the spatial model can gain precision by borrowing strength of information from neighbouring villages.

Table 1: Monte Carlo maximum likelihood estimates and their 95% confidence intervals (CI) for the model of Section 3.

	Estimate	95% CI
μ_1	-2.187	(-2.230, -2.144)
μ_2	8.258	(8.190, 8.327)
σ_1^2	0.874	(0.663, 1.152)
σ_2^2	0.146	(0.111, 0.193)
ϕ_S	17.982	(13.012, 24.850)
ϕ_T	154.520	(72.402, 329.774)
κ	0.552	(0.537, 0.568)

6 Simulation study

The objectives of this simulation study were (1) to validate the properties of the predictive inferences from the spatial model of Section 3, and (2) to quantify the effects of residual spatial correlation on the predictive performance of the non-spatial model by Schlüter et al. (2016).

We generated 10,000 data-sets under the fitted spatial model in Section 5 by simulating, at each of the observed village location, the latent variables $S_1(x)$, $S_2(x)$ and $T(x)$, and the corresponding MF counts for each of the 19,128 sampled individuals. The model parameters were fixed at the MCML estimates shown in Table 1. For each of the simulated data-sets and each of the two models, we first estimated the model parameters, then computed plug-in point predictions of $R(x)$ at each of the village locations. Finally, we summarised the results at each village location by calculating the empirical root-mean-square-error (RMSE), predictive interval length (PIL) and coverage probability (CP) at nominal levels of 90%, 95% and 99% coverage.

Table 2 reports the values of each of these summary measures averaged over all sampled villages. Both the spatial and non-spatial models give predictive intervals with CP consistent

