Modelling of spatio-temporal zero truncated patterns in infectious disease surveillance data

Oyelola A. Adegboye[1*], Denis H.Y. Leung[2], You-Gan Wang[3]

[1]Department of Mathematics, Physics and Statistics, Qatar University, Doha, Qatar
email: o.adegboye@qu.edu.qa

[2]School of Economics, Singapore Management University, Singapore
email: denisleung@smu.edu.sg

[3]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.
email: you-gan.wang@qut.edu.au

Abstract

This paper is motivated by spatio-temporal pattern in the occurrence of Leishmaniasis in Afghanistan and the relatively high number of zero counts. We hold the view that correlations that arise from spatial and temporal sources are inherently distinct. Our method decouples these two sources of correlations, there are at least two advantages in taking this approach. First, it circumvents the need to inverting a large correlation matrix, which is a commonly encountered problem in spatio-temporal analyses. Second, it simplifies the modelling of complex relationships such as anisotropy, which would have been extremely difficult or impossible if spatio-temporal correlations were simultaneously considered. We identify three challenges in the modelling of a spatio-temporal process: (1) accommodation of covariances that arise from spatial and temporal sources; (2) choosing the correct covariance structure and (3) extending to situations where a covariance is not the natural measure of association. Moreover, because the data covers a period that overlaps with the US invasion of Afghanistan, the high number of zero counts may be the result of no disease incidence or lapse of data collection. To resolve this issue, a model truncated at zero built on a foundation of the generalized estimating equations was proposed.

Keywords: generalized estimating equations; overdispersion; poisson; spatio-temporal; Leishmaniasis

1 Introduction

One of the most challenging issues in modelling spatio-temporal infectious disease data is the choice of a valid and yet flexible correlation (covariance) structure. Some examples of

doi:10.20944/preprints201703.0065.v1

correlation structures can be found in Cressie and Huang (1999); Gneiting (2002); Stein (2005) and Porcu et al. (2007), among others. The correlation structures fall into one of two types: separable in which case it is assumed that the space-time correlation can be written as a product of a correlation for the space dimension and one for the time dimension or non-separable where the space-time correlation is modelled as a single entity. Unfortunately, most of these correlation structures are either extremely complicated or infeasible to manipulate due to their high dimensions.

The motivating data is a surveillance nationally aggregated monthly counts of leishmaniasis incidence across provinces of Afghanistan between 2003 and 2009. Leishmaniasis is the third most common vector-borne disease and a very important protozoan infection Adegboye et al. (2016). The burden of the disease is overwhelming and the psychological effect can be disturbing. The impact of environmental influences on Leishmaniasis cannot be ruled out and human activities play a significant role in the dispersion of the vectors thereby changing the geographical distribution of the disease. The major goal of the study is to evaluate the influence of Satellite-derived climatic and environmental variable on the occurrence of leishmaniasis while allowing for spatio-temporal correlation in the data. Furthermore, in most previous works, the space-time correlation is considered jointly, a step that we believe is unnecessary or unrealistic. For example, it would be hard to imagine how the disease incidence in Kabul in 2003 would be in any way correlated to that in Hilmand province in 2007, because of their location and time apart. Previous study has shown significant seasonality in the occurrence of the diseaseJanuary to March and September to Decemberwith the highest peak in March, suggesting a peak in the cases of leishmaniasis in March and a through in September of each year Adegboye and Adegboye (2016).

We hold the view that correlations arising from spatial and temporal sources are inherently distinct. In this study we shall decouples these two sources of correlations, an approach that separates the modelling of the space- and time-correlations. There are at least two advantages in taking this approach. First, it circumvents the need to inverting a large correlation matrix, which is a commonly encountered problem in spatio-temporal analyses (*e.g.*, Yasui and Lele, 1997). Second, it simplifies the modelling of complex relationships such as anisotropy, which would have been extremely difficult or impossible if spatio-temporal correlations were simultaneously considered.

Our method begins with a marginal model for Leishmaniasis incidence along the time dimension. Marginal models are natural extensions of the generalized linear models and they are popular in longitudinal analysis. They are well understood and they can be easily fitted using simple modifications to existing programs. One of the most popular marginal models is the generalized estimating equations (GEE, Liang and Zeger, 1986). The standard GEE assumes longitudinal measurements within each observation are correlated but observations

doi:10.20944/preprints201703.0065.v1

are independent of each other. But in our situation, the observations are disease counts and covariates for the different spatial locations and there may be spatial dependencies.

To account for spatial dependency, we create a spatial-GEE by re-weighting the standard GEE so that locations highly correlated with each other would receive less weight. The weights are created from a semivariogram of the spatial data. Because the dimension of a semivariogram is only dependent on the number of spatial locations, it is of manageable size. Furthermore, since a semivariogram measures dissimilarity, there is no need to invert the semivariogram to create weights. The method will be illustrated using data from Leishmaniasis incident in the provinces of Afghanistan in 2009. This model makes it possible to combine the specific provincial rate with the influence of the spatial neighborhood.

