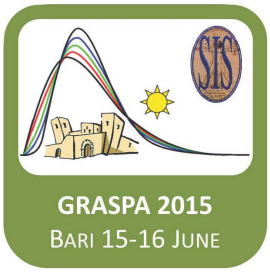

Spatio-temporal modelling of zero-truncated disease patterns

O. Adegboye[1,*], D. Leung[2] and Y-G. Wang[3]

[1]*Department of Mathematics, Statistics & Physics, College of Arts & Sciences, Qatar University; o.adegboye@qu.edu.qa,*
[2]*School of Economics, Singapore Management University, Singapore; denisleung@smu.edu.sg*
[3]*School of Mathematics and Physics, University of Queensland, Queensland, Australia; yougan.wang@uq.edu.au*
[*]*Corresponding author*

Abstract. *This paper focuses on the spatio-temporal pattern of Leishmaniasis incidence in Afghanistan. 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 (e.g., Yasui and Lele, 1997) [3]. 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. The model was built on a foundation of the generalized estimating equations (Liang and Zeger, 1986) [1]. We illustrate the method using data from Afghanistan between 2003-2009. Since 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. To resolve this issue, we use a model truncated at zero.*

Keywords. *Generalized estimating equations; Overdispersion; Poisson; Spatio-temporal*

1 Introduction

Leishmaniasis is the third most common vector-borne disease and a very important protozoan infection. The disease is contracted through bites from sand flies, which are themselves not poisonous, but the parasitic Leishmania in its saliva can result in chronic and non-healing sores. Some of the risk factors identified include household construction materials, design, density and presence of the disease in the neighborhoods and high rodent infestations. 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 present study was motivated by Leishmaniasis cases in the provinces of Afghanistan between 2003 and 2009. One of the most challenging issues in modelling spatio-temporal data is the choice of a valid and yet flexible correlation (covariance) structure. 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. Mostly, space-time correlations are considered jointly, a

step that we believe is unnecessary or unrealistic in our data.

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. 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 is based on the framework of generalized estimating equations (GEE) where the spatial dependency is accounted for by re-weighting the standard GEE so that locations that are highly correlated with each other would receive less weight. Apart from the spatial dependency in our data, the data is also characterized by a high percentage of zero disease counts which introduced over-dispersion. Since 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 often practiced to truncate the values that are bigger than a constant to overcome over-dispersion [2]. The analysis of truncated often arises from a subsidiary set of results that treat a practical problem of how data are gathered and analyzed and incompleteness of this data requires special estimators of the regression coefficients. To resolve this issue, we use a model truncated at zero.

The rest of the paper is structured as follows. Section 2, describes the materials and methods that will be used in the study. In Section 3 we shall give the results of the data analysis and conclude the paper

2 Data Sources and Methods

2.1 Data Sources

The data used in this study 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). 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). We used satellite-derived environmental data- Normalized difference vegetation index (NDVI), land surface temperature (LST) and rainfall as explanatory variables.

2.2 Model Formulation and Parameter Estimation

We begin by considering the disease counts $\mathbf{y} = (y'_{1.}, ..., y'_{S.})^T$ and observed covariates at different locations $\mathbf{X} = (x'_{1.}, ..., x'_{S.})^T$ as a set of longitudinal data over S spatial locations. Let $\mathbf{y}$ be independent and assumed to follow a Poisson model and stacked as a $S \times T$ vector. The covariance matrix of y is $\tilde{V}$ and $\tilde{v}^{-1}_{st,s't'}$ is the (st, st')-th element of $\tilde{V}^{-1}$, the dimension of $\tilde{V}$ is $ST \times ST$.

For the dataset we are working with, $S = 20$ represents the number of provinces and $T = 7$ represents the number of years with recorded data. Using the monthly data, then $T = 84$ and so $S \times T = 20 \times 84 = 1680$ and therefore $\tilde{V}$ would be a matrix that cannot feasibly be handled. Moreover, the correlation between $y(s,t)$ and $y(s',t')$ often does not have any practical meaning. For a fixed s, $v_{s,tt'}$, $t, t' = t_1, ..., t_T$ are the elements of the variance covariance matrix of disease counts between times.