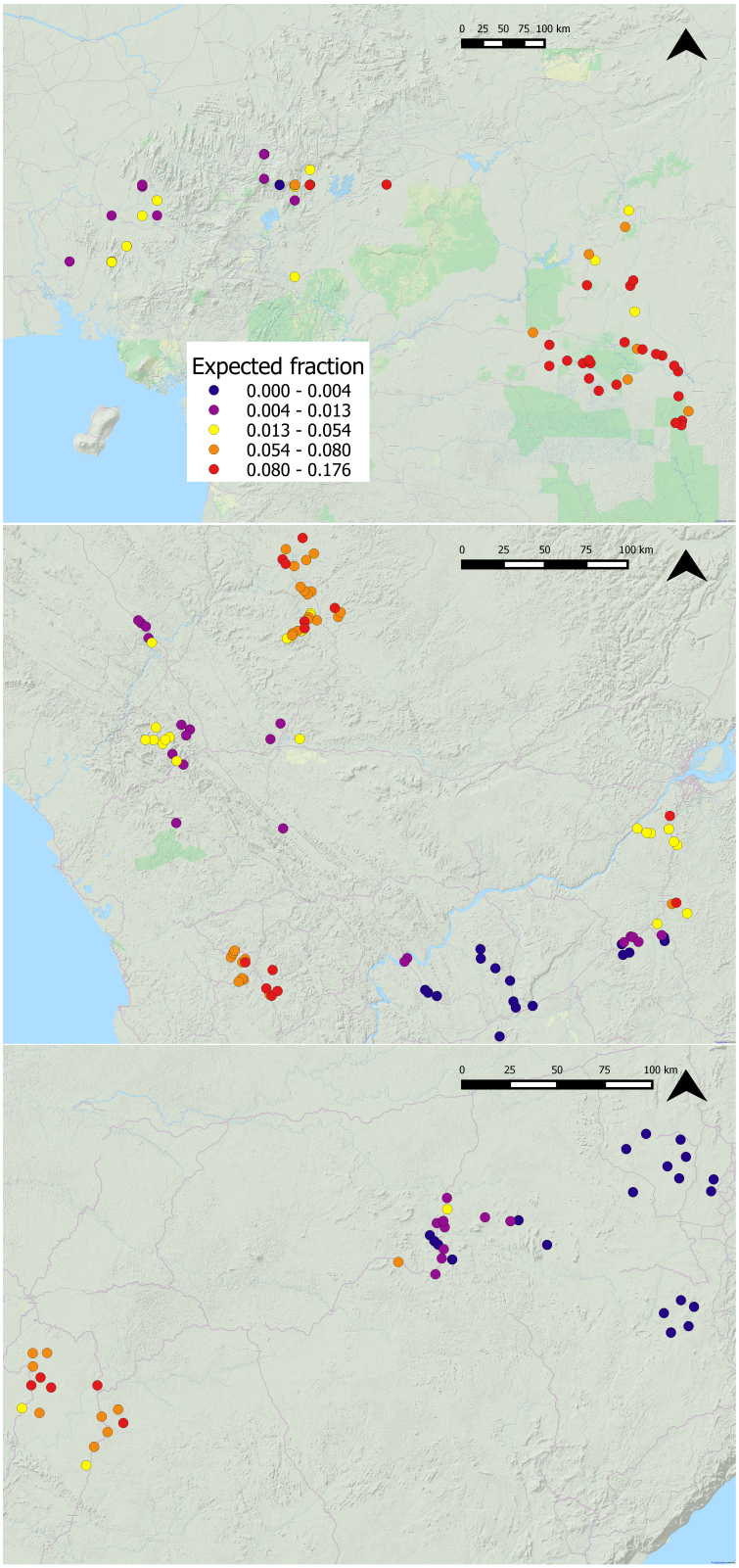

Figure 1: Expected fraction of individuals with more than 8000 MF counts per ml of blood, in each of the sampled viallges in the study sites in Cameroon (upper panel), the Republic of Congo (central panel) and the Democratic Republic of Congo (lower panel).

