academicJournals

Vol. 8(11), pp. 253-268, November 2016
DOI: 10.5897/JPHE2016.0838
Article Number: 420649261032
ISSN 2141-2316
Copyright © 2016
Author(s) retain the copyright of this article
http://www.academicjournals.org/JPHE

Journal of Public Health and Epidemiology

Full Length Research Paper

The persistence of tuberculosis in the United States: Spatial analysis and predictive modeling in the move toward elimination of tuberculosis (TB)

Ali Moradi, Sarah Elizabeth Shafer, Ricardo Izurieta, Ismael Hoare, Teresa Maria Pettersen and Benjamin George Jacob*

University of South Florida College of Public Health, Department of Global Health, USA.

Received 5 June, 2016; Accepted 19 September, 2016

Though tuberculosis (TB) prevalence has decreased dramatically in the United States, its continual presence remains a threat to those whose needs are often overlooked. Those already impacted by poverty are the most vulnerable to TB, and stand to bear the worst health impacts, should they contract this disease. Mathematical modeling and spatial analysis have become invaluable tools in TB surveillance monitoring and elimination efforts. In this contribution, we demonstrate the capability of employing a time series, interpolative, vulnerability model to forecasted, state-level TB prevalence in the United States by determining areas influenced by poverty, as well as existing TB data acquired from the Center of Disease Control (CDC). The random effects term in this orthogonal eigenvector spatial filter model was comprised of spatially structured and stochastic effects (that is, spatially unstructured) terms, which were substituted for diagnostic, remote, and clinical covariates in our model. It was assumed that random effects terms in the TB risk model had followed a Gaussian frequency distribution with a mean of zero. The estimate equations were as follows: $\hat{\xi} \sim \mathrm{N}(-0.0025,\ 0.6059^2), \mathrm{PS(S - W)} = 0.34$ **and** $\hat{p} = \frac{1}{1 + e^{10.4180 - 0.0708T + \hat{\xi}}}$**. The resulting estimated number of cases for a given state and year was** $\hat{n}_{TB} = \hat{0}.0885 + 0.9996 \times \mathrm{Population} \times \hat{p}$**. The Moran coefficient (MC) was 0.66, and its Geary Ratio (GR) was 0.35. The spatially unstructured random effects terms have only trace levels of spatial autocorrelation, with MC = 0.02, and Gr = 0.89. Thus, the assumption of non-zero spatial autocorrelation was violated. The forecast revealed possible hyperendemic transmission of TB in non-coastal, Northwestern states, as well as in some Northeastern states. As such, more intervention efforts should be directed towards these areas.**

Key words: Tuberculosis (TB), poverty, center of disease control (CDC).

INTRODUCTION

Despite promising headway towards the elimination of tuberculosis in the United States in recent decades, there

*Corresponding author. E-mail: bjacob1@health.usf.edu. Tel: 813 974 9784.

has been recognized a disconcerting stagnation in what had been an encouraging downward trend. The year 2014 marked a record low for new cases of TB in the US Kang et al., 2014). Still, the decline in incidence from 2013 to 2014 represents the least dynamic change in over ten years (Scott et al., 2015). With immigration, both documented and undocumented, as well as forced human displacement are frequently cited explanations for the persistence of US tuberculosis infections, and attention is often directed toward states with higher immigrant populations, particularly when the countries of origin for larger subpopulations are known to be of greater TB endemicity (Greenwood and Warriner, 2011; Ricks et al., 2011; Bennett et al., 2014; Davidow et al., 2015; Stennis et al., 2015). Other discussions include transmission among those experiencing homelessness (CDC, 2012; Feske et al., 2013) and outbreaks among those inhabiting long-term care facilities (Cavanaugh et al., 2012). Appropriate screening, treatment, and other interventional measures are certainly necessary in all subpopulations deemed to be at-risk, but there may yet be pockets within the US population overall which are not being reached. These pockets may include rural areas where, despite a comparatively smaller population size, crowded housing and lack of mobility are still experienced by those living in poverty and extreme poverty. Such living conditions are endured by many American Indian populations inhabiting reservations (Durand, 2015), a subject that often continues to be evaded in public discourse. In addition to those born in the US, more contemporary trends in immigrant dispersion to “new immigrant destination” sites should also be taken into account for TB transmission to be more thoroughly understood and predicted. There may also be yet risk factors that have not been identified which may allow for better prediction of higher TB prevalence. Spatially speaking, if demographic and migration patterns of today are not incorporated into predictive analysis, resources for TB intervention may be allocated around the TB challenges of a different time. Spatial analysis and mathematical modeling of tuberculosis transmission incorporating known risk factors, as well as patterns of movement using bacillus genotyping as an indicator, is already a well-recognized area of study (Ferdinand et al., 2013; France et al., 2015; Said et al., 2016).

As Houben et al. (2014) noted, mathematical modeling of interventional strategies represents a viable alternative to more expensive, time-consuming, and potentially unethical randomized control trials. These studies tend to be conducted in countries where TB is perceived as a greater threat than in the United States, and may prove quite useful in places of higher HIV prevalence, given that TB is one of the most common deadly opportunistic infections in HIV positive individuals (Houben et al., 2014). Other studies concerned themselves with issues specific to the challenges of a particular local area, including Ge et al. (2015) study of the role of regional transportation and high and low elevation in the Shandong Province of China, as well as Munch et al. (2003) investigation of the spatial distribution of TB in Cape Town, South Africa, including cluster analysis of *shebeens* (neighborhood bars), overcrowding, and unemployment (Winston and Navin, 2010). Jacob et al. (2010) and Jacob et al. (2013) investigated multi-drug resistant tuberculosis and its dispersion through the community of San Juan de Lurigancho in Lima, Peru, where prison visitation was proved key to transmission.

To our knowledge, there are no existing empirical models for the United States which geostatistically address country-wide trends. A predictive, empirical approach which encompasses space, as well as time, in prevalence estimations may allow for a more judicious allocation of resources and interventions toward the elimination of tuberculosis in the United States. To date, a promising study by Feske et al. (2011) utilized kernel density maps in ArcGIS to identify statistically significant clusters for TB transmission in Harris State, Texas. We seek to expand upon the merits of this approach in the creation of a TB transmission risk model for the United States. To accomplish this, we shall: 1) produce Poissonian and negative binomial regression models to determine frequentist pseudo R^2 estimates, and 2) construct orthogonal spatial filter eigenvectors using a decompositional algorithm for cartographically displaying predictive abundance values. When creating a risk model for tuberculosis, and other communicable diseases, several measures should be taken to ensure optimal accuracy. For a given form of statistical analysis, certain assumptions are made concerning the distribution used. Any violations of these assumptions may result in misspecifications in the model, the implication being that limited resources may be allocated less efficiently. Spatial analysis represents a special case of predictive model construction which is better able to accommodate both multicollinearity and heteroscedasticity, as well as identify response variables. We further contend that assuming normality is largely impractical for the purposes of spatial epidemiological analysis.

METHODOLOGY

Endemic, TB-related, state-level, parameterizable data including prevalence, race, gender and other socio-demographic data were acquired for each state from 2000 to 2014 from Center for Disease Control (CDC). FLEXIBLE|FLE in SAS 9.2® (Carey, North Carolina) was employed to request the flexible-beta method. The clustering methods in SAS include average linkage, the centroid method, complete linkage, density linkage (including Wong’s hybrid and *k*-th nearest-neighbor methods), ML for mixtures of spherical, multivariate, normal distributions with equal variances but possibly unequal mixing proportions, the flexible-beta method, McQuitty’s similarity analysis, the median method, single linkage two-stage density linkage, and Ward’s minimum-variance method (http://ftp.sas.com). PROC CLUSTER displayed the table of eigenvalues of the covariance matrix for the canonical variables. Generally, in a PROC CLUSTER table output the first two columns list each eigenvalue and the difference between the eigenvalue and its successor, while the last two columns display the individual and

Table 1. Parameterizable clinical, field and remote covariate samples within stratified clusters of TB data in the state study sites as entered in SAS®.

Variable	Description	Units
GCP	Ground control points	Decimal-degrees
EDUC	El	Meters
RCE	Race	Percentage
GEN	Gender	Percentage
AGE	Age	Percentage
PPOS	Previous positive cases	Numeric value
ECON	Income	Percentage

cumulative proportion of variation associated with each eigenvalue (www.sas.com). In the TB distribution model, the squared multiple correlations, pseudo R^2, was the proportion of variance accounted for by the stratified, state-level, geo-referenceable clusters. The approximate expected value of pseudo R^2 was then given in the column labelled "ERSQ". The next three columns displaced the values of the cubic clustering criterion (CCC), pseudo F (PSF), and t2 (PST2) statistics. These statistics were useful in quantitating the number of specified predictive, state-level, parameterizable, intra-cluster, covariate estimators. One method of judging the number of clusters in a dataset in PROC CLUSTER is to examine the pseudo F statistic (PSF) (www.sas.com). The CLUSTER procedure hierarchically clustered the state-level observations using SAS data. The data were then cartographically illustrated by mapping the endemic geocoordinates and squaring Euclidean distance measurements within a flexible-beta method in PROC CLUSTER.

The PROC CLUSTER statement initiated the procedure, which digitally specified a clustering method based on state-level prevalence measures, and then optionally specified each explanatory cluster covariate coefficient. The PROC CLUSTER statement specified a clustering method, and optionally specified details for clustering methods, data processing, and then displayed an output. The model estimated the number of possible transmission centers a person may encounter per state. This was calculated by multiplying the proportional prevalence probability estimates with the proportion of gridded state geolocations stratified by economics and ancestry of origin, and their human population distribution. The agglomerative, hierarchical clustering procedure then utilized geosampled, state-level observations to create geospatial clusters based on asymptotically normalized clinical TB data in a cluster by itself. Clusters were then merged to form a new cluster that replaced the two old clusters, and merging of the two closest clusters was repeated until only one cluster was left.

Beta was set at -100 for epidemiological forecasting cluster-based analyses in SAS PROC CLUSTER. The flexible-beta method began by specifying METHOD=FLEXIBLE. PROC CLUSTER then created an output, interpolative, time series, asymptotically unbiased dataset to reveal a cluster hierarchy of normalized, state-level data feature attributes based on parameterized, covariate coefficient, and estimator values. Since the explanatory estimators were deemed to be equally important, we employed the STD option in PROC CLUSTER to standardize the cluster-based predictor covariate coefficients to mean 0 with standard deviation. Covariates with large variances tended to have a greater effect on the resulting geospatialized TB clusters than variables with small variances. However, if all coefficients are considered equally important in the model, the STD option in PROC CLUSTER standardizes the geo-spatiotemporally geosampled variables.

