Journal of Public Health and Epidemiology Vol. 5(3), pp. 122-143, March 2013
Available online at http://www.academicjournals.org/JPHE
DOI: 10.5897/JPHE12.076
ISSN 2141-2316 ©2013 Academic Journals

Full Length Research Paper

A Bayesian Poisson specification with a conditionally autoregressive prior and a residual Moran's coefficient minimization criterion for quantitating leptokurtic distributions in regression-based multi-drug resistant tuberculosis treatment protocols

Benjamin G. Jacob[1]*, Fiorella Krapp[2], Mario Ponce[2], Nanhua Zhang[1], Semiha Caliskan[1], Jose Hasemann[1], Daniel A. Griffith[3], Eduardo Gotuzzo[2] and Robert J. Novak[1]

[1]Department of Global Health, College of Public Health, University of South Florida, Tampa Fl, USA 33612.
[2]School of Medicine Cayetano Heredia University, Lima, Peru.
[3]School of Economic, Political and Policy Sciences, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX, USA 75080-3021.

Accepted 17 December, 2012

In this study, we employed an eigenfunction decomposition algorithm associated with a Moran's coefficient to investigate district-level non-linearity in an empirical dataset of spatiotemporal-sampled MDR-TB parameter estimators sampled in San Juan de Lurigancho (SJL) Lima, Peru. The non-parametric technique attempted to remove the inherent autocorrelation in the model by introducing appropriate synthetic surrogate variants. We also constructed a robust Bayesian Poisson model to generate unbiased estimators for qualitatively assessing resistance to four commonly used drugs in TB treatment: isoniazid, rifampin, ethambutol, and streptomycin. Initially, data of residential addresses of individual patients with smear-positive MDR-TB were geocoded in ArcGIS. Next, the sampled data were matched automatically and interactively within the geodatabase. The MDR-TB data feature attributes were then calculated and digitally overlaid onto sub-meter resolution satellite data within a 1 km buffer of 31 georeferenced health centers using a 10 m^2 grid-based algorithm. Global autocorrelation statistics were then generated by decomposing the sampled data into positive and negative spatial filter eigenvectors using the eigenfunction decomposition algorithm. Bayesian Poisson projections were then rendered employing normal priors for each of the sampled estimators. A Residual Moran's coefficient (MC) minimization criterion was then applied to the clinical coefficients generated from the decomposition algorithm to detect any unaccounted latent autocorrelation error in the estimators. The model accounted for approximately 14% pseudo-replicated information and exhibited positive residual autocorrelation. Spatial statistics can elucidate the mechanics of MDR-TB transmission by prioritizing clinical covariates for identifying spatial distribution of high- risk populations and random heterogeneity in resistant strains.

Key words: Multi-drug resistant tuberculosis, Bayesian Poisson, residual Moran's coefficient (MC), minimization criterion, San Juan de Lurigancho (SJL) Lima, Peru.

INTRODUCTION

Multiple linear regression analysis techniques coupled with normal probability models have become standard epidemiological tools to quantitatively analyze spatiotemporal-sampled clinical and environmental covariates associated with multi-drug resistant tuberculosis (MDR-TB) for identifying high-risk populations (Smith, 1994; Johnson, 2003;

Clarke et al., 2002; Akashi et al., 1996; Barr et al., 2000). MDR-TB is defined as TB that is resistant to isoniazid (INH) and rifampicin, which most commonly develops in the course of TB treatment (Iseman, 1993). Generalized linear models (GLMs) represent a class of fixed effects regression models for several types of dependent variables (e.g., continuous, dichotomous, counts). For example, El Sahly et al., (2006) analyzed molecular epidemiological techniques of MDR-TB employing a case–control study of 2,170 patients with drug- susceptible TB in Houston and Harris County, Texas, from 1995 to 2001 using a multivariate logistic regression where drug resistance was the categorical dependent variable. Cases with various forms of resistant TB were also compared to a control group which consisted of patients with culture positive, drug susceptible TB, with respect to sociodemographic, clinical and strain-stratified genotype-dependent explanatory predictor variables using bivariate chi-square and univariate statistics. As part of the study, patients were identified as drug-resistant cases if they had a positive culture for an MDR-TB strain that was resistant to any of the following: isoniazid, rifampin, ethambutol or streptomycin. In the analyses, the variables that showed a colinearity coefficient of 0.3 or more were eliminated. Thereafter, the multivariate logistic model constructed employing the explanatory predictor variables associated with drug resistance revealed a *P* value of ≤0.1. In the final model, *P* values of ≤0.05 were considered significant. The regressed residual MDR-TB covariates revealed that the observational predictors related to human immunodeficiency virus (HIV) seropositivity, Hispanic ethnicity, Asian ethnicity and a history of past TB were associated with some parameter estimators, whereas, being foreign born having a history of past TB, and younger age were definitive estimators (i.e., P< 0.050). The model revealed that the ethnic groups may have been more affected by TB because of the propensity of HIV among these sampled populations. Moreover, the authors identified that patients with AIDS and other disseminated immunodeficiency disorders were at an increased risk of acquiring drug resistance particularly rifampin while on therapy.

Although linear mixed models are widely used in MDR-TB which can handle non-normal data by using link functions and exponential family (e.g. normal, Poisson or binomial distributions), the assumptions underpinning multiple regression necessarily impose several important constraints that may not always be satisfied or, that might at least require careful consideration when modeling time-series dependent MDR-TB clinical and /or environmental covariates. For example, commonly, the relationships between the outcome and the explanatory predictor variables in a robust spatiotemporal MDR-TB model constructed from multiple regression-based residuals are assumed to be linear and the residual error variance estimates are assumed to be the same, regardless of the value of the sampled clinical and /or environmental covariate coefficients. Also, commonly in a linear, time-series dependent MDR-TB predictive regression equation, the error residuals in the model are assumed to be normally distributed and the sampled estimators are assumed to be independent. However, this may not always be the case in spatiotemporal MDR-TB regression-based modeling since many sampled clinical and/or environmental covariate coefficients may exhibit non-linear feature attributes. As such, although the estimated regression coefficients may be unbiased in the MDR-TB model they will not express the minimum variance among all estimates. Further, the mean squared error would also tend to underestimate the variance in the model. This would lead directly to overestimation of the sampled parameter estimator significance levels which, in turn, would result in underestimation of confidence intervals thus, leading to underestimation of the test statistics for the F test. The F-test is sensitive as it is commonly quantitated by considering a decomposition of the variability in a collection of data in terms of estimable functions and their associated sum of squares (Dorman 2007).

Estimable functions are functions of model parameters (e.g. difference between two parameters, difference between a parameter and the difference of two others, etc.) that are invariant regardless of the generalized inverse employed. The GLM, VARCOMP, and other SAS/STAT procedures label the Sums of Squares associated with the various effects in the model as Type I, Type II, Type III, and Type IV (www.sas.edu). For example, in the Type I form of sum of squares (i.e., the hierarchical decomposition of the sum-of-squares method), each sampled MDR-TB term would be adjusted for only the term that precedes it in the model. Type I sums of squares could then be used for constructing a balanced ANOVA time series-dependent MDR-TB model in which the main effects in the sampled data would be specified before any first-order interaction effects are quantitated. Thereafter, any first-order interaction effects in the model would be specified before any second-order interaction effects, and the second-order interaction effects would be specified in the model before the third-order interaction effects, and so on. A polynomial time-series dependent MDR-TB regression model for any lower-order terms could also be specified before any higher-order terms are quantitated. Further, a purely nested MDR-TB model in which the first-specified effect is nested within the second-specified effect may also be determined. For defining a Type II sums of squares in a spatiotemporal MDR-TB model a method can be employed which calculates an effect that can be adjusted for by all other "appropriate" effects in the model. An appropriate effect is one that corresponds to all effects that do not contain the

*Corresponding author. E-mail: bjacob1@health.usf.edu.

effect being examined (Cressie 1993). The Type II sum-of-squares method could then be used for deriving robust unbiased estimators in a balanced ANOVA time series dependent MDR-TB model and/or any MDR-TB model with purely nested design.

Type III estimable functions for sum of squares (i.e., the default method) can also be utilized for regressing time series dependent MDR-TB for modeling clinical and environmental exploratory covariates. This method can calculate the sums of squares in a time series dependent MDR-TB model using an effect in the design matrix of the model. The Type III sums of squares have one major advantage for spatiotemporal MDR-TB modeling in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Hence, this type of sums of squares would be considered useful for an unbalanced spatiotemporal MDR-TB model with no missing cells. In a factorial design with no missing cells, this method would be equivalent to the Yates' weighted-squares-of-means technique. Today, by default, most major statistical programs perform unbalanced ANOVA based on Type III sums of squares (that is, Yates's weighted squares of means) (McPherson and Jetz 2007, Cressie 1993). The Type III sum-of-squares method could be used for any MDR-TB models listed in the aforementioned Type I and Type II specifications.

Finally, a Type IV estimable function can be designed for a situation in which there is a spatiotemporal MDR-TB model with missing cells. For example, for any effect *F* in an MDR-TB model design, if *F* is not contained in any other effect, then Type IV = Type III = Type II. When *F* is contained in other sampled MDR-TB effects, Type IV will distribute the contrasts being made among the sampled clinical and environmental parameter estimators in *F* to all higher-level effects equitably. The Type IV sum-of-squares method is commonly used for any models listed for Type I and Type II estimable functions. Fortunately, PROC GLM in SAS and the SAS regfunction in R both can calculate various F tests.

The F-test is designed to test if two population variances are equal (Homer and Lemeshew, 2000). The test does this by comparing the ratio of two variances. So, if the variances are equal in a spatiotemporal MDR-TB model, the ratio of the variances will be 1. Commonly, the F-test in one-way analysis of variance is used to assess whether the expected values of a sampled quantitative variable within several pre-defined groups differ from each other. For example, suppose that a medical trial compares four MDR-TB related treatments. The ANOVA F-test can be used to assess whether any of the treatments is on average superior, or inferior, to the others versus the null hypothesis that all four treatments yield the same mean response. This is an example of an "omnibus" test, meaning that a single test is performed to detect any of several possible differences.

Hypotheses regarding MDR-TB regression- based equality vs. inequality tests and between k expectancies $\mu_1=\mu_2=\ldots=\mu_k$ vs. $\mu_1\neq\mu_2\neq\ldots\neq\mu_k$ in ANOVA; or regarding equality between k standard deviations $\sigma 1=\sigma_2=\ldots.=\sigma_k$ vs. $\sigma_1\neq\sigma_2\neq\ldots\neq\sigma_k$ for testing equality of variances in ANOVA; or regarding the clinical and/or environmental covariate coefficients $\beta 1=\beta_2=\ldots.=\beta_k$ vs. $\beta_1\neq\beta_2\neq\ldots\neq\beta_k$ in multiple linear regression can be tested using omnibus" test (Fotheringham, 2002). Alternatively, pairwise tests could be carried out among the treatments (e.g., the MDR-TB trial example with four treatments is carried out using six tests pairs of treatments). The advantage of the ANOVA F-test for spatiotemporal MDR-TB modeling is that there is no requirement to pre-specify which treatments are to be compared, and there is no need to adjust for making multiple comparisons. The disadvantage of the ANOVA-MDR-TB related F-test is that if the null hypothesis is rejected in the time series data, the residuals would not be able to determine which treatments are significantly different from the others. If the F-test is performed at level α we cannot state that the treatment pair with the greatest mean difference is significantly different at level α (Hosmer and Lemeshew 2000). Thus, although the F-test can be used to compare nested models, in an asymptotic or approximate fashion to test the hypothesis that the simpler of the time series dependent MDR-TB models is sufficient to explain the data, for example, the residuals may have correlated error. A variance decomposition may even be performed for generating inferences for the variances in the model but, sources of variation in multilevel regression MDR-TB can still occur.

Although the F statistics may not be exact, MDR-TB researchers to date have found that the F-ratios are acceptable unless the design is highly unbalanced. The F-ratio is used to determine whether the variances in two independent samples are equal (Cressie 1993). Ideally, this ratio should be approximately 1 in a spatiotemporal MDR-TB model if the corresponding effects are zero; otherwise the expected F-ratio will exceed 1. We would expect the F-ratio to be less than 1 only in unusual models with negative within-group correlations (e.g., if the spatiotemporal-sampled MDR-TB data have been renormalized in some way, and this had not been accounted for in the data analysis). When the null hypothesis of no group differences is true, then the expected value of the numerator and denominator of the F ratio will be equal (Hosmer and Lemeshew 2000). As a consequence, the expected value of the F ratio in a spatiotemporal MDR-TB model when the null hypothesis is true is also close to one. When the null hypothesis is false in the model and there are group differences between the means, the expected value of the numerator will be larger than the denominator. As such, the expected value of the F ratio will be larger and the MDR-TB model estimates will also more likely be larger than one under the null hypothesis. However, the point is that both the numerator and denominator in the MDR-TB model would be random variables and so would be the F

ratio. If we assume the null hypothesis is true in the time series dependent MDR-TB model one distribution will be determined, and if we assume that it is false with various assumptions about effect size, sample size, and so forth another distribution would be rendered. The F ratio is drawn from a distribution (Cressie 1993). Thereafter, an F value for the MDR-TB model can be determined. Fortunately, when the null hypothesis is false in the model it would be still be possible to get an F ratio less than one.