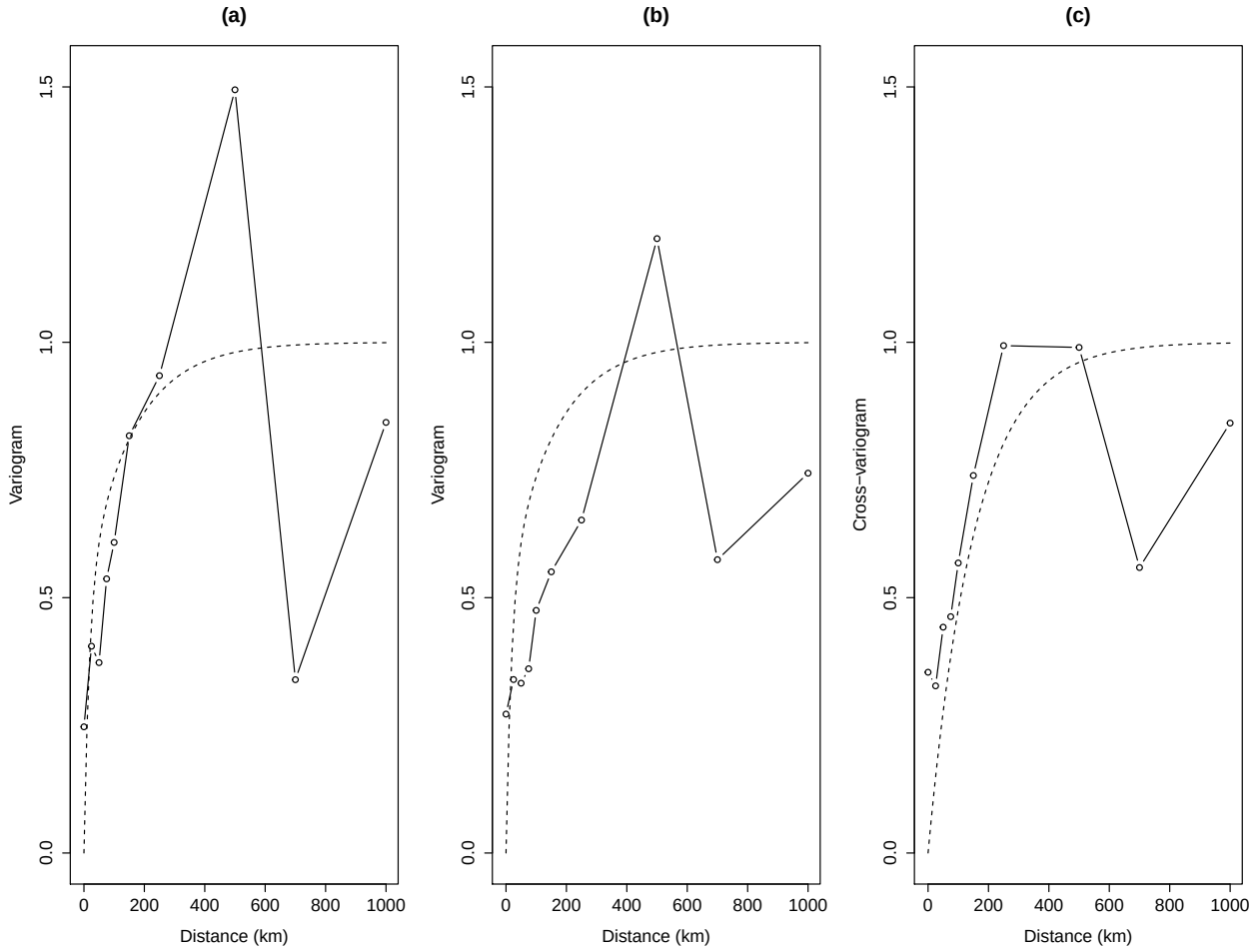

Figure 2: Empirical standardized variograms based on the predictive mean of the random effects associated with prevalence (a) and intensity (b), and their standardized cross-variogram (c), from the non-spatial model of Schlüter et al. (2016). The dashed lines represent the theoretical standardized variograms and cross-variogram from fitted model of Section 3, given by (8) and (9), respectively.