The STDIZE procedure standardized the covariate estimators in the SAS dataset by subtracting the state-level, stratified, gridded measures, and then dividing them by a scale measure. Finally, a unique identifier was incorporated for each cluster. The PLOTS option in the PROC CLUSTER statement produced plots of the cubic clustering criterion (CCC), the pseudo *F* (PSF) statistic, and the pseudo (PST2) statistic, which were then all plotted against the number of geosampled, state-level TB clusters.

In order to reduce the likelihood of chaining among the TB cluster-based dataset of predictor covariates, a partition that best represented the estimates was identified. This was performed by finding the intersection between a manageable number of state-level, cluster-based, varying and constant, explanatory covariate coefficients and then auto-probabilistically and auto-regressively quantitating them with large jumps in the normalized, Euclidean distance measurements in PROC CLUSTER. The cluster-based, covariate estimators were plotted against ArcGIS-based, Euclidean distance measurements. This revealed a clear flattening of the curve in the digitally overlain data in PROC CLUSTER, indicating that adequate separation of the parameterizable clustering covariate coefficients could not be achieved beyond a specific georeferenced capture point (for example, Veteran Administration Hospital). The number of interpolatable, geosampled TB transmission clusters in the data was also determined by preliminary evaluations with varying numbers of cluster solutions aimed at avoiding trivial error. Evaluation was done by plotting the state-level TB data in discriminant function space in PROC CLUSTER, and seeking adequate separation among group centroids. In order to compute meaningful standardized rates, the individual georeferenced state-level predictors were aggregated geographically into high-low stratified clusters in ArcGIS.

Environmental data analyses

Univariate statistics and regression models were generated by employing the data stored in PROC CLUSTER for regressively summarizing the geospatially clustered covariate coefficients. We generated a misspecification term for constructing an autoregressive, time series model in SAS. Multiple data layers were created using different coded values for the various known data feature attributes. Distance measurements and endemic transmission foci measures were then calculated by using the WV-3 data and the field-sampling information (Table 1).

Regression analyses

The relationship between the state-level, endemic TB data and each individual predictive, geospatially clustering covariate was investigated by single variable regression analysis in PROC NL MIXED. Since prevalence data are binomial fractions, a regression model was employed; as it is a standard practice for vulnerability analysis. Poisson probability regression analyses were employed to infer the relationship between the TB count data variables and the archived empirical, clinical, field and remote-specified state-level characteristics (that is, independent variables) in PROC LOGISTIC.

The regression analyses assumed independent counts (that is, *Ni*) taken at multiple geosampled, state sub-locations $i = 1, 2, \ldots, n.$ The geo-spatiotemporal-related state-level counts were then described by a set of variables denoted by matrix $\mathbf{X}_i$, where a $1 \times p$ was a vector of covariate coefficient indicator values for geosampled endemic transmission foci *i*. The expected value of these data was given by $\mu_i(\mathbf{X}_i) = n_i(\mathbf{X}_i)\exp(\mathbf{X}_i\beta)$, where β was the vector of the parameterizable, non-redundant, geosampled covariates in the epidemiological, state-level, risk model, and where the Poisson rates were 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 for each geosampled state location *i*. The dependent variable was state-level prevalence. The Poisson regression model assumed that the predictors were equally dispersed. That implied that the conditional variance equaled the condition mean. Partial correlations were then defined after introducing the concept of conditional distributions. We initially restricted ourselves to only the conditional distributions obtained from the multivariate, normalized distributions. We noted an $n \times 1$ random vector Z, which we partitioned into two random vectors X and Y, where X was an $n_1 \times 1$ vector and Y was an $n_2 \times 1$ vector in the equation $Z = (XY)$. The conditional distribution properties of the regressed, state-level, covariate coefficients were then defined. Thereafter, we partitioned the mean vector and covariance matrix in a corresponding manner: $\mu = (\mu 1\, \mu 2)$ and $\sum = (\sum_{11} \sum_{21} \sum_{12} \sum_{22})$. This way, μ_1 rendered the means for the regressed predictor variables in the set x_1, and $\sum_{11}$ along with the variances and covariances for set x_1. The matrix $\sum_{12}$ thereafter provided the covariances between the predictor variables in set x_1 and set x_2 as did matrix $\sum_{21}$. Any distribution for a subset of variables from multivariate normal, conditional on known values for another subset of variables has a multivariate normal distribution (Griffith, 2003).

It was noted that the conditional distribution of x_1 given the known values for x_2 was multivariate normal with a mean vector covariance $\mathrm{matirx} = \mu 1 + \sum_{12} \sum - 122(x_2 - \mu 2) \sum_{11} - \sum_{12} \sum - 122 \sum_{21}.$ The procedure employed ML estimation to find the operationalized, time-series, dependent, regression coefficients. The data were then log-transformed before analysis to normalize the distribution and minimize standard error. There was considerable overdispersion in the regression-based model residual forecasts, so a negative binomial model with a non-homogenous distributed mean was employed to quantitate the covariates associated with the geosampled data. Over-dispersion is often encountered when fitting very simple parametric models, such as those based on the Poisson distribution (Griffith, 2003).

A Poisson mixture model with a negative binomial distribution was employed where the mean of the Poisson distribution was itself a random variable drawn from the gamma distribution. This introduced an additional free parameter in the empirical, state-level, TB distribution model. If over-dispersion is a feature in an asymptotical, predictive risk model, an alternative model with additional free parameters may provide a better fit (Griffith, 2003). Jacob et al. (2013) employed a family of negative binomial distributions for treating over-dispersion in an MDR-TB forecasting, vulnerability model. The Poisson distribution has one free parameter and does not allow for the variance to be adjusted independently of the mean (Griffith, 2003). A parameterization technique was then employed in PROC LOGISTIC such that any two state-level had explanatory regressable variables *p* and *r* with $0 < p < 1$ and $r > 0$. Lack of fit and over-dispersion can be assessed using the Pearson and deviance statistics available in the GENMOD, LOGISTIC, and PROBIT procedures (http://support.sas.com/kb/22/630.html). Under this parameterization, the probability mass function (pmf) of the predictor variables with a NegBin (r, p) distribution took the following form:

For $k = f(k;\, r,\, p) = \begin{pmatrix} k + r - 1 \\ \\ k \end{pmatrix} \cdot p^r \cdot (1 - p)^k 0, 1, 2,$

Where

$$\begin{pmatrix} k + r - 1 \\ \\ k \end{pmatrix} = \frac{\Gamma(k + r)}{k!\,\Gamma(r)} = (-1)^k - \begin{pmatrix} -r \\ \\ k \end{pmatrix} \text{ and } \Gamma(r) = (r - 1)!.$$

Also, an alternative parameterization was employed for quantitating the state-level TB data using the mean $\lambda : \lambda = r \cdot (p^{-1} - 1)p = \frac{r}{r + \lambda}$.

The mass function then became:

$$g(k) = \frac{\lambda^k}{k!} \cdot \frac{\Gamma(r + k)}{\Gamma(r)(r + \lambda)^k} \cdot \frac{1}{\left(1 + \frac{\lambda}{r}\right)^r},$$

where λ and *r* were the parameters. Under this parameterization, $r \underset{\longrightarrow \infty}{\lim} g(k) = \frac{\lambda^k}{k!} \cdot 1 \cdot \frac{1}{\exp(\lambda)}$ was generated, which resembled the mass function of a Poisson-distributed random variable with Poisson rate (i.e., $\lambda.$). In other words, the negative binomial distribution generated from the regressed, parameterizable covariates converged to the Poisson distribution, and *r* controlled the deviation from the Poisson. This made the negative binomial habitat model suitable as a robust alternative to the Poisson regression-based framework for risk modeling the interpolatable, time-series, clinical, field and remote-specified predictors.

The negative binomial distribution of the explanatory, state-level, TB covariates arose as a continuous mixture of Poisson distributions where the mixing distribution of the Poisson rate was a gamma distribution. The mass function of the negative binomial distribution of the geosampled transmission predictor variables then was written as:

$$f(k) = \int_0^\infty Poisson(k|\lambda) \cdot Gamma(\lambda|r, (1 - p)/p)\, d\lambda$$

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

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

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

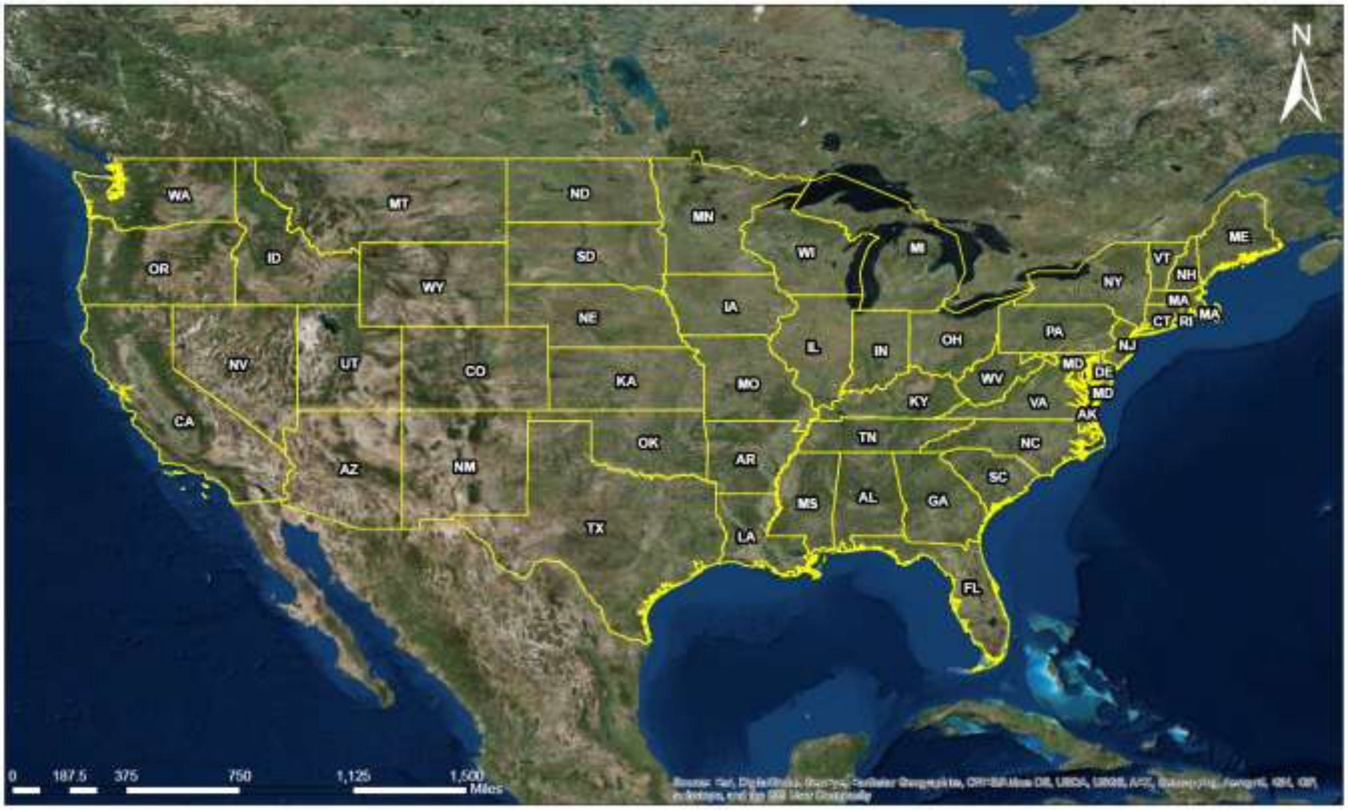
N
WA
MT
ND
MN
ME
OR
ID
SD
WI
MI
VT
NH
WY
NY
MA
CT
RI
MA
NE
IA
PA
NV
UT
CO
IL
IN
OH
NJ
MD
DE
MD
KA
MO
WV
VA
KY
AK
CA
OK
TN
NC
AR
AZ
NM
SC
MS
AL
GA
TX
LA
FL
0
187.5
375
750
1,125
1,500
Miles
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 1. The study site consisted of the contiguous United States.