However, an F-ratio based on a mean square error (MSE) in a spatiotemporal MDR-TB model will not be able to disentangle the contribution of the experimental effect (i.e. the linear component) and the degree to which the treatment effect varies across participants/covariates (i.e. the non-linear participant by experimental-effect interaction). In an analogy to standard deviation, taking the square root of MSE in a spatiotemporal MDR-TB model will yield the root mean square error or root mean square deviation (RMSE), which has the same units as the quantity being regressed for an unbiased estimator. The RMSE is the square root of the variance, known as the standard deviation. Thus, a statistically significant effect in a spatiotemporal predictive regression-based MDR-TB model could be due to one of three things: (a) a significant experimental effect, (b) significant variation in the treatment effect across participants, or (c) both of these things. Unfortunately, the F ratio would not be able to differentiate the optimal residual forecasts from such distributions.

Further, if any of these tests are performed to determine the underlying assumption of homoscedasticity (i.e., homogeneity of variance), in the spatiotemporal MDR-TB model as a preliminary step to testing for mean effects, the residuals would reveal an increase in experiment-wise Type I error rate. Therefore, significance testing for quantitating resulting confidence regions and tests of the hypotheses employing combinations of sampled MDR-TB explanatory clinical and /or environmental covariate coefficients would be critically jeopardized. Violations of linearity are extremely serious in time series dependent infectious disease models as fitting linear data attributes to non-linear algorithms would render forecasts that are erroneous especially when extrapolation occurs beyond the range of the sampled data. For example, spatio-temporal MDR-TB statistics will not follow the F-distribution, under the null hypothesis in a time series dependent model unless the sums of squares are independent, and each follow a scaled chi-squared distribution. The latter condition, however, is only guaranteed if the sampled clinical data values are independent and normally distributed with a common variance.

Another common problem in the use of linear coefficients when modeling spatiotemporal-sampled MDR-TB data is the occurrence of covariates that are not independent (i.e., non-zero correlations amongst covariates) giving rise to multicollinearity. Multicollinearity increases the standard errors of the coefficients (Hosmer and Lemeshew 2000). Increased standard errors in turn means that the spatiotemporal-sampled clinical and environmental covariate coefficients for some independent variables may be found not to be significantly different from 0. Without multicollinearity and with lower standard errors, however these same coefficients and their null findings might have been found to be significant. In other words, multicollinearity in a spatiotemporal MDR-TB model would misleadingly inflate the standard errors.

Unfortunately, since the factors associated with the emergence of MDR-TB and their effects on the epidemiology of TB are complex and multi-faceted (e.g., poor medical management, lack of direct observed treatment, limited or interrupted drug supplies, poor drug quality, widespread availability of anti- TB drugs without prescription, dissociation between public and private sector, and poorly managed national control programmes (Espinal, 2001; Farmer et al., 2001), multiple parameter estimators are commonly employed in the regression uncertainty matrix often rendering serial correlation in the residual outcome explanatory predictor covariate dataset. When more than two covariates in a model are highly correlated, multicollinearity can occur (Miles and Shelving, 2001; Pedhazur, 1997; Slinker and Glantz, 1985).Collinearity and multicollinearity can seriously distort the interpretation of a spatiotemporal linear-dependent regression model (Cohen et al., 2003; Maddala, 2001; Chatterjee and Hadi, 1988). Traditionally, the role of each sampled covariate in a spatiotemporal time-series dependent MDR-TB regression model would be to increase precision, as expressed through a reduction in residual predictive error variance covariance matrix estimates, as well as, reduced bias in the sampled coefficients. Multicollinear MDR-TB clinical and environmental-related covariate coefficients however, would be difficult to analyze as their effects on a response variable could be due to either true synergistic relationships among the sampled covariates or, confounding effects creating spurious correlations.

In some sense, the collinear MDR-TB variables would contain the same information about the dependent variable in the spatiotemporal model. If nominally "different" measures actually quantify the same phenomenon then they are redundant (Glantz and Slinker, 2001; Fotheringham et al., 2002). Alternatively, if the time series-dependent MDR-TB explanatory predictor variables are accorded different names and perhaps employ different numeric measurement scales but, continue to maintain a high correlation with each other, the residuals would still suffer from redundancy. A principal danger of spatiotemporal data redundancy is overfitting in regression model frameworks. In statistics, overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship (Dormann 2007; Homer and Lemeshew 2000; Cressie 1993; Manton and Stallard 1981). Additionally, when a spatiotemporal MDR-TB distribution model is excessively complex, such as a model with extensive parameter estimators relative to

the number of observations, biased predicted residual space-time autoregressive error estimates may be rendered by the model. Unfortunately this occurs commonly by exaggerating minor fluctuations inconspicuously in the sampled clinical and environment sampled data.

As such, separable approximations of non-separable space-time MDR-TB error covariance matrix estimates cannot be quantitated. Further, the nearest Kronecker product approximation in the time series cannot be determined especially in a MDR-TB dataset employing a Frobenius norm of a space-time error covariance matrix. The Kronker product is a generalization from vectors to matrices which renders the matrix of the tensor product (Fotheringham 2002). The Frobenius norm is the square root of the sum of the absolute squares of its elements (Cressie 1993). The tensor product, denoted by $\otimes$, may be applied in different contexts to vectors, matrices, tensors, vector spaces, algebras, topological vector spaces, and modules, among many other structures or objects (Griffith and Layne 1999, Cressie 1993). As such, in a hypothetical generalized MDR-TB bilinear operations model a function combining elements of two vector spaces (e.g., matrix multiplication) will not yield an element of a third vector space that is linear in each of its arguments. Thus, solutions preserving properties of residual space-time MDR-TB uncertainty covariance matrices, such as symmetry, positive definiteness, and other structures cannot be quantitated.

In linear algebra, a symmetric $n \times n$ real matrix M is said to be positive definite if $z^{\mathrm{T}}Mz$ is positive, for any non-zero column vector z of n real numbers; where z^{T} denotes the transpose of z. More generally, an $n \times n$ complex spatiotemporal MDR-TB matrix M would be positive definite if z^*Mz is real and positive for all non-zero complex vectors z; where z^* denotes the conjugate transpose of z. This property implies that M is an Hermitian matrix. The conjugate transpose, or adjoint matrix of an m-by-n matrix with complex entries is the n-by-m matrix A^* obtained from A by taking the transpose and then taking the complex conjugate of each entry (i.e., negating their imaginary parts but not their real parts) (Cressie 1993). The conjugate transpose would then be formally defined by where the subscripts in the MDR-TB model denotes the i,j-th entry, for $1 \le i \le n$ and $1 \le j \le m$, and where the overbar denotes a scalar complex conjugate. The complex conjugate of $a + bi$, where a and b are reals, is $a - bi$. (Cressie, 1993). This definition can also be written as $\mathbf{A}^* = (\overline{\mathbf{A}})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}$ in a spatiotemporalMDR-TB modelwhere $\mathbf{A}^{\mathrm{T}}$denotes the transpose and $\overline{\mathbf{A}}$denotes the matrix with complex conjugated entries. Thus , a Hermitian MDR-TB matrix (i.e., self-adjoint matrix) is a square matrix with complex clinical and environmental covariate entries that is equal to its own conjugate transpose – that is, the element in the i-th row and j-th column is equal to the complex conjugate of the element in the j-th row and i-th column, for all indices i and j: $a_{ij} = \overline{a_{ji}}$ .If the conjugate transpose of a matrix $A$is denoted by $A^\dagger$, then the Hermitian property can be written in a MDR-TB model concisely as $A = A^\dagger$. for efficient predictive residual uncertainty quantification. Unquantitated hidden latent correlation error coefficients in spatiotemporal datasets of time series dependent covariate coefficients cangenerate misspecified estimates (Griffith, 2008).

Additionally, since one of the features of multicollinearity is that the standard errors of the affected regression residual coefficients tend to be large (Glantz and Slinker, 2001; Glantz and Amrhein, 1997), the test of the hypothesis that the sampled explanatory covariate coefficients would be equal to zero in a spatiotemporal clinical/environmental–oriented MDR-TB regression-based equation would then subsequently lead to a failure to reject the null hypothesis. In such circumstances, if the linear-dependent observational explanatory predictors are estimated, a covariate would still be found to be significant; specifically, a TB analyst will reject the hypothesis that the coefficient is zero. In statistics, simple linear regression is the least squares estimator of a linear regression model with a single predictor variable (Dutilleul 1993; Hosmer and Lemeshew, 2000). A simple linear regression fits a straight line only through the set of n points in such a way that makes the sum of squared residuals of a model robust that is, vertical distances between the points of the spatiotemporal-sampled dataset and the fitted line are as small as possible (Fotheringham, 2002). As such, in the presence of multicollinearity, a TB analyst might falsely conclude that there is no linear relationship between an independent and a dependent variable in a spatiotemporal MDR-TB regression-based predictive risk-based model.

So long as the underlying specification is correct, however, multicollinearity will not actually bias spatiotemporal MDR-TB regression model residuals; it will just produce large standard errors in the related inde-pendent variables. If, however, there are other problems such as omitted variables which introduce bias in the model, multicollinearity can multiply the effects of that bias in the residuals by orders of magnitude within spatially autoregressive uncertainty dependent frameworks. Importantly, the common use of regression in spatiotemporal MDR-TB modeling exercises is to take sampled explanatory covariate coefficients rendered from the model residuals and then apply them to other non-linear higher order autoregressive matrices (e.g., block kriging). Kriging is a group of geostatistical techniques commonly employed to interpolate the value of a random field (e.g., the elevation, z, of the landscape as a function of a geographic sampled MDR-TB-related point) at an unobserved location from observations of its value at nearby locations (Fotheringham 2002). Thus, if the new MDR-TB data generated from a stochastic interpolation based algorithm, for example, differs in any way from the

linear dependent data that was initially fitted, large residual error coefficients will be introduced in the forecasts as the pattern of multicollinearity between the independent variables would be very different in the simulated MDR-TB data. Consequently, linear coefficients, based on collinear and multicollinear variables, can bias time series dependent MDR-TB explanatory clinical and/or environmental covariate coefficients yielding unstable, non-normal parameter estimators and unreliable autoregressive significance tests.

Further, the data regularization framework in such an interpolator may not recover well-behaved functional representations of the time series-dependent input MDR-TB data. Although the procedure would split the interpolation operator into a discrete deconvolution followed by a discrete convolution, misspecifications will still arise in the stochastic matrix within the probabilistic weighting scheme. As such, connections to radial basis functions will also be erroneous. Since the radial basis function is a real-valued function whose value depends only on the distance from the origin, so that $\phi(\mathbf{x}) = \phi(\|\mathbf{x}\|)$; or alternatively on the distance from some other sampled point *c*, so that $\phi(\mathbf{x}, \mathbf{c}) = \phi(\|\mathbf{x} - \mathbf{c}\|)$ (Cressie 1993), any function ϕ that satisfies the property $\phi(\mathbf{x}) = \phi(\|\mathbf{x}\|)$ in a spatiotemporal regression-based MDR-TB model is a radial function. Therefore, it would be difficult to posit a general framework for linking spatiotemporal MDR-TB statistical data analysis with approximation methods that are built on non-negative operators.

In mathematics, on a finite-dimensional inner product space, a self-adjoint operator is an operator that is its own adjoint, or, equivalently, one whose matrix is Hermitian(Cressie 1993). By the finite-dimensional spectral theorem, such operators can be only associated in a spatiotemporal MDR-TB model when employing an orthonormal basis of the underlying space in which the operator is represented as a diagonal matrix constructed from the covariate entries. In linear algebra and functional analysis, the spectral theorem is any of a number of results about linear operators or about matrices (Hazewinkle 2001). In broad terms the spectral theorem provides conditions under which an, operator or, a time series dependent MDR-TB matrix can be diagonalized. This concept of diagonalization would be relatively straightforward for operators on finite-dimensional spaces, however this would require some modification for operators on infinite-dimensional spaces. In general, the spectral theorem will identify a class of MDR-TB linear operators that can be modeled by multiplication operators. In more abstract language, the spectral theorem is a statement about commutative C*-algebra.