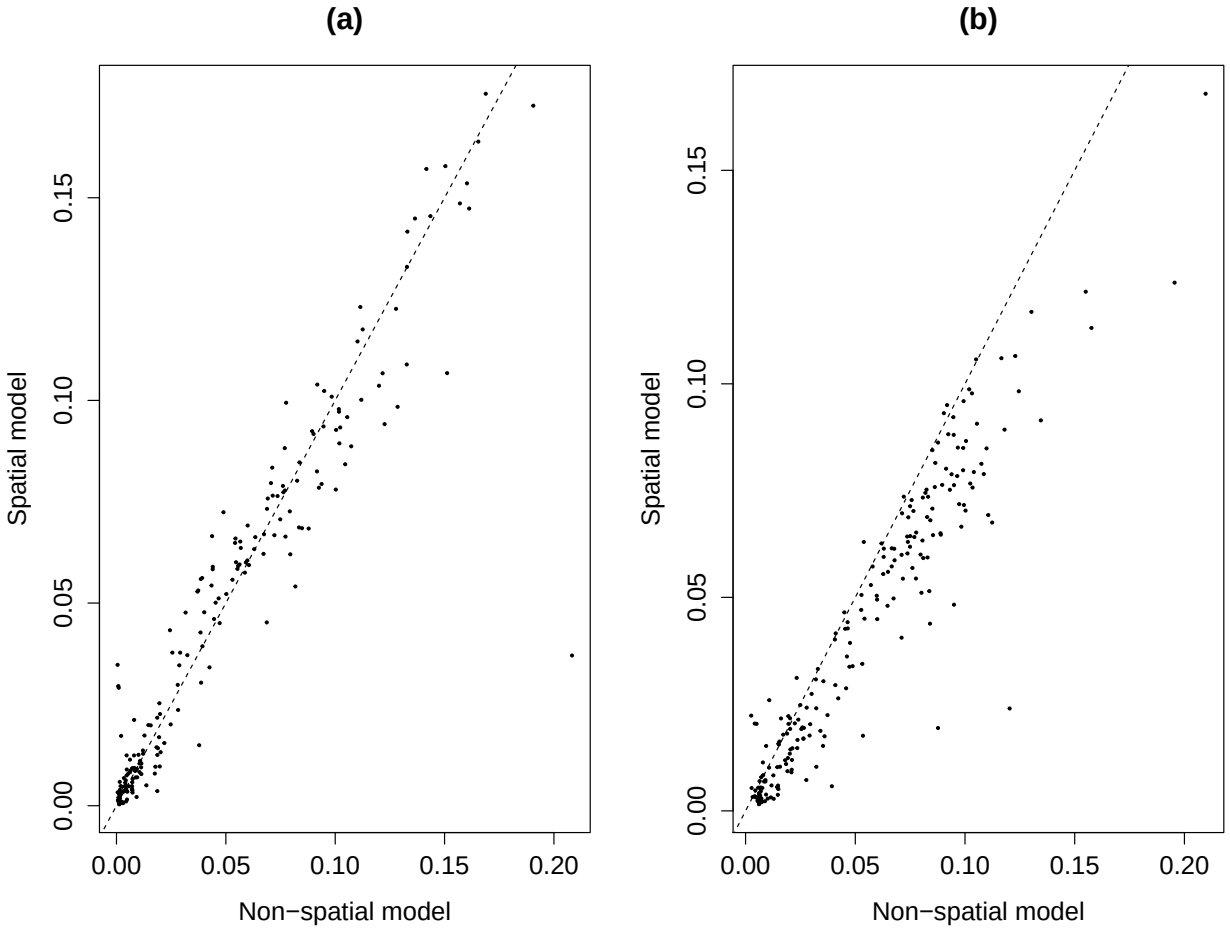

Figure 3: Scatter plot of the point estimates (a) and length of the 95% predictive intervals (b) for the prediction target, defined in (13) with $c = 8000$, from the non-spatial model of Schlüter et al. (2016) and the spatial model of Section 3.

with their nominal levels, but the spatial model shows a better predictive performance overall, as it gives accurate coverage alongside smaller average values of RMSE and PIL

Table 2: Simulation study summaries: root-mean-square-error (RMSE), length of the predictive interval with coverage with coverage ($\text{PIL}_{100\times\alpha}$) and their coverage probability ($\text{CP}_{100\times\alpha}$), with nominal level $\alpha = 0.90, 0.95, 0.99$, for the predictor of (13) from the spatial and non-spatial model, averaged across all the sampled villages.

Model	RMSE	PIL_{90}	PIL_{95}	PIL_{99}	CP_{90}	CP_{95}	CP_{99}
Spatial	0.013	0.033	0.040	0.053	0.895	0.947	0.989
Non-spatial	0.015	0.041	0.049	0.066	0.901	0.949	0.988

7 Discussion

We have developed a geostatistical zero-inflated Weibull-mixture model in order to study the relationship between prevalence and intensity of *Loa loa* infections. By re-analysing extensive *Loa loa* data previously used by Schlüter et al. (2016) and conducting a simulation study, we have shown that both the non-spatial model used in Schlüter et al. (2016) and our proposed spatial model deliver predictions of the proportion of highly infected individuals in a community that have the correct coverage properties, but the spatial model leads to shorter predictive intervals. These results support the notion, often implicitly assumed, that spatial correlation should be exploited whenever it is present, so as to make the best possible use of the information in the data.

The main advantages of the non-spatial model over the spatial model concern its practical utility in low resource settings. Once model parameters have been estimated with sufficient precision that parameter uncertainty is an order of magnitude smaller than prediction uncertainty, the non-spatial model can be applied to data from a newly sampled community with minimal computational and data-storage requirements; at the time of writing, the non-spatial model is being field-tested in rural west African communities. From this perspective, it is important that in the presence of spatial autocorrelation the non-spatial model leads to predictive intervals with approximately correct coverage properties, at the cost of some loss of precision.

