Geospatial Health 4(2), 2010, pp. 201-217

Accounting for autocorrelation in multi-drug resistant tuberculosis predictors using a set of parsimonious orthogonal eigenvectors aggregated in geographic space

Benjamin G. Jacob[1], Fiorella Krapp[2], Mario Ponce[2], Eduardo Gotuzzo[2], Daniel A. Griffith[3], Robert J. Novak[1]

[1]School of Medicine, Division of Infectious Disease, University of Alabama at Birmingham, Birmingham, AL 35294-2170, USA; [2]Instituto de Medicina Tropical Alexander Von Humboldt-Universidad Peruana Cayetano Heredia, Lima, Peru; [3]School of Social Sciences, The University of Texas at Dallas, P.O. Box 830688 Richardson, TX 75083-0688, USA

Abstract. Spatial autocorrelation is problematic for classical hierarchical cluster detection tests commonly used in multi-drug resistant tuberculosis (MDR-TB) analyses as considerable random error can occur. Therefore, when MDR-TB clusters are spatially autocorrelated the assumption that the clusters are independently random is invalid. In this research, a product moment correlation coefficient (i.e. the Moran's coefficient) was used to quantify local spatial variation in multiple clinical and environmental predictor variables sampled in San Juan de Lurigancho, Lima, Peru. Initially, QuickBird (spatial resolution = 0.61 m) data, encompassing visible bands and the near infra-red bands, were selected to synthesize images of land cover attributes of the study site. Data of residential addresses of individual patients with smear-positive MDR-TB were geocoded, prevalence rates calculated and then digitally overlaid onto the satellite data within a 2 km buffer of 31 georeferenced health centres, using a 10 m^2 grid-based algorithm. Geographical information system (GIS)-gridded measurements of each health centre were generated based on preliminary base maps of the georeferenced data aggregated to block groups and census tracts within each buffered area. A three-dimensional model of the study site was constructed based on a digital elevation model (DEM) to determine terrain covariates associated with the sampled MDR-TB covariates. Pearson's correlation was used to evaluate the linear relationship between the DEM and the sampled MDR-TB data. A SAS/GIS® module was then used to calculate univariate statistics and to perform linear and non-linear regression analyses using the sampled predictor variables. The estimates generated from a global autocorrelation analyses were then spatially decomposed into empirical orthogonal bases, using a negative binomial regression with a non-homogeneous mean. Results of the DEM analyses indicated a statistically non-significant, linear relationship between georeferenced health centres and the sampled covariate elevation. The data exhibited positive spatial autocorrelation and the decomposition of Moran's coefficient into uncorrelated, orthogonal map pattern components which revealed global spatial heterogeneities necessary to capture latent autocorrelation in the MDR-TB model. It was thus shown that Poisson regression analyses and spatial eigenvector mapping can elucidate the mechanics of MDR-TB transmission by prioritizing clinical and environmental-sampled predictor variables for identifying high risk populations.

Keywords: multi-drug resistant tuberculosis, geographical information system, digital elevation model, Poisson regression analyses, spatial eigenvector mapping, Peru.

Corresponding author:
Benjamin G. Jacob
School of Medicine, Division of Infectious Disease
University of Alabama at Birmingham
Birmingham, AL 35294-2170, USA
Tel. +1 205 996 7894 Fax +1 205 934 5600
E-mail: bjacob@uab.edu

Introduction

Multiple linear regression analysis techniques, coupled with normal probability models, have become standard epidemiological tools to examine predictor variables associated with multi-drug

resistant tuberculosis (MDR-TB) for identifying covariates associated with high-risk populations (Smith, 1994; Clarke et al., 2002; Johnston, 2003). MDR-TB is defined as TB resistant at least to isoniazid (INH) and rifampicin, which commonly develops in the course of TB treatment (Iseman, 1993). Multiple regression model outputs are often used to establish the significant level and, thus, the relative predictive importance of a set of sampled MDR-TB predictors. The regression estimates can then be used to construct maps using geographical information systems (GIS) for determining predictor variables that are associated with MDR-TB for further statistical analysis. The assumptions underpinning multiple regression, however, necessarily impose several important constraints that may not always be satisfied, or that might require careful consideration when mapping MDR-TB explanatory parameters. For example, the relationships between the outcome and the sampled predictor variables in a MDR-TB model, generated from multiple regression analyses, are assumed to be linear, and the variance of the residual errors are assumed to be the same, regardless of the value of the covariate measurements. If there is non-linearity, serial correlation, heteroscedasticity and/or non-normality in a model, the forecasts, confidence intervals, and insights yielded by a regression model may be seriously biased or misleading (Hastie and Tibshirani, 1990). Another problem in the use of regression coefficients is the occurrence of predictor variables that are not independent, i.e. non-zero correlations amongst covariates (Miles and Shevlin, 2001), giving rise to collinearity (Gantz, 1997). When more than two covariates in a model are highly correlated, multicollinearity can occur (Slinker and Glantz, 1985; Pedhazur, 1997), which can seriously distort the interpretation of a MDR-TB model. Multicollinear explanatory variables are difficult to analyse, as their effects on a response variable can be due to either true synergistic relationships among the variables, or confounding effects creating spurious correlations (Glantz and Slinker, 2001; Maddala, 2001; Fotheringham et al., 2002). Consequently, linear coefficients, based on collinear and multicollinear variables, can bias sampled covariate measurements yielding unstable parameter estimates and unreliable significance tests in a MDR-TB model.

Since the role of each predictor variable in a MDR-TB model is to increase precision, the effect of covariate measurement error on maximum likelihood (ML) and quasi-maximum likelihood (MQL) estimates of regression parameters must be considered. Obtaining estimates that are unbiased; however, has proven to be difficult when random effects are incorporated into a generalized linear model (GLM), which normally uses a common algorithm for the estimation of the ML and MQL (Chatterjee, 1988). GLMs can be used to model spatial distribution by relating the response variable (abundance, or presence/absence) and spatially referenced covariates but such models ignore unmeasured covariates. Generalized linear mixed models (GLMMs), a natural outgrowth of both linear mixed models and GLMs, enable the accommodation of non-normally distributed responses and specification of a non-linear link between the mean of the response and the predictors, and they can model random effects which also can account for unmeasured covariates and overdispersion. For example, a penalized quasi-likelihood method can be used for fitting GLMM (Fotheringham et al., 2002), which can account for overdispersion in sampled data which arises through the omission from the regression models of important variables, existence of outliers, and the use of inappropriate link functions (Hastie and Tibshirani, 1990). Although ML and variants are standard for both linear mixed models, e.g. restricted ML (REML) and GLMs, its use in infectious disease modeling has been limited to simple models due to the need to numerically evaluate high-dimensional integrals. High-dimensional integrals are usually solved with Monte Carlo algorithms and quasi-Monte Carlo algorithms. However, residual-based diagnostics for multivariate heteroscedasticity from previously constructed models using Bayesian statistics has

revealed that errors in variance uncertainty estimation are common and can substantially alter numerical predictions of a model by inflating the value of test statistic thereby, increasing the chance of a type I error, i.e. incorrect rejection of the null hypothesis (Jacob et al., 2009d).