The data sets is also characterized by a high percentage of zero disease counts. The data covers a period that overlaps with the US invasion of Afghanistan, the zero counts may be the result of no disease incidence or lapse of data collection. It is a common practice in large survey to use zero (0) as missing value. Faraway (2004) argued that although this is not a good choice since it is a valid value for some of the variables and not mentioning it in their data description, unfortunately this act is common particularly with data sets of any size or complexity .

Apart from the spatial dependency in the data, this study also presents additional challenge. It is very difficult to distinguish between ”true” and ”imputed” zeros, because of the reporting mechanism of disease in Afghanistan (due to security, technical and logistics issues). These problems prompted us to consider the option of discarding the zeros and model the non-zero data using a Poisson model conditional on greater zero. We make the assumption that ”imputed” zeros are a random event. It is often practiced to truncate the values that are bigger than a constant to overcome over-dispersion (Saffari et al., 2011). The analysis of truncated data often arises from a subsidiary set of results that treat a practical problem of how data are gathered and analyzed (Greene, 2005) and incompleteness of this data requires special estimators of the regression coefficients (Karlsson and Lindmark, 2014). To resolve this issue, we use a model truncated at zero. Lee and Kim (1998) provides detail review and a comparison of properties of estimators for regression models under truncated data.

The rest of the paper is structured as follows. In Section 2, we describe the data collection method and the variables that will be used in the study. Section 3 describes the method. Section 4, we give the results of the data analysis, while Section 5 concludes the paper.

2 Methods

2.1 Model specification

Let $\mathbf{y} = \{y(s,t), s = s_1, ..., s_S, t = t_1, ..., t_T\}$ where $y(s,t) \equiv y_{st}$ denotes the count of disease at spatial location s and time t. Suppose associated with (s,t) are covariates $\mathbf{x}(s,t) \equiv \mathbf{x}_{st}$ that record the spatial location and time as well as other information that might affect disease counts. Furthermore, let $\mathbf{X} = \{\mathbf{x}(s,t), s = s_1, ..., s_S, t = t_1, ..., t_T\}$ and the disease incidence at each location is model as a Poisson count.

To model spatio-temporal correlation (or equivalently covariance) and overdispersion, we assume there is a non-negative weakly stationary latent process e_{st} such that conditional on the e's, the y's are independent and are assumed to follow a log-linear model given by

$$\mathrm{E}(y_{st}|e_{st}) = \exp(\mathbf{x}_{st}^{\tau}\boldsymbol{\beta})e_{st}, \quad \text{and} \quad \mathrm{var}(y_{st}|e_{st}) = \phi\mathrm{E}(y_{st}|e_{st}),$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters that captures the association between incidence and the covariates. We assume $\mathrm{E}(e_{st}) = 1$ so $\exp(\mathbf{x}_{st}\boldsymbol{\beta})$ represents the marginal mean of y_{st}. The latent process e_{st} is assumed to have a variance of σ^2 and the covariance between e_{st} and $e_{s't'}$ is given by

$$\mathrm{cov}(e_{st}, e_{s't'}) = \sigma^2\rho(\mathbf{z}_{st}, \mathbf{z}_{s't'}, \boldsymbol{\alpha})$$

where $\mathbf{z}_{st}, \mathbf{z}_{s't'}$ are covariates from $(s,t), (s',t')$ that jointly induce spatio-temporal correlation and $\boldsymbol{\alpha}$ are unknown parameters. Depending on the context, the covariates $\mathbf{z}_{st}$ and $\mathbf{x}_{st}$ may be distinct, may share some components or may be the same. This model was considered by Zeger (1988) to model discrete time series data. Under these assumptions, it can easily be shown that

$$\begin{aligned} \mathrm{E}(y_{st}) &= \exp(\mathbf{x}_{st}^{\tau}\boldsymbol{\beta}) \equiv \mu_{st}(\boldsymbol{\beta}), && (1)\\ \mathrm{var}(y_{st}) &= \mu_{st}(\boldsymbol{\beta}) + \mu_{st}(\boldsymbol{\beta})^2\sigma^2, && (2)\\ \mathrm{corr}(y_{st}, y_{s't'}) &= \rho(\mathbf{z}_{st}, \mathbf{z}_{s't'}, \boldsymbol{\alpha})[\{1 + (\sigma^2\mu_{st}(\boldsymbol{\beta}))^{-1}\}\{1 + (\sigma^2\mu_{s't'}(\boldsymbol{\beta}))^{-1}\}]^{\frac{1}{2}}. && (3) \end{aligned}$$

