International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-23
Available Online at http://acascipub.com/Journals.php

Research article

Regressively distinguishing the determination of coefficient of heterogeneous socio-demographic and clinical diagnostics human immunodeficiency virus covariates on Tuberculosis prevalence in a georeferenced grid-stratified zip-code, county level polygon.

Jeegan U. Parikh[a], Toni Panaou[a], Benjamin G. Jacob[a*]

[a]Department of Global Health, College of Public Health, University of South Florida, Tampa, Florida, United States

[*]Corresponding author. Email: bjacob1@health.usf.edu.
E-mail: jeeganparikh@health.usf.edu, apanaou@mail.usf.edu

Abstract:

The re-emergence of Tuberculosis (TB) and development of multidrug resistant tuberculosis (MDR-TB) is a perfect example of the failure of the global public health community in control of a disease which was once considered curable. The re-emergence of the disease has been associated with HIV/AIDS epidemic; however, no studies previously have been done which signifies the strength of the association of TB to HIV/AIDS. The delay in diagnosis and treatment of those suffering from TB has resulted in rise of MDR-TB cases which has raised alarm bells throughout the world. The association of HIV/AIDS and TB and development of drug resistance has also been researched but results have been inconclusive till now (Centers for Disease Control and Prevention 2008). The spread of the disease is through airborne transmission and thus various, heterogeneous, sociodemographic explanatorial factors affect the transmission of the disease (Centers for Disease Control and Prevention 2016a). The socioeconomic factors may put population of certain race/ethnic groups at higher risk of getting TB. The goal of the study is to distinguish how the different sociodemographic factors and

STDs affect the prevalence/incidence of TB. The incidence rate of TB in US is 3 per 100,000 persons in 2015, an increase over previous years(Centers for Disease Control and Prevention 2017). The county with high incidence of TB cases in Florida (Miami-Dade) was selected and various sociodemographic factors (African-American population, Hispanic population, land use land cover, population density and median household income) and number of HIV cases were modelled. A multiple linear regression model was created to determine the strength of association of these variables. The model rendered results which suggest a strong HIV association ($p<0.0001$) at 95% confidence interval with TB. In the model, the heterogeneous sociodemographic explanators which were found non-significant were rendered significant without the HIV variable, thus revealing the operational power of the HIV variable. Thus, HIV is a significant marker for TB stratification. The results of the model can be extrapolated for risk of TB in Hillsborough County and zip-codes could be classified into risk zones based on the number of HIV cases in zip-codes. The risk map created using Geographic Information System software can be employed for optimal surveillance, screening and targeting of interventions for prevention and control of TB.

Keywords: Tuberculosis; HIV; sociodemographic factors; multiple linear regression; GIS

1. Introduction

Tuberculosis is one of the re-emerging diseases along with malaria which the world had thought of having brought under control but has re-emerged due to the evaluation of the bacteria, mycobacterium tuberculosis (TB). The re-emergence of TB has been associated with trickling down of resistance among mycobacterium tuberculosis bacilli to various anti-tubercular drugs. The outbreaks of multidrug resistant tuberculosis (MDR-TB) and extensively drug resistant tuberculosis (XDR-TB) has raised alarm bells throughout the world and counters all the gains made in the control of tuberculosis. MDR-TB is defined as strain of mycobacterium tuberculosis resistant to isoniazid and rifampicin and XDR-TB is a strain resistant to isoniazid, rifampicin, fluoroquinolones and any of the second line of anti-tubercular injectable drug (kanamycin, amikacin or capreomycin) (Centers for Disease Control and Prevention 2016b).The prevalence of drug resistance is much higher among previous treated cases of tuberculosis than the new transmission of disease (Streicher et al. 2012,Moonan et al. 2013, Rifat et al. 2014 and Sahebi et al. 2016). People diagnosed with tuberculosis but did not complete the treatment or lost to follow up are at more risk of developing drug resistance (Shringarpure et al. 2015). Studies on association of HIV/TB co-infection in development of drug resistant tuberculosis has been

inconclusive till now with high prevalence areas for both disease showing significant association while low prevalence areas not finding any association (Suchindran et al. 2009, Sethi et al. 2013 and Lee et al. 2016.). Most of the cases found in the developed world are in foreign born or migrant population from higher endemic regions of TB (Haar et al. 2007, Hattori et al. 2016 and Sotgiu et al. 2017). Tuberculosis is spread by airborne transmission and various factors like socioeconomic status, density of area and immunity status may affect the transmission pattern. Migrant population usually lives in the densely population areas and thus are at a higher risk of transmission of tuberculosis (Bojorquez et al. 2013). Identifying population employing may identify at risk for tuberculosis, early diagnosis and treatment not only improves prognosis of disease but helps in control of TB epidemic and transmission of drug resistant tuberculosis.

In 2015, 9,557 cases of TB were reported in United States with an incidence rate of 3.0 per 100,000 persons, which is an increase of 1.6% over the cases reported in 2014(Centers for Disease Control and Prevention 2017). In 2014-2016, Florida ranked 9th in the United States with 1,836 cases of TB. The counties in Florida with the largest number of cases are Miami-Dade (371), Broward (201), Orange (186), Duval (147) and Hillsborough (133). Hillsborough County has an incidence rate of 3.3 per 100,000 persons which is higher than national incidence rate of 3.0 per 100,000; thus an area of interest for geolocating clusters and developing a cost-effective TB control strategy (Florida Department of Health 2017a).

Jacob et al. (2014) determined that a newer predictive spatial statistics algorithm is a vigorous tool for geolocating areas with high prevalence areas geographically for targeting resource allocation to those areas for more cost effective control of tuberculosis. Further, this research discovered that the spatiotemporal clinical regression model residual outputs can be affected by arbitrary variation attributable to population variability leading to a loss of statistical power when cases are assigned to subgroups leading to an awkward situation when mapping and analyzing incidence, time series, TB data with conventional hierarchical statistical approaches. Jacob et al. (2014) employed a compound Poisson approach for detection of residual clustering of varying and constant, georeferenced, explanatory, time series, covariate coefficient estimates by testing individual areas pooled with their neighbours to reduce the standard deviation within subject. Besag and Newell (1991) had proposed a hierarchical, explanatory, cluster-based, regression model to screen for collections of childhood leukaemia cases in northern England; whereby each georeferenced classified sub-location was based on the number of neighbours that had to be combined in order to contain a minimum number of cases (i.e., cluster sampled size). This method scanned the data for collections of cases that appeared to be unusual clusters.

Spatial statistics may prove to be useful in quantitating local clustering in explanatory, spatiotemporal, clinical, endemic TB transmission data (Jacob et al. 2014). Previously research has been done to develop algorithms using Statistical Analysis Software (SAS)/ArcGIS to locate high discrepancy regions geographically with methods ranging from fast heuristics in special cases to digitized grid based metrics (Gandhi et al. 2006). Using this spatial scan statistics and explanatory time series approximation cluster based error-detection algorithms, the factors affecting computational studies of spatiotemporal TB transmission can be created based on sampled clinical and explanatory, observational, covariate estimates (Jacob et al. 2014). Using explanatory, residual-based, cluster-based, error diagnostics for determining multivariate heteroscedastic parameters, like from hierarchical explanatory intra-cluster-based regression model covariates for identifying distribution of those at high risk can be identified (Jacob et al. 2014). This article proposed a method to address the problem that can arise when covariates in a forecast, vulnerability, endemic, TB, regression model setting are not Gaussian, which may give rise to approximately mixture-distributed errors, or when a true mixture of regressions produced the data. The authors began with a non-Gaussian, mixture-based, marginal, variable screening, followed by fitting a full but relatively smaller mixture regression model to the selected TB data with help of a new penalization scheme. Under certain regularity conditions, the new screening procedure was shown to possess a sure screening property even when the population was heterogeneous. The authors further prove that there exists a capture point in the associated regression screen plot which results in a consistent estimator of the set of active covariates in the model. By simulations, the authors demonstrated that the spatiotemporally dependent procedure can substantially improve the performance of the existing procedures in the context of variable screening and TB data clustering.