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

Spatial analyses

This is represented by Figure 1. Initially, a misspecification perspective for the asymptotically normalized estimation models was generated in SAS/GIS 9.2 assuming that the geo-spatiotemporal, risk model parameter fit was $y = X\beta + \varepsilon^*$ (that is, regression equation). The primary function of the model generation was for quantitating the auto-correlated disturbances ε^* in the residually forecasted, regression-based derivatives. The SAS/Graph mapping functionality allowed us to create choropleth, prism, block, and surface maps. This included the GMAP, GREMOVE, GREDUCE, GPROJECT and MAPIMPORT procedures. Three key techniques were highlighted to deliver SAS/GRAPH auto-regressable, predictive risk maps: Map data, Annotate and the Output Delivery System (ODS). The latent autocorrelation coefficients were decomposed into a white-noise component, ε, and a set of unspecified and/or misspecified model outputs that had the structure $y = XB + \underbrace{E\gamma + \varepsilon}_{=\varepsilon^*}.$ White noise is a univariate or multivariate, discrete-time, stochastic process, whose terms are independent and identically distributed (i.i.d.) with a zero mean (Jacob et al., 2013).

The Annotate facility enabled generating a special dataset of graphics commands from which the state-level TB graphic output was created. The annotate output combined with the PROC GMAP output generated multiple customized surface maps. The misspecification term was $S(T) = S(t)\exp$. Quantification of the topographic patterns rendered from the distribution of the regressed, predictive covariates was required to describe independent key dimensions of the underlying spatial processes in the geosampled data for heuristically defining a pattern in the misspecification term. SAS/GIS software provided an efficient interactive tool for organizing and analyzing the state-level TB, clinical, field and remote-sampled data that was referenced spatially.

A geospatialized, time series, autoregressive model was generated employing a predictive, specified variable Y as a function of a nearby geosampled variable Y in the SAS/GIS autoregressive model. A covariate coefficient indicator value *I*, an autoregressive response, and the residual of Y were treated as a function of a nearby geosampled Y residuals, as a spatially autoregressive (SAR) or spatial error specification. For TB transmission risk modeling, the SAR model furnishes an alternative specification that frequently is written in terms of matrix W (Griffith, 2003). As such, the spatial covariance of the geosampled dataset was a function of the matrix $(\mathbf{I}-\rho\mathbf{C}\mathbf{D}^{-1})(\mathbf{I}-\rho\mathbf{D}^{-1}\mathbf{C}) = (\mathbf{I}-\rho\mathbf{W}^{T})(\mathbf{I}-\rho\mathbf{W}),$ where *T* denoted the matrix transpose. The resulting matrix was symmetric and was considered a second-order specification as it included the product of two spatial structure matrices $(\text{i.e.}, \mathbf{W}^T\mathbf{W}).$ This matrix restricted positive values of the autoregressive parameter to the more intuitively interpretable range of $0 \le \hat{\rho} \le 1$.

Distance between the predictive covariate coefficients was defined in terms of an *n*-by-*n* geographic weights matrix, C, whose

c_{ij} values were 1 if the geosampled locations *i* and *j* were deemed nearby, and 0 otherwise. Adjusting this matrix by dividing each row entry by its row sum then rendered C1, where 1 was an *n*-by-1 vector of ones, and subsequently converted the time series regression-based matrix to matrix W. The resulting SAR model specification with no parameterizable predictor covariates the pure spatial autoregression specification, then took on the form $\mathbf{Y} = \mu(1-\rho)\mathbf{1} + \rho\mathbf{WY} + \varepsilon,$ where μ was the scalar conditional mean of Y, and ε was an *n*-by-1 error vector with parameters independently and identically distributed (normally random variates). The spatial covariance matrix for analyzing the state-level covariate coefficients was thereafter expressed employing $E[(\mathbf{Y} - \mu\mathbf{1})'(\mathbf{Y} - \mu\mathbf{1})] = \Sigma = [(\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W})]^{-1}\sigma^2,$ where $E(\bullet)$ denoted the calculus of expectations, I was the *n*-by-*n* identity matrix denoting the matrix transpose operation and σ^2 was the asymptotical, stochastically/deterministically-related error variance.

The TB predictive model was written as: $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t,$ where $\varphi_1, \ldots, \varphi_p$ were the empirical covariate estimators, *c* was a constant and ε_t was the white noise. When coupled with regression and the normal probability model, an autoregressive specification results in a covariation term characterizing spatial autocorrelation by denoting the autoregressive parameter with ρ at a conditional autoregressive covariance specification (Griffith, 2003). This specification involved the matrix $(\mathbf{I} - \rho\mathbf{C})$ where I was an *n*-by-*n* identity matrix. In an autoregressive expression, however, the optimal response variable is on the left-side of the equation, while the spatial lagged version of the variable is on the right side (Griffith, 2003). Therefore, one of the main objectives in this research was to bring the spatially unlagged, TB predictor variable *y* exclusively to the left-hand side of the regression equation in order to decorrelate the normalized state-level TB covariate coefficients. This was accomplished by expanding the weighted regression coefficient matrix term: $(I - \rho V)^{-1} = \sum_{k=0}^{\infty} \rho^k V^k$ as an infinite power series, which was feasible only under the assumption that the underlying spatial process in the normalized, state-level TB datasets was stationary. The autoregressive, interpolatable, forecasting error model was then rewritten as $y - \rho V y = X\beta - \rho V X\beta + \varepsilon.$ in AUTOREG. Substituting this transformation rendered:

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

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

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

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

$$y = X\beta + \underbrace{\sum_{k=1}^{\infty} \rho^k V^k \varepsilon}_{misspecificationterm} + \varepsilon.$$

The misspecification term $\sum \rho^k V^k \varepsilon \ (k = 1, \ldots, \infty)$ remained uncorrelated with the explanatory variable *X*, as the standard OLS assumption of the disturbances ε was uncorrelated with the geosampled residualized variables generated from the parameter estimation process. The spatial lag model was expressed as $(I - \rho V)y = X\beta + \varepsilon.$ Substituting the transformation generated:

$$y = \sum_{k=0}^{\infty} \rho^k V^k (X\beta + \varepsilon) \text{ and } y = X\beta + \underbrace{\sum_{k=1}^{\infty} \rho^k V^k (X\beta + \varepsilon)}_{misspecificationterm} + \varepsilon.$$

The misspecification term $\sum \rho^k V^k (X\beta + \varepsilon)$ $(k = 1, \ldots, \infty)$ then included the exogeneous variables *X*. Consequently, the state-level variables were correlated with the misspecification term. Under this condition, standard OLS results for the basic regression model $y = X\beta + \varepsilon^*$ generated from the regressed covariate coefficients provided biased estimates $\hat{\beta}$ of the underlying regression parameters β.

Eigenvector analyses

The correlation, or lack thereof, between the predictor variables and the misspecification terms in the autoregressive risk model were utilized to design spatial proxy variables so the properties of either model could be satisfied. Unfortunately, misspecification of the main exposure variable, as well as other covariates, is not uncommon in regression models (Jacob et al., 2013).

Functional forms can adversely affect tests of the association between the exposure and response variables (Jacob et al., 2013). In regression analyses, the process of developing a regression model consists of selecting an appropriate functional form for the model and then choosing which variables to be included in the regression procedure (Griffith, 2003). A function shall be defined for our purposes as a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output (Griffith, 2003). An example in transmission risk modeling is the function utilized by Jacob et al. (2014) when constructing the MDR-TB explanatory covariate coefficient *x* to its square x^2. In the model, the output of a function *f* corresponded to an input *x* which then was denoted by *f*(*x*). The input variable(s) are sometimes referred to as the argument(s) of the function (Griffith, 2003). The first step for constructing a robust TB model in a SAS covariance matrix is to specify the model (Jacob et al., 2014). If an estimated covariate coefficient model is misspecified it will be biased and inconsistent (Jacob et al., 2013). In regression-based risk models, the term misspecification covers a broad range of modeling errors including measurement errors and discretizing continuous, normalized explanatory variables (Griffith, 2003).