If $\rho(\mathbf{z}_{st}, \mathbf{z}_{s't'}, \boldsymbol{\alpha}) = 0$, then we have a Poisson model with overdispersion. Furthermore, if $\sigma^2 = 0$, then we have a standard Poisson model at each spatial location and time. For convenience, we define $\mathbf{y}_{s\cdot} = (y_{s1}, ..., y_{sT})^{\tau}$ as the vector of counts taken at times $1, ..., T$ at spatial location s, $\mathbf{y}_{\cdot t} = (y_{1t}, ..., y_{St})^{\tau}$ as the vector of counts taken at locations $1, ..., S$ at time t and we use similar definitions for $\mathbf{x}_{s\cdot}$, $\mathbf{z}_{s\cdot}$, $\boldsymbol{\mu}_{s\cdot}$ and $\mathbf{x}_{\cdot t}$, $\mathbf{z}_{\cdot t}$, $\boldsymbol{\mu}_{\cdot t}$.

We begin by considering the data $\mathbf{y} = (\mathbf{y}_{1\cdot}^{\tau}, ..., \mathbf{y}_{S\cdot}^{\tau})^{\tau}$ and $\mathbf{X} = (\mathbf{x}_{1\cdot}^{\tau}, ..., \mathbf{x}_{S\cdot}^{\tau})^{\tau}$ as a set of longitudinal data over S spatial locations. If we temporarily treat the observations between spatial locations to be independent of each other, then the GEE (Liang and Zeger, 1986) is given by:

$$\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \sum_{s=s_1}^{s_S} \mathbf{D}_s^{\tau} \mathbf{V}_s^{-1} (\mathbf{y}_{s\cdot} - \boldsymbol{\mu}_{s\cdot}) = 0, \tag{4}$$

where $\mathbf{D}_s = \partial \boldsymbol{\mu}_{s\cdot} / \partial \boldsymbol{\beta}^{\tau}$ and $\mathbf{V}_s$ is the covariance matrix of $\mathbf{y}_{s\cdot}$. The matrix $\mathbf{V}_s$ can be expressed as $\mathbf{A}_s^{1/2} \mathbf{R}_s(\boldsymbol{\alpha}) \mathbf{A}_s^{1/2}$, where $\mathbf{A}_s = \text{diag}[\mu_{s1}(\boldsymbol{\beta}) + \mu_{st}(\boldsymbol{\beta})^2 \sigma^2, ..., \mu_{sT}(\boldsymbol{\beta}) + \mu_{st}(\boldsymbol{\beta})^2 \sigma^2]$ and $\mathbf{R}_s(\boldsymbol{\alpha})$ is a correlation matrix, with the (t, t')-th element representing the correlation between times t and t' at location s. If we let $v_{s,tt'}^{-1}$ as the (t, t') element of $\mathbf{V}_s^{-1}$, then (4) can be written as

$$\sum_{s=s_1}^{s_S} \sum_{t=t_1}^{t_T} \sum_{t'=t_1}^{t_T} \frac{\partial \mu_{st}}{\partial \boldsymbol{\beta}^{\tau}} v_{s,tt'}^{-1} \{y_{st'} - \mu_{st'}\} \equiv \sum_{s=s_1}^{s_S} \sum_{s'=s_1}^{s_S} \sum_{t=t_1}^{t_T} \sum_{t'=t_1}^{t_T} \frac{\partial \mu_{st}}{\partial \boldsymbol{\beta}^{\tau}} w_{ss'} v_{s,tt'}^{-1} \{y_{st'} - \mu_{st'}\} = 0, \tag{5}$$

where $w_{ss'} = 1, s = s'$ and $w_{ss'} = 0, s \neq s'$. We can compare (5) to a set of estimating equations that considers space-time correlation simultaneously. Let the data $\mathbf{y}$ be stacked as a $S \times T$ vector. Then a set of estimating equations can be written as

$$\mathbf{D}^{\tau} \tilde{\mathbf{V}}^{-1} \{\mathbf{y} - \boldsymbol{\mu}\} \equiv \sum_{s=s_1}^{s_S} \sum_{s'=s_1}^{s_S} \sum_{t=t_1}^{t_T} \sum_{t'=t_1}^{t_T} \frac{\partial \mu_{s't'}}{\partial \boldsymbol{\beta}^{\tau}} \tilde{v}_{st,s't'}^{-1} \{y_{st} - \mu_{st}\} = 0, \tag{6}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_{1\cdot}^{\tau}, ... \boldsymbol{\mu}_{s\cdot}^{\tau})^{\tau}$ and $\mathbf{D} = \partial \boldsymbol{\mu} / \partial \boldsymbol{\beta}^{\tau}$ and $\tilde{\mathbf{V}}$ is the covariance matrix of $\mathbf{y}$ and we let $\tilde{v}_{st,s't'}^{-1}$ be the (st, st')-th element of $\tilde{\mathbf{V}}^{-1}$. We can interpret (6) as a linear combination of $\partial \mu_{s't'} / \partial \boldsymbol{\beta}^{\tau} \{y_{st} - \mu_{st}\}$ with coefficients given by $\tilde{v}_{st,s't'}^{-1}$. Comparing (6) to (5), we observe that a standard GEE is also a linear combination but it replaces $\tilde{v}_{st,s't'}^{-1}$ with $w_{ss'} v_{s,tt'}^{-1}$.