To solve these problems, statistical methods can be applied to control for interaction, among sampled predictor variables when mapping MDR-TB-related data. One of these alternatives is to incorporate localized interaction terms into a spatial statistical algorithm using the sampled predictor variables together with their interactions and an autocovariate term, i.e. Moran's index (Moran's *I*). The autocovariate term enables different measures and their approximations in autologistic models to be compared with respect to aggregated patterns caused by different processes (Anselin, 1995). The performance of approximations for the autocovariate strongly depends on the cause of spatial aggregation in the data (Griffith, 2003). Inclusion of a spatial autocovariate term has important effects on model selection. For example, autocovariate terms can be used to calculate autonormal, autopoisson or autologistic regression, to capture spatial autocorrelation in a model originating from endogenous processes, such as contagious population growth and movement of censused individuals between sampling sites (Griffith, 2003). Autocorrelation is the degree to which a set of features tend to be clustered together, i.e., positive spatial autocorrelation (PSA), i.e. when similar attribute values aggregate or when the data are unevenly dispersed (negative spatial autocorrelation (NSA)) over the earth's surface (Cliff and Ord, 1973).

Autocorrelation is a very general statistical property of explanatory variables observed across geographic space. Its most common forms are patches and gradients. Spatial autocorrelation presents a problems for standard testing as autocorrelated data violates the assumption of independence of most standard statistical procedures (Griffith, 2003). Since the presence of spatial autocorrelation can violate the ordinarily stated assumption of stochastic independence among observations, on which statistical inference from most classical statistical models is based, it is important to identify whether different measures can account for different spatial autocorrelation patterns in MDR-TB parameters. Ignoring spatial autocorrelation in MDR-TB modeling distributions can introduce uncertainty in model fit.

Autocovariate models may address residual spatial autocorrelation components among sampled explanatory variables by estimating the co-variation of a response variable at any one sampled site based on the response values at surrounding sites (Chatterjee, 1988). In autologistic regression models employed in the analysis of spatial distributions, an additional explanatory variable, the autocovariate, is used to correct the effect of spatial autocorrelation (Griffith, 2003). While this approach has been widely used over the last 10 years in biogeographical analyses, it has not been assessed for its validity and performance against simulation data generated from infectious disease processes. Furthermore, since autologistic regression models consistently underestimate the effect of the environmental variable in a disease model and give biased estimates compared to a non-spatial logistic regression (Jacob et al., 2008a), a model generated with alternative methods available may reveal that autologistic regression is more biased and less reliable and should be used only in concert with other reference methods.

Recent quantitative geographical analysis methods have supplemented mapping georeferenced explanatory data by decomposing the Moran's *I* into synthetic variates, whose linear combinations constitute a spatial filter logistic model, with a GLM specification to determine residual autocorrelation (Griffith, 2003). The eigenvector filtering approach is a non-parametric technique that removes the inherent autocorrelation from generalized linear regression models by treating it as a missing variable (i.e. first order) effect. The spatial filtering then converts the variables that are spatially autocorrelated into spatially independent variables in an ordinary

least squares (OLS) regression framework. The aim of a non-parametric spatial filtering is to control for autocorrelation with a set of proxy variables rather than to identify a global autocorrelation parameter for a spatial process (Griffith, 2003). The basis for this procedure is the decomposition of the Moran's *I* into orthogonal and uncorrelated map pattern components. This decomposition makes orthogonal the latent spatial correlation represented by the geographic configuration of locations described by a given spatial weights matrix. "The procedure can be used to account for redundant locational information by generating the eigenvectors a special set of vectors associated with a linear system of equations (i.e. a matrix equation)". Unexplained MDR-TB data clustering may be only artefactual as a result of differential case reporting, unknown demographic changes or duplication of case data (Godoy, 2004). These corresponding eigenvectors can then be used as predictor variables in a regression equation for determining covariates associated with specific MDR-TB parameters.

This study was carried out to identify geographic areas with on-going MDR-TB transmission by generating spatial eigendecomposition models within a SAS/GIS® (SAS Institute Inc.; Cary, NC, USA) database. GIS combined with robust statistical methods and software may assist clinicians, epidemiologists and programme managers by adding descriptive images that are systematically created according to proper scientific protocol for evaluations of potential MDR-TB covariates. For example, high-resolution terrain data generated from digital elevation models (DEMs) orthophoto mosaics, or multispectral satellite imagery combined with linear and/or non-linear correlation statistics in GIS may facilitate regional analysis of MDR-TB data by identifying landscape characteristics associated with georeferenced environmental-sampled variables. In this research, spatial indices were generated based on DEM statistics in ArcGIS® (Redlands, CA, USA) and Poisson probability models generated in SAS/GIS®, using multiple clinical and environmental MDR-TB predictor variables sampled in San Juan de Lurigancho (SJL), Lima, Peru. Our research objectives were to:

(i) perform Poisson regression analyses to determine covariates affecting MDR-TB prevalence;
(ii) generate global autocorrelation statistics for evaluating spatial dependence among the sampled data; and
(iii) determine latent autocorrelation components in model output, using a stepwise negative binomial regression analysis with a gamma distributed mean for identifying MDR-TB epicenters in SJL.

Generating GIS cluster models based on MDR-TB parameters, using GLMMs, autocovariate regression and eigenvector mapping, may elucidate the mechanics of MDR-TB transmission for optimising existing management programs by spatially targeting high-risk populations.

Materials and methods

Study site

The work was carried out in SJL, the largest district in the northeast of Lima, Peru (Fig. 1). With a current population exceeding one million people and a total surface area of 131.3 km^2, constituting 4.91% of the total area of the province of Lima, it is the country's most populous district. SJL is bordered by the districts of Carabayllo and San Antonio in the Huarochirí province to the north, by the Comas, Independencia and Rímac districts to the west and by Lurigancho to the east. The Rímac River marks the district's border with downtown Lima and El Augustino to the south. The most important urban areas in the district are Mangomarca, Zárate, Las Flores, Canto Grande and Bayovar. One of the first urban areas in SJL is Caja de Agua, located at the entrance of the district surrounded by the San Cristobal and Santa Rosa hills from south to west (Fig. 2). The altitude of SJL ranges from 2,240 m above mean sea level (AMSL) at the peaks of Cerro Colorado Norte to 200 m AMSL at the level of the Rimac River. The

Fig. 1. General base map of Peru.

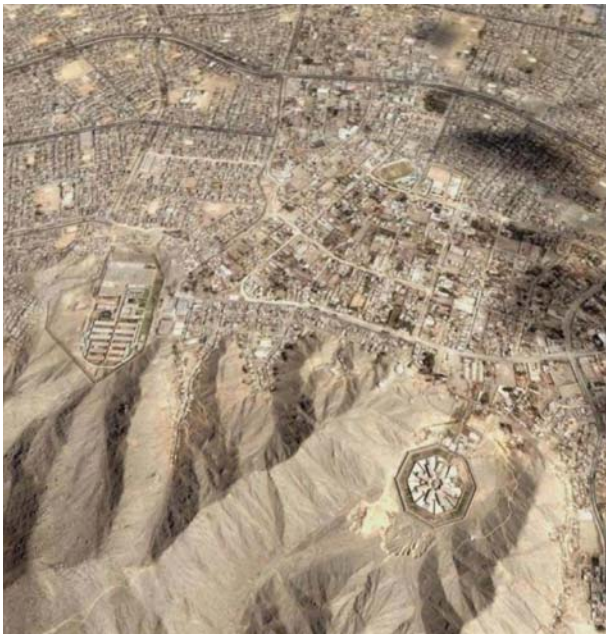
Fig. 2. Caja de Agua and other urban areas in the San Juan de Lurigancho, study site.

urban areas have been developed in a longitudinal direction from the river border up to 350 m AMSL. Lima has a mild climate, although it is situated in the tropics. The weather in the SJL study site is influenced by the cold offshore Humboldt Current, which ensures that summer temperatures hover around 16-20°C; only a few degrees lower in June and July. Humidity in the city is very high and fog is often present, especially between May and November much like in many parts of the country.