Another issue inherent to the additional complexity of the spatial model is that it requires large amounts of data in order to estimate all model parameters with high precision. Our strategy for dealing with a flat likelihood surface was to reduce the number of parameters that determine the spatial covariance structure of the data. An alternative, if informative piors can be justified, would be to use Bayesian methods of inference.

The methods presented in this paper can be applied more widely to problems that involve spatially structured zero-inflation. The authors are currently applying these methods to the mapping of other NTDs, such as Onchocerciasis, soil-transmitted helminths and Lymphatic Filariasis. Another area of application is to the modelling of zero-inflated longitudinal data; see, for example, Olsen & Schafer (2001) on healthcare utilization, Rose et al. (2006) on

adverse events post-vaccination, Buu et al. (2012) on substance abuse and Min & Agresti (2005) on treatment comparison.

Acknowledgements

We thank the authors of Schlüter et al. (2016) for permission to re-use data analysed in that paper. Emanuele Giorgi holds an MRC Career Development Award in Biostatistics (MR/M015297/1).

References

Agarwal, D. K., Gelfand, A. E. & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics* **9**, 341–355.

Amek, N., Bayoh, N., Hamel, M., Lindblade, K. A., Gimnig, J., Laserson, K. F., Slutsker, L., Smith, T. & Vounatsou, P. (2011). Spatio-temporal modeling of sparse geostatistical malaria sporozoite rate data using a zero inflated binomial model. *Spatial and Spatio-temporal Epidemiology* **2**, 283–290.

Boussinesq, M., Gardon, J., Gardon-Wendel, N., Kamgno, J., Ngoumou, P. & Chippaux, J. (1999). Three probable cases of loa loa encephalopathy following ivermectin treatment for onchocerciasis. *Am J Trop Med Hyg* **58(4)**, 461–9.

Boussinesq, M., Gardon, J., Kamgno, J., Pion, S., Gardon-Wendel, N. & Chippaux, J. (2001). Relationships between the prevalence and intensity of loa loa infection in the central province of cameroon. *Ann Trop Med Parasitol* **95(5)**, 495–507.

Buu, A., Li, R., Tan, X. & Zucker, R. (2012). Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in medicine* **31**, 4074–4086.

Carme, B., Boulesteix, J., Boutes, H. & Puruehnce, M. (1991). 5 cases of encephalitis during treatment of loiasis with diethylcarbamazine. *Am J Trop Med Hyg* **44(6)**, 684–90.

CDC (2013a). Paraistes: Lymphatic filariasis. `https://www.cdc.gov/parasites/lymphaticfilariasis/epi.html`. Accessed: 28 October 2016.

CDC (2013b). Paraistes: Onchocerciasis. `http://www.cdc.gov/parasites/onchocerciasis/epi.html`. Accessed: 28 October 2016.

CDC (2015). Parasites: Loiasis. `https://www.cdc.gov/parasites/loiasis/disease.html`. Accessed: 28 October 2016.

Cressie, N. (1993). *Statistics for spatial data*. Wiley, New York.

Diggle, P., Thomson, M., Christensen, O., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J., Boussinesq, M. & Molyneux, D. (2007). Spatial modelling and the prediction of loa loa risk: decision making under uncertainty. *Annals of Tropical Medicine and Parasitology* **101**, 499–509.

Diggle, P. J. & Giorgi, E. (2016). Model-based geostatistics for prevalence mapping in low-resource setting (with discussion). *Journal of the American Statistical Association* DOI: 10.1080/01621459.2015.1123158.

Firth, D. (1988). Multiplicative errors: Log-normal or gamma? *Journal of the Royal Statistical Society. Series B (Methodological)* **50**, 266–268.

Gardon, J., GardonWendel, N., DemangaNgangue, D., Kamgno, J., Chippaux, J. & Boussinesq, M. (1997). Serious reactions after mass treatment of onchocerciasis with ivermectin in an area endemic for loa loa infection. *Lancet* **350(9070)**, 18–22.