2.2 Parameter estimation

Consider the following; suppose we remove all $y_{st} = 0$, then conditioned on $y_{st} > 0$, (1)-(2) become

$$\text{E}(y_{st} | e_{st}) = c\mu_{st}(\boldsymbol{\beta}) e_{st}, \qquad \text{var}(y_{st} | e_{st}) = [c\mu_{st}(\boldsymbol{\beta}) + c(1 - c)\mu_{st}(\boldsymbol{\beta})^2] e_{st}$$

where $c = 1/[1 - \exp(-\mu_{st})]$, leading to

$$\begin{aligned} \text{E}(y_{st}) &= c\mu_{st}(\boldsymbol{\beta}) \equiv \phi_{st}(\boldsymbol{\beta}), & (7) \\ \text{var}(y_{st}) &= c\mu_{st}(\boldsymbol{\beta}) + c(1 - c)\mu_{st}(\boldsymbol{\beta})^2 + c^2 \mu_{st}(\boldsymbol{\beta})^2 \sigma^2. & (8) \end{aligned}$$

Let $\mathbf{d} = \{d(s, t) = d_{st}\}_{S \times T}$ be a matrix of indicators such that $d_{st} = 1$ if $y_{st} > 0$ and $d_{st} = 0$ otherwise. Note that $y_{st} = 0$ could mean the count was zero or count was not taken. To resolve the missing counts, we assumed counts were missing completely at random. The problem of counts not missing completely at random can be handled by adding an extra model for the propensity of $d_{st} = 1$. However, we wanted to illustrate the idea of a zero truncated spatio-temporal GEE and so we chose to minimise any distraction to this main idea. For a particular set of variance covariance matrix $v_{s,tt'}$ and spatial weight $\tilde{w}_{ss'}$, the spatio-temporal GEE conditioned only on those observations with $y_{st} > 0$ can be written as

$$\tilde{\mathbf{U}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \sum_{s=s_1}^{s_S} \sum_{s'=s_1}^{s_S} \sum_{t=t_1}^{t_T} \sum_{t'=t_1}^{t_T} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} \tilde{w}_{ss'} v_{s,tt'}^{-1} \{y_{st'} - \phi_{st'}\} = 0, \tag{9}$$

where $v_{s,tt'}$ is the t,t-th element of $\mathbf{V}_s$, the covariance matrix of $\mathbf{y}_{s\cdot}$. The matrix $\mathbf{V}_s$ can be expressed as $\mathbf{A}_s^{1/2}\mathbf{R}_s(\boldsymbol{\alpha})\mathbf{A}_s^{1/2}$, where $\mathbf{A}_s = \text{diag}[c\mu_{s1}(\boldsymbol{\beta}) + c(1-c)\mu_{s1}(\boldsymbol{\beta})^2 + c^2\mu_{s1}(\boldsymbol{\beta})^2\sigma^2, ..., c\mu_{sT}(\boldsymbol{\beta}) + c(1-c)\mu_{sT}(\boldsymbol{\beta})^2 + c^2\mu_{sT}(\boldsymbol{\beta})^2\sigma^2]$ and $\mathbf{R}_s(\boldsymbol{\alpha})$ is a matrix with its (t,t')-th element representing the correlation between times t and t' at location s.

Our primary interest lies in the parameters $\boldsymbol{\beta}$ but we also must deal with the nuisance parameters $\boldsymbol{\alpha}$. Let $\mathbf{R}(\boldsymbol{\alpha})$ be a 84×84 matrix where $\boldsymbol{\alpha}$ contains the parameters (θ) estimated via weighted least square method as described in Section 3.2. The parameters are estimated via a Newton-Raphson iteration method. To solve for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ jointly, we employed the method of (Prentice, 1988). Let $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\alpha}}_k$ be the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ at the k-th iteration. We first fitted a GEE with an independence working correlation structure, we then solve the estimating equation for $\boldsymbol{\alpha}$, and we then iterate until convergence. This step gives the values $v_{s,tt'}$. Denoting $\sum_{s,s',t,t'} \equiv \sum_{s=s_1}^{s_S} \sum_{s'=s_1}^{s_S} \sum_{t=t_1}^{t_T} \sum_{t'=t_1}^{t_T}$, we estimate an initial estimate $\hat{\boldsymbol{\beta}}_0$ using (5) by assuming an identity matrix for $\mathbf{R}_s(\boldsymbol{\alpha})$, equivariance, *i.e.*, $v_{s,tt'}^{-1} = 1$ and, spatial weight. Then at iteration k,