Subjects and setting

This was a prospective, multi-centre, observational study comparing the use of several investigational techniques with standard methods to assess the *in vitro* antimicrobial susceptibility of *Mycobacterium tuberculosis,* either directly from patient specimens or from culture isolates. Data acquired from a retrospective study of a cohort of patients diagnosed with pulmonary TB and MDR-TB over an 18-month period in the SJL study site was used. All patients underwent drug susceptibility testing for first-line drugs for TB treatment. Overall, 1250 adults with pulmonary tuberculosis cultures were confirmed. After collection of baseline samples and completion of initial measurements, including susceptibility testing by conventional and research methods, all subjects started anti-TB chemotherapy as dictated by the standard of care at the site of enrollment. Subjects were recruited, among patients presenting with smear-positive pulmonary TB at diagnostic and treatment sites in the following health centres: San Fernando, La Huayrona, Canto Grande, Jose Carlos Mariátegui, Huáscar XV, Huáscar II, Ganímedes, Cruz de Motupe, Piedra Liza, Bayóvar, Jaime Zubieta, San Juan, San Benito, Mangomarca, San Hilarion, Campoy, 15 de Enero, La Libertad, Juan Pablo II, Ascarruz Alto, 10 de Octubre, Sta Fe de Totoritas, Proyectos Especiales, Santa Rosa, Ayacucho, Zarate, Medalla Milagrosa, Campoy Alto, Montenegro, Santa Maria, Tupac Amaru II and Caja de Agua.

After confirmation of the sputum-smear microscopy results, the subjects were screened for the presence of productive cough. Patients with positive sputum-smears are the ones with the capacity to spread the infection (Zignol et al., 2006). Eligible subjects received an explanation of the study and provided written consents to participate. Initial data from the screening included past medical history, basic socio-demographic descriptors (age, sex, occupation, address, etc.) and detailed symptom-oriented history with physical examination. Drug susceptibility testing for INH, rifampin, ethambutol and streptomycin was performed on the initial sputum culture isolates of all enrolled subjects. Those subjects with initial drug-resistant *M. tuberculosis* isolates were confirmed using a treatment regimen with a duration deemed appropriate by the Committee of the National Tuberculosis Control Programme (NTCP) and the Committee for Evaluation of Retreatment (CER). All information collected was recorded on standardised data collection forms, labeled with the date and the subject's name and study number, edited as needed, and entered into data files for further analysis. Case report forms were developed to record baseline, clinical and socio-demographic information as well as laboratory data such as human immunodeficiency virus (HIV) testing results, mycobacterial smear and culture results.

Geographic mapping

Field sampling was conducted from July 2005 to July 2007. Thirty-one health centres in the study site were mapped and classified using a differentially corrected global positioning systems (DGPS) Max receiver from CSI-Wireless (Calgary, Alberta, Canada). This remote technology relies on the OmniStar L-Band satellite signal yielding a positional error of X.179m (+/- 0.392 m) (Jacob et al., 2009a).

Remote sensing data

QuickBird (www.digitalglobe.com) images were acquired on March 11, 2008 for the SJL study site. QuickBird multispectral products provided four discrete non-overlapping spectral bands covering a range from 0.45 µm to 0.72 µm, with a spatial resolution of 0.61 m. QuickBird imagery was classified using the iterative self-organizing data analysis technique (ISODATA), unsupervised routine in ERDAS *Imagine* v.8.7™ (Earth Resource Data Analysis System; Atlanta, GA, USA). The images were co-registered manually, using gathered ground control point and georectified images from the QuickBird data. The satellite images were co-registered by applying a first order polynomial algorithm with a nearest neighbour resampling method and the universal transverse mercator (UTM) zone 37S datum WGS-84 was used for the projection of the spatial sampled datasets.

Environmental parameters

Variables recorded included MDR-TB prevalence rates, distance between individual health centres, population data and land surface elevation and slope per sampled site in the SJL study site. Distance measures were recorded in ArcInfo 9.2® (ESRI; Redlands, CA, USA) with QuickBird data and field sampling. The distances between health centers were categorized into numerous classes (e.g. 1 = 0-5 km, 2 = 5-10 km, and so on) and the number of MDR-TB cases at each individual health centre was recorded.

Grid-based algorithm

A 10 x 10 m grid-based algorithm was overlaid on the base maps of the study site in ArcInfo 9.2® (ESRI) to generate spatial sampling units. A 2 km buffer was placed around each health centre and a unique identifier was placed in each gridded buffer (Fig. 3). The level of house spacing, road types and networks, community water sources and access to utilities were also noted within each buffer. Information contained in Census and district Development Reports for the SJL study site, as well as environmental descriptions from previ-

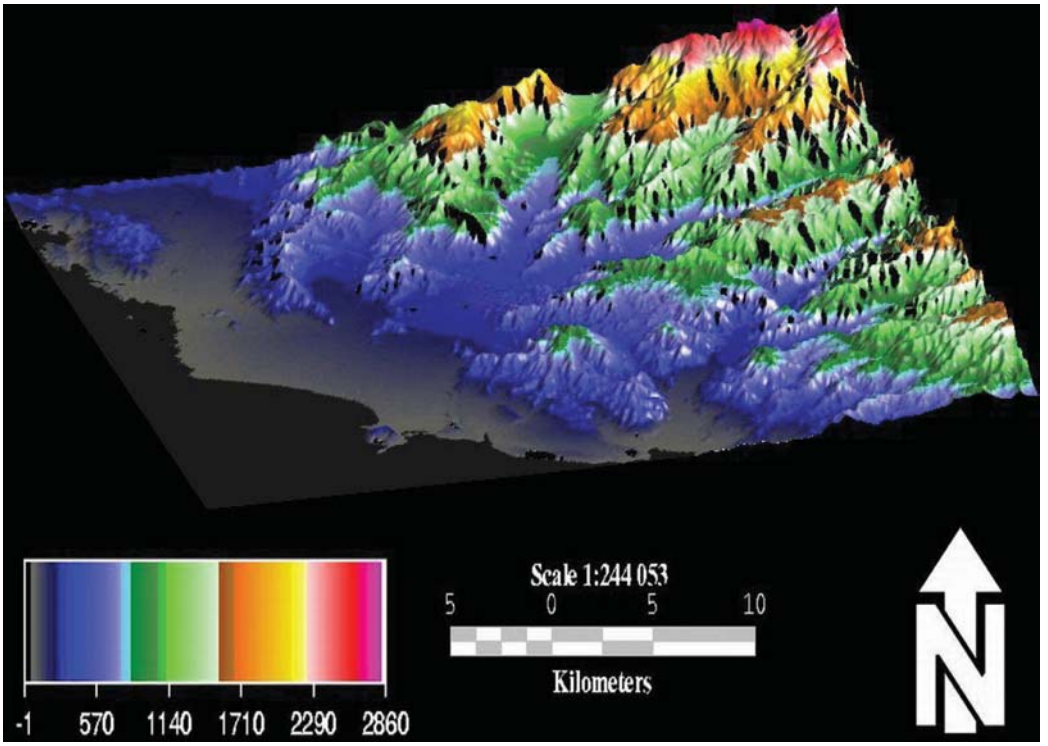

Fig. 3. Digital elevation model (DEM) generated from covariates sampled in the San Juan de Lurigancho, study site.

ous field and topographical maps were used to assist with the stratification process. The boundaries of selected grid cells were located in the field using the hand-held DGPS navigational units and base maps with landmarks/paths/roads. Latitude and longitude readings were taken at the corners and centre of each selected grid cell to confirm the location and extent of the grid cell boundaries.