Two different projection matrices, $M_{(1)} \equiv I - 1\left(1^T 1\right)^{-1} 1^T$ and $M_{(X)} \equiv I - X\left(X^T X\right)^{-1} X^T.$ in AUTOREG were considered for autoregressing the state-level TB data. The projection matrix $M_{(1)}$ is a special case of the more general projection matrix $M_{(X)}$ [ww.sas.edu]. The general projection matrix $M_{(X)}$ in the TB model included a constant unity vector 1, as well as additional TB-transmission explanatory variables. A set of eigenvectors $\{e_1, \ldots, e_n\}_{SAR}$ was then extracted from the regressed quadratic form

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

which was designed orthogonal to the exogeneous variable *X*. The projection matrix $M_{(X)}$ imposed this constraint. In contrast, the set of operationizable eigenvectors $\{e_1, \ldots, e_n\}_{Lag}$ was extracted from

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

These two different sets of eigenvectors established a basis for constructing a robust, regression-based distribution model, with both expressions solely defined in terms of the regressed exogeneous information. This model feature in AUTOREG enabled us to employ the eigenvector spatial filtering approach for predictions of the regressed, endogeneous variable *y*. The associated sets of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}_{Lag}$ and $\{\lambda_1, \ldots, \lambda_n\}_{SAR}$, with $\lambda_i \geq \lambda_{i+1}$, range were then employed in AUTOREG for properly standardizing adjacent link matrices *V* that related to irregular spatial tessellations generated from the regressed, predictive state-level TB covariate coefficients. The components of each eigenvector, e_i, were then mapped in SAS/GIS onto an underlying spatial tessellation which exhibited a distinctive topographic pattern ranging from positive spatial autocorrelation (PSA), or similar values of log-transformed count data aggregating in space, for $\lambda_i > E(I)$ to negative spatial autocorrelation NSA, which is the aggregation of dissimilar log-values in space for $\lambda_i > E(I)$. Each eigenvector was mapped where $E(I)$ was the expected value of Moran's *I* under the assumption of (a) spatial independences and (b) as outputs from related projection matrices $M_{(1)}$ or $M_{(X)}$, respectively.

It was noted that the associated Moran's *I* autocorrelation coefficient of each eigenvector e_i generated from the risk model was equal to its associated eigenvalue $\lambda_i = \left[e_i^T\left(V + V^T\right)e_i\right] / \left(2e_i^T e_i\right)$, but only if *V* was scaled to satisfy $\left[1^T\left(V + V^T\right)1\right]/2 = n.$. Moran's autocorrelation is often denoted as *I*, which is an extension of Pearson's product moment correlation coefficient, a commonly used measure of the amount of autocorrelation in regressed, empirical, multivariate, estimators (Griffith, 2003). In previous research, Jacob et al. (2014) employed the Pearson's correlation coefficient for spatially summarizing a dataset of autocovariance terms quantitated between multiple empirical, geosampled predictor variables to define the covariance of multivariate parameterized covariates divided by the product of their standard deviations using $\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}.$ In this research the formula defined the normalized, state-level population, correlation coefficient. Substituting estimates of the covariances and variances derived from the auto-regressed dataset of covariate coefficients provided the sample correlation coefficient, denoted by:

$$r : r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

An equivalent expression rendered the correlation coefficient as the mean of the products of the standard scores in AUTOREG. Based on paired normalized, spatial data feature attributes $(\text{i.e.}, X_i, Y_i)$, the sample Pearson correlation coefficient was:

$$r - \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{Y_i - \overline{Y}}{s_X}\right),$$

where $\frac{X_i - \overline{X}}{s_X}$ and $\overline{X}$ were the standard score sample mean and the sample standard deviation, respectively.

The eigenvectors yielded distinct, predictive, geo-spatiotemporal, map pattern descriptions of latent spatial autocorrelation in the empirical, geosampled data. This was interpreted as synthetic map variables that represented specific natures (positive or negative) and degrees (negligible, weak, moderate and strong) of potential spatial autocorrelation.

For the covariates, two counteracting spatial autocorrelation effects were conceptualized (that is, common factors leading to PSA, and competitive factors leading to NSA materializing) at the same time, with a possible net effect being global detection of near-zero spatial autocorrelation. If a parsimonious set of eigenvectors is to be selected for, eigenvectors depicting near-zero spatial autocorrelation should be avoided, as such a set of latent vectors associated with a matrix equation will fail to capture any geographic information (Griffith, 2003).

The eigenvector spatial filtering approach added a minimally sufficient set of eigenvectors as proxy-variables to the set of linear predictors in our predictive model by inducing mutual independence in the covariate estimators. The regression residuals represented geo-spatiotemporally independent, state-level, predictor variable components. The spatial pattern in the eigenvectors was synthetic. In the state level TB model, positive global autocorrelation in the local patterns of the parameters exhibited only positive local autocorrelation and vice versa for negative global autocorrelation. The eigenvectors e_i and e_j within each set of eigenvectors were then mutually orthogonal, as the symmetry transformation $\frac{1}{2}\left(V + V^T\right)$ was a quadratic form as revealed in Equations (2.1) and (2.2).

As mentioned previously, the eigenvectors of specification (2.1) were orthogonal to the time-series, exogeneous, variables *X* of the regression TB forecast model constructed in AUTOREG employing the georeferenced, explanatory covariates. Conversely, the eigenvectors of specification (2.2) were orthogonal only to the constant unity vector 1 in *X*. This quantifiable orthogonality had implications for modeling the geospatialized misspecification terms in the risk model which allowed each collection of eigenvectors to be linked to its specific autoregressive model residual. This was accomplished by letting E_{SAR} be a matrix whose vectors were subsets of $\{e_1, \ldots, e_n\}_{SAR}$. Within-group estimation of higher-order autoregressive panel models were also considered with exogenous regressors and fixed effects, where the lag order was possibly misspecified. Even when disregarding the misspecification bias, the fixed-effect bias formula regressed differently from the correctly specified case, though its asymptotic order remained the same under stationary conditions. A linear combination of this subset was approximated by employing the misspecification term of the autoregressive version of the state-level, TB predictive, risk model which was expressed as

$$(E_{SAR}\gamma \approx \sum_{k=1}^{\infty}\rho^k V^k \varepsilon). \tag{3}$$

The linear combination $E_{SAR}\gamma$ remained orthogonal to exogeneous variables *X*, so the estimated predictor variables $\hat{\beta}$ were unbiased. Also, as a property of the OLS estimator, the estimated term $E_{SAR}\gamma$ was orthogonal to the residuals $\hat{\varepsilon}$. The model $y = X\hat{\beta} + E_{SAR}\hat{\gamma} + \hat{\varepsilon}$ decomposed the endogeneous predictor variable *y* into a systematic trend component, a stochastic signal component and white-noise residuals. The term $E_{SAR}\hat{\gamma}$ removed variance inflation in the mean square error (MSE) term attributable to spatial autocorrelation in dataset of covariate coefficients.

Alternatively, for the spatial lag model (Equation 3), a risk model was constructed employing E_{Lag}, a matrix of those eigenvectors which were a subset of $\{e_1, \ldots, e_n\}_{Lag}$. The approximation of the misspecification term became $E_{Lag}\gamma \approx \sum_{k=0}^{\infty} \rho^k V^k (X\beta + \varepsilon)$. Since $E_{Lag}\gamma$ was correlated with the exogeneous variables *X*, its incorporation into the state-level, TB risk model corrected the bias of estimated plain OLS parameters $\hat{\beta}$ in the analysis of latent spatial lag. The model $y = X\hat{\beta} + E_{Lag}\hat{\gamma} + \hat{\varepsilon}$ was generated from covariates, which were a decomposition of the spatial lag model o, a systematic trend component, a stochastic signal component, and a dataset of white-noise residuals. For the risk model, it was now noted that the trend and the stochastic/deterministic, time-series, signals were no longer uncorrelated and the mean square error (MSE) was deflated.

The set of eigenvectors $\{e_1, \ldots, e_n\}_{Lag}$ of the spatial lag model (Equation 3) was then calculated in AUTOREG independently of the exogeneous, state-level, predictor variables *X*. This calculation was dependent on the underlying spatial link matrix *V*. It was found that this filtering approach was more adaptable to a specification search of the relevant exogeneous variables and spatial predictions with the regressed temporal shifting predictor variable values in our risk model in AUTOREG. In contrast, for the simultaneous autoregressive model (Equation 2), the eigenvectors $\{e_1, \ldots, e_n\}_{SAR}$ depended through the projection of $M_{(X)}$ on the exogenous variables *X*. Thus, any change in the underlying model structure required a recalculation of the eigenvectors to generate more robust tessellations. Thereafter, spatial filtering of either the spatial lag model or the simultaneous autoregressive model with a common factor constraint only required identification of one set of selected eigenvectors, namely E_{SAR} or E_{Lag},. The relevant set of eigenvectors was applied to all the TB predictor covariates in both models. For the generic autoregressive model (Equation 1), however, spatial filtering was applied individually to each covariate coefficient. The generic specification of autoregressive spatial models then associated a specific spatial lag factor with the endogeneous *y* variable and other lag factors for each additional exogeneous variable. The eigenvectors $\{e_1, \ldots, e_n\}_{Lag}$ were employed to filter spatial autocorrelation in the generic, autoregressive vulnerability model employing each geosampled covariate estimator.

The next step was identification of appropriate, parsimonious subsets of eigenvectors E_{SAR} or E_{Lag} from either risk model explanatory specification (Equation 1) or (Equation 2). A particular subset of eigenvectors was deemed suitable if the optimizable residuals $\hat{\varepsilon}$ of the resulting spatially filtered model becomes stochastically independent with respect to the underlying sampled spatial structure *V* (Griffith, 2003). Thereafter, parsimony in model estimation was defined as the smallest possible subset of eigenvectors leading to geospatial independence in the residually forecasted derivatives of the TB model being identified. It was noted that geospatial patterns of different eigenvectors expressed independent and filter autocorrelation of the derivatives of the regression model as formalized by a georeferenced vector. Similar methodology has been employed for extrapolation of predictor covariates associated with hyperendemic transmission foci in other contexts.

RESULTS

Initially, a Poisson regression model was constructed in PROC LOGISTIC using temporospatial TB covariate coefficient measurement values. The Poisson process in our analysis was provided by the limit of a binomial distribution of the sampled state-level explanatory predictor covariate coefficient estimates using:

$$P_p(n \mid N) = \frac{N!}{n!\,(N-n)!}\, p^n\,(1-p)^{N-n} \tag{4}$$

The distribution was viewed as a function of the expected number of state-level count variables using the sample size *N* for quantifying the fixed *p* in Equation 1, which was then transformed into the linear equation:

$$P_{\frac{v}{N}}(n|N) = \frac{N!}{n!(N-n)!}\left(\frac{v}{N}\right)^n\left(1-\frac{v}{N}\right)^{N-n}.$$

Based on the sample size *N*, the distribution as it approached $P_v(n)$ was:

$$\lim_{n\to\infty} P_p(n|N) = \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{n!}\frac{v^n}{N^n}\left(1-\frac{v}{N}\right)^N\left(1-\frac{v}{N}\right)^{-n} = \lim_{N\to\infty}\frac{N(N-1)\cdots(N-n+1)}{N^n}\frac{v^n}{n!}\left(1-\frac{v}{N}\right)^N\left(1-\frac{v}{N}\right)^{-n} = 1\cdot\frac{v^n}{n!}\cdot\ e^{-v}\cdot 1 = \frac{v^n e^{-v}}{n!}.$$

The PROC LOGISTIC procedure then fit a generalized linear model to the sampled data by maximum likelihood estimation of the parameter vector β. The PROC LOGISTIC procedure estimated the seasonal-sampled parameters of each state-level TB model numerically through an iterative fitting process. The dispersion parameter was then estimated by the residual deviance and by Pearson's chi-square divided by the degrees of freedom (df). Covariances, standard errors, and *p*-values were computed for the sampled covariate coefficients based on the asymptotic normality derived from the maximum likelihood estimation.