Previous research done found an association between HIV and TB however, the strength of association has never been determined. Along with clinical diagnostics we employed various heterogeneous sociodemographic, explanatory covariates for association with TB. Linear and non-linear, analytical, regression analysis can be done to find if a statistically significant association exists (Jacob et al. 2010). The objectives of the project is to develop a multiple regression model using multiple predictor variables, find the strength of association and develop TB control strategies by determining the covariate coefficient estimates based on data (Jacob et al. 2014). We assumed that the model formed will help in distinguishing the significance that various heterogeneous sociodemographic and clinical diagnostics (HIV) covariates have on TB.

2. Materials and Methods

2.1 Study site

Florida is a state located in the south-eastern region of the United States, bounded by Gulf of Mexico in west, Alabama and Georgia in north, Atlantic Ocean in the east and Straits of Florida and Cuba in south. The most populous urban area in Florida is Miami metropolitan area. The city of Tallahassee is the state capital.

A peninsula between the Gulf of Mexico, the Atlantic Ocean, and the Straits of Florida, it has the longest coastline in the contiguous United States, approximately 2,170 km, and is the only state that borders both the Gulf of Mexico and the Atlantic Ocean. Much of the state is at or near sea level and is characterized by sedimentary soil. The climate varies from subtropical in the north to tropical in the south

To quantitate hyperendemic foci for TB on the basis of landscape heterogeneity in Hillsborough County, landscape heterogeneity for clusters of tuberculosis in the Miami-Dade County, Florida was examined. Miami-Dade County (see Figure 2(a)) is located on the south-eastern side of Florida with a population of 2,496,435 (U.S Census 2010).

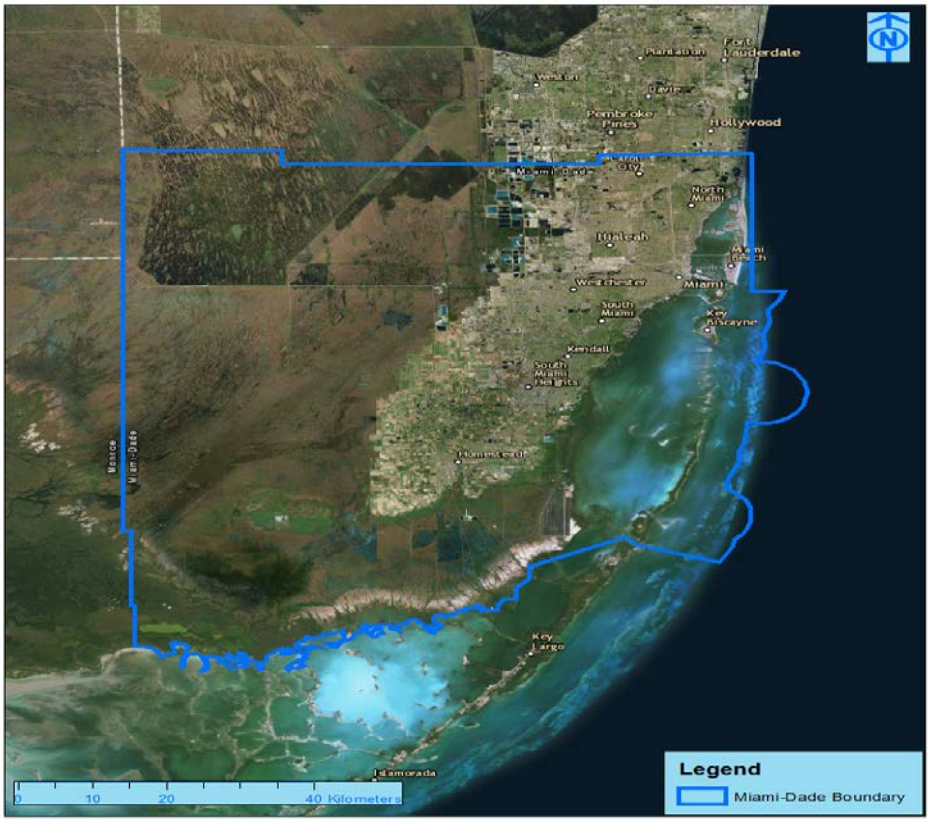

Figure 2(a) Miami-Dade County

During the 2014-2016, the county had the highest number of cases of TB in a county in Florida (371) with an incidence rate of 4.7 per 100,000 population, which is significantly higher than the incidence rate for TB in Florida (3 per 100,000 population) (Florida Department of Health 2017a). Hillsborough county (see Figure 2(b)) is the fourth largest county of Florida with population of 1,229,226 (U.S Census 2010). It had an incidence rate of 3.3 per 100,000 with 133 cases of TB in 2014-2016 (Florida Department of Health 2017a).