Digital elevation model (DEM)

A three-dimensional model of the study site was constructed, based on the DEM, using the ArcScene extension of ArcGIS® (ESRI). The DEM used was a raster representation of a continuous surface, originating from the Shuttle Radar Topography Mission (SRTM) which has a spatial resolution of 92 m. The purpose of the DEM construction was to extract topographic parameters that may have been associated with the georeferenced MDR-TB predictor variables sampled in the SJL study site. Data from SRTM version 2 (i.e. the finished version) was downloaded from http://srtm. usgs.gov/. The MDR-TB predictor variables were defined by geocoordinates in a tiled format.

Regression analysis

All sampled parameters were entered into Excel files and analysed using SAS/GIS®. The first stage of these analysis used Poisson regression to determine the relationship between the clinical and environmental-sampled MDR-TB covariate measurements. The Poisson regression assumed that each independent count estimate (i.e. n_i), recorded at a health centre location "i" = 1, 2, … n, was from a Poisson distribution. These data were described by a set of predictor variables denoted by matrix X_i, a $1 \times p$ vector of covariate indicator values for a georeferenced health centre location i in the SJL study site. The expected value of these data was given by

$\mu_i(X_i) = n_i(X_i) \exp(X_i\beta)$, where β was the vector of non-redundant parameters, and the Poisson rates were parameter given by $\lambda_i(X_i) = \mu_i(X_i)/n_i(X_i)$; the rates parameter $\lambda_i(X_i)$ was both the mean and the variance of the Poisson distribution (McCullagh and Nelder, 1989) for a sampled health centre location. The regression analyses were performed in SAS/GIS® using a 95% confidence level. The data was log-transformed before analysis to normalise the distribution and minimize the standard error.

There was considerable overdispersion in the model. Thus, we used a negative binomial model to evaluate the sampled MDR-TB covariates as negative binomial models fitted by the ML method are considered to be convenient and practical for handling overdispersion in remote-sampled covariates. This approach allowed the likelihood ratio and other standard ML tests to be implemented, permitting the fitting procedure to be carried out by using an iterative weighted least squares regression similar to those of the Poisson (Jacob et al., 2005, 2009c). In this research, the fitting of overdispersed Poisson models was performed using:

$$f(k) = \int_0^\infty \text{Poisson}\ (k \mid \lambda) \bullet \text{Gamma}\ \ (\lambda \mid r,(1\text{-}p)/p)\mathrm{d}\lambda$$

$$= \int_0^\infty \frac{\lambda^k}{k!} \exp\ (\text{-}\lambda) \bullet \frac{\lambda^{k\text{-}1} \exp\ (\text{-}\lambda p/(1\text{-}p))}{\Gamma\ (r)\ ((1\text{-}p)/p)^r}\ \mathrm{d}\lambda$$

$$= \frac{1}{k!\ \Gamma\ (r)} p^r \frac{1}{(1\text{-}p)^r} \int_0^\infty \lambda^{(r+k)\text{-}1} \exp\ (\text{-}\lambda p/(1\text{-}p))\mathrm{d}\lambda$$

$$= \frac{1}{k!\ \Gamma\ (r)} p^r \frac{1}{(1\text{-}p)^r} (1\text{-}p)^{r+k}\ \Gamma\ (r+k)$$

$$= \frac{\Gamma\ (r+k)}{k!\ \Gamma\ (r)} p^r\ (1\text{-}p)^r$$

in *Arc*, a computer programme described in Glantz and Slinker (2001) and obtained at the follwing site: http://www.stat.umn.edu/arc.

Autoregressive spatial models

We restricted our attention in this research to the simultaneous autoregressive Gaussian spatial process and the autoregressive Gaussian response model (i.e. spatial lag model). We assumed that a common spatial structure V applied to all terms in the autoregressive spatial models, and that either autocorrelation tests or the spatial eigenvectors were developed with this spatial structure as their underlying basis. An "*n x n*" spatial structure matrix V was used to specify the hypothetical pairwise spatial similarity relationships among the MDR-TB observations. By definition, the matrix diagonal elements were zero and the notation "V" was used generically for different coding schemes. For empirical convenience, we used a topological adjacency specification to denote these spatial relationships.

In this research, the generic specification of autoregressive spatial models used for analysing the sampled MDR-TB clinical and environmental-sampled covariates was:

$$y = \rho_y V_y + (I - \rho_1 V)\ x_1\beta_1 + \ldots + (I - \rho_k V)\ x_k\beta_k + \varepsilon \qquad (2.1)$$

where $\varepsilon = N(0, \sigma^2 I)$ which involved the estimation of $(k+1)$ spatial autocorrelation coefficients. Model (2.1) allowed us to deduce a nested sequence of more commonly used models of autoregressive spatial processes. Depending on constraints imposed on the spatial autocorrelation in the sampled MDR-TB parameters, ρ_y, ρ_1, ..., ρ_k, different autoregressive models were specified by setting, $\rho \equiv \rho_y = \rho_1 = \ldots = \rho_k$, which led to a simultaneous autoregressive (SAR) spatial model:

$$y = \rho V_y + (I - \rho_1 V)\ X\beta + \varepsilon \qquad (2.2)$$

For the spatial lag model, the autocorrelation parameters in the sampled MDR-TB clinical and environmental predictor variables were further con-

strained by the spatial lag factors of exogenous variables, which in this research became zero (i.e. $\rho_1 = ... = \rho_k = 0$,); whereas, for the endogenous variable, we had $\rho \equiv \rho_y$. This process generated the model:

$$y = \rho V_y + X\beta + \varepsilon \tag{2.3}$$

The misspecification perspective for spatial regression models assumed that the basic regression model, $y = X\beta + \varepsilon^*$, had spatially autocorrelated disturbances ε^*, which was decomposed into a specific white-noise component ε (i.e. a stationary time-series or a stationary random process with zero autocorrelation) and a set of unspecified and/or misspecified models which had the structure:

$$y = X\beta + \underbrace{E\gamma + \varepsilon}_{=\varepsilon^*} \tag{2.4}$$

where $E\gamma$ was the misspecification in the MDR-TB model term. This misspecification perspective for spatial autocorrelation was not directly comparable with the spatial process models (2.2) and (2.3), which were based on spatial relationships in the random components y and/or ε. Nevertheless, specific terms were isolated in the spatial process models, where the structure was similar to that of the misspecification term. This resemblance established an indirect link between the MDR-TB models (2.2) and (2.3) and misspecified model (2.4). Furthermore, the unknown misspecification term was approximated by a set of spatial proxy variables. Spatial proxy variables are characterized by strong components, such as a spatially autocorrelated patterns (Griffith, 2003).