Note that the sample size *N* completely dropped out of the probability function, which had the same functional form for all the sampled state-level parameter estimator indicator values (that is, v). As expected, the Poisson distribution was normalized so that the sum of probabilities equaled 1. The ratio of probabilities was then determined by:

$$\sum_{n=0}^{\infty} P_v(n) = e^{-v}\sum_{n=0}^{\infty}\frac{v^n}{n!} = e^{-v}e^{v} =$$

which was then expressed as

$$\frac{P_v(n=i+1)}{P(n=i)} = \frac{\frac{v^{i+1}e^{-v}}{(i+1)!}}{\frac{e^{-v}v^i}{i!}} = \frac{v}{i+1}.$$

Our model was generalized by introducing an unobserved heterogeneity term for each sampled state-level observation i. The TB weights were then assumed to differ randomly in a manner that was not fully accounted for by the other covariates. This state-level process was formulated as $E(y_t \mid x_t, \tau_t) = \mu_t \tau_t = e^{x_t' \beta - \varepsilon_t}$, where the unobserved heterogeneity term $\tau_i = e^{\varepsilon_i}$ was independent of the vector of regressors x_i. The distribution of y_i was conditional on x_i and had a Poisson specification with conditional mean and variance $\mu_i \tau_i : f(y_i \mid x_i, \tau_i) = \frac{\exp(-\mu_i \tau_i)(\mu_i \tau_i)^{y_i}}{y_i!}$. We then let $g(\tau_i)$ be the probability density function of τ_i. At this point, the distribution $f(y_i \mid x_i)$ was no longer conditional on τ_i. Instead it was obtained by integrating $f(y_i \mid x_i, \tau_i)$ with respect to $g(\tau_i)$: $f(y_i \mid x_i) = \int_0^\infty f(y_i \mid x_i, \tau_i) g(\tau_i) d\tau_i$. It was found that an analytical solution to this integral existed in our state-level model when τ_i was assumed to follow a gamma distribution. The model also revealed that y_i was the vector of the sampled predictor covariate coefficients while x_i was independently Poisson-distributed with $P(Y_i = y_i \mid x_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, y_i = 0, 1, 2 \ldots$, and the mean parameter (that is, the mean number of state-level sampling events per temporospatial period) was given by $\mu_i = \exp(x_i' \beta)$, where β was a $(k+1) \times 1$ parameter vector. The intercept in the model was β_0 and the coefficients for the k regressors were $\beta_1, \ldots, \beta_k$. Taking the exponential of $x_i' \beta$ ensured that the mean parameter μ_i was nonnegative. Thereafter, the conditional mean was provided by $E(y_i \mid x_i) = \mu_i = \exp(x_i' \beta)$. The state-level parameter estimators were then evaluated using $\ln[E(y_i \mid x_i)] = \ln(\mu_i) = x_i' \beta$ in PROC LOGISTIC. Note, that the conditional variance of the count random variable was equal to the conditional mean (i.e., equidispersion) in our model $V(y_i \mid x_i) = E(y_i \mid x_i) = \mu_i$]. In a log-linear model the logarithm of the conditional mean is linear (Haight 1970). The marginal effect of any state-level regressor in the TB model was provided by $\frac{\delta E(y_i \mid x_i)}{\delta x_{ji}} = \exp(x_i' \beta) \beta_j = E(y_i \mid x_i) \beta_j$. Thus, a one-unit change in the jth regressor in the model led to a proportional change in the conditional mean $E(y_i \mid x_i)$ of β_j. The standard estimator for our Poisson model was the maximum likelihood estimator. Since the state-level observations were independent, the log-likelihood function in the model was then:

$$= \sum_{i=1}^{N} (-\mu_i + y_i \ln \mu_i - \ln y_i!) = \sum_{i=1}^{N} (-e^{x_i' \beta} + y_i x_i' \beta - \ln y_i!)$$

Given the sampled dataset of state-level parameter estimators *θ* and an input vector *x*, the mean of the predicted Poisson distribution was provided by $E(Y \mid x) = e^{\theta' x}$. This way, the Poisson distribution's probability mass function was then rendered by:

$$p(y \mid x; \theta) = \frac{e^{y(\theta' x)} e^{-e^{\theta' x}}}{y!}$$.

The probability mass function in our targeted TB risk model is the primary means for defining a discrete probability distribution. As such, functions could exist for either scalar or multivariate field-sampled random variables, given that the distribution is discrete (Jacob et al., 2015). Since the geosampled, state-level TB dataset consisted of *m* vectors $x_i i \in \mathbb{R}^{n+1}, i = 1, \ldots, m,$, along with a set of *m* values $y_1, \ldots, y_2 \in \mathbb{R}$, the sampled estimators *θ*, the probability of attaining this particular set of the sampled observations was provided by:

$$p(y_1, \ldots, y_m \mid x_1, \ldots, x_m; \theta) = \prod_{i=1}^{m} \frac{e^{y_i(\theta' x_i)} e^{-e^{\theta' x_i}}}{y_i!}$$

Consequently, the set of *θ* that made this probability as large as possible in the model estimates was obtained. The equation was first rewritten as a likelihood function in PROC LOGISTIC in terms of θ:

$$p(y_1, \ldots, y_m \mid x_1, \ldots, x_m; \theta) = \prod_{i=1}^{m} \frac{e^{y_i(\theta' x_i)} e^{-e^{\theta' x_i}}}{y_i!}$$

Note the expression on the right hand side in our model had not actually changed. Next, we used a log-likelihood

$$\ell(\theta \mid X, Y) = \log L(\theta \mid X, Y) = \sum_{i=1}^{m} (y_i(\theta' x_i) - e^{\theta' x_i} - \log(y_i!)))]$$.

Because the logarithm is a monotonically increasing function, the logarithm of a function achieves its maximum value at the same points as the function itself, and, hence, the log-likelihood can be used in place of the likelihood in maximum likelihood estimation and related techniques (Hosmer and Lemeshew 2011). Finding the maximum of a function often involves taking the derivative of a function and solving for the parameter estimator being maximized; this is often easier when the function being maximized is a log-likelihood rather than the original likelihood function (Jacob et al., 2012). We

Table 2. Summary of backward elimination.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Black	21	0	0.5154	21.002	0	0.968
2	Multiple_Races	20	0	0.5154	19.017	0.02	0.902
3	Age_under_5	19	0	0.5153	17.036	0.02	0.889
4	Native_Born	18	0.0025	0.5128	16.239	1.22	0.271
5	Long_Care	17	0.0027	0.5101	15.517	1.29	0.257
6	Age_25_44	16	0.0034	0.5067	15.146	1.65	0.201
7	Age_45_64	15	0.0043	0.5024	15.185	2.06	0.153
8	Age_65_More	14	0.0041	0.4983	15.147	1.97	0.162
9	Age_5_14	13	0.0041	0.4942	15.12	1.97	0.162
10	NonInject_Drug	12	0.0046	0.4896	15.31	2.18	0.141
11	Alcohol	11	0.0049	0.4847	15.648	2.32	0.129
12	Jail	10	0.0049	0.4798	15.992	2.31	0.13

noted the parameters *θ* only appeared in the first two terms of each term in the summation. Therefore, given that we were only interested in finding the best value for *θ* in the state-level predictive TB regression model, we dropped the $y_i!$ and simply wrote:

$$\ell(\theta \mid X,Y)=\sum_{i=1}^{m}\left(y_i(\theta' x_i)-e^{\theta' x_i}\right)$$

Thereafter, to find a maximum, we solved an equation $\frac{\partial\ell(\theta\mid X,Y)}{\partial\theta}=0$ which had no closed-form solution. However, the negative log-likelihood (LL) $-\ell(\theta\mid X,Y)$ was a convex function, and so standard convex optimization was applied to find the optimal value of *θ*. It was found that, given the Poisson process in our regression model, the limit of a binomial distribution was:

$$p_p(n\mid N)=\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}.$$

Viewing the distribution as a function of the expected number of successes $\nu \equiv Np$ in the model, rather than the sample size *N* for fixed P, rendered the equation (2.1), which became:

$$p_{\nu/N}(n\mid N)=\frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

Our model revealed that as the sample size *N* became larger, the distribution approached *P* when the following equations aligned:

$$\lim_{N\to\infty} p_p(n\mid N) \quad \lim_{N\to\infty}\frac{N(N-1)\ldots(N-n+1)}{n!}\frac{\nu^n}{N^n}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n} \quad ; \quad \lim_{N\to\infty}\frac{N(N-1)\ldots(N-n+1)}{N^n}\frac{\nu^n}{n!}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n}$$

$$1\cdot\frac{\nu^n}{n!}\cdot e^{-\nu}\cdot 1 \text{and} \frac{\nu^n e^{-\nu}}{n!}.$$

We then considered the Euler product $\zeta(s)=\prod_{k=1}^{\infty}\frac{1}{1-\frac{1}{p_k^s}}$ where $\zeta(s)$ was the Riemann zeta function and p_k was the k the prime. $\zeta(1)=\infty$. Thereafter, by taking the finite product up to k=n in our TB regression model and pre-multiplying by a factor $1/\ln p_n$, it was then possible to employ $n\to\infty$, which rendered $\lim_{n\to\infty}\frac{1}{\ln p_n}\prod_{k=1}^{n}\frac{1}{1-\frac{1}{p_k}}=e^{\gamma}$, or 1.431912. To check for non-normalities (for example, heteroskedascity, multicolineraity) in the regression forecasts a stepwise backward model validating procedure in PROC LOGISTIC was employed (Table 2).