$$\hat{\boldsymbol{\beta}}_{k+1} = \hat{\boldsymbol{\beta}}_k - \left[\sum\nolimits_{s,s',t,t'} \frac{\partial \phi_{st}(\hat{\boldsymbol{\beta}}_k)}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} \tilde{w}_{ss'} v_{s,tt'}^{-1} (\hat{\boldsymbol{\beta}}_k) \frac{\partial \phi_{st}(\hat{\boldsymbol{\beta}}_k)}{\partial \boldsymbol{\beta}^{\tau}}\right]^{-1} \left[\sum\nolimits_{s,s',t,t'} \frac{\partial \phi_{st}(\hat{\boldsymbol{\beta}}_k)}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} \tilde{w}_{ss'} v_{s,tt'}^{-1} \{y_{st'} - \phi_{st'}(\hat{\boldsymbol{\beta}}_k)\}\right] \tag{10}$$

Here we take the slope of the linear regression of $\log(\hat{r}_{st}^k \hat{r}_{st'}^k)$ on $\log(|t - t'|)$ as $\hat{\boldsymbol{\alpha}}_k$. We then iterate between (9) and (10) until convergence.

The standard errors for the parameter estimates can be obtained via the large-sample properties (Liang and Zeger, 1986; Leung et al., 2009; Paul et al., 2013). Under mild regularity conditions, $K^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{GEE} - \boldsymbol{\beta})$ is asymptotically multivariate Gaussian with zero mean and covariance matrix given by:

$$\lim_{K \to \infty} K \left[\sum\nolimits_{s,s',t,t'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} \tilde{w}_{ss'} v_{s,tt'}^{-1} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}}\right]^{-1} \left[\sum\nolimits_{s,s',t,t'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} \tilde{w}_{ss'} v_{s,tt'}^{-1} cov(y_{st'}) \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}}\right] \left[\sum\nolimits_{s,s',t,t'} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}} d_{st'} \tilde{w}_{ss'} v_{s,tt'}^{-1} \frac{\partial \phi_{st}}{\partial \boldsymbol{\beta}^{\tau}}\right]^{-1} \tag{11}$$

However, in our experience, the sandwich formula does not work very well. Instead, we use resampling by blocked jackknife (see, e.g., Künsch, 1989 and Sherman, 2011, chapter 10). Let $\hat{\beta}_{-k}$ be an estimate of β with data from the k-th province removed and let $\bar{\beta} = 1/K \sum_{k=1}^{K} \hat{\beta}_{-k}$. We then define the resampled estimate of $\text{var}(\hat{\beta})$ by

$$\widehat{\text{var}(\hat{\beta})} = \frac{K-1}{K} \sum_{k=1}^{K} (\hat{\beta}_{-k} - \bar{\beta})^2. \tag{12}$$

doi:10.20944/preprints201703.0065.v1

3 Application to Afghanistan leishmaniasis incidence data

3.1 Data description

The above methodology was applied to the analysis of leishmaniasis incidence data in Afghanistan between 2003 and 2009. The data sets were monthly cases of leishmaniasis reported to the Afghanistan Health Management Information System (HMIS) under the National Malaria and Leishmaniasis Control Programme (NMLCP) of the Ministry of Public Health (MoPH). Leishmaniasis infections were confirmed clinically or calibrated ocular micrometer supported binocular light microscopy of Leishmania parasites. The data consists of 148,945 new cases of Leishmaniasis from 20 provinces in Afghanistan between 2003 and 2009 (of these, 41,072 occurred in 2009)(Figure 1). Satellite-derived environmental and climatic data such as accumulated rainfall, land surface temperature and Wind were obtained from the National Aeronautics and Space Administration-NASA Earth Observations (NEO) [http://earthobservatory.nasa.gov

Figure 2 presents the distribution of the disease incidence across the 20 provinces with available data sets. A striking feature of the data is the high number of zero incidence for many locations. Many of the provinces have counts of zeros for months, then a sudden jump to a few hundreds or thousands, then back to zero. Between 2003 and 2006, most of the provinces reported no cases of Leishmaniasis; this claim cannot be verified because this period coincides with the US led war in Afghanistan and disease reporting may only be possible in a relatively safe environment.

3.2 Model fitting

The modelling is a 2-step process, we need to find the spatial weight, $\tilde{w}_{ss'}$, then the variance-covariance matrix, $v_{s,tt'}$ that induce will the temporal dependency. Recall the dimension of $\tilde{\mathbf{V}}$ is $ST \times ST$. For the current data set, $S = 20$ represents the number of provinces and $T = 7$ represents the number of years with recorded data. If we use the monthly data, which would allow us to study how seasonality affects the transmission (incidence) of the disease, then $T = 84$ and so $S \times T = 20 \times 84 = 1680$ and therefore $\tilde{\mathbf{V}}$ would be a matrix that cannot feasibly be handled. Furthermore, as we argued in Section 1, the correlation between $y(s,t)$ and $y(s',t')$ often does not have any practical meaning so (6) does not seem to be a route that we should follow.