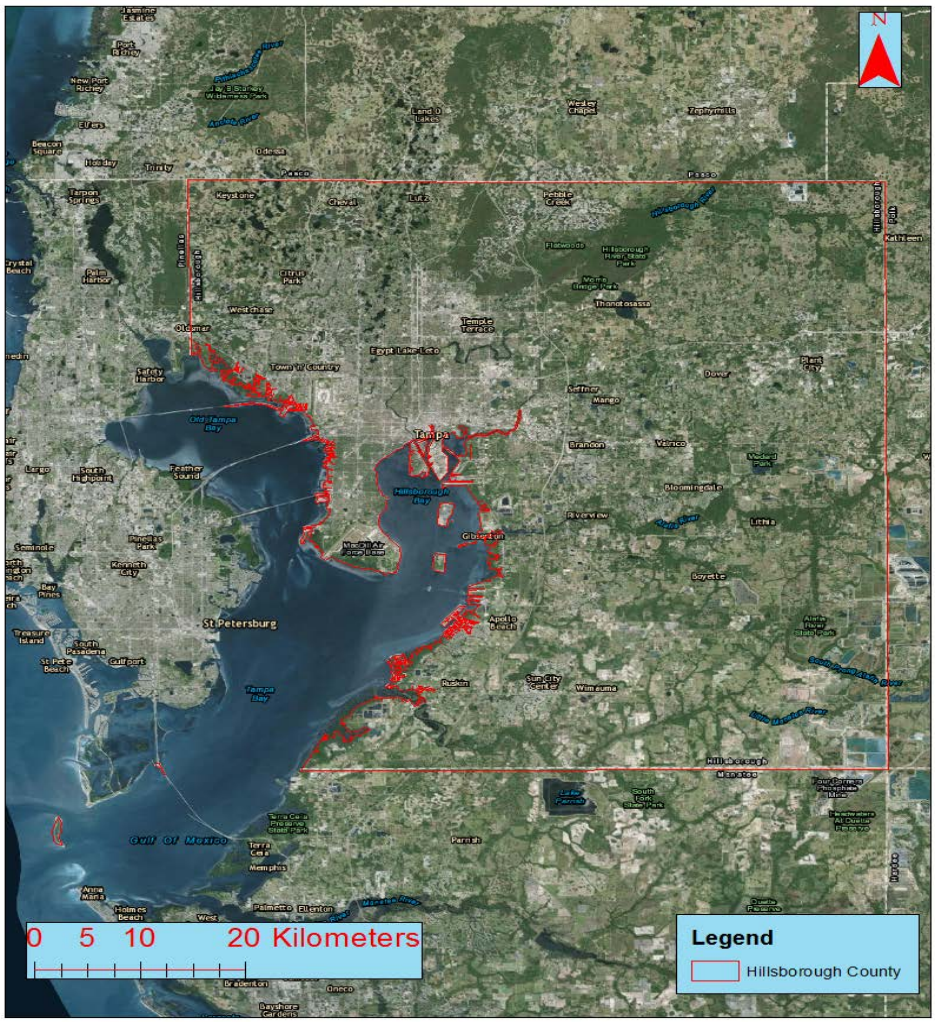

Figure 2(b) Hillsborough County

2.2 Subject and setting

This is an observational study comparing the clusters of TB in high prevalent area with the land cover usage, socioeconomic indicators and clinical diagnostics (HIV); using the result to develop an effective TB control strategy for Hillsborough County. Data for the study is acquired from the Florida Health charts (http://www.flhealthcharts.com/charts/default.aspx) while demographic data was obtained from 2010 US Census. For the study, county wise data of the HIV cases during the same timeframe was obtained from the

Florida Health charts. Miami-Dade County had 3,692 cases of HIV with an incidence of 47 per 100,000 population (Florida Department of Health 2017b). The TB cases and HIV cases in the zip-codes were obtained by multiplying the population in zip-code with the incidence of TB/HIV in the county.

TB is an AIDS defining illness and thus HIV cases can be positive indicator to look for endemic TB (Centers for Disease Control and Prevention 2008). TB is transmitted by airborne method and thus population density is used as a predictor variable to look for an association (Centers for Disease Control and Prevention 2016a). The reason for selecting African-American population, Hispanic population, population density or average income is to look for an association of socioeconomic status and incidence of TB cases.

2.3 Demographics

According to the 2010 Census, Miami-Dade County has a population of 2,496,435 with 73.8% being White, 18.9 % African American, 1.5% Asian, and 0.2% American Indian. 65% of the population is Hispanic (U. S. Census Bureau 2010). The demographics (total population, African American) were obtained on the basis of zip-codes in Miami-Dade County. Population density (average number of people living in the house) and average income of the household were also obtained on the basis of zip-codes and all these were used as predictor variables to create a multiple linear regression model.

2.4 Land use land cover

A land use land cover (LULC) map of Miami-Dade county was created using ESRI ArcMap extension of ArcGIS software. A land usage map of Florida was obtained and localized to the County (Florida Department of Environmental Protection Geospatial open data, 2017). The LULC map for the county was divided into various polygons based on the zip-codes (from ArcGIS online) in the Miami-Dade County. Every zip-code (polygon) was classified as Urban Residential, Urban Commercial and Agriculture as shown in figure 2(c). The zip-codes were classified into the 3 land cover areas based on the highest proportion of land use in that zip code e.g. if a zip-code had 70% Urban Residential, 20% Urban Commercial and 10% Agriculture it was classified as Urban Residential. The LULC data was inserted into Excel dataset to then extract hyperendemic foci of TB associated land cover.

International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-23
Available Online at http://acascipub.com/Journals.php

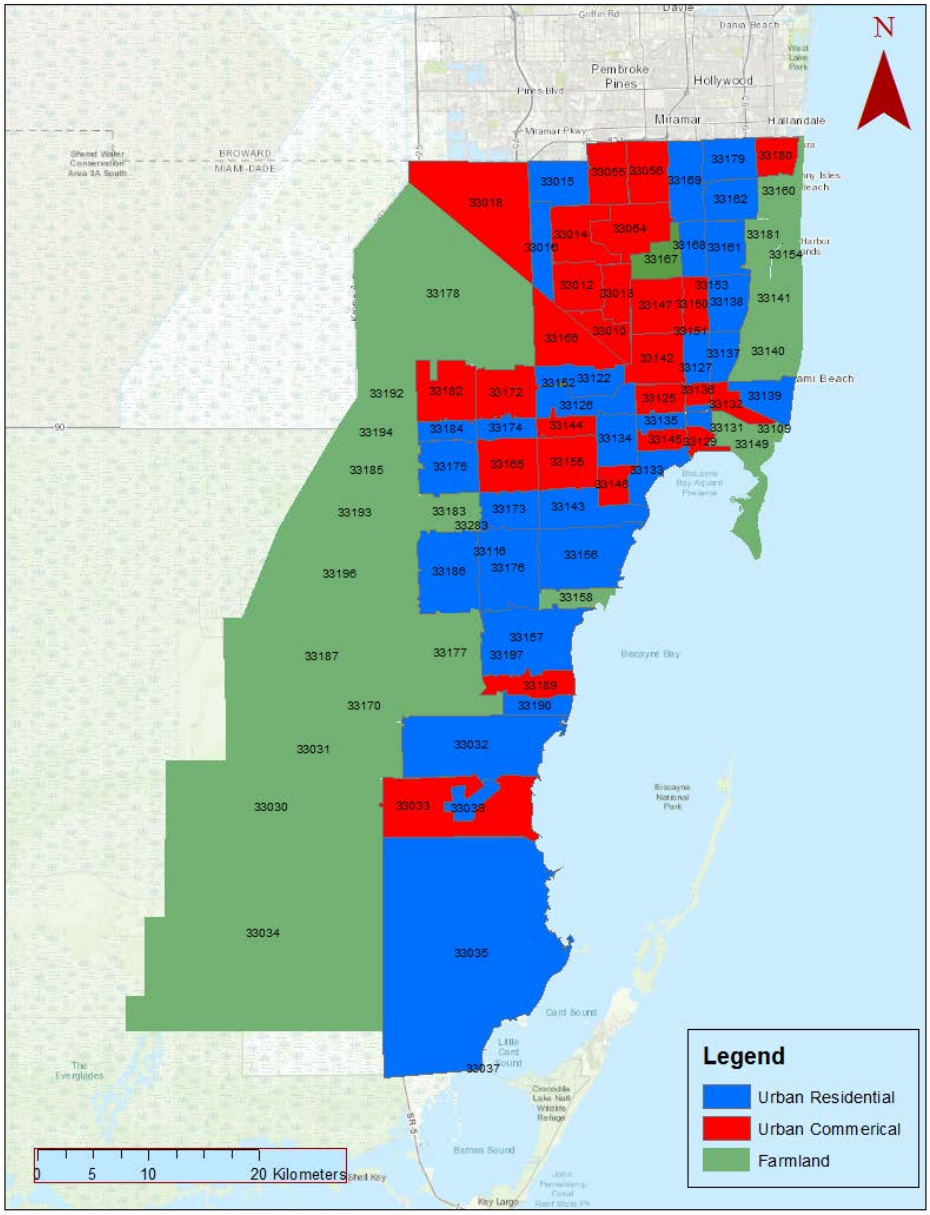

Figure 2(C) Land Use Land Cover in Miami-Dade County (zip-code level)

2.5 Regression analysis

All the data collected were analysed with SAS (Statistical Analytical Software). The relationship between the geosampled, zip-codelevel, endemic sociodemiographic, covariates was investigated by single variable regression analysis in PROC REG. A regression model was employed, as is a standard practice for the analysis of the county level data.