In linear algebra, an orthonormal basis for an inner product space *V* with finite dimension is a basis for *V* whose vectors are orthonormal (Griffith 2003). For example, the standard basis for a Euclidean space R^n is orthonormal in a robust spatiotemporal MDR-TB which would then represent a model where the relevant inner product would be the dot product of vectors. In mathematics, the dot product, or scalar product or sometimes inner product in the context of Euclidean space, is an algebraic operation that takes two equal-length sequences of numbers, usually coordinate vectors and returns a single number which then can be defined either algebraically or geometrically. The coordinate representation or coordinate vector of a vector is the unique tuple of numbers that describes the vector in terms of a particular ordered basis (Cressie 1993). Thus, the spatiotemporal-sampled clinical and environmental explanatory covariate coefficient coordinates would always be specified relative to an ordered basis. Bases and their associated coordinate representations would then enable realization of vector spaces and linear transformations concretely as column vectors, row vectors, and matrices, in the MDR-TB model . In three dimensional space the dot product would contrast with the cross product of two vectors, which then would produce a pseudovector as result in the model. A vector-like object which is invariant under inversion is called a pseudovector, or an axial vector (Hosmer and Lemeshew 2000). The cross product A × B is a pseudovector, whereas the vector triple product A× (B×C) is a polar vector. The term "polar vector" is used to refer to a representation of a vector magnitude (that is, length) and angle, which is equivalent to specifying endpoints (i.e., polar coordinates). In contrast, pseudovectors (i.e., axial vectors) do not reverse sign when the coordinate axes are reversed. Examples of polar vectors include the velocity vector, momentum, and force. The cross product of two polar vectors is a pseudo-vector (Cressie 1993). Polar vectors and pseudovectors are interrelated in the following ways under application of the cross product:

$$[\text{pseudovector}] \times [\text{pseudovector}] = [\text{pseudovector}]$$
$$[\text{vector}] \times [\text{pseudovector}] = [\text{vector}].$$

The dot product is directly related to the cosine of the angle between two vectors in Euclidean space of any number of dimensions (Cressie 1993).

Thus, the image of the standard basis under a rotation or reflection or any orthogonal transformation in the MDR-TB model then would also be orthonormal, and every orthonormal basis for R^n would thus arise in a similar fashion. The natural basis for a polar coordinate system is orthogonal (Cressie 1993). Since For a general inner product space *V*, an orthonormal basis can be used to define normalized orthogonal coordinates on *V*, the inner product in the MDR-TB model would then become a dot product of vectors. Thus, the presence of an orthonormality in the model would reduce the study of a finite-dimensional inner product space to the study of R^n under dot product. Further, since every finite-dimensional inner product space

space has an orthonormal basis (Griffith 2003), the MDR-TB distribution may be obtained from an arbitrary basis using the Gram–Schmidt process.

In mathematics, particularly linear algebra and numerical analysis, the Gram–Schmidt process is a method for orthonormalizing a set of vectors in an inner product space, most commonly the Euclidean space R^n. The Gram–Schmidt process takes a finite, linearly independent set $S = \{v_1, \ldots, v_k\}$ for $k \leq n$ and generates an orthogonal set $S' = \{u_1, \ldots, u_k\}$ that spans the same k-dimensional subspace of R^n as S (Cressie 1993). Unfortunately, certainty principles for orthonormal bases has not been spatially quantitated within a time series. As such, subspace sampling frame employing any form of relative-error matrix approximations for quantitating spatially dependent uncertainty has never been performed for an empirical dataset of MDR-TB explanatory clinical and environment covariate coefficients.

Additionally, violations of normality in a hierarchical linear spatiotemporal-sampled district-level MDR-TB model can also compromise the predictive estimation of clinical and/or environmental covariate coefficients and the calculation of confidence intervals. Generally, the error distribution in a time series-dependent MDR-TB infectious disease model is skewed by the presence of a few large outliers (Chatterjee and Hadi,1988). Scenes can calculate how symmetric the data is, in other words, if there a tendency for the data to be positive or negative (Fotheringham, 2002). Therefore, quantitating spatiotemporal MDR-TB regression-based covariates would simply require measuring the difference between the average and median of the sampled data (Smith, 1994; Johnston, 2003; Clarke et al., 2002; Akashi et al., 1996; Barr et al., 2000). The median measures the midpoint of the data, the value for which half the points are greater and half are smaller (Homser and Lemeshew, 2000).

Therefore, for a robust symmetrical MDR-TB distribution, like the normal, the median then would be the spatiotemporally tabulated averages and, as such, the quantitated skewness would be zero. Further, if the skewness is negative in the model then there would be more negative values indicating the presence of outliers. An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs and are often indicative either of measurement error, or that the population has a heavy-tailed distribution (Hosmer and Lemeshew, 2000; Manton and Stallard 1988). In probability theory, heavy-tailed distributions are distributions whose tails are not exponentially bounded: that is, they have heavier tails than the exponential distribution (Asmussen, 2003). If the skewness is positive in a spatiotemporal MDR-TB regression model then there are more positive values indicating the long tail generated by the explanatory covariate coefficients is on the positive side of the peak (i.e., "skewed to the right"). Spatiotemporal parameter estimation is based on the minimization of squared error; however, a few extreme observations can exert a disproportionate influence on sampled estimators (Griffith, 2003). For example, if the error distribution is significantly non-normal, in a time series-dependent MDR-TB regression-based model, the confidence intervals may be too wide or too narrow. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution (Hosmer and Lemeshew 2000). Further, since kurtosis is a measure of the extreme observations in a spatiotemporal model (Hosmer and Lemeshew, 2000), the sign of skewness would also indicate if the sampled explanatory covariate coefficients was kurtotic.

Kurtosis is a descriptive statistics based on a relative concentration of scores in the center, the upper and lower ends (that is, tails), and the shoulders of a distribution (Fotheringham et al., 2002). As such, higher kurtosis in a spatiotemporal-sampled MDR-TB regression-based model constructed from an empirical dataset of clinical explanatory covariates coefficients for example, would indicate more of the variance in the residuals generated from the model was due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations in the sampled covariate coefficients. Environmental-related data that has more kurtosis than the normal is sometimes called fat-tailed as its extremes extend beyond that of the normal (Piorecky and Prescott 2006, Wintle and Bardos 2006, Miller 2007, He et al. 2003,Hoeting et al. 2000). Ideally, a TB predictive risk modeler would prefer a distribution with low kurtosis (i.e., predictive residuals not far away from the mean). However, for spatiotemporal MDR-TB distribution to be normalized, the sampled explanatory covariate coefficients would have to exhibit an excess kurtosis equal to 0. Alternatively, a MDR-TB regression-based distribution with positive kurtosis in a spatiotemporal model would have to exhibit a peak in the middle and fat tails versus a normal distribution. Fat-tailed distributions have values of kurtosis that are greater than 3.0 (Fotheringham, 2002). Thus, the extreme values would be positive in a spatiotemporal MDR-TB regression–based model. However, this is only possible when the skewness in the model is positive. Further, the skewness is negative in the MDR-TB model combined with the impact of a high excess kurtosis would adversely affect causing extreme mispecified negative explanatory predictor covariate coefficient values in the residual error variance.

Frequently, adjusted version of Pearson's kurtosis has been used to quantitate the excess kurtosis and to provide a comparison of the shape of a given MDR-TB model distribution, to that of the normal distribution. Pearson (1905) introduced kurtosis as a measure of how flat the top of a symmetric distribution was when compared to a normal distribution of the same variance. He referred to more flat-topped distributions ($\gamma_2 < 0$) as "platykurtic," less flat-topped distributions ($\gamma_2 > 0$) as "leptokurtic," and equally flat-topped distributions as "mesokurtic" ($\gamma_2 \approx 0$).

Kurtosis is actually more influenced by scores in the tails of the distribution than scores in the center of a distribution (Hosmer and Lemeshew, 2000). Accordingly, it is often appropriate to describe a MDR-TB leptokurtic distribution as “fat in the tails” and a MDR-TB platykurtic distribution as “thin in the tails”. Distributions with negative or positive excess kurtosis are called leptokurtic distributions, respectively (Hosmer and Lemeshew, 2000). Leptokurtic distributions are identified by peaks that are thin and tall (Fotheringham 2002, 2000). Platykurtic curves, on the other hand, have shorter ‘tails’ than the normal curve of error and leptokurtic longer ‘tails’. Skewed distributions are always leptokurtic (Hopkins and Weeks, 1990). Pearson’s measure of kurtosis, however, has been often criticized as it does not focus adequately on the central part of a distribution. Although never proposed, an alternative measure of kurtosis for spatiotemporal MDR-TB regression-based modeling is one which adjusts the measurement of kurtosis by removing the effect of skewness using autocorrelation statistics.

Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional surface, introducing a deviation from the independent observations assumption of classical statistics (Griffith, 2003). Since spatially structured infectious disease data always violate the assumption of independence (Legendre 1993), residual serial autocorrelation oriented statistics would enhance predictive autoregressive MDR-TB risk mapping based on sampled georeferenced explanatory covariate coefficients. Identification of the presence of positive autocorrelation (i.e., aggregation of similar values in geographic space) in residual predictive error variance-covariance matrices always leads to underestimation of standard errors and inflated Type I errors, when employing standard methods based on ordinary least squares (OLSs) (e.g. ANOVA, correlation, and regression) to test statistical hypotheses (Cliff and Ord, 1981; Legendre, 1993). Lennon (2000) argued that autocorrelation renders inflated Type I errors, and had a systematic bias towards particular predictive estimators with greater autocorrelation. These autocorrelation-related components may be illustrated by a density graph which can reveal the leptokurtic nature of a time series-dependent MDR-TB distribution rendered from a robust spatial autocorrelation matrix while simultaneously revealing the associated thicker tails compared to a normal density using an autocovariate term.

Traditionally, an autocovariate analysis is indexed with a Moran Coefficient (MC; a product moment correlation coefficient type of spatial autocorrelation index) (Griffith, 2002). The simplest and most straightforward null hypothesis, on which to test the significance of the MC, would assume spatial autocorrelation in an empirical spatiotemporal-sampled dataset of the MDR-TB-related explanatory covariate coefficients, for example, from which a sample is drawn to be zero. Two assumptions about the sample can then be made: the covariate coefficient values are drawn from a normally distributed population; or, the sample values represent one random arrangement of the attribute values from all the possible arrangements that could occur. MC indices may be tested using analytical expectations and variances from a non-linear estimation model based largely on the neighborhood structure assumed in a spatially weighted uncertainty-oriented matrix. The sampling distributions of MC rendered from a spatiotemporal MDR-TB regression-based model may then be quantitated as a dataset of asymptotically normally distributed standard errors of the estimators which may then be valid for summarizing virtually any type of non-normal factor analysis or, for certain structural equation model construction.

Further, recent quantitative geographical analysis methods have supplemented spatial statistics with an approach to quantify latent autocorrelation error coefficients, by decomposing the MC into synthetic variates whose linear combinations constitute a spatial filter model specification. This eigenvector filtering approach is a non-parametric technique that removes the inherent autocorrelation from generalized linear regression models by treating them as a missing variables (i.e., first order) effect. The aim of this non-parametric approach is to control spatial autocorrelation by introducing appropriate synthetic variables that serve as surrogates for serially correlated missing origins and destination variables (Griffith, 2003). This shift in focus leads to spatial filter variants of the classical spatial interaction model. Further, by so doing, the non-parametric spatial filtering may control for autocorrelation and heteroskedastic error components in a time-series dependent MDR-TB model with a set of spatial proxy predictor variables, rather than identify a global error autocorrelation parameter for a spatial process in the model. In time-series infectious disease models, heteroscedasticity (that is, uncommon variance) often arises due to the effects of inflation perhaps magnified by a multiplicative seasonal pattern (Griffith, 2005).The basis for this procedure is the decomposition the MC into orthogonal and uncorrelated map pattern components. As such, a MDR-TB-oriented spatial filter analyses can be used to account for an empirical dataset of regressed pseudo-replicated explanatory covariate coefficients by generating eigenvectors which may exhibit a distinct spatial topographic pattern while simultaneously rendering a given residual autocorrelation level.

The goal of this study was to identify geographic areas with on-going MDR-TB transmission in San Juan de Lurigancho (SJL), a district in Lima, Peru by performing a residual spatial autocorrelation analysis within an SAS database to derive simulation models. Our assumption was that the residuals from these models would reveal how departures from normality affect the performance of exact confidence intervals for a population mean and variance within a time series- dependent empirical

dataset of spatiotemporally-sampled clinical end environmental MDR-TB explanatory covariates. SAS PROC REG can calculate univariate statistics, and perform robust parsimonious linear and non-linear regression analyses using spatiotemporal-sampled data (www.sas.edu). In this research estimates generated from a global autocorrelation analyses were spatially decomposed into empirical orthogonal bases using a negative binomial regression with a non-homogeneous, gamma distributed mean. Thereafter, we proposed a test of goodness of fit for the time-series dependent models based on the sum of the squared residual partial autocorrelations. The test statistic was asymptotically X^2. The residual times-series autocorrelation estimation performance was thereafter studied through a Monte Carlo experiment. Monte Carlo experiments are a broad class of computational algorithms used in optimization and numerical intergration for generation of samples from a probability distribution (Cressie 1993). Another of our assumption in this research, was that the decomposition of Moran's coefficient into uncorrelated, MDR-TB orthogonal mapping components could reveal global spatial heterogeneities necessary to capture latent autocorrelation in a spatiotemporal regression-based model for implementing control strategies in the SJL study site.

Geographically based screening and treatment could be an effective method for MDR-TB control programs to identify high-risk populations (WHO, 2009).