Spatial error autocorrelation were included in the regression specification by bringing the spatially unlagged endogenous variable y exclusively to the left-hand side of the regression equation. In an autoregressive expression, the response variable is on the left-side of the equation, while the spatial lagged version of the variable is on the right side (Glantz and Slinker, 2001). This statistical adjustment was accomplished by expanding the matrix term:

$$(I - \rho V)^{-1} = \sum_{k=0}^{\infty} \rho^k V^k \tag{2.5}.$$

The simultaneous autoregressive error model (2.2) was then rewritten as $y - \rho Vy = X\beta - \rho VX\beta + \varepsilon$. The disturbances ε were assumed to be white-noise. Substituting the transformation (5) rendered:

$$y = (I - \rho V)^{-1} [X\beta - \rho V (X\beta) + \varepsilon],$$

$$y = \sum_{k=0}^{\infty} \rho^k V^k (X\beta - \rho VX\beta + \varepsilon),$$

$$y = \sum_{k=0}^{\infty} \rho^k V^k X\beta - \sum_{k=0}^{\infty} \rho^{k+1} V^{k+1} (X\beta) + \sum_{k=0}^{\infty} \rho^k V^k \varepsilon,$$

$$y = X\beta + \underbrace{\sum_{k=1}^{\infty} \rho^k V^k X\beta - \sum_{k=1}^{\infty} \rho^k V^k (X\beta)}_{=0} + \sum_{k=0}^{\infty} \rho^k V^k \varepsilon,$$

$$y = X\beta + \underbrace{\sum_{k=1}^{\infty} \rho^k V^k \varepsilon}_{misspecification\ -\ term} + \varepsilon$$

The model generated implied that the estimated regression parameters $\hat{\beta}$ were unbiased for the basic regression equation, i.e. $y = X\beta + \varepsilon^*$, where ε^* incorporated the misspecification term and the white-noise disturbances. However, the standard errors of the regression parameters were biased. Therefore, we used the spatial lag model (3), but expressed as $(I - \rho V)\, y = X\beta + \varepsilon$. Substituting the transformation (5) rendered:

$$y = \sum_{k=0}^{\infty} \rho^k V^k (X\beta + \varepsilon),$$

$$y = X\beta + \underbrace{\sum_{k=1}^{\infty} \rho^k V^k (X\beta + \varepsilon)}_{misspecification\ -\ term} + \varepsilon.$$

In this case, the misspecification term $\sum \rho^k V^k (X\beta + \varepsilon)(k = 1, ..., \infty)$ included the exogenous variables "X". Consequently, the exogenous variables were correlated with the misspecification term. Under this condition, the standard OLS results for the basic regression model generated from the sampled MDR-TB parameters, i.e. $y = X\beta + \varepsilon^*$, provided biased estimates $\hat{\beta}$ of the underlying regression parameters β.

In this research, we used Moran's *I* in ArcGIS® defined as:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (X_i - \bar{X}) (X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

where "*N*" was the number of MDR-TB covariates indexed by *i* and *j*, *X* the variable of interest (i.e. prevalence rate), $\bar{X}$ the mean of *X*, and w_{ij} was a matrix of spatial weights. The expected value of Moran's *I* under the hypothesis of no spatial autocorrelation was:

$$E(I) = \frac{-1}{N-1}$$

where the variance in the MDR-TB model was generated using:

$$Var(I) = \frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)(\sum_i \sum_j w_{ij})^2}$$

and where all spatial autocorrelation components in the MDR-TB model were quantified using:

$$S_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2$$

$$S_2 = \frac{\sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2}{1}$$

$$S_3 = \frac{N^{-1} \sum_i (x_i - \bar{x})^4}{(N^{-1} \sum_i (x_i - \bar{x})^2)^2}$$

$$S_4 = \frac{(N^2 - 3N + 3) S_1 - NS_2 + 3 (\sum_i \sum_j w_{ij})^2}{1}$$

$$S_5 = S_1 - 2NS_1 = \frac{6 (\sum_i \sum_j w_{ij})^2}{1}$$

Values range from -1 (perfect dispersion indicating NSA) to +1 (perfect correlation or PSA) in the sampled clinical and environmental explanatory variables. Zero values in the MDR-TB models indicate a random spatial pattern.

Results

Our main result is the development and implementation of approximate normality for the detection of MDR-TB clusters of correlated events. We assumed that a buffered health centres region was divided into separate, non-overlapping, administrative areas. The georeferenced health centre was selected in ArcGIS® which was the representative middle point (i.e. centroid). The total number of health centres in the study region was denoted by *I*. We labeled each buffered health centre area i_p as the *p*-th closest health centre to a neighbouring health centre *i*, $p \in \{1, \ldots, I-1\}$, and $i_0 = i$. We let N_i be the population size of the *i*-th health centre, thus the total patient population was N = ∑i = 1INi. We then let C_i and C_{ix} be the number of sampled clinical and environmental covariates with exactly *x* in the *i*-th cell, respectively, with observed MDR-TB prevalence values of c_i and c_{ix}. We also had $C_i = \sum_x C_{ix}$, and the random variable $V_i = \sum_x xC_{ix}$ denoting these attribute values in a health centre *i* with the observed value of v_i. We assumed that $C = \sum_i C_i$ and $V = \sum_i V_i$ denoted the total number of covariates for the SJL study site, respectively, with observed values of *c* and *v*.

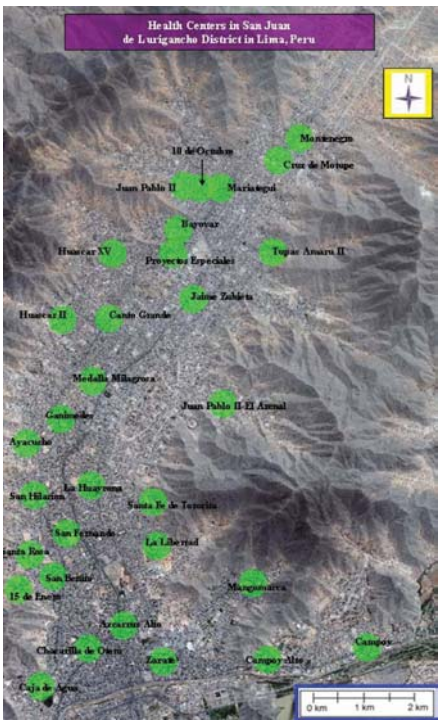

Fig. 4. Base-map of the study site in San Juan de Lurigancho, Lima, Peru.

We generated a broad-scale quantification of topography in the study site using a spatial hydrological model (Fig. 4). Results of the analyses, using Pearson's correlation determined the linear relationship between the sampled MDR-TB covariates which indicated a statistically non-significant linear relationship between georeferenced health centres and the covariate elevation (m) (r = 0.423; P <0.001; n = 31), with a standard deviation (SD) = 104.6 for the sampled MDR-TB covariates and SD = 23.0 for elevation (m) (Table 1).

The field-sampled covariates were then linked with the satellite data in ArcGIS® order to query spatial proximities of the sampled habitats in SAS/GIS®. The SAS Bridge for ESRI alleviated the need for customized MDR-TB data transfers by providing the ability to exchange spatial data between ArcGIS® and SAS®. A model was developed for allowing a parsimonious, but flexible, representation of the covariance matrix of a multivariate model generated from the MDR-TB predictors. Poisson regression analyses were created from the sampled data. An examination of the coefficient estimates from a Poisson model specification indicated; however, that significant overdispersion was present. Thus, to remove the effects of overdispersion and provide more accurate estimates of standard error, a negative binomial with a gamma distributed mean was used to model the MDR-TB parameters.