The regression analyses assumed independent counts (i.e., *Ni*), taken at multiple, geosampled, georeferenced, zip-code level sub-geolocations $i = 1, 2, \ldots, n.$ The spatiotemporal-related, zip-code level, discrete integer 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 a interpretively geosampled, endemic transmission-oriented, explanative 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 parameterizable non-redundant covariates in the endemic, transmission-oriented, interpretively interpolative, time-series, operationalizable, epidemiological, prognosticative, zip-code level, risk model. The dependent variable was zip-code level cases. The regression analysis was performed in the SAS using a 95% confidence interval. There was considerable overdispersion in the model. Thus, we ran a regression model backwards to remove the covariates not showing significant association. To determine the degree of significance of the clinical diagnostics (HIV), a regression analysis was run taking the clinical, diagnostic,regressor out of the model. A regression analysis was also run using county level georeferenced, socioeconomic and clinical diagnostics data to establish the strength of the model.

3. Results

The association between county level cases and each explanatory, individual, potential, endemic, TB transmission-oriented, socio-demographic regressors was investigated using a single variable regression analysis in PROC NL MIXED. The first line of the code started the PROC REG command. The second line of code specific each, endemic, transmission-oriented, socio-demographic and zip-code based risk model [i.e. the model without random intercept, value]. The third line created a value which was equal to the fixed part of the model and a random intercept term *u*. The model statement precisely said that the parameterizable covariates estimators were distributed normally with a mean of xb and variance *s2*. This statement defined the random effected *u* at the same time of quantitating the normally, distributed, operationalizable, time series data with a mean of zero and a variance term. The *s2u* was solved optimally by doing this. The units in level 2 were identified by subject = id in PROC REG.

The last two lines of code in PROC REG were predictive statements. Usually only a single set of predicted values are created when constructing a time-series, clinical, and remote-specified, endemic TB, georeferenceable, forecasting, epidemiological, zip-code level, socio-demographic, probabilistic, risk models but we generated two predicted values. The predict statement in SAS rendered the explanatorial, zip-code level, time-series, predicted values for the model. The model identified xb and also created an output dataset called output-fixed. The second predict statement generated the regressed, endemic, zip-code level TB, transmission-

oriented, predicted values which not only included the quantified, fixed portion, randomized estimates but also estimate of the random intercept.

We employed the regression line $(\mathbf{y}_i - \bar{\mathbf{y}}) = (\hat{\mathbf{y}}_i - \bar{\mathbf{y}}) + (\mathbf{y}_i - \hat{\mathbf{y}}_i)$ to generate a R^2 value where the first term was a total variation in the response y (zip-code level TB cases) and the second term was the variation in mean response based on the asymptotical, normalized, parameterizable, socio-demographic, covariate estimators. The third term was the rendered residually forecasted, elucidative regressed endemic, transmission-oriented, derivative values in the operationalizable, time-series, endemic, transmission-oriented, and risk model,interpolative derivatives. Squaring each of these terms and adding over all the observations generated the equation $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$. This equation was written as SST = SSM + SSE, where SS was notation for sum of squares and *T*, *M*, and *E* were the notation for total quantized model error estimates. The square of the sample correlation was then equal to the ratio of the estimates while the sum of squares was related to the total sum of squares: r^2 = SSM/SST. This formalized the interpretation of R^2 for explaining the fraction of variability in the epidemiological, zip-code level data explained by the regression model. The sample variance s_y^2 was equal to $\sum\frac{(y_i - \bar{y})^2}{n-1}$, which in turn was equal to the SST/df, the total sum of squares divided by the total df.

A regression equation was constructed by employing the mean square model (i.e., MSM) = $\sum\frac{(\hat{y}_i - \bar{y})^2}{l}$, which was equal to the SSM/df. The corresponding MSE was $\sum\frac{(y_i - \hat{y}_i)^2}{n-2}$ which was determined to be equal to SSE/df and the quantitated, time-series, operationalizable, zip-code level, TB endemic, transmission-oriented, explicatory, georeferenceable estimate of the variance about the regression line (i.e., σ^2). The MSE is an estimate of σ^2 for determining whether or not the null hypothesis is true (Draper and Smith 1981).

For robustly, parsimoniously, quantizing, the operationalizable, transmission-oriented, explanatory, interpolative, endemic, zip-code level TB, georeferenecable prognosticators (*p*)a DFM, was generated which we noted was equal to *p* and the error degrees of freedom (DFE). This product was also equal to $(n - p - 1)$, and the total degrees of freedom (DFT) which was subsequently equal to $(n - 1)$. The sum of DFM and DFE was determined. The relationship between the mean of the response variable (i.e., zip-code level case count) and the level of the

explantorial, parameterizable, zip-code, covariate coefficients in the regression equation were assumed to be approximately linear (i.e., straight line). The table 1 classified each time series, clinical, field and remote asymptotical, unbiased, covariate estimator in SAS.

Table 3(a) The time series regressed endemic TB regression-based model parameter estimates

Source	Sum of Squares	Formula
Model	$\sum(\hat{\mathbf{y}}_i - \overline{\mathbf{y}})^2$	SSM/DFM
Error	$\sum(y_i - \hat{\mathbf{y}}_i)^2$	SSE/DFE
Totzl	$\sum(y_i - \overline{\mathbf{y}})^2$	SST/DFT

In the multiple regression analyses, the test statistic MSM/MSE had a $F(p, n - p - 1)$ distribution. The null hypothesis was $\beta_1 = \beta_2 = \ldots = \beta_p = 0,$ and the alternative hypothesis was evaluated by encompassing the zip-code level, geosampled, endemic, transmission-oriented, predictive, epidemiological, socio-demographic risk parameters $\beta_j \neq 0,\ j = 1, 2, \ldots, p.$ The *F* test did not indicate which of the parameters $\beta_j \neq 0$ nor, which was not equal to zero only that at least one of them was linearly related to the response variable.

The ratio SSM/SST $= R^2$ (i.e., squared multiple correlation coefficient) was thereafter the proportion of the variation in the response variable that was explained by the zip-code TB polygonized data. The square root of R^2 (i.e., the multiple correlation coefficient) was the correlation between the explanatorily, time-series, empirical observations (i.e., y_i) and the fitted values (i.e., $\hat{\mathbf{y}}_i$). Additionally, from the sampling distribution generated from the *t* parameters, the probability of obtaining an *F* was calculated. There were only two means to compare, the *t*-test and the *F*-test, which coincidentally were equivalent. The relation between ANOVA and *t* was then given by $F = t^2.$ Thereafter; significant differences by ANOVA were noted for the quantitated mean numbers of explicative, endemic, transmission-oriented, operationalizable, time series, iteratively interpolative, asymptotically normalized, georeferenceable, data, feature attributes captured throughout the sampling frame $(F = 44.7,\ DF = 1).$

We then generated a stepwise backward regression model to tease out any propagation probabilistic uncertainties (heteroskedastic noisy parameters) in the zip-code level, socio-demographic, epidemiological, forecast vulnerability, and endemic model. In each step, a zip-code level, socio-demographic, explanatorial prognosticator was considered for addition to or subtraction from the set of the diagnostic variables based on some pre-specified criterion.