In this research, we also complemented the autocovariate logistic parameter estimation model using a Bayesian Poisson matrix in SAS for formally hypothesis-testing the spatiotemporal-sampled MDR-TB drug resistant parameter estimators at the SJL study site. SAS/STAT software now provides Bayesian analysis including Bayesian zero-inflated Poisson models for zero-inflated count data employing a Markov Chain Monte Carlo (MCMC) algorithm in downloadable, experimental versions of three procedures for SAS 9.1.3 on Windows: GENMOD, LIFEREG, and PHREG (www.sas.edu). Markov Chain is a mathematical system that undergoes transitions from one state to another, between a finite or countable number of possible states which can be characterized as random "memoryless" process (Cressie 1993). In recent years MCMC has revolutionized the practicability of Bayesian inference methods allowing a wide range of posterior distributions to be simulated and their parameters to be quantitated numerically in time series-dependent infectious disease modeling. In Bayesian statistics, the posterior probability of a random event or an uncertain proposition in a time series dependent MDR-TB model would be the conditional probability that is assigned after the relevant evidence is taken into account (Cressie 1993). Similarly, the posterior probability distribution would be the MDR-TB distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey.

In this research specifically we used a Bayesian Poisson model to estimate the risks of resistance to four commonly used drugs at the SJL study site in TB treatment: isoniazid, rifampin, ethambutol, and streptomycin. A Bayesian Poisson vector autoregression model can characterize endogenous infectious disease dynamic count data with no restrictions on the contemporaneous correlations (Griffith, 2005). Bayesian linear regression techniques can also be used when the variance is assumed to be a function of the mean (Cressie, 1993). Therefore, it would be possible in some cases to amend the problem of propagated residual uncertainty in a time series dependent MDR-TB model by applying a transformation to the response variable (e.g., fitting the logarithm of the response variable using a linear regression model) which would then imply that the response variable has a log-normal distribution rather than a normal distribution. In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed (Hosmer and Lemeshew 2000).

Thus, if X is a random variable in a spatiotemporal Bayesian generalized hierarchical MDR-TB spatiotemporal model with a normal distribution, then $Y = \exp(X)$ will have a log-normal distribution; likewise, if Y is log-normally distributed in the model then $X = \log(Y)$ has a normal distribution. The log-normal distribution would then be the spatiotemporal MDR-TB model distribution of the sampled random variables with only positive real values.

Further, in this research, the decomposition of the residual forecast errors were illustrated in the Bayesian Poisson model residuals for quantitating the effects of exogenous covariate shocks. We then spatially decomposed uncertainty values to quantify the effects of exogenous-sampled explanatory covariate coefficients related to special resistant strains. Since drug resistance is very common in tuberculosis treatment (Orenstein et al., 2009), we assumed that robust Bayesian Poisson model outputs could quantitate interactions between the clinical and environmental sampled parameter estimators (i.e., resistant strain data) and time series-dependent MDR-TB district-level indices at the SJL study site. It is well known that drug resistance of TB is unevenly distributed and, therefore, MDR resistance can be perceived as problems of local rather than global importance (Dye et al., 2002). Although there have been a few studies on the mechanism of drug resistance in tuberculosis (Al-Orainey, 1989; Crofton et al., 1997), the reasons why tuberculosis is resistant to a certain treatment is largely unknown. Therefore, correctly estimating the drug resistance at a local level may have important implications for control and treatment of MDR-TB. As such, in this paper, we used a flexible Bayesian Poisson regression model to estimate the risk of drug resistance at the SJL study site. Further, since independent marginal distributions are necessary for non-

normal probability analyses in a predictive autoregressive risk model framework (Griffith 2003), we assumed that synthetic spatiotemporal MDR-TB map patterns based on specific disease transmission data (e.g., distribution of resistant strains) would produce robust pseudo-likelihood estimates with high predictive power.

In this paper we also considered both low and high-dimensional predictive residual uncertainty covariance matrix estimation problems and present asymptotic properties of sample MDR-TB -related covariances and covariance matrix estimates. In particular, we provide spatially quantitated asymptotic uncertainties for high dimensional covariance matrices in the time series, and a consistency result for the MDR-TB-related error covariance matrix estimation for regressing the spatiotemporal dataset of clinical and environmental explanatory covariate coefficients. The problem of high -dimensional covariance matrix latent error estimation often arises when estimate unknown parameters that are associated with a time series (Cressie 1993).

Additionally, we generated a residual Moran's coefficient (MC) minimization criterion for permitting a more detailed interpretation of latent autocorrelation in the MDR-TB data sampled at the SJL study site by allowing explicit visualization of inconspicuous negative spatial autocorrelation (NSA) patterns in the georef-erenced clinical and environmental parameter datasets. Negative spatial autocorrelation naturally materializes with competitive locational processes, negative spatial externalities, the spectrum (e.g., eigenvalues) of a geographic weights matrix, the calculation of linear regression residuals, and the computation of local indicator of spatial autocorrelation (LISA) statistics (Griffith, 2008; Anselin, 1995). To date spatial analyses of infectious disease data commonly have employed only a first conditional autoregressive model or, a second-order, that is, a simultaneous autoregressive with spatial lag covariance matrix for determining hidden NSA attributable to model misspecifications. Although these models have performed extremely well across a myriad of georeferenced attributes, higher order spatial covariance matrix specifications may be needed to capture NSA in an autoregressive spatiotemporal predictive risk MDR-TB model. Failure to posit the correct order of a spatial covariance matrix can constitute a prominent form of model error (Griffith, 2003). Thus, we assumed that qualitatively assessing residual time series dependent autocorrelation error coefficients may improve present MDR-TB control strategies at the SJL study site by revealing how hidden NSA furnishes a diagnostic in a predictive autoregressive risk model misspecification.

Since the prediction error is the expected quadratic loss incurred by the difference of observed event status and by the model predicted event probabilities (Cressie 1993), it may be shown that the prediction error is minimal in a MDR-TB if, and only if, the true probabilities are accurately spatially quantitated.

In this paper we considered both low and high-dimensional predictive residual uncertainty covariance matrix estimation problems and present asymptotic properties of sample MDR-TB-related covariances and covariance matrix estimates in GIS using QuickBird data. Raster representations of thematic and numerical spatial attributes of MDR-TB can be spatially quantitated in a GIS environment for computational simulation and analysis of spatial processes (Jacob et al. 2010). This paper addresses the problem of MDR-TB-related predictions and their uncertainty assessment for creating GIS raster representations created from a set of sample points of spatial attributes. Spatial mapping in GIS using sub-meter resolution remote sensing data [e.g., QuickBird visible and near infra-red (NIR) 0.61 m pixels] may be an alternative tool in MDR-TB control, in the SJL study site for aiding in the assessment of transmission dynamics for optimizing existing management programs. Accounting for the autocorrelation between neighboring districts, thereafter, and studying whether other spatiotemporal-sampled georeferenced district-level clinical and environmental covariate coefficients are related to drug resistance also may also develop and implement robust MDR-TB control strategies in the SJL study site. Therefore, our research objectives were: (1) to perform Poisson regression analyses to determine explanatory covariates affecting MDR-TB incidence rates; (2) to construct a flexible Bayesian regression model to esti-mate the risks of district-level resistance to four common drugs: rifampin, isoniazid, ethambutol and streptomycin; (3) to generate global autocorrelation statistics for evaluating spatial dependence and kurtosis among the data feature attributes while quantifying all residual error autocorrelation components in the model output; (4) to generate a Bayesian Poisson model for evaluating distri-bution of district-level resistant strains for identifying epicenters for MDR-TB and, (5) to use a Residual MC minimization criterion for detecting and quantitating non-conspicuous NSA in a dataset of clinical explanatory predictor variables spatiotemporally-sampled in SJL, Lima, Peru.

MATERIALS AND METHODS

Study Site

San Juan de Lurigancho (SJL) is the largest district in Lima, located in the Northeast area of the province of Lima. With a current population exceeding one million people, it is the country's most populous district, with a total surface area of 131.3 km^2 constituting 4.91% of the total area of the province of Lima. On the north, SJL is bordered by the districts of Carabayllo and San Antonio, which is in the Huarochirí Province. San Juan de Lurigancho is bordered by Comas, Independencia and Rímac on the west; and Lurigancho on the east. The Rímac River marks the district's border with downtown Lima and El Augustino on the south. The most important urban areas in the district are Mangomarca, Zárate, Las Flores, and Canto Grande and Bayovar. One of the first urban areas in SJL is Caja de Agua; which is located at the entrance of the district. Caja

de Agua is surrounded by San Cristobal and Santa Rosa hills from south to west. The altitude of SJL ranges from 2,240 meters above sea level (m.a.s.l.) at the peaks of Cerro Colorado Norte, to 200 m.a.s.l., at the level of the Rimac river. Urban areas have been developed in a longitudinal direction from the river border up to 350 m.a.s.l. The mean temperature ranges between 17 and 19ºC throughout the year.

Subjects and setting

This research used the data acquired from a retrospective study of a cohort of 1,571 patients diagnosed with pulmonary TB and MDR-TB enrolled over an 18 month period in the district of SJL in Lima, Peru.

Patient selection and enrollment

In this research all participating patients underwent a complete evaluation, including drug susceptibility for first line drugs. This was a prospective multi-center observational study comparing the use of several investigational techniques with standard methods to assess the *in vitro* antimicrobial susceptibility of *M. tuberculosis,* either directly from patient specimens or from culture isolates. One thousand two hundred and fifty adults with pulmonary tuberculosis cultures were confirmed with $\geq$10 colonies of *M. tuberculosis*. 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 tuberculosis, to diagnostic and treatment sites in the following Health Centers: 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 sputum smear microscopy results, subjects were screened for the presence of productive cough for eligibility in the study. Patients with positive sputum smear**s** are those with the capacity to spread infection (Godoy et al., 2004). Eligible subjects received an explanation of the study and were asked to provide written informed consent to participate. Initial data collected during screening included: a past medical history, collection of basic socio-demographic descriptors (age, sex, occupation, address, etc.) and a detailed symptom-oriented history with physical examination.

Drug susceptibility testing (isoniazid, rifampin, ethambutol and streptomycin) were performed by Gold Standard method on the initial sputum culture isolates of all enrolled subjects. Those subjects with initial drug resistant *M. tuberculosis* clinical isolates were determined using a treatment regimen with a duration deemed appropriate by a Committee of the National Tuberculosis Control Program (NTCP) and Committee for Evaluation of Retreatment (CER). All information collected was recorded on standardized 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 then developed to record baseline clinical and socio-demographic data, HIV testing results, mycobacterial smear and culture results.

Geographic mapping

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

Data from the characterization of each epidemiological village was then recorded on a Mobile Vector Control Management System (VCMS™) electronic data recording device. The field sampling was extended to a 5 km distance from the external boundary of a sampled MDR-TB-related site. Specific environmental explanatory variables of the georeferenced data were recorded. Individual georeferenced Health Centers and their associated land cover attributes identified from the satellite imagery were then entered into a VCMS relational database software product. The VCMS database supported a mobile field data acquisition component module (Mobile VCMS) utilizing an industry standard Microsoft Windows Mobile™ device and an add-on DGPS connection. In this research, Mobile VCMS™ and its FieldBridge Server middleware component were used to support wireless synchronization of the clinical and environmental MDR-TB data collected at the SJL study site directly into a centralized database repository. Additional geocoding and spatial display of the clinical and environmental sampled data was handled in the embedded VCMS GIS Interface Kit™. This was developed using ESRI's MapObjects™ 2 technology. The VCMS database with the DGPS information, supported exporting all data in a spatial format; whereby, any individual Heath Center data and supporting MDR-TB covariates were described in an ESRI shapefile format for use in GIS. The database displayed this information on a user-defined field base map.

Remote sensing data

QuickBird (www.digitalglobe.com) images were acquired in March 11th 2008 for the SJL study site. QuickBird multispectral products provided four discrete non-overlapping spectral bands covering a range from 0.45 to 0.72 µm, with an 11-bit collected information depth with a spatial resolution of 0.61 m (Figure 1)

The QuickBird imagery was then classified using the Iterative Self-Organizing Data Analysis Technique (ISODATA) unsupervised routine in ERDAS *Imagine* v.8.7™. The images were co-registered manually, using gathered ground control point (GCPs) and georectified images from the QuickBird data. The satellite images were co-registered by applying a first order polynomial algorithm with a nearest neighbor resampling method and the GCPs.. The Universal Transverse Mercator (UTM) Zone 37S datum WGS-84 projection was used for all of the spatial datasets.

Environmental parameters

Variables recorded included, MDR-TB prevalence rates, distance between individual Health Centers, population data, and aspects of catchment-related ecohydrological land-surface covariates in the SJL study site such as elevation and slope per sampled site. Distance measures were recorded in ArcGIS 9.2® with QuickBird data and by field sampling. The distance between Health centers was categorized into numerous classes (e.g., 1: 0 to 5 km; 2: 5 to 10 km, and so on). The number of individuals cases of MDR-TB at each individual Health Center was then calculated and recorded (Table 1).

Regression analyses

All sampled parameters were entered in Excel files and analyzed using SAS 9.1.3® (SAS Inc. Cary, North Carolina). The first stage of

Figure 1. QuickBird visible and near infra-red data of the San Juan de Lurigancho study site.

Table 1. Clinical and environmental MDR-TB data sampled in the San Juan de Lurigancho study site.