Firstly, our idea is to consider the spatial correlations that are omitted by the standard GEE. Therefore, we propose using a set of weight $w_{ss'}$ different from those in (5). The weights are derived from a commonly used measure of spatial correlation, the semivariogram (see, *e.g.*, Cressie, 1993). For the data in this paper, we define an empirical semivariogram

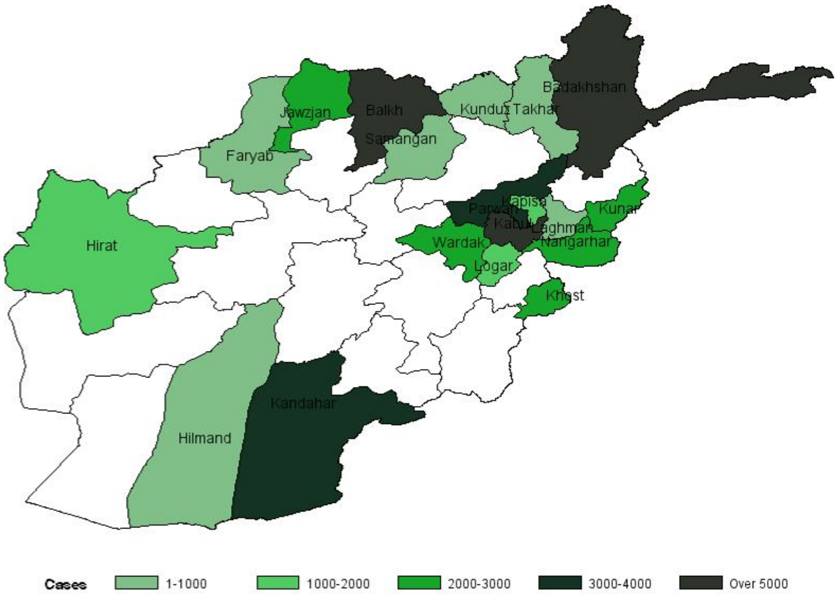

Figure 1: Map showing the distribution of new cases of Leishmaniasis in Afghanistan in 2009

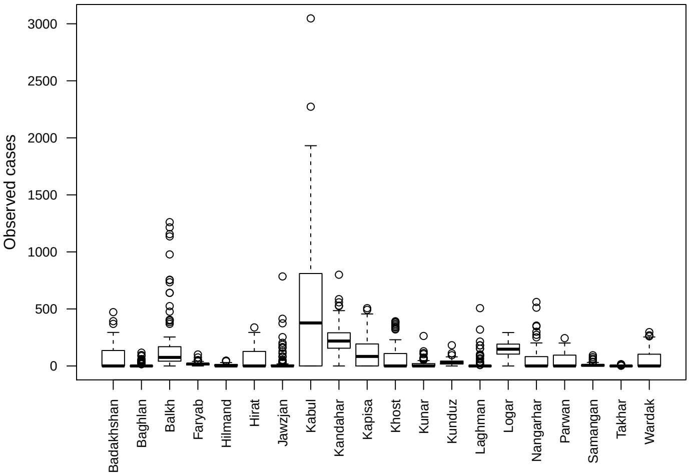

Figure 2: Distribution of total cases of Leishmaniasis at provincial level in Afghanistan (2003-2009)

as

$$\hat{\gamma}(h) = \frac{1}{2N(h) \times T} \sum_{t=t_1}^{t_T} \sum_{s,s' \in N(h)} (y_{st} - y_{s't})^2, \tag{13}$$

where h is a lag distance between spatial locations s and s' and $N(h)$ is the number of pairs of spatial locations separated by no more than h. The empirical variogarms is computed at different time scale and then averaged over the same spatial lags. Unlike covariance or correlation, which measure similarity, semivariogram measures dissimilarity. Hence, if we use semivariogram to create weights, we do not have to carry out any matrix inversion. This empirical semivariogram (13) can be used to fit a parametric semivariogram model, *e.g.*, nugget, exponential or Gaussian, for illustration, we have chosen powered exponential;

$$\gamma(h, \alpha) = \tau^2 + \sigma^2 \left(1 - e^{-|h/\phi|^q}\right), h > 0 \tag{14}$$

where $0 < q \leq 2, \phi > 0$ and $\alpha = (\tau^2, \sigma^2, \phi)$. The quantities τ^2, σ^2 and ϕ represent the nugget, sill, and range, respectively. This semivariogram includes as special cases the exponential $(q = 1)$ and Gaussian $(q = 2)$. The corresponding correlation function has the form

$$\rho(h, \alpha) = \frac{\sigma^2}{\sigma^2 + \tau^2} e^{-|h/\phi|^q}, h > 0. \tag{15}$$

An exponential model was used to obtain the parameters that were used in the construction of the spatial weights matrix.