The model for overdispersion was then with a likelihood ratio test. This test quantified the equality of the mean and the variance imposed by the Poisson distribution against the alternative that the variance exceeded the mean. For the negative binomial distribution, the variance= mean + k mean^2 (k≥ 0, the negative binomial distribution reduces to Poisson when k=0) (Jacob et al., 2013). For this study, the null hypothesis was H_0: k=0 and the alternative hypothesis was H_a : k>0. To carry out the test, we used the following steps initially and then ran the model using negative binomial distribution and a record log-likelihood (LL) value. We then recorded LL for the Poissonized TB model. The likelihood ratio (LR) test was employed, and the LR statistic was computed, -2(LL (Poisson) – LL (negative binomial). The asymptotic distribution of the LR statistic had probability mass of one half at zero and one half – chi-square distribution with 1 df. To test the null hypothesis further, the critical value of chi-sq distribution corresponding to significance level 2 was used. That is, H_0 was rejected if LR statistic > 2 $_{(1-2,\ 1\ df)}$.

Next, our predictor covariate coefficient estimates were assumed to be based on the log of the mean, which was a linear function of independent variables, log() = intercept + b1*X1 +b2*X2 + ...+ b3*Xm. This log-transformation implied that the exponential function of independent variables equaled exp(intercept + b1*X1 +b2*X2 + ...+ b3*Xm). Instead of assuming, as we did before, that the distribution of the state-level covariate coefficients Y was Poissonian, a negative binomial distribution was assumed. The implications meant relaxing the generalized linear Poisson regression specification assumption concerning equality of the mean and variance, since it was found in our model that the variance of negative binomial was equal to + k2, where k ≥ 0 was a dispersion parameter. The maximum likelihood method was then used to estimate k as well as the parameter estimators of the model for log(). The SAS syntax for running negative binomial regression is very similar to the syntax for Poisson regression. The only change is the dist option in the MODEL statement is used instead of dist = poisson,dist = nb. The probability mass function of the negative binomial distribution with a gamma distributed mean in the predictive TB model was then expressed using the sampled covariate coefficients estimates as:

$f(k) \equiv \Pr(X=k) = \binom{k+r-1}{k}(1-p)^r p^k$ for the variables $k = 0, 1, 2, \ldots$.

In this equation, the quantity in parentheses was the binomial coefficient, and was equal to:

$$\binom{k+r-1}{k} = \frac{(k+r-1)!}{k!\,(r-1)!} = \frac{(k+r-1)(k+r-2)\cdots(r)}{k!}.$$

This quantity was alternatively written as:

$$\frac{(k+r-1)\ldots(r)}{k!} = (-1)^k \frac{(-r)(-r-1)(-r-2)\ldots(-r-k+1)}{k!} = (-1)^k \binom{-r}{k}$$

for explaining the negative binomial qualities in our regression model (Jacob et al., 2013).

Results from both Poissonian and negative binomial model residuals revealed that the covariate coefficient estimates were highly significant, but furnished virtually no predictive power. Inclusion of indicator variables denoting the time sequence and the district location spatial structure were articulated with ArcGIS Thiessen polygons, which also failed to reveal meaningful covariates. Perhaps the presence of noise in the geosampled state –level TB data was attributable for this misspecification. Thus, the state-level TB data were adjusted and quantified for space-time consistency. Next, an Autoregressive Integrated Moving Average (ARIMA) analysis of individual district time-series was conducted in PROC ARIMA. Given our temporospatial data X_t, where t was an integer index and the X_t the values, an ARIMA model was built using:

$$\left(1 - \sum_{i=1}^{p} \alpha_i L^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$$

where L was the lag operator, α_i were the parameters of the autoregressive portion of the model, θ_i were the parameters of the moving average part, and ε_t were error terms. ARIMA models are, in theory, the most general class of models for forecasting a time series which can be made stationary by transformations such as differencing and logging (Griffith, 2003). The easiest way to think of ARIMA models is as fine-tuned versions of random-walk and random-trend models: the fine-tuning consists of adding lags of the differenced series and/or lags of the forecast errors to the prediction equation, as needed to remove any last traces of autocorrelation from the forecast errors (Griffith 2003). The error terms ε_t were generally assumed to be independently sampled from a normal distribution with zero mean: ε_t~ N(0,σ2), where σ2 was the variance. Thereafter, a random effects term was specified with the time series, state–level data. This random effects specification revealed a non-constant, variable mean across states, which mathematically represented a state–level constant across time. The random effects specification also represented a state-specific intercept term, a random deviation from the overall intercept term, which was based on a draw from a normal frequency distribution. This random intercept represented the combined effect of all explicative state-level predictor covariate coefficients, which caused some states to be prone to greater TB prevalence than others. Inclusion of a random intercept assumed random heterogeneity in the states' propensity or underlying risk of TB prevalence which persisted throughout the entire duration of the time sequence under study. Values were procured for this random effects term, and state-level for prevalence regressed on predicted prevalence rates. The Poisson mean response specification was mu = exp[a + re+ LN(population)], Y ~Poisson(mu).

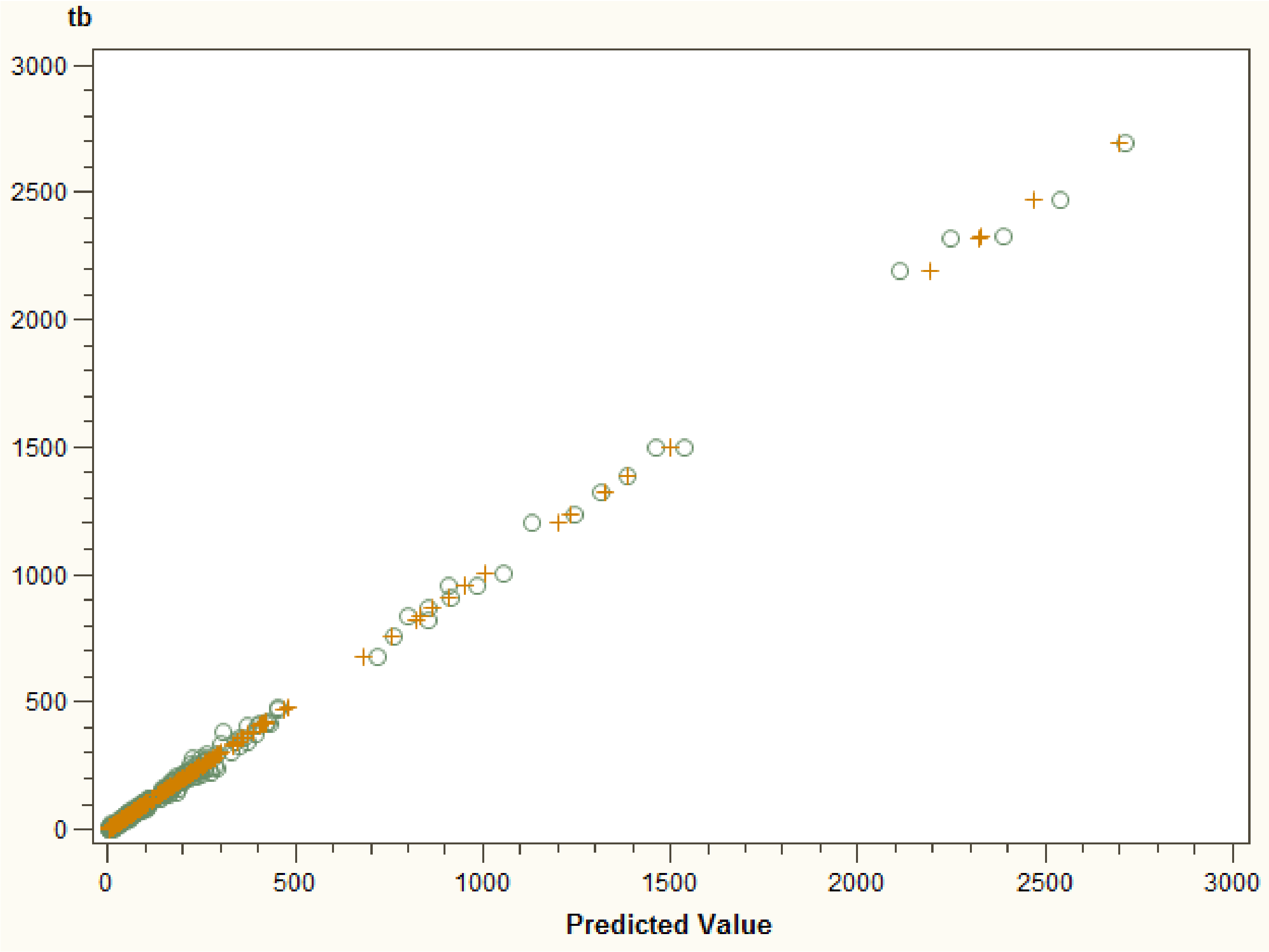

Figure 2. Scatterplot of the predicted and the observed TB rates by state and time period.

A simple space-time binomial mixed model was then estimated, with the random effects term furnishing a common factor through time beyond the simple time sequence fixed component. This random effects term comprised spatially structured and spatially unstructured state-level components. The time sequence covariate alone accounts for roughly 2% of the variation in TB rates across the space-time series. Its combination with the random effects term accounts for roughly 99% of this variation. The deviance statistics for the excess binomial variation is 53.4. Figure 2 portrays the relationship between the predicted and the observed rates. The estimate equations are as follows:

$$\hat{p} = \frac{1}{1 + e^{10.4180 - 0.0708T + \hat{\xi}}}$$

$$\hat{\xi} \sim N(-0.0025,\ 0.6059^{2}), PS(S\text{ - }W) = 0.34$$

The resulting estimated number of cases is for a given state for a given year is:

$$\hat{n}_{TB} = 0.0885 + 0.9996 \times \text{Population} \times \hat{p}$$

The random effects term has both a spatially structured and a spatially unstructured component (Figure 3). The spatially structured random effects term contains five eigenvectors representing non-trivial levels of positive spatial autocorrelation, and accounts for roughly 50% of the variation in the random effects term. Its Moran Coefficient (MC) is 0.66, and its Geary Ratio (GR) is 0.35. The spatially unstructured random effects terms has only trace levels of spatial autocorrelation, with MC = 0.02, and Gr = 0.89. All three components closely conform to a bell-shaped curve (Figure 4).