Variable in database	Description of variable
ESTAB	Health care center
FENAC	Birth date
EdadA	Age
SEXO	Sex
TIPOVIV	Home
NUMHAB	Number of bedrooms
MATVIV	Building material
NUMPER	Number of persons living in the house
ELECTRIC	Electricity supply at home
AGUAPOT	Home access to potable water
DESAGUE	Wastepipe connected to the public network
ECIVIL	Marital status
OCUPA	Ocupation
TRAESTS	Do you work in any health care center?
TIEMTRA	Time of employment
INGMEN	Salary/Income per month
LJINH	Sensitivity test to isoniazid in LJ medium
LJIRIF	Sensitivity test to rifampin in LJ medium
LJIETB	Sensitivity test to ethambutol in LJ medium
LJISTM	Sensitivity test to streptomycin in LJ medium
MDR	Multidrug resistant

this analysis utilized Poisson regression to determine the relationship between the MDR-TB sampled clinical and environmental covariates. Poisson regression is one special case of the Generalized Linear Model (GLM) which allows one to fit models to a dependent variable that is a member of the exponential distribution family for linear quantitation of covariate variabilities. (Pielou, 1969). Our MDR-TB GLM was characterized with three components: the distribution of the dependent variable, a linear function of a set of

independent variables, and a link function between the dependent variable and its expectation as expressed by the linear function of independent variables. When the logarithm was applied as a link function, the Poisson regression had a log-linear form. Poisson regression is estimated based on the likelihood function that is constructed under the independence assumption (Haight 1967). Poisson distribution predicts non-negative integers in data analyses, where the mean and variance are equal (Kaiser and Cressie, 1997).

Next, non-linearity in the relationship between MDR-TB resistant infection rates and their explanatory predictor variables, were explored by adding polynomial terms and then grouping the values of continuous variables into categorical ones. Variable selection for the multiple regression models was carried out by a combination of automatic (stepwise) procedures and goodness-of-fit criteria and by selecting the covariates that explained the prevalence of MDR-TB cases and distribution in the SJL study site. A Poisson regression with statistical significance, determined by a 95% confidence level was then constructed to ascertain whether the proportions of sampled explanatory predictor variables differed by individual MDR-TB Health Centers.

The Poisson regression assumed that each independent count value (that is. n_i), recorded at a Health Center location i=1,2,…n, from a sampled covariate was from a Poisson distribution. These data were described by a set of predictor variables denoted by matrix $\mathbf{X}_i$, a 1×p vector of covariate values for a Health Center location i. The expected value of these data was given by $\mu_i(\mathbf{X}_i)=n_i(\mathbf{X}_i)\exp(\mathbf{X}_i\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ was the vector of non-redundant parameters, and the Poisson rates parameter was given by $\lambda_i(\mathbf{X}_i)=\mu_i(\mathbf{X}_i)/n_i(\mathbf{X}_i)$; the rates parameter $\lambda_i(\mathbf{X}_i)$ was both the mean and the variance of the Poisson distribution as in McCullagh and Nelder, (1989) for sampled Health Center location i. The regression analyses were performed in SAS PROCREG. The sampled data was log-transformed before analyses to normalize the distribution and minimize standard error.

Thereafter, we used a Bayesian Poisson model to estimate the risks of resistance to each of the four common drugs in TB treatment. We then fit a Bayesian Poisson regression model for the frequency of the strains with density using MDR-TBi $\text{Poisson}(\lambda_i)\log(\lambda_i)=\mathbf{X}_i\boldsymbol{\beta}$ (2.1) for the $i=1,\ldots,18$ plates, where $\boldsymbol{\beta}$ represented the regression parameters and $\mathbf{X}_i$ was the vector of covariates The likelihood function for each of the corresponding MDR-TB sampled explanatory covariates was $p(\text{SMDR}/\mathbf{X}_i\boldsymbol{\beta}=\text{Poisson}(\lambda_i)$ where $p(\cdot|\cdot)$denoted a conditional probability mass function. The Poisson density was then evaluated with a corresponding mean parameter λ_i. The three parameters, β_1, β_2, and β_3, corresponded to an intercept, the positive and the negative effect of the strain respectively. The following prior distributions were then placed on the spatiotemporal-sampled MDR-TB parameter estimators, where $\pi(\cdot)$ indicated a prior distribution: $\pi(\beta_1),\pi(\beta_2),\pi(\beta_3)=\text{normal}(0,\sigma^2=1000)$. The diffuse normal prior expressed lack of knowledge about the regression parameters.

Using Bayes' theorem, the likelihood function and prior distributions determined the posterior distribution of β_1,β_2, and β_3. The goodness-of-fit Pearson chi-square statistic χ^2_P was then derived as in McCullagh and Nelder (1989).By so doing we were able to assess model fit which in this research was achieved employing MDR-TBi-E(MDRi)]^2:/ V (MDR-TBi). We let $E(\cdot)$ represent an expectation for a Poisson likelihood E(MDR-TBi= V(MDR-TBi) where λ_i was defined in Equation 2.1 If there is no overdispersion, the Pearson statistic approximately equals the number of observations in the data set minus the number of parameters in the model.(Fotheringjam 2002)

The parameter μ was interpreted as rates (e.g., the average number of new TB cases per 1,000 population). If Y is the number of occurrences, its probability distribution can be written as:

$$f(y)=\frac{\mu^y e^{-\mu}}{y!},$$

for $y=0,1,2,\ldots$

where μ was the mean number of occurrences (Kaiser and Cressie, 1997). We then used y_i to denote the number of MDR-TB patients who were resistant to a specific drug in a georeferenced health center i. We let Ni denote the population size of health center i. We assumed a Poisson model for the spatial count data as follows:

$$\sim(a\,bi)$$

$$i\,i\,y\,Poisson\,N\,e+,i=1,2,\ldots,M.$$

where

$i\,b$

was the spatial random effect for the ith georeferenced health center, controlling whether the risk is above or below the average.

We further modeled the spatial random effect b_i using a conditionally autoregressive (CAR) prior (see Hodges et al., 2003). In this research under the CAR prior,

$$b_i \sim N\left(\frac{\sum_{j\in\partial_i} b_j}{m_j},\frac{\sigma^2}{m_j}\right), \quad i=1,2,\ldots,M.$$

where ∂_i was the index set of neighboring districts of the i^{th} district, m_j was the number of neighboring districts to the district i, and σ^2 was the unknown variance parameter. We used noninformative priors for other assessing additional parameter estimators which were represented as a flat prior for N and a conjugate inverse gamma prior for σ^2. In Bayesian probability theory, the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$. Thereafter, the prior and posterior were the conjugate distributions, and the prior was a conjugate prior for the likelihood in the MDR-TB model.

The Gaussian family is conjugate to itself (that is, *self-conjugate*) with respect to a Gaussian likelihood function in a spatiotemporal model if the likelihood function is Gaussian, (Fotheringham 2002). In this research, choosing a Gaussian prior over the mean ensured that the posterior distribution was also Gaussian. Further, the Gaussian distribution was a conjugate prior for the likelihood which was also Gaussian in the model. Conjugate priors are analogous to eigenfunctions in operator theory, in that they are distributions on which the "conditioning operator" acts in a well-understood way, thinking of the process of changing from the prior to the posterior as an operator.

Spatial analyses of MDR-TB covariates using Moran's *I*

Spatial autocorrelation was evaluated among the sampled clinical and environmental covariates at the SJL study site using Moran's *I*. In statistics, Moran's *I* is a measure of spatial autocorrelation (Griffith, 2003) In this research Moran's *I* was defined as

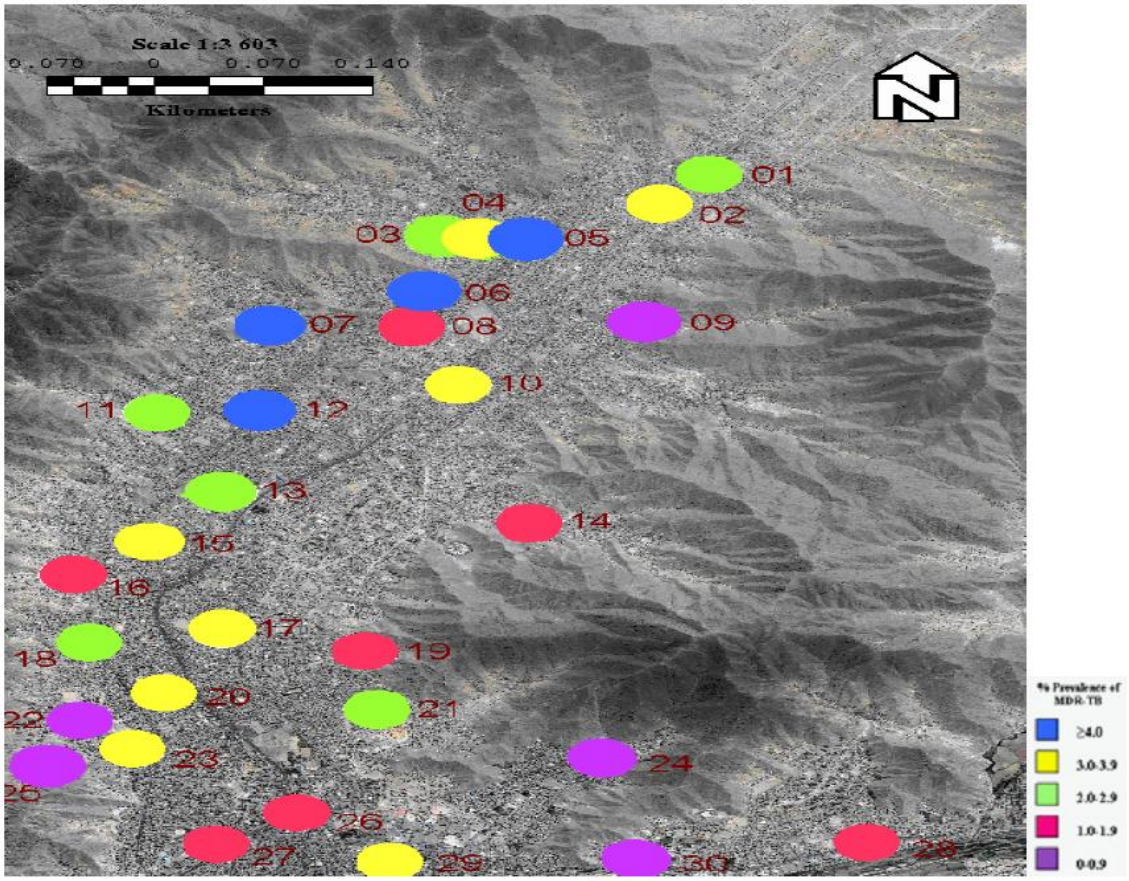

Figure 2. Geographical clusters of Health Centers in San Juan de Lurigancho study site.

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

where N was the number of georeferenced health centers indexed by i and j; X was the MDR-TB incidence rates; $\bar{X}$ was the mean of X; and w_{ij} was an element of a matrix of spatial weights. The expected value of Moran's *I* under the null hypothesis of no spatial autocorrelation was then:

$$E(I) = \frac{-1}{N-1}$$. Its variance thereafter was equal to:

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

Where

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

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

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

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

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

For statistical hypothesis testing, the Moran's *I* values were then transformed to *Z*-scores where values greater than 1.96 or smaller than −1.96 indicated spatial autocorrelation that was significant at the 5% level.

We also used the Geary's coefficient (that is, Geary's C) which is inversely related to Moran's *I*,. Moran's *I* is a measure of global spatial autocorrelation, while Geary's C is more sensitive to local spatial autocorrelation (Griffith, 2003). In this research Geary's C was defined as

$$C = \frac{(N-1) \sum_i \sum_j w_{ij}(X_i - X_j)^2}{2W \sum_i (X_i - \bar{X})^2}$$

where N was the number of health centers indexed by i and j; X where the MDR-TB incidence rates ; $\bar{X}$ was the mean of X; w_{ij} was a matrix of spatial weights; and W was the sum of all w_{ij}. The value of Geary's C lies between 0 and 2. Geary's C is inversely related to Moran's *I*, but it is not identical (Cliff and Ord 1971). Moran's *I* is a measure of global spatial autocorrelation, while Geary's C is more sensitive to local spatial autocorrelation (Griffith, 2003). Neighboring georeferenced health centers were then identified based on MDR-TB resistant prevalence values (Figure 2)

We analyzed the *n*-by-1 vector $x = [x_1 \cdots x_n]^T$ containing the MDR-TB covariates for *n* spatial units and n-by-n symmetric spatial weighting matrix W using Moran's Indices. The usual formulation for Moran's index of spatial autocorrelation (Griffith, 2003) is

$$I(x) = \frac{n\sum_{(2)} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{(2)} w_{ij}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

The values w_{ij} where the spatial weights based on the sampled clinical and environmental MDR-TB variables stored in the matrix **W** where $\sum_{(2)} = \sum_{i=1}^{n}\sum_{j=1}^{n}$ with $i \neq j$ which had a null diagonal $(w_{ii} = 0)$. This symmetric matrix revealed W*ij* = W*ji* was then generalized to a non-symmetric matrix W by using W=(W*+W*T/2) .