The modelling is a 2-step process, we first needed to find the variance covariance matrix, $v_{s,tt'}$ and spatial weight, $\tilde{w}_{ss'}$. We compute empirical temporal variograms at different spatial locations and then average all temporal variograms with the same temporal lag. We applied the empirical semivariogram based on the Pearson residuals and fitted a parametric semivariogram models. 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:

$$C(t,t') = C^0_T(t) \tag{1}$$

where $C_T^0(t) = e^{(-\theta t)}$ is the temporal covariance function with months apart. The parameters $\theta = \tau^2, \sigma^2, \phi$ represents the nugget, sill, and range, respectively.

In order to model spatial correlation and overdispersion, we assume there is a nonnegative weakly stationary latent process e and conditioned on this process, the y's are independent and follow a log-linear model given below. Consider the following; suppose we remove all $y_{st} = 0$, then conditioned on $y_{st} > 0$, we have $E(y_{st}|e_{st}) = c\mu_{st}(\beta)e_{st}$, and $var(y_{st}|e_{st}) = [c\mu_{st}(\beta) + c(1-c)\mu_{st}(\beta)^2]e_{st}$ where $c = 1/[1-\exp(-\mu_{st}(\beta))]$, leading to

$$E(y_{st}) = c\mu_{st}(\beta) \equiv \phi_{st}(\beta), \tag{2}$$

$$var(y_{st}) = c\mu_{st}(\beta) + c(1-c)\mu_{st}(\beta)^2 + c^2\mu_{st}(\beta)^2\sigma^2. \tag{3}$$

where β are unknown parameters. We assume $\mathrm{E}(e_{st})$ to be 1 so that $\mu_{st}(\beta)$ represents the marginal mean of y_{st}.

Let $\doteq\{d(s,t) = d_{st}\}_{S\times T}$ 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. For a particular set of spatial weight $\tilde{w}_{ss'}$, the spatial GEE conditioned only on those observations with $y_{st} > 0$ can be written as

$$\tilde{U}(\beta,\alpha) \equiv \sum_{s=s_1}^{s_S}\sum_{s'=s_1}^{s_S}\sum_{t=t_1}^{t_T}\sum_{t'=t_1}^{t_T}\frac{\partial\phi_{st}}{\partial\beta^T}d_{st'}\tilde{w}_{ss'}v_{s,tt'}^{-1}\{y_{st'} - \phi_{st'}\} = 0, \tag{4}$$

where $v_{s,tt'}$ is the t,t-th element of V_s, the covariance matrix of $y_{s\cdot}$. The matrix V_s can be expressed as $A_s^{1/2}R_s(\alpha)A_s^{1/2}$, where $A_s = \mathtt{diag}[c\mu_{s1}(\beta) + c(1-c)\mu_{s1}(\beta)^2 + c^2\mu_{s1}(\beta)^2\sigma^2, ..., c\mu_{sT}(\beta) + c(1-c)\mu_{sT}(\beta)^2 + c^2\mu_{sT}(\beta)^2\sigma^2]$ and $R_s(\alpha)$ 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 β but we also must deal with the nuisance parameters α. Let $R(\alpha)$ be a 84×84 matrix where α contains the parameters (θ) estimated via weighted least square method. The parameters are estimated via a Newton-Raphson iteration method. To solve for (α,β) jointly. Let $\hat{\beta}_k$ and $\hat{\alpha}_k$ be the estimates of β and α at the k-th iteration. We first fitted a GEE with an independence working correlation structure, we then solve the estimating equation for α, 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 $\hat{\beta}_0$ using (4) by assuming an identity matrix for $R_s(\alpha)$, equivariance, *i.e.*, $v_{s,tt'}^{-1} = 1$ and, spatial weight.