Secondly, we also need to construct the variance matrix, $\mathbf{V}_s$ to induce the temporal dependency. Recall that for fixed a s, $v_{s,tt'}$, $t, t' = t_1, ..., t_T$ are the elements of the variance covariance matrix of disease counts between times. The data set consists of monthly disease counts for each province that were captured over 7 years, from 2003-2009, with up to 84 observations per province. However, as mentioned before, year is an artificial variable that is not of interest. On the contrary, there might be two different types of temporal correlations: (1) Between months that are nearby and (2) Between the same month in different years (seasonality). A simple temporal correlation function may be of the form $\boldsymbol{\alpha} = \{\alpha_{t,t'}, t, t' = 1, ..., 12\}$, such that $\alpha_{t,t'} = \alpha^{|t-t'|}$, $0 < \alpha < 1$ and $t = 1, ..., 12$ be indicator for months of the year, and $|t - t'|$ is the time lag. But this seems too simple, for example, supposed January 2003 and May 2003 are correlated because they are only 4 months apart, whereas January 2003 and February 2004 are correlated because they are from the same season (Winter in Afghanistan) but different years, on the other hand, January 2003 and May 2003 probably are not likely to be correlated. In order to account for the presence of seasonality in the data i.e. temporal correlation between the same month in different years (Afghanistan is characterized by four seasons namely; winter, spring, summer and autumn), we used dummy (binary) variable to capture and quantify the seasonal effects.

Similarly, a semivariogram was used to parameterize the correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$, with its (t, t')-th element representing the correlation between times t and t' at location s. We compute the empirical variograms at different spatial locations and then average the variograms over temporal lags.

For two different times, say t, t', that are $t = |t - t'|$ months apart, the correlation between the two times, t, t' could be written as:

$$C(t, t') = C_T^0(t) \tag{16}$$

where $C_T^0(t)$ represent an temporal exponential function with parameters $\boldsymbol{\alpha} = \tau^2, \sigma^2, \phi$.

We defined a 84×84 matrix $\mathbf{R}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ contains the parameters (τ^2, σ^2, ϕ) that can be estimated by using a graphical display of $\hat{\gamma}(h)$ at $h = h_1, ..., h_K$. Another method is to use weighted least squares, i.e., minimise

$$\sum_{k=1}^{K} w_k \{\hat{\gamma}(h_k) - \gamma(h_k, \alpha)\}^2 \tag{17}$$

with respect to α for some weights w_k's. Following Cressie (1985), we use $w_k = |N(h_k)|$.

4 Results

We applied the spatio-temporal zero truncated model to study the effects of environmental variables on Leishmaniasis, while allowing for different dependencies in the data. Population size was added as offset, the environmental variables used as covariates were average monthly temperature (Celsius), average monthly accumulated rainfall (Inches), average monthly wind speed (Knots) and altitude (Metres). In an attempt to investigate the effects of spatial and temporal interaction, the model incorporated both the temporal variance covariance matrix $v_{s,tt'}$ and spatial weight $\tilde{w}_{ss'}$.

The results from the spatio-temporal truncated model is given in Table 1. The environmental parameters are significant risk factors for leishmaniasis in Afghanistan. There appears to be a negative effect of altitude, temperature, wind and accumulated rainfall as predictors for leishmaniasis incidence.

5 Conclusion

The technique used is this article allow for correct specification of correlation structures to improve the efficiency of the GEE method. The leishmaniasis data presented several problems with modelling issues, ranging from correlation/covaraince specification to issues with "imputed" or "non true" zeros. The high percentage of zero disease counts may be the

Table 1: Parameter estimates (and standard errors) of the zero-truncated model for Leishmaniasis data

Risk factors	Estimate	Standard error
Intercept	-8.847	0.331
Altitude	-0.011	0.001
Temperature	-0.004	0.001
Rainfall	-0.032	0.013
Wind	-0.027	0.005
Season: Winter	0.256	0.062
Season: Spring	0.128	0.096
Season: Autumn	-0.135	0.206

result of no disease incidence or lapse of data collection. Moreover, the dependency in the data may be a result of spatial variation, temporal or both. To resolve this issue, a renowned method was used to address the many issues that the data presented in a very novel way. A model truncated at zero was fitted while allowing for dependency in the data via a working correlation matrix using the technique of GEE.

The results from this study are similar to that of (Adegboye and Kotze, 2012; Rajesh and Sanjay, 2013; Thompson et al., 2002; Valderrama-Ardila et al., 2010; Karagiannis-Voules et al., 2013). The model confirms the significant influence of environmental factors on the incidence of Leishmaniasis. The model indicates that high temperatures are associated with a lower incidence of Leishmaniasis; this is similar to the findings of (Rajesh and Sanjay, 2013). The survivability of the sand fly (Leishmaniasis vector) has been reported to reduce during high temperatures (Rajesh and Sanjay, 2013). A negative association between accumulated rainfall and incidence of Leishmaniasis has been found; this is not surprising as extreme rainfall may have a negative effect on the vector such as flooding (Thompson et al., 2002). The negative effect of temperature and rainfall is also in line with what was observed in the exploratory analysis. Two peaks were observed in the disease occurrence between 2003 and 2009 – January to March and September to December – which coincide with the cold period, while July is the hottest month and March is the wettest month. The results also indicate that low altitudes are associated with an increase in the cases of Leishmaniasis, whereas an increase in the wind speed has a negative effect on the disease.