Moran's *I* was then rewritten using matrix notation as:

$$I(x) = \frac{n}{1^t W 1}\frac{x^T HHWHHx}{x^T HHx} = \frac{n}{1^T W 1}\frac{x^T HWHx}{x^T Hx} \quad (2)$$

SAS/GIS® (http://www.sas.com/products/gis/) was then used to perform the spatial filter analysis on the sampled MDR-TB data while SAS PROC GENMOD was used to build Poisson models with a gamma-distributed mean. In the study site, positive spatial autocorrelation (PSA) and NSA eigenvectors were selected by the stepwise negative binomial regression procedure. To expand the inferential basis with a random effect, a GLMM was used to account for latent non-spatial residual correlation time series dependent MDR-TB data. The GLMM estimation was computed using SAS PROC NLMIXED.

Spatial eigenvector mapping

Global indicators of spatial autocorrelation were then calculated from the ground-based and remotely-sensed ecological databases. Box-Cox type of power transformation was employed for normal approximation analysis purposes so that the frequency distributions of the georeferenced Health Centers in the SJL study site better approximated a bell-shaped curve. The spatial filter construction methodology transformation procedure was then used, as proposed by Griffith (2003), which depended on the eigenfunctions of a spatially weighted matrix.

To identify spatial clusters that can be uncovered with spatial filtering, Thiessen polygon surface partitionings were generated to construct geographic neighbor matrices, each denoted by the spatially weighted matrix which also was used in the spatial autocorrelation analysis. Entries in matrix were 1, if two health centers shared a common Thiessen polygon boundary and 0 otherwise. Next, the linkage structure for each surface was edited to remove unlikely geographic neighbors to identify pairs of health centers sharing a common Thiessen polygon boundary (Liang and Zeger, 1986; Griffith and Peres-Neto, 2006; Pielou, 1969; McCullagh and Nelder, 1989; Fotheringham, 1993; Wintle and, Bardos 2006). Eigenvectors of a modified version of the spatially weighted matrix was then used to furnish synthetic variates to determine distinct MDR-TB map patterns representing the full range of autocorrelation possibilities. Attention was restricted to those map patterns associated with at least a minimum level of spatial autocorrelation, which, for implementation purposes, was defined by $|MC_j/MC_{max}| > 0.25$, where MC_j denoted the *j*th value and MC_{max}, the maximum value of MC. This threshold value allowed two candidate sets of eigenvectors to be considered for substantial positive and substantial negative spatial autocorrelation respectively.

Extending the findings of de Jong et al. (1984) and Tiefelsdorf and Boots (1995) we established a set of MC values that was related to matrix (**I** _ **11**T/*n*)**C**(**I** _ **11**T/*n*), where **C** was a 0/1 binary geographic connectivity weights matrix, **I** was an *n*-by-*n* identity matrix, **1** was an *n*-by-1 vector of ones, T was the matrix transpose, and, vector **Y** was the pre-multiplied georeferenced data matrix (**I** _ **11**T/*n*). In practice, these MC values are related to the binary geographic connectivity matrix **C** itself, after the principal eigenvalue has been replaced with 0 (Griffith and Amrhein, 1997). The decomposition discussed by Tiefelsdorf and Boots furnished a basis for the eigenfunction decomposition approach outlined here. In this research the decomposition expressed a given MI value as a weighted sum of the eigenvalues of matrix (**I** _ **11**T/*n*)**C**(**I** _ **11**T/*n*). Additionally, our model revealed that the upper and lower bounds for the spatial matrix generated using MC was rendered by $\lambda_{\max}(n/1^T W 1)$ and $\lambda_{\min}(n/1^T W 1)$ where $\lambda_{\max}$ and $\lambda_{\min}$ which were the extreme eigenvalues of $\Omega = HWH$. Hence, in this research, the eigenvectors of Ω were vectors with unit norm maximizing MC. The eigenvalues of this matrix were equal to MC of spatial autocorrelation post-multiplied by a constant. Eigenvectors associated with high positive (or negative) eigenvalues have high positive (or negative) autocorrelation (Griffith, 2003).

The diagonalization of the spatial weighted matrix generated from the clinical and environmental-sampled MDR-TB explanatory covariates coefficients consisted of finding the normalized vectors , stored as columns in the matrix $U = [u_1 \cdots u_n]$, which satisfied: $\Omega = HWH = U\Lambda U^T = \sum_{i=1}^{n}\lambda_i u_i u_i^T$ where $\Lambda = diag(\lambda_1 \cdots \lambda_n)$, $u_i^T u_i = \|u_i\|^2 = 1$ and $u_i^T u_j = 0$ for $i \neq j$ (Griffith, 2003). The double centering of Ω implied that the eigenvectors u_i generated from the sampled MDR-TB covariates were centered and that at least one eigenvalue was equal to zero. Introducing these eigenvectors in the original formulation of MC led to:

$$I(x) = \frac{n}{1^T W 1}\frac{x^T HWHx}{x^T Hx} = \frac{n}{1^T W 1}\frac{x^T U\Lambda U^T x}{x^T Hx} = \frac{n}{1^T W 1}\frac{\sum_{i=1}^{n}\lambda_i x^T u_i u_i^T x}{x^T Hx} \quad (3)$$

Considering the centered vector $z = Hx$ and using the properties of idempotence of H, Equation (2.3) was equivalent to:

$$I(x)=\frac{n}{1^TW1}\frac{\sum_{i=1}^{n}\lambda_i z^T u_i u_i^T z}{z^T z}=\frac{n}{1^TW1}\frac{\sum_{i=1}^{n}\lambda_i \left\|u_i^T z\right\|^2}{\left\|z\right\|^2} \quad (4)$$

As the eigenvectors u_i generated from the eigendecomposition of the spatially weighted matrix and the vector z were centered, Equation (2.4) was then rewritten as:

$$I(x)=\frac{n}{1^TW1}\frac{\sum_{i=1}^{n}\lambda_i cor^2(u_i,z)\mathrm{var}(z)n}{\mathrm{var}(z)n}=\frac{n}{1^TW1}\sum_{i=1}^{n}\lambda_i cor^2(u_i,z) \quad (5)$$

In this research r was the number of null eigenvalues of $\Omega(r\geq 1)$. These eigenvalues and corresponding eigenvectors were removed from Λ and U respectively. Equation (2.5) was then equivalent to:

$$I(x)=\frac{n}{1^TW1}\sum_{i=1}^{n-r}\lambda_i cor^2(u_i,z) \quad (6)$$

Moreover it was demonstrated that MC for a given eigenvector u_i generated from the clinical and environmental sampled MDR-TB covariates was equal to $I(u_i)=(n/1^TW1)\lambda_i$ so the equation was rewritten:

$$I(x)=\sum_{i=1}^{n-r}I(u_i)cor^2(u_i,z) \quad (7)$$

The term $cor^2(u_i,z)$ represented the part of the variance of z that was explained by u_i in the spatiotemporal MDR-TB model using $z=\beta_i u_i+\varepsilon_i$. Estimation of covariance matrices is needed in the construction of confidence regions for unknown parameters, hypothesis testing, principal component analysis, prediction, discriminant analysis among others (Cressie 1993).This quantity was equal to $\beta_i^2/n\,\mathrm{var}(z)$. By definition the eigenvectors u_i were orthogonal and therefore regression coefficients of the MDR-TB model was verified employing $z=\beta_i u_i+\varepsilon_i$ were those of the multiple regression model were quantified by $z=U\beta+\varepsilon=\beta_i u_i+\cdots+\beta_{n-r}u_{n-r}+\varepsilon$.

The distribution of the error residuals in the autocovariance matrix of the spatiotemporal MDR-TB was then quantified**.** The maximum value of I was obtained by all of the variation of z as explained by the eigenvector u_1 which corresponded to the highest eigenvalue λ_1 in the autocorrelation error matrix. In this research, $cor^2(u_i,z)=1$ (and $cor^2(u_i,z)=0$ for $i\neq 1$) and the maximum value of I, was deduced for Equation (2.7), which was equal to $I_{\max}=\lambda_1(n/1^TW1)$. The minimum value of I in the error matrix was obtained as all the variation of, z was explained by the eigenvector u_{n-r} corresponding to the lowest eigenvalue λ_{n-r} rendered from the MDR-TB model. This minimum value was equal to $I_{\min}=\lambda_{n-r}(n/1^TW1)$. If the clinical and environmental sampled predictor variable was not spatialized, the part of the variance explained by each eigenvector was equal, on average, to $cor^2(u_i,z)=1/n-1$. Because the clinical and environmental-sampled MDR-TB explanatory covariates in z were randomly permuted, it was assumed that we would obtain this result. In this research the set of $n!$ random permutations, revealed that

$$\mathrm{E}_R(I)=\frac{n}{1^TW1(n-1)}\sum_{i=1}^{n}\lambda_i=\frac{n}{1^TW1(n-1)}trace(\Omega)$$

It was easily demonstrated that $trace(\Omega)=-\frac{1^TW1}{n}$ and it followed that $\mathrm{E}_R(I)=-\frac{1}{n-1}$.

We also used a Residual MC Minimization criterion suggested by van Tiefelsdorf and Griffith (2007) to further decompose the MC generated from the spatial decomposition of the sampled MDR-TB predictor variables to detect hidden NSA in the clinical and environmental data. The MC expected value for residuals from a linear spatial filter analyses was constructed with the eigenvectors from the MDR-TB data analyses using:

$$-\frac{n}{1^TC1}\frac{TR\left[\left(X^TX\right)^{-1}X^TCX\right]}{n-P-1}=\frac{n}{1^TC1}\frac{TR\left\{\left[(1E_P)^T(1E_P)\right]^{-1}(1E_P)^TC(1E_P)^T\right\}}{n-P-1}$$

$$=\frac{n}{1^TC1}\frac{1^TC1/n+\sum_{j=1}^{P}\lambda_j}{n-P-1}$$

$$=\frac{n}{1^TC1}\frac{1^TC1/n+\sum_{j=1}^{H}\lambda_j}{n-P-1}-\frac{n}{1^TC1}\frac{\sum_{j=H+1}^{K}\lambda_j}{n-P-1}+\frac{n}{1^TC1}\frac{\sum_{j=K+1}^{P}\lambda_j}{n-P-1} \quad (2.8)$$

where X=1E*p* was a covariate matrix, 1 was an n-by-1 vector of ones, **C** was the binary geographic weights connectivity matrix when , c*ij* = 1 but only if, georeferenced health centers *i* and *j* were adjacent, and c*ij* = 0 otherwise; $\mathbf{1}^T\mathbf{C1}$ counted the number of ones in the spatially weights matrix, **T** denoted matrix transpose, **TR** denoted the matrix trace operator, $\mathbf{E}_p$ was the *n*-by-P matrix of selected eigenvectors, *kj* was the eigenvalue corresponding to the *j*th eigenvector appearing in the SF $\left(MC_i=\frac{n}{1^TC1}\lambda_j\right)$; H was the number of selected eigenvectors portraying PSA, *K – (H* + 1*)* were the number of selected eigenvectors classified as compensatory and P – K were the number of selected eigenvectors portraying NSA in the MDR-TB model residuals generated from the spatial filter analyses. The right-hand side of Equation 2.8 then contained three terms. The first represented the expected value of MC for the PSA uncovered with the restricted candidate set of PSA eigenvectors; the second represented the expected value of MC for the additional PSA eigenvectors selected to counterbalance selection of NSA eigenvectors; and, the third represented the expected value of MC for the eigenvectors capturing hidden NSA. Equation 2.9 then indicated that when the residual MC value was positive and a hidden NSA spatial filter moved the corresponding residual MC expected value back toward zero, but at a rate

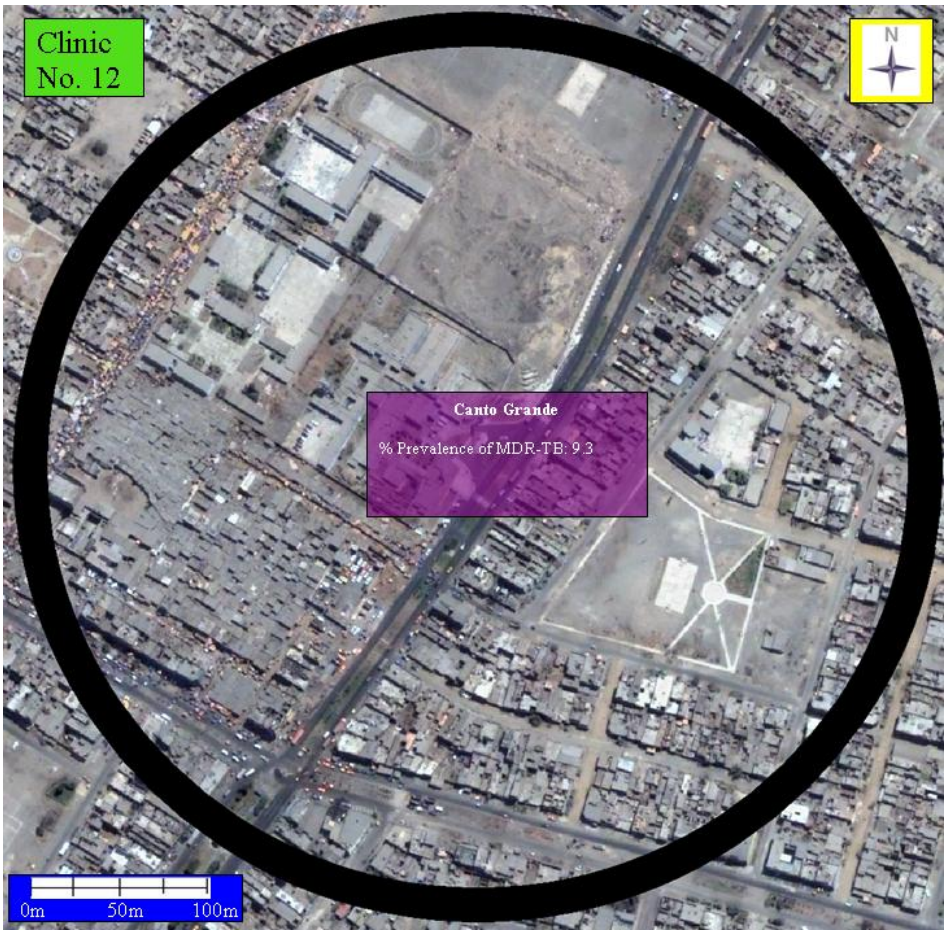

Figure 3. 1 km grid-based algorithm for Canto Grande Health Center with display of MDR-TB prevalence rate.

discounted by the denominator adjustment (that is, the additional subtraction of P – K). Meanwhile, the residual MC for a spatial filter was given by:

$$MC_Y - MC_{\hat{Y}} = MC_Y - \frac{n}{1^T C 1} \frac{\sum_{j=1}^{P} b_j^2 \lambda_j}{\sum_{j=1}^{P} b_j^2} \quad (9)$$

where MC_Y denoted the MC for the georeferenced response variable **Y,** $MC_{\hat{Y}}$ denoted the MC for a constructed spatial filter, and *bj* denoted the linear regression coefficient of the *j*th eigenvector.