Limitations

As with most predictive analyses, there were limitations in

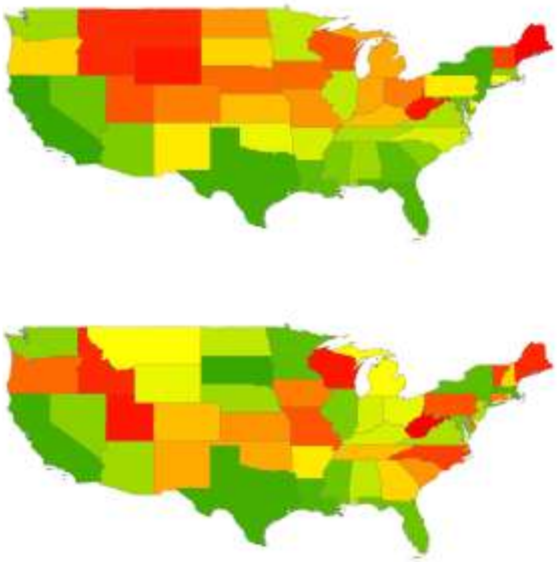

Figure 3. Top left (a): random effects term; MC = 0.33, and GR = 0.55. Top right (b): spatially structure random effects term; MC = 0.66, and GR = 0.35. Bottom left (c): spatially unstructured random effects term; MC = 0.02, and GR = 0.89.

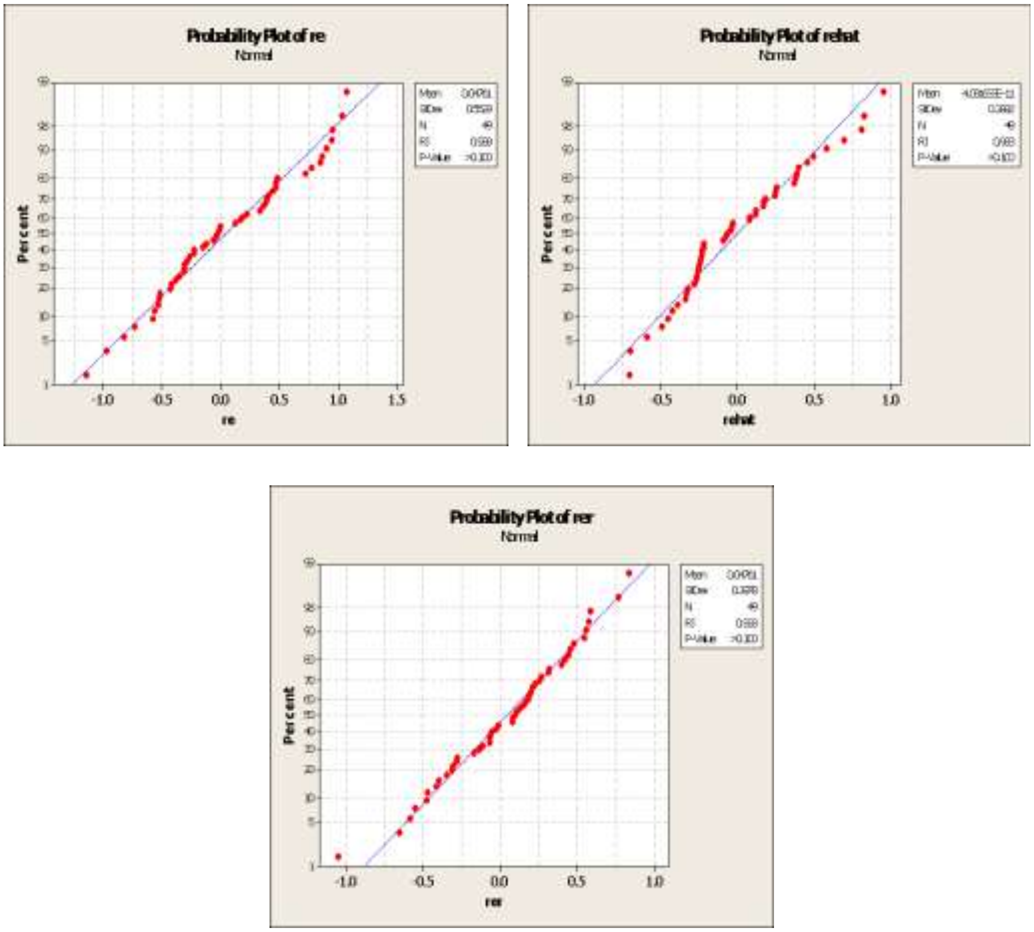

Figure 4. Top left (a): normal quantile plot for the random effects term. Top right (b): normal quantile plot for the spatially structure random effects term. Bottom left (c): normal quantile plot for the spatially unstructured random effects term.

the proposed model which may lead to varying outcomes. The model may have too few variables for the scope of the problem that we are attacking. As stated before, investigation into other potential risk factors, including diabetes is warranted for construction of more robust predictive models. A Bayesian approach may also provide superior methodology for finding local and residual autocorrelation that more traditional frequentist methods may be less sensitive to. An autoregressive integrated moving average (ARIMA) model may also provide output more sensitive to temporality.

DISCUSSION

The clustering of like tendencies, according to this predictive analysis and shown by a Moran's coefficient of 0.66 for our spatially structured random effects model, produced results that, to say the least, would be considered anomalous with conventional wisdom. States that have been historically considered as hubs for immigrant inflow, including California, Texas, Illinois, Florida and New York, do not project higher rates of TB transmission according to our model. In light of these findings, more contemporary patterns of human movement must be considered to fully understand the nature of TB transmission in the United States. Thanks much in part to economic growth in various regions through the 1990s and early 2000s, coupled with hostile policy toward less affluent immigrant populations in states such as California, the modern landscape for immigrant dispersion within the US began to shift dramatically (Ellis et al., 2014). The Migration Policy Institute, citing the Office of Immigration Statistics of the Department of Homeland Security, states that the number of foreign-born individuals grew by roughly 50 percent or more in the states of Alabama, Arkansas, Delaware, Georgia, Idaho, Indiana, Kentucky, Mississippi, Nevada, North Carolina, South Carolina, South Dakota, Tennessee, and Wyoming through the time period of 2000 to 2009 (Terrazas, 2011). These, of course, are not states that were readily equated with heavier inward flow of immigrant populations through much of the last century.

Very recent findings may also indicate an overall decline in tuberculosis in the foreign-born population in the US, some of which can be attributed to demographic changes. In their analysis, Baker et al. (2016) cite three reasons for this shift: 1) changes in the proportion of foreign-born persons through continued movement; 2) changes in the distribution of countries of origin, and 3) actual changes in TB rates for the countries of origin. True decreases in TB case rates in recent immigrants from China, India, and the Philippines were cited for the reason for the decline in these sub-populations, while a decrease in the Mexican sub-population size in the US was cited as the principal reason for the decline in this group. This study is fairly new but nevertheless points to the importance of considering dynamic demographics in the study of TB transmission. Further, as TB is associated with lifestyle, employing an anthropological approach with ethnographic research concerning known risk factors may lend deeper insight into how these variables relate to transmission, as well as allow for existing transmission models to be individually tailored with unique explanatory variables for a given locale. Munch et al. (2003) employed such an anthropological approach in Cape Town, South Africa in an already cited study, as did Ge et al. (2015) with local transportation and population dispersion. Such an approach may not be as useful at the national level, but it may allow for local governments and health departments to better detail individualized intervention and funding needs to their respective states.

Still another set of challenges arises when we consider American Indian peoples, who continue to suffer higher rates of mortality and hospitalization for infectious diseases, including tuberculosis, as compared even with immigrant groups more often implicated in transmission in most research (Bloss et al., 2011; Cheek et al., 2014; Holman et al., 2011; Reilley et al., 2014; Schneider, 2005). Though there is continued decline in rates of tuberculosis among American Indians, as well, they still continue to experience infectious disease transmission disproportionately as compared with other US populations. In South Dakota in 2015, 11 of the 17 reported TB cases (65%) were identified as Native American, with Native Americans maintaining almost consistently higher rates of mortality due to TB (South Dakota Department of Health, 2015). Montana has also continued to see higher rates of TB for American Indian populations over time (Montana Department of Public Health, 2015). It is not uncommon knowledge that many common risk factors for TB, such as poverty, homelessness, and higher levels of alcohol consumption continue to endure in some American Indian communities. These are colonial legacies that have been carried from the fairly recent past of a young nation's history, and may remain as long as apathy remains. It would seem that as much as the general US population seems to diversify and shift, the indigenous people are static by comparison. Further challenges to quantifying the needs for tuberculosis treatment among American Indians lie with lack of coordination of patient records among care providers in several of these communities and use of coding systems that emphasize billing over surveillance (Podewils et al., 2014), as well as lack of adequate funding for the Indian Health Service.

Other factors that must be taken into account are the distributions of other comorbid infections, as well as non-communicable, chronic diseases that have risen alarmingly in the United States. Potential comorbidities with other diseases are often cited as risk factors for TB. These include infectious diseases such as HIV, as well as the more behavioral and psychosocial issues of IV drug use and alcoholism. Coinfection with HIV contributes to higher rates of mortality from tuberculosis, and is already a well-established risk factor for TB, being one of the most common opportunistic infections in those of HIV positive status. Currently, diabetes is being investigated as another potential risk factor for tuberculosis (Demlow et al., 2015; Suwanpimolkul et al., 2014; Gil-Santana et al., 2016). Though more research is necessary, and does not appear at this point in time to affect the transmission of tuberculosis, it does appear to complicate existing cases (Gil-Santana et al., 2016). This may increase overall healthcare costs and personal hardship for those enduring tuberculosis comorbid with diabetes, resulting in greater hardship as well as a greater general burden for healthcare providers. The financial implications alone make diabetes worth investigating as a risk factor. In communities where high rates for diabetes and tuberculosis overlap, more targeted interventions should be considered.

CONCLUSION AND RECOMMENDATIONS

Predictive modeling using GIS, remote sensing and geostatistics, of course, represents only the first step taken. Toward intervention, community-based approaches are already well known and may provide more specialized solutions closer to home for communities more heavily impacted by the persistence of TB. Ecological interventions are best applied when individualized to the needs and wishes of the community in question. Identification of specific target areas within a larger area, potential contact investigations, weak points in community structure, and concerns of individuals specific to a particular locale are more efficient when members of the community are collaborated with by interventionists or researchers. In this vein, the two-phase PRECEDE-PROCEED Model (PPM) may be of use toward TB elimination in and out of the US. The first grand phase encompasses bottom-up assessments of individual behaviors (proximal elements that directly affect an individual), social context and environment (mid-level elements that an individual may), administrative policy (distal elements), as well as potential interplay of these elements (Gielen et al., 2008). The second grand phase is comprised of implementation and evaluation with continuous improvement and modifications as needed (Gielen et al., 2008). For this reason, the PPM need not be applied in a linear manner. General community-based participatory research methods may also show promise in foreign-born populations, who may be prone to reactivation of latent TB. For recent immigrants seeking classes in English as a second language upon arrival to the United States, community centers where these courses are held may be useful spaces for examination of attitudes and knowledge about TB; referrals for screening may also be a possibility (Wieland et al., 2011, 2012, 2013).