In this paper we also concentrated on standard regression models $y = X\beta + \epsilon$, where y was an (n x 1) vector of the endogenous variable for the n sampled MDR-TB parameters, X an (n x k) matrix of k exogenous variables, including an (n x 1) unity vector 1, b the (k x 1) vector of regression parameters, and ε an (n x 1) vector of random disturbances. We assumed that spatial autocorrelation among regression disturbances was induced by exogenous spatially autocorrelated factors, which were not incorporated into the model. This led to a model misspecification by shifting parts of the relevant information from the mean response $X\beta$ (or first-order component) into an (n x n) covariance structure of the disturbances [or second-order component cov(ϵ)].

Table 1. Pearson correlation for georeferenced MDR-TB health centre data and the predictor variable elevation in the SJL study site.

Predictor variables	Statistical tests	MDR-TB health centre data	Elevation (m)
MDR-TB health centre data	Pearson correlation	1	0.423
	Sig. (2-tailed)		<0.001
	N	31	31
Elevation (m)	Pearson correlation	0.423	1
	Sig. (2-tailed)	<0.001	
	N	31	31

The correlation, or lack thereof, between the exogenous variables and the misspecification terms of the MDR-TB models were used to design spatial proxy variables so that the properties of either model could be satisfied. We considered two different projection matrices, $M_{(1)} \equiv I - 1(1^T 1)^{-1} 1^T$ and $M_{(X)} \equiv I - X(X^T X)^{-1} X^T$. The projection matrix $M_{(1)}$ was a special case of the more general projection matrix $M_{(X)}$ (Griffith, 2003). The general projection matrix $M_{(X)}$ included, in addition to the constant unity vector 1, additional exogenous variables. The set of eigenvectors $\{e_1, \ldots, e_n\}_{SAR}$ was extracted from the quadratic form

$$\{e_1, \ldots, e_n\}_{SAR} \equiv evec\left[M_{(X)} \frac{1}{2}(V + V^T) M_{(X)}\right]$$

and designed orthogonal to the exogenous variable X. The projection matrix $M_{(X)}$ imposed this constraint. In contrast, the set of eigenvectors that was extracted from

$$\{e_1, \ldots, e_n\}_{Log} \equiv evec\left[M_{(1)} \frac{1}{2}(V + V^T) M_{(1)}\right].$$

These two different sets of eigenvectors established a basis for a spatial regression model. Both expressions were solely defined in terms of exogenous information. This model feature enabled us to also use the eigenvector spatial filtering approach for predictions of the endogenous variable y. The associated sets of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}_{Log}$ and

$\{\lambda_1, \ldots, \lambda_n\}_{SAR}$, with a $\lambda_i \geq \lambda_{i+1}$, range, were used for properly standardizing adjacent link matrices V that were related to irregular spatial tessellations, generated from the MDR-TB parameters. The components of each eigenvector e_i when mapped onto an underlying spatial tessellation, exhibited a distinctive topographic pattern ranging from PSA for $\lambda_i > E(I)$, to NSA for, $\lambda_i > E(I)$.

Each eigenvector was mapped where $E(I)$ was the expected value of Moran's under the assumption of (i) spatial independences and (ii) use of the related projection matrix $M_{(i)}$ or $M_{(X)}$, respectively. The associated Moran's I autocorrelation coefficient, of each eigenvector e_i generated, was equal to its associated eigenvalue $\lambda_i = [e_i^T (V + V^T) e_i] / (2 e_i^T e_i)$, if V was scaled to satisfy $[1^T(V + V^T)1] / 2 = n$. The spatial pattern in the eigenvectors was, - somewhat synthetic for positive global autocorrelation in that the local patterns of the MDR-TB parameters exhibited only positive local autocorrelation, but not negative local autocorrelation (and vice versa for negative global autocorrelation). Finally, the eigenvectors e_i and e_j within each set of eigenvectors, were mutually orthogonal, as the symmetry transformation

$$\frac{1}{2}(V + V^T)$$

was a quadratic form.

Estimation results, for these models, appear in Table 3. Although the reported positive and negative spatial autocorrelation spatial filter component pseudo-R^2 values did not exactly sum for the complete spatial filter, they was very close to their corresponding totals, suggesting that any induced multicollinearity was quite small. Multicollinearity is a term to denote the presence of linear relationships or near linear relationships among sampled predictors and explanatory, independent, or concomitant variables in a model (Hastie and Tibshirani, 1990). Positive spatial autocorrelation and NSA spatial filter component pseudo-R^2 values are reported. GLMM estimation results appear in Table 4. These spatial autocorrelation components suggest the presence of roughly 14% redundant information in the sampled datasets.

Table 2. Global spatial analyses of MDR-TB prevalence rates by health centers in the Lurigancho study site.

Study site	Sampled number	Transformation	MC	sMC	GR
San Lurigancho	31	LN(count + 1.5)	0.58	0.06	0.81

LN = Natural logarithm; MC = Moran's coefficient; sMC = standard error of the MC; GR = Geary ratio.

Table 3. Poisson spatial filtering model results for MDR-TB prevalence rates by health centres in the San Lurigancho study site.

Spatial statistics	Model output
SF: # of eigenvectors	7
SF: MC	0.03
SF: GR	0.68
SF pseudo-R^2	0.32
Positive SA SF: # of eigenvectors	2
Positive SA SF: MC	0.90
Positive SA SF: GR	0.06
Positive SA SF pseudo-R^2	0.04
Negative SA SF: # of eigenvectors	3
Negative SA SF: MC	-0.48
Negative SA SF: GR	0.63
Negative SA SF pseudo-R^2	0.29
Deviance statistic	1.03
Dispersion parameter	0.11

MC = Moran's coefficient; GR = Geary's ratio; SF = spatial filter; SA = spatial autocorrelation. A pseudo-R^2 is the squared correlation between observed and GLM-predicted counts.

Table 4. Poisson spatial filter (SF) generalized linear mixed model (GLMM) random effects for MDR-TB prevalence rates by health centres in the San Lurigancho study site.

Statistics	Model output
Mean	0.03
Standard deviation	0.31
MC	0.14
GR	0.78
Pseudo-R^2	0.86
Changes in significance (using a 0.10 level) of eigenvectors	none

MC = Moran's Coefficient; GR = Geary Ratio; SA = spatial autocorrelation.

Discussion

The spatial hydrological model generated from the clinical and environmental-sampled MDR-TB predictor variables revealed that elevation was not an important variable in the DEM model. The accuracy of our DEM; however, may have been limited by spatial resolution (i.e. 92 m). Several factors can play a role in the quality of DEM-derived products (i.e. terrain roughness, sampling density, interpolation algorithm and terrain analysis algorithm), but spatial resolution is the most important for topographic feature detection and extraction (http://eros.usgs.gov/). Spatial resolution of DEMs is vital for generating MDR-TB models with less simulated errors while allowing for higher quality orthoimagery production, hydrologic modeling, view-shed determination, slope/aspect analyses, and three-dimensional surface visualization. More DEMs should be generated for the SJL study site at varying resolutions for further clarification of elevation and other terrain covariates and their association with georeferenced MDR-TB parameters. For example, a 10 m drainage-enforced DEMs compiled using both the hypsography contour and hydrography elements present in 7.5-minute topographic quadrangle maps may be useful for modeling MDR-TB parameters. Ten m DEMs have the same vertical accuracy as 30 m level 2 products, but their 1/3-arc-second profile supplies a much improved representation of features of the actual landscape. Local differences among DEM grid cells are often analysed to calculate slope and other land surface parameters using the relative vertical accuracy, or point-to-point accuracy on the surface of the elevation model, and the absolute accuracy which determines the quality of such parameters derived from local differencing operations (http://eros.usgs.gov/). Thus, the resolution of a DEM used for quantifying MDR-TB covariates at each pixel (i.e. absolute accuracy) and the accurateness of the sampled clinical and environmental-sampled data represented (i.e. relative accuracy) may be vital for generating a geometrically correct reference frame for validating parameters such as elevation associated with MDR-TB parameters sampled in the SJL study site.