In this research all spatially filtered MDR-TB data in SAS/GIS® were integrated with SAS® application, using SAS/EIS. SAS/GIS® allowed for the creation and modification of the MDR-TB maps, as well as interactive feature selection and exploration. Typically, SAS/GIS® application sessions, driven from SAS/EIS®or SAS/AF®, provide powerful SAS Component Language (SCL) components and data step processing capabilities for manipulating data, such as theme datasets utilized in disease mapping (Jacob et al. 2010a). The SAS/GIS® module allowed for the creation and modification of the MDR-TB maps to accurately display results, as well as interactive feature selection and exploration of each georeferenced health center. Spatial information, of each individual health center was imported interactively and in a batch mode.

Additionally**,** Proc MAPIMPORT was used to import the shapefile data created from the MDR-TB data into a SAS/Graph as map datasets. The geographic tables generated, however, had to be processed to identify the coordinates of each health center, with attribute tables being joined to the sampled MDR-TB explanatory covariates for statistical analyses and cartographic display. Additionally, the SAS/GIS® program action linked each table generated with a subset of key predictor variables associated to each sampled Health Center. Thematic map layers were then used to provide more detail for each table. In this research SAS/GIS® used SAS/SHARE to open all datasets, allowing GIS applications to simultaneously read and update all data generated.

RESULTS

A grid-based algorithm and a 1 km buffer generated in an ArcGIS® geodatabase, overlaid on the QuickBird visible and NIR data identified all health centers in the SJL study site. Each grid cell within the matrix contained an attribute value (MDR-TB covariate coefficient value), as well as location coordinates. The spatial location of each cell was implicitly contained within the ordering of the matrix. The health center with the highest MDR-TB prevalence rate was Canto Grande (9.3), while the lowest

Table 2. Global spatial analyses of MDR-TB prevalence rates by Health Centers in the San Lurigancho study site.

Study site	n	Transformation	MC	s_{MC}	GR
San Lurigancho	120	LN(count + 1.5)	0.58	0.06	0.81

LN, Natural logarithm; MC, Moran coefficient; s_{MC}, the standard error of the MC; GR, geary ratio.

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

Spatial statistics	Model output
SF: No. of eigenvectors	7
SF: MC	0.03
SF: GR	0.68
SF pseudo-R^2	0.32
Positive SA SF: No. of eigenvectors	2
Positive SA SF: MC	.899
Positive SA SF: GR	0.06
Positive SA SF pseudo-R^2	0.04
Negative SA SF: No. 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 Centers in the San Lurigancho study site

Statistics	Model ouput
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, the Moran coefficient; GR, the geary ratio; SA, spatial autocorrelation.

MDR-TB resistant rate was Campoy Altos (0.5) (Figure 3).

An examination of the model output from the Poisson regression analyses indicated that significant over-dispersion was present in the sampled MDR-TB data. Therefore, a negative binomial was used to model the overdispersed Poisson data. Negative binomial regression models estimate a dispersion parameter that can be used to remove the effects of overdispersion and provide more accurate estimates of standard error (Kaiser and Cressie, 1997). The negative binomial was derived as a Poisson–gamma mixture and as a GLM. PROC GENMOD expresses the variance of the response for the negative binomial as variance(y) = μ + kμ2, as opposed to the more common notation, (y) = μ + μ2/vμ (Pielou, 1969). In this research the difference in notation was trivial (k = 1/v).

The straightforward derivation of the linear MDR-TB model, from the negative binomial probability distribution function, did not, equate with the Poisson–gamma mixture-based version of the negative binomial. Rather, canonical link and inverse canonical link were converted to log form. A GLM-based negative binomial was produced that yielded identical parameter estimates based on the sampled MDR-TB covariates to those calculated by the mixture-based model. As a non-canonical linked model, however, the standard errors did differ slightly from the mixture model. A maximum likelihood estimator used an observed information matrix to produce standard errors. The GLM algorithm produced standard errors, based on the expected information matrix using the difference in standard errors in the negative binomial analyses. The GLM negative binomial algorithm was amended to allow production of standard errors based on the sampled MDR-TB data. The amended GLM-based negative binomial produced identical estimates and standard errors to that of the mixture-based negative binomial analyses. The log-negative binomial data was then imported into an ArcGIS® database, using the spatial analytical tools in SAS/GIS®.

The spatial autocorrelation analysis rendered the results included in Table 2. Results indicated that negligible PSA was detected in the geographic distribution of the clinical and remote-sampled MDR-TB predictor variables. Estimation results from SAS PROC GENMOD for these models appear in Table 3. Positive spatial autocorrelation and NSA spatial filter component pseudo-R^2 values are reported. These values did not exactly sum for the complete spatial filter; however, the values were very close to their corresponding totals, suggesting that any induced multicollinearity was quite small.

Rather than switching from a Poisson to a negative binomial probability model, the GLMM was extended to account for latent non-spatial correlation effects, as well as to allow inferences to be drawn for a much wider range of geographic sampling configurations. The GLMM included a random effect, which was specified in this research as a random intercept that was assumed to be normally distributed with a mean of zero, a constant

Table 5. A Residual MC minimization of the spatially filtered MDR-TB covariates in the SJL study site.

Criterion	Positive eigenvectors only		Positive and negative eigenvectors	
	# Eigenvectors	Residual	# Eigenvectors	Residual$\|z_{MC}\|$
Min-Max	7	0	7	0.4

variance, and zero spatial autocorrelation. This varying intercept term compensated for the non-constant mean associated with a negative binomial model GLMM specification. The spatial structuring of random effects was than implemented with a conditional autoregressive model which was generated with a spatial filter.

The GLMM estimation results from SAS PROC NLMIXED appear in Table 4. Notably, an extremely strong linear correlation existed between the negative binomial dispersion parameter estimate reported in Table 4 and the random effects variance estimate reported in Table 3. These spatial autocorrelation components suggested the presence of roughly 14% redundant information in the sampled datasets.

The Residual MC minimization criterion analyses rendered the same set of PSA eigenvectors from the spatial decomposition of the Moran's *I* statistic in a stepwise outcome but in a different order. Spatial filters corresponding to the tabulation of eigenvectors appear in Table 4. No compensatory eigenvectors appeared in the residual MC minimization selection criterion. There were no eigenvectors portraying NSA in the model output (Table 5).

DISCUSSION

In this research, we demarcated NSA spatial filters in a MDR-TB regression-based model using a Residual MC minimization criterion and a candidate set of eigenvectors from an eigenfunction decomposition algorithm. A Thiessen polygon surface was constructed for partioning the sampled MDR-TB data in ArcGIS using the MC criterion based on the spatial configuration of the health centers at the study site. Spatial filters were constructed from linear combination of eigenvectors calculated from the connectivity matrix representing a surface partitioning for a spatial dataset. In our spatial filtering analyses of the clinical and environmental MDR-TB data, synthetic variates from a set of eigenvectors were extracted with the matrix $(\mathbf{I} - \mathbf{11}^T/n)\ \mathbf{C}\ (\mathbf{I} - \mathbf{11}^T/n)$ which appeared in the numerator of the MC index. This matrix decomposed the Moran's *I* statistic generated using the sampled MDR-TB explanatory covariate coefficients for generating a robust Poisson spatial filtering GLMM. 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 then derived in terms of the eigenfunction spatial filter. The eigenvectors described the full range of all possible mutually orthogonal MDR-TB map patterns based on the spatiotemporal-sampled clinical and environmental covariate coeffcients. The ratio of the areas of the Thiessen polygons to the gridded areas of their corresponding georeferenced health centers were then evaluated for global and local negative dependencies. When the ratios of the actual-to-Thiessen-polygon area ratio were spatially quantified no NSA was detected in the model.

The pioneering nature and the conceptualization of our analysis presented in this research alludes to many themes meriting future spatiotemporal MDR-TB research in the SJL study site. For example, hidden NSA may be detected and qualitatively assessed in a spatiotemporal MDR-TB model which may signify something beyond the more obvious model misspecifications. For example, seasonal MDR-TB model misspecifications may be associated with some anthropogenic population-concentration mechanism at the SJL study site (e.g., rural-to-urban migration) that may require further quantitative monitoring and thereafter inputting as an independent covariate in a robust regression-based inference model. For instance, as people move into areas with little access to piped water as in shantytowns, there may be wider communal use of living quarters at the SJL study site. Additional socio-geographic dependent explanatory covariate coefficients therefore, may add more precision to a predictive spatial autoregressive MDR-TB transmission-oriented model.

Overall, general findings in this research suggest several rules that should help guide a TB researcher in modeling clinical and environment sampled explanatory covariate coefficients in urban environments. Foremost, switching between spatial and non-spatial regression model specifications should yield similar intercept values. Second, non-normal sampled MDR-TB data are best described with non-normal probability models. Third, a Gaussian approximation spatial filter model can be used to quickly explore whether both PSA and NSA components underpin a MDR-TB map; a spatial filter model specification enables a detailed understanding of latent spatial autocorrelation. And, fourth, a Residual MC minimization criterion can be used to determine if hidden NSA furnishes a diagnostic for spatiotemporal MDR-TB model misspecification.

Further, it is important to note that an autocorrelation graph can be employed to determine if a leptokurtic distribution is symmetrical in shape and similar to a normal distribution, while simultaneously quantitating if

the center peak is much higher; that is, if there is a higher frequency of the sampled MDR-TB clinical and environmental covariate coefficients values near the mean. Moran scatterplot and prediction intervals can capture movements from a platykurtic to leptokurtic profile (Anselin, 1995). Leptokurtic distributions in robust spatiotemporal MDR-TB data would then be indicated in the model by higher central peak and larger tails than a normal distribution that persists over time. Theoretically, this output would be counter to the predictions for random walks in homogeneous time series-dependent MDR-TB population database as the central limit theorem (CLT) would predict that the distribution of the distances moved by infected individuals which would approach normality with repeated draws (e.g., seasonal samplings), if the draws are from the same population. In probability theory, the CLT states that, given certain conditions, the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed (Rice, 1995). Okubo (1980) and Skalski and Gilliam (2000) proposed a population heterogeneity hypothesis to explain leptokurtic distributions, drawing from the fact that leptokurtic distributions can be generated as the composite of two or more normal distributions with similar means and contrasting variances. Heterogeneity in infected MDR-TB-related population movement behavior (e.g., from residence to employment sites and primary school locations) as revealed by scatterplots based on leptokurtic distributions could then help derive and quantitate important differences among sexes, age, or social status and disease transmission vulnerability. For example, other clinical explanatory covariate coefficients representing behavioral or psychological variables (e.g., homelessness, alcoholism) and/or more environmental proxy variables associated to MDR-TB transmission (e.g. Euclidean distance measurements to prison) may also reveal differences in MDR-TB transmission-oriented variables within a lagged scatterplot. Robustness testing differences in the variances of the normal distributions of any spatiotemporal-sampled MDR-TB covariates influenced by infected population movement distances, for example, could produce, leptokurtic patterns when plotted together.

Interestingly, the spatial analyses in this research initially produced platykurtic distributions, but the autocorrelation died off exponentially and converged to a Gaussian relatively fast. Population heterogeneity produces leptokurtic distributions of distance moved when a subset of the individuals consistently move longer distance than others (Skalski and Gilliam, 2000; Fraser et al., 2001). When the heterogeneity is in the landscape, not in the individuals, the departures from a Gaussian will eventually be washed out because a particular individual will switch its movement behavior as it encounters patches of different sampled covariates (Betts, 2009).