Though we have come very far in lowering the prevalence of tuberculosis in the US, further efforts should not fall by the wayside simply because of a leveling-off in decreasing prevalence rates. With more rigorous screening of incoming individuals, including not only immigrants, but any individuals such as students who have spent a significant amount of time in a place of known hyperendemicity) and further research into risk factors that may indicate higher rates of transmission or complication of existing cases of TB, continued decrease in prevalence rates may be possible.

Conflict of Interests

The authors have not declared any conflict of interests.

REFERENCES

Baker BJ, Winston CA, Liu Y, France AM, Cain KP (2016). Abrupt Decline in Tuberculosis among Foreign-Born Persons in the United States. PloS one 11(2):e0147353.

Bennett RJ, Brodine S, Waalen J, Moser K, Rodwell TC (2014). Prevalence and treatment of latent tuberculosis infection among newly arrived refugees in San Diego State, January 2010–October 2012. Am. J. Publ. Health 104(4):e95-e102.

Bloss E, Holtz TH, Jereb J, Redd JT, Podewils LJ, Cheek JE, McCray E (2011). Tuberculosis in indigenous peoples in the US, 2003-2008. Publ. Health Reports pp. 677-689.

Cavanaugh JS, Powell K, Renwick OJ, Davis KL, Hilliard A, Benjamin C, Mitruka, K (2012). An outbreak of tuberculosis among adults with mental illness. Am. J. Psychiatry 169(6):569-575.

Centers for Disease Control and Prevention (2012). Notes from the field: tuberculosis cluster associated with homelessness--Duval State, Florida, 2004-2012. *MMWR.* Morbidity mortality weekly report 61(28):539.

Cheek JE, Holman RC., Redd JT, Haberling D, Hennessy TW (2014). Infectious Disease Mortality among American Indians and Alaska Natives, 1999–2009. Am. J. Publ. Health 104(S3):S446-S452.

Davidow AL, Katz D, Ghosh S, Blumberg H, Tamhane A, Sevilla A, Reves R (2015). Preventing Infectious Pulmonary Tuberculosis Among Foreign-Born Residents of the United States. Am. J. Publ. Health 105(9):e81-e88.

Demlow SE, Oh P, Barry PM (2015). Increased risk of tuberculosis among foreign-born persons with diabetes in California, 2010–2012. BMC Publ. Health 15(1):1.

Durand C (2015). *Native American Housing: Federal Assistance, Challenges Faced and Efforts to Address Them*. Hauppauge: Nova Science Publishers, Inc.

Ellis M, Wright R, Townley M (2014). The allure of new immigrant destinations and the great recession in the United States. Int. Migr. Rev. 48(1):3.

Ferdinand S, Millet J, Accipe A, Cassadou S, Chaud P, Levy M, Rastogi N (2013). Use of genotyping based clustering to quantify recent tuberculosis transmission in Guadeloupe during a seven years period: analysis of risk factors and access to health care. BMC infect. Dis. 13(1):1.

Feske ML, Teeter LD, Musser JM, Graviss EA (2011). Including the third dimension: a spatial analysis of TB cases in Houston Harris State. Tuberculosis 91:S24-S33.

Feske ML, Teeter LD, Musser JM, Graviss EA (2013). Counting the homeless: a previously incalculable tuberculosis risk and its social determinants. Am. J. Publ. Health 103(5):839-848.

Ge E, Lai PC, Zhang X, Yang X, Li X, Wang H, Wei X (2015). Regional transport and its association with tuberculosis in the Shandong province of China, 2009–2011. J. Transport Geogr. 46:232-243.

Gielen AC, McDonald EM, Gary TL, Bone LR (2008). Using the precede-proceed model to apply health behavior theories. Health behavior and health education: Theory, research, and practice 4:407-429.

Gil-Santana L, Almeida-Junior JL, Oliveira CA, Hickson LS, Daltro C, Castro S, Andrade BB (2016). Diabetes Is Associated with Worse Clinical Presentation in Tuberculosis Patients from Brazil: A Retrospective Cohort Study. PLoS One 11(1).

Greenwood MJ, Warriner WR (2011). Immigrants and the spread of tuberculosis in the United States: A hidden cost of immigration. Population Res. Pol. Rev. 30(6):839-859.

Griffith DA (2003). Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization. Springer-Verlag Berlin Heidelberg. New York.

Haight J (1970). A linear set of infinite measure with no two points having integral ratio. Mathematika 17(01):133-138.

Holman RC, Folkema AM, Singleton RJ, Redd JT, Christensen KY, Steiner CA, Cheek JE (2011). Disparities in infectious disease hospitalizations for American Indian/Alaska Native people. Publ. Health Reports pp 508-521.

Hosmer D, Lemeshew S (2011). Applied Logistic Regression. A Wiley-Interscience Publication, New York.

Houben RM, Dowdy DW, Vassall A, Cohen T, Nicol MP, Granich RM, White RG (2014). How can mathematical models advance tuberculosis control in high HIV prevalence settings?: the official journal of the International Union against Tuberculosis and Lung Disease. Int. J. Tuberculosis lung Dis. 18(5):509-514.

Jacob BG, Krapp F, Ponce M, Gotuzzo E, Griffith DA, Novak RJ (2010). Accounting for autocorrelation in multi-drug resistant tuberculosis predictors using a set of parsimonious orthogonal eigenvectors aggregated in geographic space. Geospatial health 4(2):201-217.

Jacob BG, Krapp F, Ponce M, Zhang N, Caliskan S, Hasemann J, Griffith DA, Gotuzzo E, Novak RJ (2013). 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. J. Publ. Health Epidemiol. 5(3):122-143.

Jacob BG, Novak RJ, Toe L, Sanfo MS, Afriyie AN, Ibrahim MA, Griffith DA, Unnasch TR (2012). Quasi-likelihood techniques in a logistic regression equation for identifying *Simulium damnosum* s.l. larval habitats intra-cluster covariates in Togo. Geo-spatial Inform. Sci. 15(2):117-133.

Jacob BG, Mendoza D, Ponce M, Caliskan S, Moradi A, Gotuzzo E, Griffith DA, Novak RJ (2014). Pseudo R2 Probability Measures, Durbin Watson Diagnostic Statistics and Einstein Summations for Deriving Unbiased Frequentistic Inferences and Geoparameterizing Non-Zero First-Order Lag Autocorvariate Error in Regressed Multi-Drug Resistant Tuberculosis Time Series Estimators. Am. J. Appl. Math. Stat. 2(5):252-301.

Kang J, Zhang N, Shi R (2014). A Bayesian nonparametric model for spatially distributed multivariate binary data with application to a multidrug-resistant tuberculosis (MDR-TB) study. Biometrics 70(4):981-992.

Montana Department of Public Health (2015). Tuberculosis Surveillance Report—Montana, 2015.

Munch Z, Van Lill SW, Booysen CN, Zietsman HL, Enarson DA, Beyers N (2003). Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. Int. J. Tuberculosis Lung Dis. 7(3):271-277.

Podewils LJ, Alexy E, Driver SJ, Cheek JE, Holman RC, Haberling D, Brett M, McCray EU, Redd JT (2014). Understanding the burden of tuberculosis among American Indians/Alaska Natives in the US: a validation study. Publ. Health Reports 129(4):351-360.

Reilley B, Bloss E, Byrd KK, Iralu J, Neel L, Cheek J (2014). Death rates from human immunodeficiency virus and tuberculosis among American Indians/Alaska Natives in the United States, 1990–2009. Am. J. Publ. Health 104(S3):S453-S459.

Ricks PM, Cain KP, Oeltmann JE, Kammerer JS, Moonan PK (2011). Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the US, 2005-2009. PloS one 6(11):e27405.

Said HM, Kushner N, Omar SV, Dreyer AW, Koornhof H, Erasmus L, Gardee Y, Rukasha I, Shashkina E, Beylis N, Kaplan G (2016). A Novel Molecular Strategy for Surveillance of Multidrug Resistant Tuberculosis in High Burden Settings. *PLoS ONE* 11(1).

Schneider E (2005). Tuberculosis among American Indians and Alaska Natives in the United States, 1993-2002. Am. J. Publ. Health 95(5):873-880.

Scott C, Kirking HL, Jeffries C, Price SF, Pratt R (2015). Tuberculosis Trends-United States, 2014. MMWR: Morbidity mortality weekly report 64(10):265-269.

South Dakota Department of Health (2015). Tuberculosis Control Program Annual Report.

Stennis N, Trieu L, Perri B, Anderson J, Mushtaq M, Ahuja S (2015). Disparities in Tuberculosis Burden Among South Asians Living in New York City, 2001–2010. Am. J. Publ. Health 105(5):922-929.

Suwanpimolkul G, Grinsdale JA, Jarlsberg LG, Higashi J, Osmond DH, Hopewell PC, Kato-Maeda M (2014). Association between diabetes mellitus and tuberculosis in United States-born and foreign-born populations in San Francisco. PLoS ONE 9(12):e114442.

Terrazas A (2011). “Immigrants in New Destination States.” The Online Journal of thMigration Policy Institute. Accessed 3/5/2015 Available at: http://www.migrationpolicy.org/article/immigrants-new-destination-states .

Wieland ML, Weis JA, Olney MW, Alemán M, Sullivan S, Millington K, O'Hara C, Nigon JA, Sia IG (2011). Screening for tuberculosis at an adult education center: Results from a community-based participatory process. Am. J. Publ. Health 101:1264-1267.

Wieland ML, Weis JA, Yawn BP, Sullivan SM, Millington KL, Smith CM, Bertram S, Nigon JA, Sia IG (2012). Perceptions of tuberculosis among immigrants and refugees at an adult education center: a community-based participatory research approach. J Immigrant Minority Health 14(1):14-22.

Wieland ML, Nelson J, Palmer T, O'Hara C, Weis JA, Nigon JA, Sia IG (2013). Evaluation of a Tuberculosis Education Video Among Immigrants and Refugees at an Adult Education Center: A Community-Based Participatory Approach. J. Health Commun. 18(3):343-353.

Winston CA, Navin TR (2010). Birth cohort effect on latent tuberculosis infection prevalence, United States. BMC infect. Dis. 10(1):1.