In the regression probabilistic paradigm, constructed in PROC REG, we were available to distinguish statistical significance of various heterogeneous, socio-demographic and clinical, diagnostics covariates of tuberculosis. Interestingly, the HIV predictor was found to be statistically significant at 95% confidence interval [see Figure 3 (a-e)]. The model did not violate any of the regression assumptions (the variance was homoscedastic). In the HIV-TB interaction model, the covariates that were seemed significant without the sexually transmitted disease (STD) forecastable, vulnerability variable were found to then be rendered non-significant [see Figure 3(a-h)]. This deterioration of statistical significance amongst the socio-demographics and clinical iteratable regressors demonstrated the operational power of the HIV explanatory variable. The residual regression dataset revealed that HIV is a significant marker for county-level TB stratification.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	39.24365	6.54061	255202	<.0001
Error	67	0.00172	0.00002563		
Corrected Total	73	39.24536			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.00238	0.00403	0.00000890	0.35	0.5577
HIV_cases	0.09919	0.00022131	5.14870	200892	<.0001
AA	-0.00002213	0.00006050	0.00000343	0.13	0.7157
Hispanicpop	-1.42234E-8	1.425656E-7	2.551009E-7	0.01	0.9208
Popdensity	0.00200	0.00154	0.00004343	1.69	0.1975
Income	1.220996E-8	3.714055E-8	0.00000277	0.11	0.7434
LULC	-0.00098893	0.00081533	0.00003771	1.47	0.2294

Figure 3(a) Regression output for TB cases (zip-code level)

International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-23
Available Online at http://acascipub.com/Journals.php

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Hispanicpop	Hispanicpop	5	0.0000	1.0000	5.0100	0.01	0.9208
2	Income	Income	4	0.0000	1.0000	3.2314	0.22	0.6370
3	AA	AA	3	0.0000	1.0000	1.8818	0.67	0.4168
4	LULC	LULC	2	0.0000	1.0000	0.9042	1.05	0.3080
5	Popdensity	Popdensity	1	0.0000	1.0000	0.2547	1.39	0.2421

Figure 3(b) Stepwise backward regression summary of backward elimination for TB cases (zip-code level)

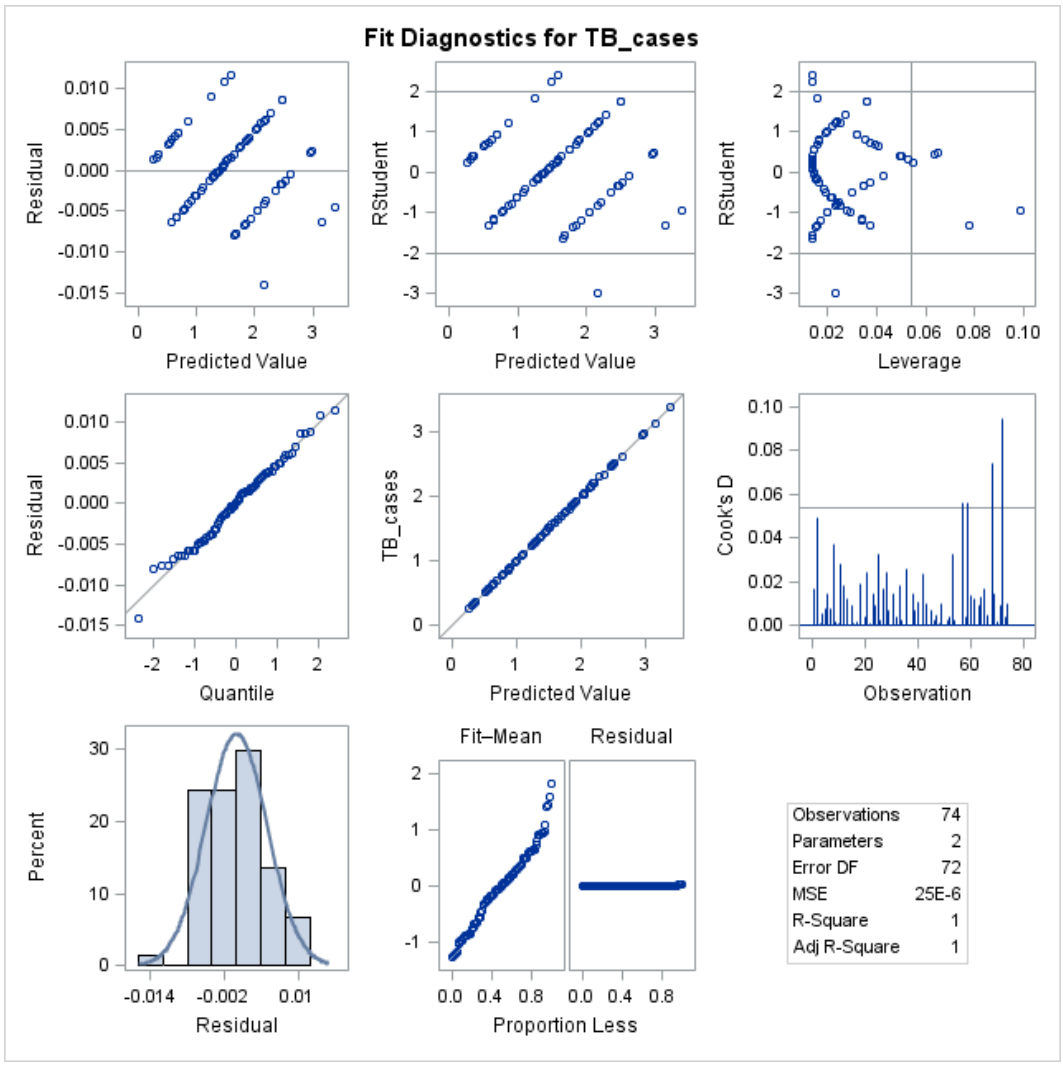

Figure 3(c) Fit diagnostics for TB cases (zip-code level)

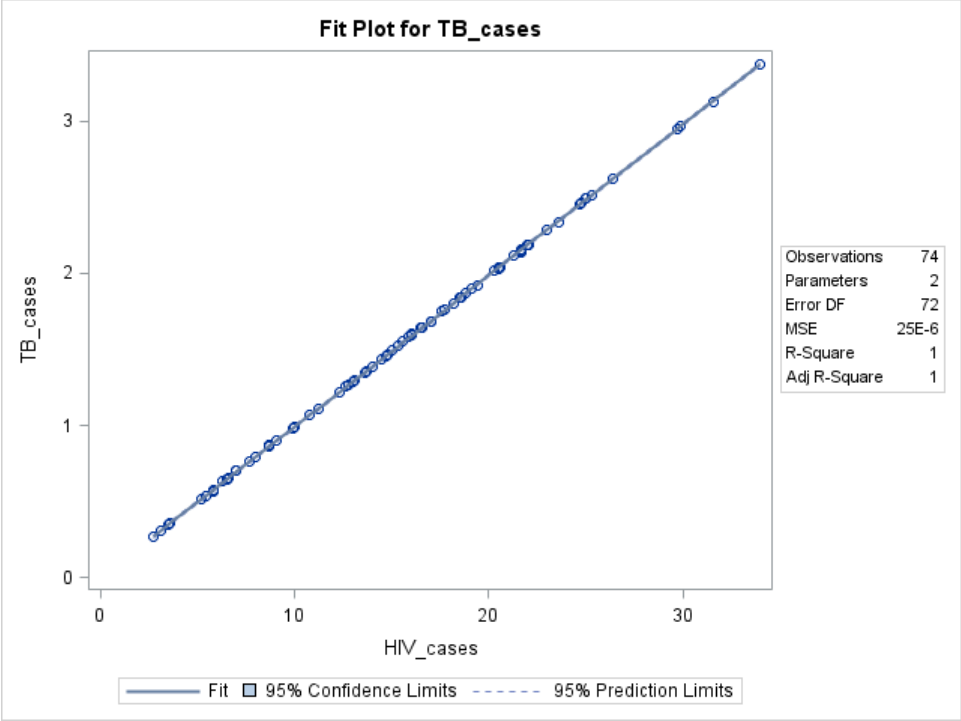

Figure 3(d) Fit Plot for TB cases (zip-code level)