Geary, T. (2012). A step toward eradication of human filariases in areas where loa is endemic. *MBio* .

Giardina, F., Gosoniu, L., Konate, L., Diouf, M. B., Perry, R., Gaye, O., Faye, O. & Vounatsou, P. (2012). Estimating the burden of malaria in Senegal: Bayesian zero-inflated binomial geostatistical modeling of the MIS 2008 data. *PLoS ONE* **7**, e32625.

Keating, J., Yukich, J., Mollenkopf, S. & Tediosi, F. (20014). Lymphatic filariasis and onchocerciasis prevention, treatment, and control costs across diverse settings: A systematic review. *Acta Trop* **135**, 86–95.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Mackenzie, C., Homeida, M., Hopkins, A. & Lawrence, J. (2012). Elimination of onchocerciasis from africa: possible? *Trends Parasitol* **28(1)**, 16–22.

Min, Y. & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**, 1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.

Olsen, K. & Schafer, L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.

Oluwole, A. S., Ekpo, U. F., Karagiannis-Voules, D.-A., Abe, E. M., Olamiju, F. O., Isiyaku, S., Okoronkwo, C., Saka, Y., Nebe, O. J., Braide, E. I., Mafiana, C. F., Utzinger, J. & Vounatsou, P. (2015). Bayesian geostatistical model-based estimates of soil-transmitted helminth infection in nigeria, including annual deworming requirements. *PLoS Negl Trop Dis* **9**, 1–15. DOI: 10.1371/journal.pntd.0003740.

Pion, S., Filipe, J., Kamgno, J., Gardon, J., Basanez, M. & Boussinesq, M. (2006). Microfilarial distribution of loa loa in the human host: population dynamics and epidemiological implications. *Parasitology* **133**, 101–9.

Rose, C. E., Martin, S. W., Wannemuehler, K. A. & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics* **16**, 463–481.

Schlüter, D. K., Ndeffo-Mbah, M. L., Takougang, I., Ukety, T., Wandji, S., Galvani, A. P. & Diggle, P. J. (2016). Using community-level prevalence of *loa loa* infection to predict the proportion of highly-infected individuals: Statistical modelling to support lymphatic filariasis and onchocerciasis elimination programs. *PLoS Neglected Tropical Diseases.* In press.

Takougang, I., Meremikwu, M., Wanji, S., Yenshu, E., Aripko, B., Lamlenn, S., Ekka, B., Enyong, P., Meli, J., Kale, O. & Remme, J. (2002). Rapid assessment method for prevalence and intensity of loa loa infection. *Bull World Health Organ* **80(11)**, 852–8.

Taylor, M., Hoerauf, A. & Bockarie, M. (2010). Lymphatic filariasis and onchocerciasis. *Lancet* **376(9747)**, 1175–85.

Thomson, M., Obsomer, V., Kamgno, J., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Remme, J., Molyneux, D. & Boussinesq, M. (2004). Mapping the distribution of Loa loa in Cameroon in support of the African Programme for Onchocerciasis Control. *Filaria Journal* **3**, 7. DOI:10.1186/1475-2883-3-7.

Wanji, S., Akotshi, D., Kankou, J., Nigo, M., Tepage, F., Ukety, T., Diggle, P. & Remme, J. (2012). The validation of the rapid assessment procedures for loiasis (RAPLOA) in the Democratic Republic of Congo: health policy implications. *Parasites and Vectors* **5**, 25 doi:10.1186/1756--3305--5--25.

WHO (2012). Provisional strategy for interrupting lymphatic filariasis transmission in loiasis-endemic countries; report of the meeting on lymphatic filariasis,malria ans integrated vector management. Tech. rep., WHO.

WHO (2016a). Fact sheet: Lymphatic filariasis. http://www.who.int/mediacentre/factsheets/fs102/en/. Accessed: 18 October 2016.

WHO (2016b). Fact sheet: Onchocerciasis. http://www.who.int/mediacentre/factsheets/fs374/en/. Accessed: 28 October 2016.