The speed of convergence is related to how fast the individuals “forget” their previous direction (Fotheringham, 2002). In this research, the distribution of step vectors generated from the regressed MDR-TB covariate coefficients not only affected the rate of convergence but also the way in which convergence was achieved.

A formal treatment of the rate of convergence to a Gaussian in heterogeneous landscapes such as the SJL study site is beyond the scope of this research, but an inspection of the simulation results revealed that the decay of kurtosis with time may be described using our model framework. Thus, the rate of convergence to a Gaussian will also be affected by skewness in spatiotemporal MDR-TB-related movement vectors. In this work, the occurrence of skewed distributions of distance movements at the SJL study site was minimized since there was no external bias in movement direction. The Bayesian Poisson model estimated rates of resistance to each drug by characterizing the endogenous counts, which was thereafter classified by the sampled health center data. For pathogens that must be treated with combinations of antibiotics and acquire resistance through genetic mutation, knowledge of the order in which drug-resistance mutations occur may be important for determining treatment policies. (Reichman et al., 1979) Our Bayesian approach fit branching tree models which revealed that isoniazid and rifampicin were important for MDR-TB treatment in the SJL study site. The standard "short" course treatment for TB-related diseases is isoniazid along with pyridoxal phosphate to obviate peripheral neuropathy caused by isoniazid, rifampicin, pyrazinamide, and ethambutol for two months, then isoniazid and rifampicin alone for a further four months (Iseman 1993).

The residual output from the model alludes to many Bayesian themes for future predictive spatiotemporal MDR-TB research in the SJL study site. For example, once a robust Bayesian probabilistic estimation matrix renders an autoregresssive unbiased estimator it may be kriged using a deterministic interpolator (e.g., inverse distance weighting matrix) which may be employed for time series multivariate prediction of sampled clinical and environmental explanatory covariate coefficients. Since kriging can also be as a form of Bayesian inference (Griffith 2003), a TB analyst could hypothetically begin with a prior distribution over the functions rendered from regressed seasonal-sampled explanatory covariate coefficients. This prior would then be made to take the form of a Gaussian process in the spatiotemporal MDR-TB model. Thus, *N* samples from a function in the model would be normally distributed, whereas, the covariance between any two of the samples would be the covariance function or kernel of the Gaussian process evaluated at a spatial location (for example, georeferenced health center) where the points were sampled, Next, a set of values would then be quantified whereby each value would be associated with the spatial location.

Thereafter, a new sampled clinical value can be predicted at any new spatial location, by combining the Gaussian prior with a Gaussian likelihood function for each of the observed MDR-TB –related Bayesian values. The resulting posterior distribution would also be Gaussian, with a mean and covariance that would then be simply computed from the observed values, their variance, and the kernel matrix derived from the prior.

In conclusion, the spatial analyses of the clinical and environmental covariates sampled in the SJL study site revealed PSA in all models tested; similar log-MDR-TB prevalence rates of the health centers aggregated in geographic space. Our spatial filter model specification enabled an eigenfunction decomposition of the regression residuals, to yield eigenvectors with latent spatial autocorrelation in the sampled data. The orthogonal parameter estimation algorithm allowed each parameter in the non–linear difference equation model to be estimated sequentially and independently of the other explanatory covariates in the model. 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: (1) a spatial filter variate capturing latent spatial dependency, that otherwise would have remained in the response residuals, and (2) a non-spatial variate that was free of spatial dependence. These spatial autocorrelation components suggested the presence of roughly 14% redundant information in the clinical and environmental sampled data. The residual MC minimization criterion analyses found no evidence to suggest that there were negative dependencies present in the model residuals. The algorithm, however, provided unbiased estimates in the presence of correlated noise and provided an indication of which terms to include in the final model. Linear mixed models, autocovariate regression, spatial eigenvector mapping and a residual MC Minimization criterion can be used for qualitatively assessing latent autocorrelation error coefficients in empirical datasets of spatiotemporal-sampled MDR-TB clinical and environmental explanatory covariate coefficients. A lagged-scatterplot can then allow the autocorrelation error coefficients to be displayed .This information can be used for analyzing clinical and environmental sampled MDR-TB data and for implementing control strategies in the SJL study site.

REFERENCES

Al-Orainey IO, Saeed ES, El-Kassimi FA, Al-Shareef A (1989). Resistance to antituberculosis drugs in Riyadh, Saudi Arabia. Tubercle 70:207-10.

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

Asmussen S (2003). Applied Probability and Queues. Springer-Verlag, USA.

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.

Betts M (2009). The ecological importance of space in species distribution models: a comment on Dormann et al. Ecography 32:1–5.

Chatterjee S, Hadi A (1998). Sensitivity analysis in linear regression. Wiley, New York.

Clarke SE, Bough C, Brown RC, Walgreen GE, Thomas CJ, 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.

Cliff AD, Ord JK (1981). Spatial Processes. Pion, London.

Cohen A, Madigan D, Sackrowitz HB (2003). Effective directed tests for models with ordered categorical data. Aust. NZ. J. Stats 45(3): 285-300.

Cressie NAC (1993). Statistics for Spatial Data Revised Edition. New York: John Wiley & Sons, Inc.

Crofton J, Chaulet P, Maher D, Grosset J, Harris W, Horne N, Iseman M, Watt B (1997). Guidelines on the management of drug-resistant tuberculosis. WHO/TB/96.210.

de Jong P, Sprenger C, van Veen F (1984). On Extreme values of Moran's I and Geary's C. Geo Anal 16(1):1-8.

Dormann CF (2007). Assessing the validity of autologistic regression. Ecol. Model. 207:234-242.

Dutilleul P (1993). Modifying the t-test for assessing the correlation between two spatial processes. Biometrics 49:305-314.

Dye C, William BG, Espinal MA, Raviglione MC (2002). Erasing the world's slow stain: strategies to beat multidrug-resistant tuberculosis. Sci. 295(5562):2042-2046.

El Sahly HM, Teeter LD, Pawlak RR, Musser JM, Graviss EA (2006). Drug-resistant tuberculosis: a disease of target populations in Houston, Texas. J. Infect. 53:5-11.

ERDAS Imagine v.8.7™ (Atlanta, USA)

Espinal MA, Laszlo L, Simonsen F, Boulahbal F, Kim SJ (2001). Global trends in resistance to antituberculosis drugs: World Health organization-international union against tuberculosis and lung disease working group on anti-tuberculosis drug resistance surveillance. N. Engl. J. Med. 344:1294-1303.

Iseman MD (1993). Treatment of multidrug-resistant tuberculosis. NEJM 11:784-791.

Farmer P, Le´andre F, Mukherjee JS (2001). Communitybased approaches to HIV treatment in resource-poor settings. Lancet 358: 404-409.

Fotheringham AS, Brunsdon C, Charlton M (2002). Geographically weighted regression: The analysis of spatially varying relationships. John Wiley & Sons Ltd., Sussex, England.

Fraser DF, Gilliam JF, Daley MJ, Le AN and Skalski GT (2001). Explaining Leptokurtic Movement Distributions:Intrapopulation Variation in Boldness and Exploration. Am. Nat. 158(2):124-135.

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

Glantz SA, Slinker BK (2001). A Primer of Applied Regression and Analysis of Variance-New York: McGraw-Hill.

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. Let.t 58:245-251

Griffith DA (2003). Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization. Springer-Verlag, Berlin.

Griffith DA (2005). A comparison of six analytical disease mapping techniques as applied to West Nile Virus in the coterminous United States. Int. J. Health Geogr. 4:18.

Griffith DA (2006). Beyond the Bell-shaped curve: Poisson Models in Spatial Data Analysis. Geo. Anal. 38(2):iii-iv.

Griffith DA (2008). Spatial filtering-based contributions to a critique of geographically weighted regression (GWR). Environ. Plan A. 40(11):2751-2769.

Griffith DA, Amrhein CG (1997). Multivariate Statistical Analysis for Geogrpahers. Prentice Hall, New Jersey.

Griffith DA, Layne LJ (1999). A Casebook for Spatial Statistical Data

Analysis: A Compilation of Analyses of Different Thematic Datasets. New York: Oxford University Press.
Griffith DA, Peres-Neto PR (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses in exploiting relative location information. Ecol. 87:2603-2613.
He F, Zhou J, Zhu H (2003). Autologistic regression model for the distribution of vegetation. J. Agric. Biol. Environ. Stats. 8(2):205-222.
Hopkins WL, Weeks DL (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. Educ. Psychol. Meas. 50:717-729.
Hoeting JA, Leecsater M, Bowden D (2000). An improved model for spatially correlated binary responses. J. Agric. Bio. Environ. Stat. 5:102-114.
Hosmer DW, Lemeshow S (2000). Applied logistic regression 2nd edn. John Wiley & Sons, New York.
Jacob BG, Muturi E, Mwangangi J, Wanjogu RK, Mpanga E, Funes J, Halbig P, Shililu J, Githure J, Regens JL and Novak RJ (2007). Land use land cover change on Anopheles arabiensis (Diptera:Culicidae) aquatic habitats in Karima village, Mwea Rice Scheme, Kenya. J. Am. J. Trop. Med. Hyg. 76(1):73-80.
Jacob BG, Krapp F, Ponce M, Gotuzzo E, Griffith DA, Novak RJ (2010a). Accounting for autocorrelation in multi-drug resistant tuberculosis predictors using a set of parsimonious orthogonal eigenvectors aggregated in geographical space. Geospatial Health 4, 2, 201-217.
Johnson RT (2003). Emerging viral infections of the nervous system. J. Neurobiol. 9:140-147.
Kaiser M, Cressie N (1997). Modeling Poisson variables with positive spatial dependence. Stat. Probab Lett. 35:423-432.
Legendre P (1993). Spatial autocorrelation: trouble or new paradigm? Ecol. 74:1659-1673.
Lennon JJ (2000). Red-shifts and red herrings in geographical ecology. Ecography 23:101-113.
Kung-Yee L, Zeger S (1986). Longitudinal data analysis using generalized linear models. Biometrika 73(1):13-22
Maddala GS (2001). Introduction to Econometrics, John Wiley & Sons Ltd, Miles, J.N.V.
Manton KG, Stallard E (1981). Methods for Evaluating the Heterogeneity of Aging Processes in Human Populations Using Vital
Statistics Data (1981). Explaining the Black/White Mortality Crossover by a Model of Mortality Selection. Hum. Biol. 53:47-67.
McPherson JM, Jetz W (2007). Effects of species' ecology on the accuracy of distribution models. Ecography 30:135-151.
McCullagh P, Nelder JA (1989). Generalized Linear Models. Chapman and Hall, London.
Miles JNV, Shevlin ME (2001). Applying regression and correlation: a guide for students and researchers. Sage Publications, London, UK.
Miller J (2007). Incorporating spatial dependence in predictive vegetation models. Ecol. Model. 202:225-242.
Okubo A (1980). Diffusion and ecological problems: mathematical models. Springer, New York.
Orenstein EW, Basu S, Shah SN, Andrews JR, Friedland GH, Moll AP, Gandhi NR, Galvani AP (2009). Treatment outcomes among patients with multidrug-resistant tuberculosis: systematic review and met-analysis. Lancet Inf. Dis. 9(3):153-161.
Pearson K (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder. Biometrika, 4: 169-212.
Pedhazur EJ (1997). Multiple regression in behavioral research: Explanation and prediction. Orlando, FL: Harcourt Brace.
Pielou EC (1969). An Introduction to Mathematical Ecology. Wiley, New York.
Piorecky MD, Prescott DRC (2006). Multiple spatial scale logistic and autologistic habitat selection models for Northern pygmy owls, along the Eastern slopes of Alberta's Rocky Mountains. Biol. Conserv. 129: 360-371.
Reichman LB, Felton CP, Edsall JR (1979). Drug dependence, a possible new risk factor for tuberculosis disease. Int. Med. 139:337-339.
Rice J (1995). Mathematical StTAatistics and Data Analysis (Second ed.). Duxbury Press.
Skalski GT, Gilliam JF (2000). Modeling diffusive spread in a heterogeneous population: a movement study with stream fish. Ecol. 81:1685-1700.
Slinker BK, Glantz SA (1985). Multiple regression for physiological data analysis: the problem of multicollinearity. Am. J. Physiol. 249:1-12.
Smith PA (1994). Autocorrelation in logistic regression modeling of species distributions. Global Ecol. Biogeogr 4: 47-61.
Tiefelsdorf M, Boots B (1995). The exact distribution of Moran's I. Environ. Plan. A. 27(6):985-999
Tiefelsdorf M, Griffith DA (2007). Semi-parametric Filtering of Spatial Autocorrelation: The Eigenvector Approach. Environ. Plan. A 39:1193-1221.
van Teeffelen AJA, Ovaskainen O (2007). Can the cause of aggregation be inferred from species distributions? Oikos 116:4-16.
World Health Organization (2000). Global Tuberculosis Control Report. WHO/CDS/TB/2000.275.
Wintle BA, Bardos DC (2006). Modeling species-habitat relationships with spatially autocorrelated observation data. Ecol. Appl. 16:1945-1958.