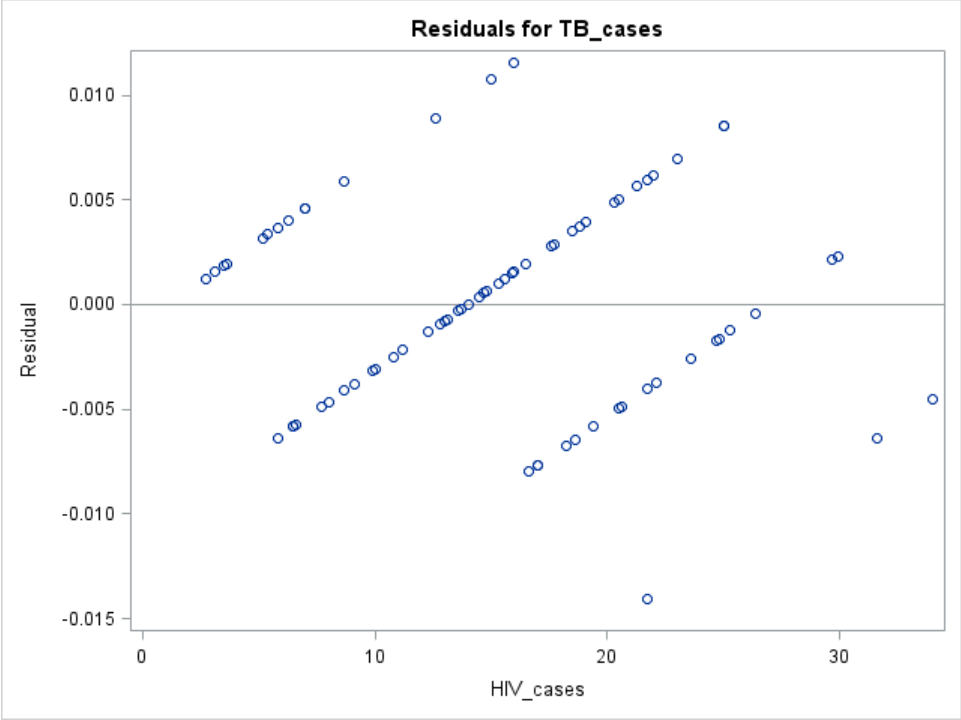

Figure 3(e) Residuals for TB cases (zip-code level)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	34.09495	6.81899	90.03	<.0001
Error	68	5.15041	0.07574		
Corrected Total	73	39.24536			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.39887	0.21372	0.26383	3.48	0.0663
AA	0.02008	0.00221	6.25998	82.65	<.0001
Hispanicpop	0.00005788	0.00000328	23.56140	311.08	<.0001
Income	0.00000685	0.00000184	1.04836	13.84	0.0004
LULC	-0.06092	0.04372	0.14703	1.94	0.1681
Popdensity	-0.25186	0.07785	0.79281	10.47	0.0019

Figure 3(f) Regression output for TB cases (zip-code level) without STD

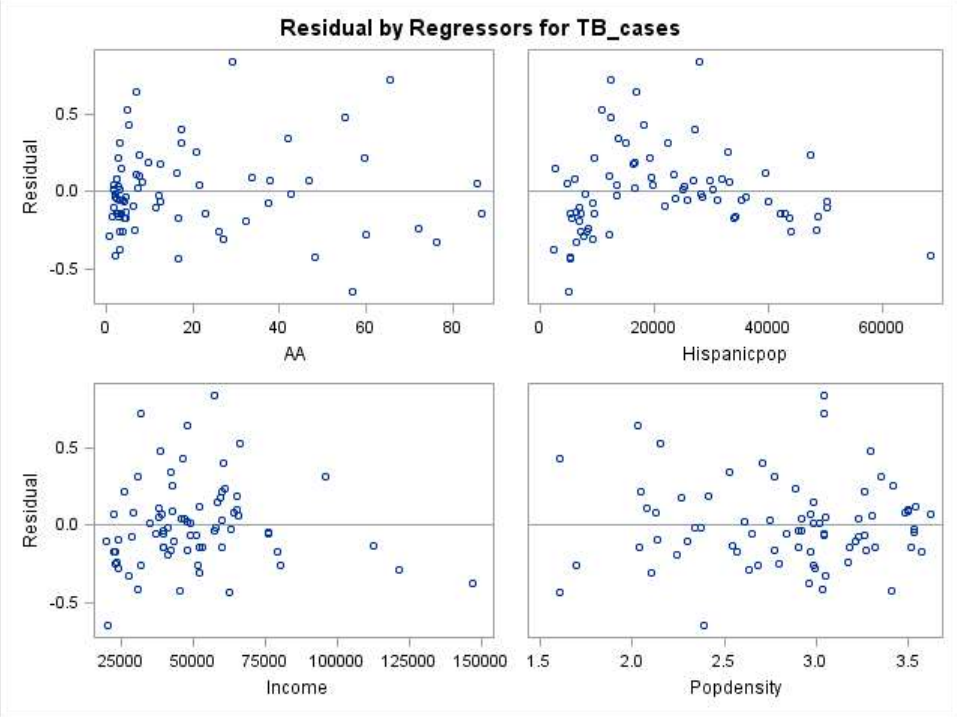

Figure 3(g) Residual by regressors for TB cases (zip-code level) without STD

International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-23
Available Online at http://acascipub.com/Journals.php