Acknowledgements

Denis Leung's research is funded by the Research Center at Singapore Management University. You-Gan Wang's research is funded by ARC Discovery Grant DP130100766. A short version of this paper was presented at the 2015 conference of the Italian Research Group for

doi:10.20944/preprints201703.0065.v1

Environmental Statistics (GRASPA-SIS).

References

Adegboye, O., Al-Saghir, M., LEUNG, D., 2016. Joint spatial time-series epidemiological analysis of malaria and cutaneous leishmaniasis infection. Epidemiology & Infection , 1–16.

Adegboye, O.A., Adegboye, M., 2016. Spatially correlated time series and ecological niche analysis of cutaneous leishmaniasis in afghanistan .

Adegboye, O.A., Kotze, D., 2012. Disease mapping of Leishmaniasis outbreak in Afghanistan: Spatial hierarchical Bayesian analysis. Asian Pacific Journal of Tropical Disease 2, 253–259.

Cressie, N., 1985. Fitting variogram models by weighted least squares. Journal of the International Association for Mathematical Geology 17, 563–586.

Cressie, N., 1993. Statistics for Spatial Data. New York: Wiley.

Cressie, N., Huang, H.C., 1999. Classes of nonseparable, spatio-temporal stationary covariance functions. Journal of American Statistical Association 94, 1330–1340.

Faraway, J., 2004. Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Gneiting, T., 2002. Nonseparable, stationary covariance functions for space-time data. Journal of American Statistical Association 97, 590–600.

Greene, W.H., 2005. The Handbook of Econometrics: Vol. 1 Theoretical Econometrics. Palgrave, London. chapter Censored Data and Truncated Distributions. 20, pp. 695–726.

Karagiannis-Voules, D., Scholte, R., Guimara, L., Utzinger, J., P, V., 2013. Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. PLOS Neglected Tropical Diseases 7.

Karlsson, M., Lindmark, A., 2014. truncSP: An R package for estimation of semi-parametric truncated linear regression models. Journal of Statistical Software 57, 1–19. URL: `http://www.jstatsoft.org/v57/i14/`.

Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. Annals of Statistics 17, 1217–1241.

Lee, M., Kim, H., 1998. Semiparametric Econometric Estimators for a Truncated Regression Model: A Review with an Extension. Statistica Neerlandica 52, 200–225.

Leung, D., Wang, Y.G., Zhu, M., 2009. Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. Biostatistics 10, 436–445.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Paul, S., Zhang, X., Xu, J., 2013. Estimation of regression parameters for binary longitudinal data using GEE: Review, extension and an application to environmental data. Journal of Environmental Statistics 4, 1–12.

Porcu, E., Mateu, J., Bevilacqua, M., 2007. Covariance functions that are stationary or nonstationary in space and stationary in time. Statistica Neerlandica 61, 358382.

Prentice, R.L., 1988. Correlated binary regression with covariates specific to each binary observation. Biometrics 44, 1033–1048.

Rajesh, K., Sanjay, K., 2013. Change in global climate and prevalence of visceral leishmaniasis. International Journal of Scientific and Research Publications 3.

Saffari, S.E., Adnan, R., Greene, W., 2011. Handling of over - dispersion of count data via truncation using poisson regression model. Journal of Computer Science and Computational Mathematics 1.

Sherman, M., 2011. Spatial Statistics and Spatio-Temporal Data. Wiley, Chichester.

Stein, M.L., 2005. Space-time covariance functions. Journal of American Statistical Association 100, 310–321.

Thompson, R., De Oliveira Lima, J., Maguire, J., Braud, D., Scholl, D., 2002. Climatic and demographic determinants of american visceral leishmaniasis in northeastern brazil using remote sensing technology for environmental categorization of rain and region influences on leishmaniasis. American Journal of Tropical Medicine and Hygiene 67, 648–655.

Valderrama-Ardila, C., Alexander, N., Ferro, C., Cadena, H., Marn, D., Holford, T., Munstermann, L., Ocampo, C., 2010. Environmental risk factors for the incidence of american cutaneous leishmaniasis in a sub-andean zone of colombia (chaparral, tolima). American Journal of Tropical Medicine and Hygiene 82, 243–250.

Yasui, Y., Lele, S., 1997. A regression method for spatial disease rates: An estimating function approach. Journal of the American Statistical Association 92, 21–32.

Zeger, S.L., 1988. A regression model for time series of counts. Biometrika 75, 621–629.