A negative binomial regression analyses identified nearest neighboring health centre distance as significantly influencing the sampled MDR-TB data. On visual inspection of the cluster maps it was clear that there is relative clustering among all MDR-TB parameters as evident by the steep rise in prevalence estimates at small distances. Physical distance between social networks is among the most important facilitators of MDR-TB transmission (Godvoy, 2004). Monitoring changes in local case numbers using distance based measurements could help target health services to specific regions in the SJL study site with the highest disease burden. Descriptive MDR-TB maps of the spatio-temporal patterns of sampled clinical and environmental covariates should be generated, using different statistical tests based on distance- based measurements using space–time interaction models (e.g. the Knox test and *k*-nearest neighbor test) and a cluster-detection algorithm (e.g. the space–time scan statistic).

Spatial autocorrelation indices based on log-transformed MDR-TB prevalence rates sampled in each health centre, revealed PSA. The use of prophylactic treatment and other MDR-TB control measures may tend to have demographic dimensions with spatial expressions in the SJL study site. Socio-economic factors may also impact contagion diffusion, inducing PSA in sampled MDR-TB covariates. For example, neighborhoods in the SJL study site were composed of clustered households with similar attributes among specific sociodemographic characteristics (e.g. distance to the nearest health centre). Communities with overcrowded housing experience a higher prevalence of latent MDR-TB infection and/or risk factors for progression from MDR-TB infection to disease (Manton and Stallard, 1981; Godoy, 2004). Other risk factors for MDR-TB may include substance abuse, and insufficient nutrition (Reichman et al., 1979), which may be more prevalent in communities with socio-economic disadvantages such as unemployment and homelessness (Concato and Rom, 1994). Furthermore, the PSA

found in the spatial distribution of health centres in the SJL study site may reflect the effects of changes in spatial patterns and the effectiveness of local public health programmes that attempt to minimise the size of MDR-TB infected populations.

In our spatial filtering analyses of the clinical and environmental MDR-TB data, synthetic variates appeared in the numerator of Moran's *I*. Eigenvectors were extracted from a transformed spatial link matrix which exhibited distinctive spatial patterns with associated spatial autocorrelation levels. This matrix decomposed the Moran's *I* statistic generated using the sampled MDR-TB covariates for constructing a Poisson spatial filtering GLMM. One advantage of a spatial filter approach is that it also enables use of a GLM specification which for disease mapping purposes is based upon the binomial, Poisson, or negative binomial probability models depending upon whether a disease map is expressed in terms of a binary, a percentage or a count variable (Griffith, 2005). The regression residuals represented spatially independent variable components. Mean, variance and statistical distribution characterizations and descriptions of the georeferenced random variables and their interrelationships were derived in terms of the eigenfunction spatial filter. The eigenvectors described the full range of all possible mutually orthogonal MDR-TB map patterns, in the SJL study site, based on the clinical and environmental-sampled covariates.

The spatial dependency in our models suggested the presence of spatially pseudo-replicated data in the MDR-TB observations due to the presence of latent autocorrelation. In this research, redundant information in the MDR-TB model was most probably attributable to the locational arrangements of sampled health centres in the SJL study site, which caused the observations to be dependent, rather than independent. As redundant information, spatial autocorrelation in MDR-TB data may be linked to missing value estimation and interpolation, as well as the notions of effective sample size and spatial configuration of georeferenced data (Dutilleul, 1993). Spatial autocorrelation devices are constructed from geographic weights matrices, which are used to capture the covariation among values of one or more random variables that are associated with the configuration of areal units (Griffith, 2003). In future analyses of MDR-TB parameters, in the SJL study site, an eigenfunction spatial filter formulation should be used to reduce sampling variability in accordance with the degree of redundant information quantified based on latent autocorrelation estimates. Other Gaussian regression models, such as moving average models, conditional autoregressive models or autoregressive models without a common factor constraint, may also be accommodated by the eigenvector spatial filtering approach for quantifying spatially pseudo-replicated MDR-TB data.

In this paper, we demonstrate that the eigenvector spatial filtering approach can be embedded into a semiparametric statistical framework using MDR-TB parameters. Although the eigenvector spatial-filtering approach is statistically and numerically robust, for spatial tessellations generated from spatio-temporal sampled datasets of MDR-TB parameters, it does require specific statistical restrictions. For example, since eigenfunction decomposition yields n eigenvectors, an MDR-TB researcher needs to restrict attention to only those eigenvectors describing substantive PSA and NSA (e.g. MC > 0.25, a value that tends to relate to about 5% of the variance being attributable to redundant information arising from latent spatial autocorrelation, in our areal unit neighborhood configuration). Generating statistical limitations reduces the candidate set to a more manageable number for describing a given MDR-TB map. Supervised stepwise selection from a set of such eigenvectors is a useful and effective approach to identifying the subset of eigenvectors that best describes latent spatial autocorrelation in an MDR-TB map. This procedure begins with only the intercept included in a regression specification. Next, at each step an eigenvector is considered for additional model specification. For the stepwise linear Gaussian model, commonly the eigenvector having the largest partial correlation

is selected, but only if its corresponding F-ratio achieves or surpasses a prespecified level of significance; this is the criterion used to establish statistical importance of an eigenvector for describing the full range of all possible mutually orthogonal MDR-TB map patterns which may be also interpreted as synthetic map variables that are analogous to the residual spatial variables. In stepwise generalized linear modeling regression, the eigenvector that produces the greatest reduction in the log-likelihood function chi-square test statistic is selected, but only if it produces at least a prespecified minimum reduction; as before, this is the criterion used to establish statistical importance of an eigenvector (Griffith, 2003). In each statistical procedure, at each step all eigenvectors, generated from MDR-TB parameters, previously entered into a spatial filter equation are reassessed, with the possibility of removal of vectors added at an earlier step. The forward/backward stepwise procedure terminates automatically when some prespecified threshold values (respectively for F-ratios and chi-square statistics) are encountered for entry and removal of all candidate eigenvectors. The ultimate inclusion criterion is determined by the Moran's *I* coefficient value of the residuals, which should indicate an absence of spatial autocorrelation. Satisfying this condition sometimes requires supervised backward elimination of marginally selected eigenvectors because their inclusion can force the residual Moran's *I* coefficient value to decrease too far below zero. This final stopping criterion for the linear Gaussian model is relatively easy to implement as Moran's *I* coefficient distributional theory is known for linear regression residuals; a corresponding stopping rule for GLM regression is far more difficult to implement because of a lack of such distributional theory.