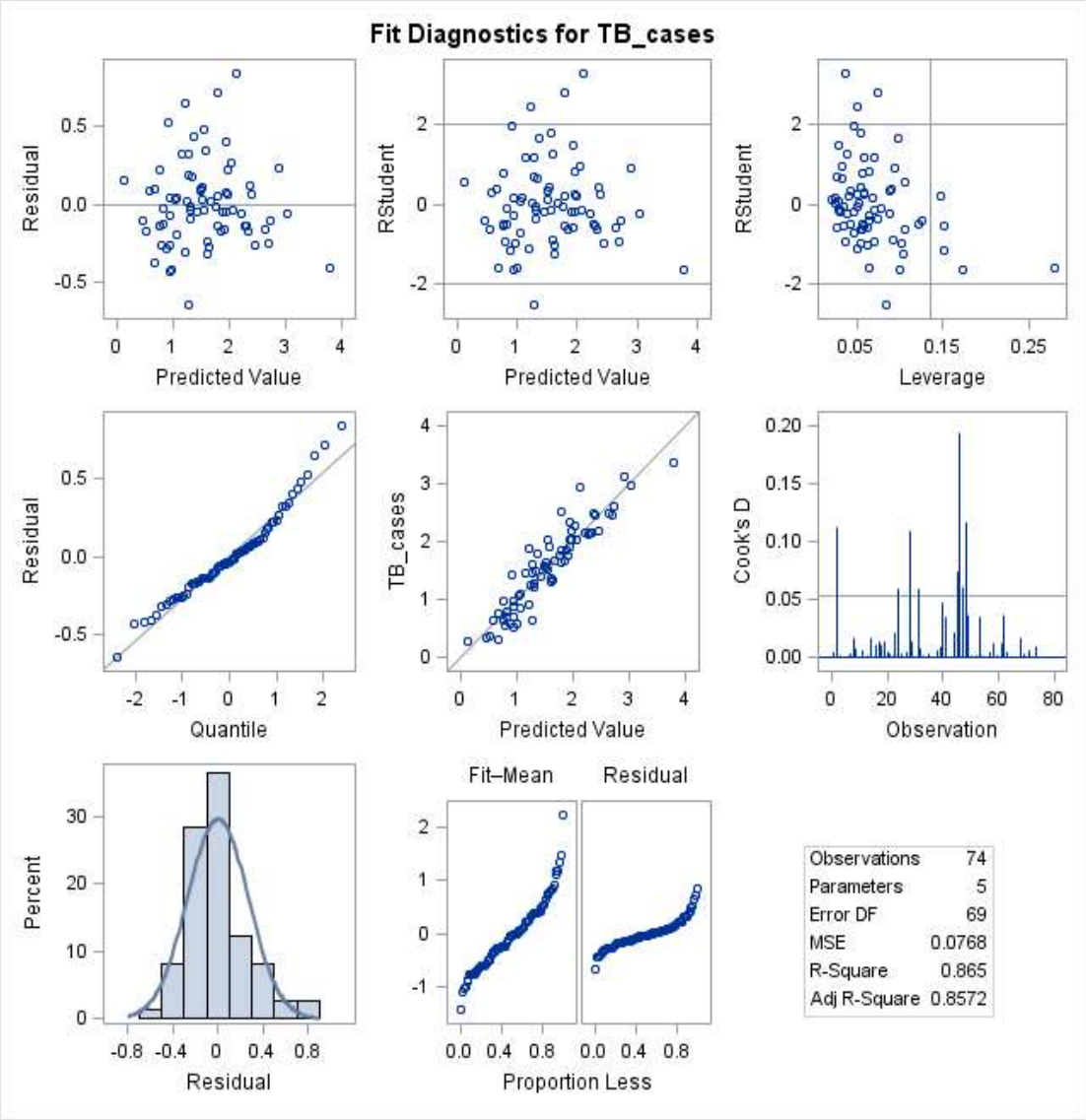

Figure 3(h) Fit diagnostics for TB cases (zip-code level) without STD

At the confidence interval of 95%, the zip-code level, TB cases had a very strong association with the number of HIV cases ($p < 0.0001$). The association with other socio-demographic explanators(African-American population, Hispanic population, Population density, Median income) to the number of tuberculosis cases was not found to be significant. Similarly, the result of multiple regression did not suggest any hyperendemic foci of Tuberculosis based on land usage. However, the p-value of 0.2294 suggests a future epidemiological study may be needed to investigate the effect of geomorphological land cover has on the tuberculosis cases.

To validate the model further, the county -level TB cases were regressed against HIV cases and socio-demographic indicators. At 95% confidence interval HIV cases revealed a strong association ($p < 0.0001$) (see

Figure 3(i-k)). Amongst the socio-demographic indicators, only median income of household showed some significance ($p < 0.0255$).

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-7.29848	8.45350	19.07132	0.75	0.3913
HIV_cases	0.10142	0.00441	13555	529.79	<.0001
Aapop	0.10370	0.07264	52.14039	2.04	0.1585
Hispop	-0.01443	0.08092	0.81367	0.03	0.8591
Popdensity	-0.47660	2.91438	0.68425	0.03	0.8706
Medianincome	0.00020399	0.00008909	134.11963	5.24	0.0255

Figure 3(i) Regression output for TB cases (county level)

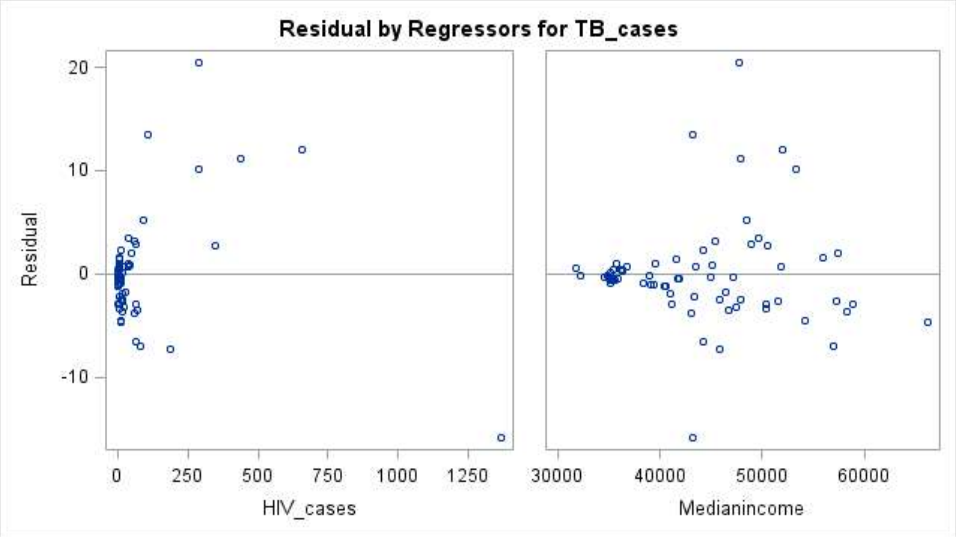

Figure 3(j) Residual regressors for TB cases (county level)

International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-23
Available Online at http://acascipub.com/Journals.php