Then at iteration k,

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \left[\sum_{s,s',t,t'}\frac{\partial\phi_{st}(\hat{\beta}_k)}{\partial\beta^\tau}d_{st'}\tilde{w}_{ss'}v_{s,tt'}^{-1}(\hat{\beta}_k)\frac{\partial\phi_{st}(\hat{\beta}_k)}{\partial\beta^\tau}\right]^{-1}\left[\sum_{s,s',t,t'}\frac{\partial\phi_{st}(\hat{\beta}_k)}{\partial\beta^\tau}d_{st'}\tilde{w}_{ss'}v_{s,tt'}^{-1}\{y_{st'} - \phi_{st'}(\hat{\beta}_k)\}\right]. \tag{5}$$

Since we are using an AR(1) correlation structure, we take the slope of the linear regression of $\log(\hat{r}_{st}^k\hat{r}_{st'}^k)$ on $\log(|t-t'|)$ as $\hat{\alpha}_k$. We then iterate between (4) and (5) until convergence.

The standard errors for the $\beta's$ were obtained using large-sample properties.

3 Illustration

We shall illustrate our method using the Leishmaniasis cases data reported to the Afghanistan Health Management Information System (HMIS) of the Ministry of Public Health (MoPH) between 2003 and 2009. We observe higher disease incidence around the Kabul area (North Eastern). Similar patterns were observed in 2003-2008 (maps not shown here but are available on request). The monthly profile of cases of Leishmaniasis revealed two peaks in the disease occurrence in Afghanistan between 2003 and 2009 – January to March and September to December – which coincide with the cold period while July is the

Table 1: Parameter estimates together with the standard errors from GEE with different correlation structures of Leishmaniasis incidence in Afghanistan

Risk factors	$GEE_{Spatial}$	$GEE_{Temporal}$	$GEE_{Spatio-temporal}$
Intercept	-0.52289 (0.07342)	-9.09746 (0.08526)	-9.09818 (0.02206)
Altitude (*m*)	-0.00012 (0.00023)	0.00026 (0.00001)	0.00026 (0.00001)
Temperature (°C)	-0.42460 (0.00022)	-0.00118 (0.00035)	-0.00113 (0.00017)
Precipitation (*Inches*)	1.58830 (0.00785)	-0.03920 (0.00066)	-0.03895 (0.00210)
Wind (*Knot*)	0.53639 (0.00528)	0.02089 (0.01566)	0.02078 (0.00112)
2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$	69.33	87.054	19.38
AIC	103.112	179.39	46.511

hottest month and March is the wettest month. The time series plot for the number of Leishmaniasis cases reveal upward trend and regularly repeating patterns of highs and lows related to the months of the year which suggests seasonality in the data. The variogram of space-time autocorrelation is obtained by considering time as discrete. This method models the cross-variograms between data with time replication (months/years) and captures the variability in space and time. We hold the view that correlations arising from spatial and temporal sources are inherently distinct. Our method makes it possible to combine the specific provincial rate with the influence of the spatial neighborhood. Three different models were fitted namely; spatial only, temporal only and spatio-temporal model. In Table 1, perhaps the most distinctive results are from the model with spatial correlation; the model parameter estimates are remarkably different from others. The result may not be surprising as it has been assumed that the correlation remains the same across time. This also suggests that spatial correlation only may not be sufficient for the data, because it involves the specification of spatial correlation across time. The results have shown that the specified spatio-temporal function is more suitable and appropriate for this data (smaller 2 trace$(\hat{\Sigma}_I^{-1}\hat{\Sigma}_R)$). Moreover, the model with the spatio-temporal correlation function significantly improves the model fit when compared to other specifications, as judged by the smaller AIC. Although the parameter estimates from both temporal and spatio-temporal models are similar, significant differences can be observed in their precision estimation. The technique used is this study 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 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.

References

[1] Leung, D., Wang, Y.-G., and Zhu, M. (2009). Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics* **10**, 436–445.

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

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