In conclusion, a DEM revealed that elevation was not significantly associated with the sampled MDR-TB data. A negative binomial regression analyses identified the independent variable distance between health centres as significantly influencing sampled data. We then decorrelated the MDR-TB observations using a spatial filter analyses which revealed PSA in all models tested; similar log-MDR-TB prevalence rates of the health centres aggregated in geographic space. The spatial filtering analyses transformed all variables containing spatial dependence into covariates free of spatial dependence by partitioning the original georeferenced attribute variable into two synthetic variates: (i) a spatial filter variate capturing latent spatial dependency, that otherwise would have remained in the response residuals, and (ii) a nonspatial variate that was free of spatial dependence. These latent autocorrelation estimates suggested the presence of roughly 14% spatially pseudo-replicated data in the clinical and environmental-sampled explanatory variables. Poisson regression models and a distance weighted error autocovariate matrix can be used for analyzing and prioritizing clinical and environmental-sampled MDR-TB covariates. Furthermore, eigenvector mapping can be used for resource allocation and for implementing MDR-TB control strategies in urban environments.

References

Akashi SS, Hawker J, Ali S, 1996. Tuberculosis mortality in notified cases from 1989-1995 in Birmingham AJPH 112, 165-168.

Anselin L, 1995. Local indicators of spatial association-LISA. Geogr Anal 27, 93-115.

Augustin NH, Mugglestone MA and Buckland ST, 1996. An autologistic model for the spatial distribution of wildlife. J Appl Ecol 33, 339-347.

Barr RG, Dies-Roux AV, Kirsch CA, Pablos-Méndez A, 2000. Neighborhood poverty and the resurgence of tuberculosis in New York city, 1984-1992. AJPH 9,1487-1493.

Chatterjee S, Hadi A, 1988. Sensitivity analysis in linear regression. Wiley, New York, USA.

Clarke SE, Bough C, Brown RC, Walgreen GE, Thomas CJ, and Lindsay SW, 2002. Risk of malaria attacks in Gambian children is greater away from malaria vector breeding sites. Trans R Soc Trop Med Hyg 96, 499-506.

Cliff AD, Ord JK, 1973. Spatial autocorrelation. Pion, London, UK.

Concato J, Rom WN,, 1994. Endemic tuberculosis among

homeless men in New York city. Arch Int Med 154, 2069-2073.

Dutilleul P, 1993. Modifying the t test for assessing the correlation between two spatial processes. Biometrics 305-314.

ENVI.4.5. http://www.ittvis.com/

ERDAS Imagine v.8.7™, Atlanta, USA.

Farrar D, Glauber R,1967. Multicollinearity in regression analysis: the problem revisited. Rev Econ Stat 49, 92-107.

Fotheringham AS, Brunsdon C, Charlton M, 2002 Geographically weighted regression: the analysis of spatially varying relationships. Wiley.

Glantz SA, Slinker BK, 2001. A primer of applied regression and analysis of variance. McGraw-Hill, New York, USA.

Glantz S, 1997 Primer of biostatistics (4th Ed.). McGraw-Hill New York, USA.

Godoy P, Domínguez A, Alcaide J, Camps N, Jansà JM, Minguell S, Pina JM, Díez M, 2004. The working group of the Multicentre Tuberculosis Research Project (MTRP): characteristics of tuberculosis patients with positive sputum smear in Catalonia, Spain. 14, 71-75.

Griffith DA, 2002. A spatial filtering specification for the auto-Poisson model. Stat Prob Lett 58, 245-251.

Griffith DA, 2003 Spatial autocorrelation on spatial filtering. Springer.

Griffith DA, 2006. Hidden negative spatial autocorrelation. J Geogr Sys 8, 335-355.

Griffith DA, Layne LJ, 1999. A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Datasets. Oxford University Press, New York, USA.

Hastie T, Tibshirani R, 1990 Generalized Additive Models. Published by Chapman and Hall, New York, USA, 335 pp.

Iseman MD, 1993.Treatment of multidrug-resistant tuberculosis. NEJM 11,784-791.

Johnson RT, 2003. Emerging viral infections of the nervous system. J Neurobio 9, 140-147.

Jacob BG Lampman RL Ward MP Muturi E Funes J Morris JA Novak RJ, 2009a. Geospatial variability of *Culex pipiens* and *Culex restuans* aquatic habitats in urban Champaign, Illinois. Int J Remot Sens 30, 2005-2019.

Jacob BG, Griffith DA, Gunter JT, Muturi EJ, Caamano EX, Githure JI, Regens JL, Novak RJ, 2009b Quantifying stochastic error propagation in Bayesian parametric estimates using non-linear parameters of *Anopheles gambiae* s.l. habitats. Int J Remot Sens, in press.

Jacob BG, Gu W, Muturi EJ, Caamano EX, Morris JM, Lampman R, Novak, RJ, 2009c. Developing operational algorithms using linear and non-linear least squares estimation in ArcGIS® and Python® for identification of *Culex pipiens* and *Culex restuans* aquatic habitats in a mosquito abatement district (Cook County, Illinois). Geospat Health 3, 157-176.

Jacob BG, Griffith DA, Novak RJ, 2008a. Decomposing malaria mosquito aquatic habitat data into spatial autocorrelation eigenvectors in a SAS/GIS® module. Trans GIS 12, 341-364.

Jacob BG, Griffith DA, Gunter JT, Muturi EJ, Caamano EX, Shililu JI, Githure JI, Regens JL, Novak RJ, 2008c. Spatial filtering specification for an auto-negative binomial model of *Anopheles arabiensis* aquatic habitats. Trans GIS 12, 243-259.

Jacob BG, Muturi E, Mwangangi J, Funes J, Shililu J, Githure J, Novak RJ, 2008a. Hydrological modeling of geophysical parameters of arboviral and protozoan disease vectors in internally displaced people camps in Gulu, Uganda. Int J Health Geog 7, 11-18.

Johnson RT, 2003. Emerging viral infections of the nervous system. J Neurobio 9, 140-147.

Kaiser M, Cressie N, 1997. Modeling Poisson variables with positive spatial dependence. Stat Probab Lett 35, 423-432.

Maddala GS, 2001. Introduction to Econometrics, John Wiley & Sons Ltd.

Miles JNV, Shevlin ME, 2001. Applying regression and correlation: a guide for students and researchers. Sage Publications, London, UK.

Manton KG, Stallard E, 1981. Methods for evaluating the heterogeneity of aging processes in human populations using vital statistics data: explaining the black/white mortality crossover by a model of mortality selection. Hum Biol 53, 47.

McCullagh P, Nelder JA, Generalized Linear Models. Chapman and Hall, London, UK.

Pedhazur EJ, 1997. Multiple Regression in Behavioral Research (3rd Ed). Harcourt Brace, Orlando, USA.

Pielou EC, 1969. An Introduction to Mathematical Ecology. New York, USA.

Reichman LB, Felton CP, Edsall JR. Drug dependence, a possible new risk factor for tuberculosis disease. Int Med 139,

337-339.

Slinker BK, Glantz SA, 1985. Multiple regression for physiological data analysis: the problem of multicollinearity. Amer J Physiiol 249, 1-12.

Smith PA, 1994. Autocorrelation in logistic regression modeling of species distributions. Global Ecol Biogeogr 4, 47-61.

QuickBird http://www.digitalglobe.com/ product/index.shtml

Yang ZH, Rendon A, Flores A, Medina R, Ijaz K, Llaca J, Eisenach KD, Bates JH, Villarreal A, Cave MD, 2001. A clinic-based molecular epidemiologic study of tuberculosis in Monterrey, Mexico. IJTLD 5, 313-320.

Zignol M, Hosseini MS, Wright A, Weezenbeek CL, Nunn P, Watt CJ, 2006. Global incidence of multidrug-resistant tuberculosis. J Infect Dis 194, 479-485.