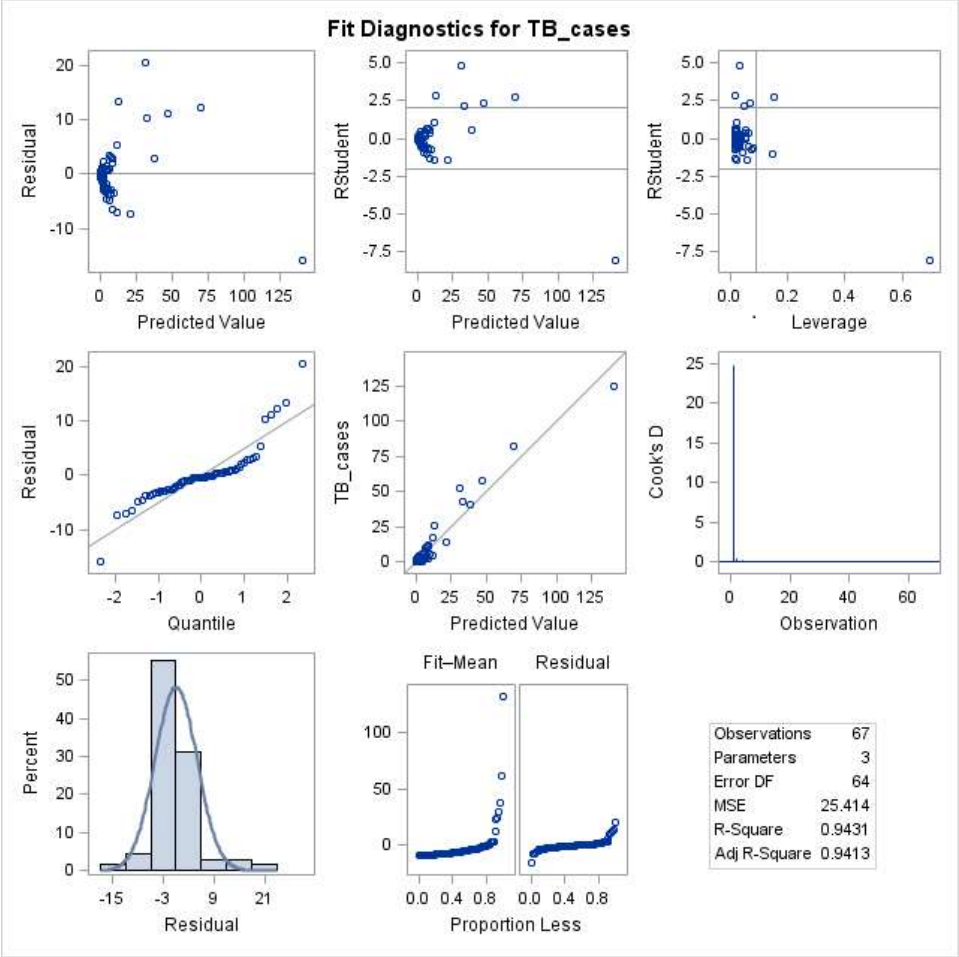

Figure 3(k) Fit diagnostics for TB cases (county level)

Hence, a zip-code level model created can be employed to geolocate TB clusters in Hillsborough County by cluster stratified of HIV. The zip-code with highest number of TB cases in Miami was used to deduce the predictor value of 10 HIV cases for 100% chance of 1 TB case. The Hillsborough County was than divided into 4 zones based on probability of risk in each zip-code. Even probability of 1 case was counted as high risk, 50% as moderate, 0-50% as mild and no risk (see Figure 3(l)). The one with high and moderate risks can be specifically targeted initially for TB screening and therapy for an effective TB control strategy.

International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-23
Available Online at http://acascipub.com/Journals.php

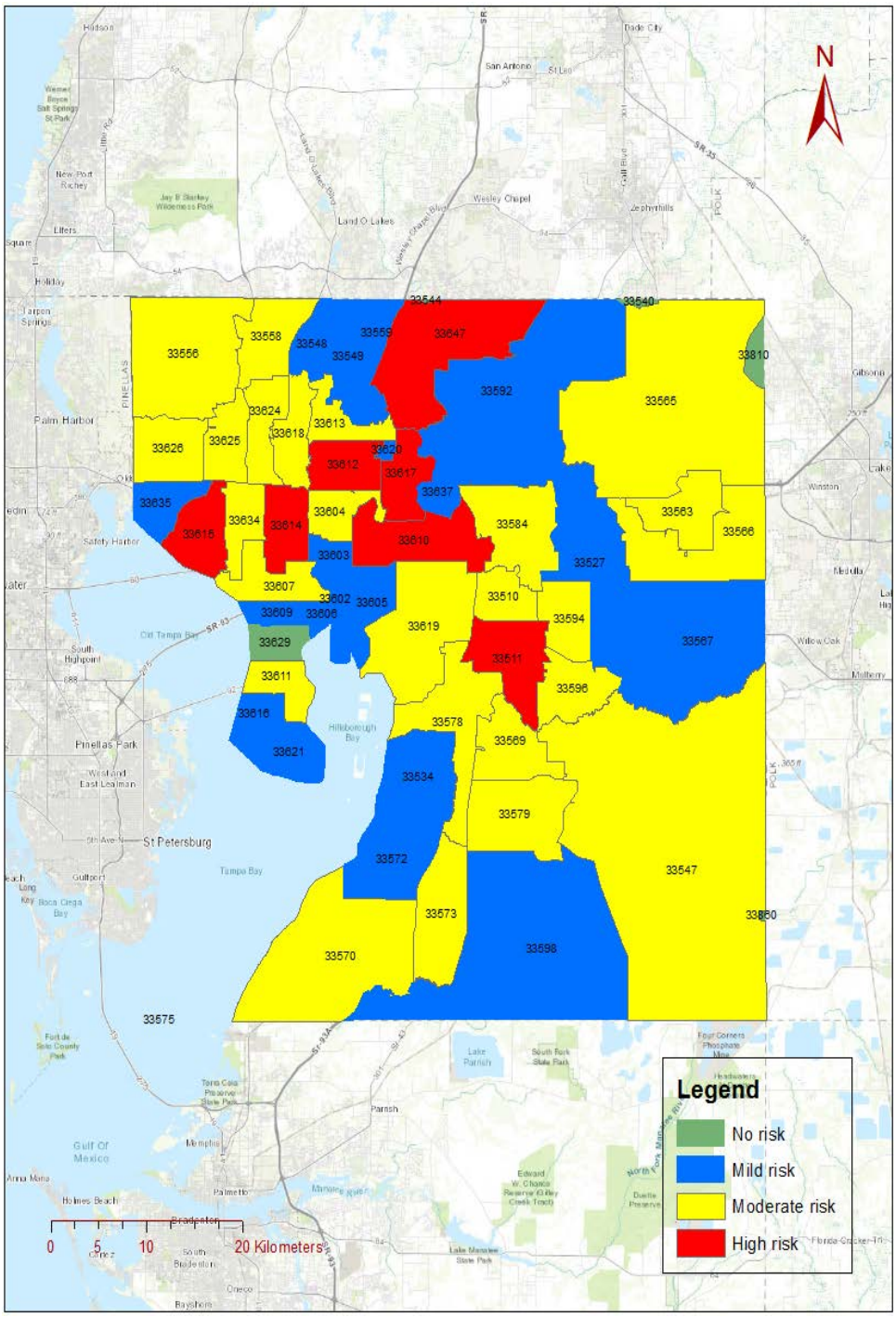

Figure 3(l) Risk for tuberculosis in Hillsborough County (zip-code) based on HIV.

The linear regression residuals indicated an inappropriate model fit due to over dispersion caused by outliers. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution (i.e., probability distributions whose tails are not exponentially bounded)(Draper and Smith 1981). Although the linear probabilistic paradigm was robust for the examined TB regression covariates it may be advisable to utilize spatial autoregressive models to determine the

actual weight of the HIV prognosticator. Although, Jacob et al. 2014 constructed multiple autoregressive spatially dependent forecast vulnerability models, the authors never examined an HIV covariate. Hence, it would be appropriate to use autocorrelation models and Bayesian evidential properties to quantitate the weight of the HIV variable within a TB paradigm.

Bayesian inferencial treatment has the ability to properly account for high variance of estimates in geographic areas and clarify overall spatial trends and patterns, regardless of distribution of data (Hastie and Tibshirani 1990). There is a great deal of literature on Bayesian model averaging for non-spatial regression models. For example, work by Fernandez et al. (2001) considered cases where the number of possible models was sufficiently large so that calculation of posterior probabilities for all models was difficult or infeasible. A Markov chain Monte Carlo (MCMC) model comparison methodology proposed by Madigan et al. (1995) has gained popularity in the mathematical statistics and econometrics literature. An extension to spatial autoregressive regression models is provided by Lesage and Parent (2007). Lesage and Fischer (2008) include simultaneous comparison of models based on both alternative explanatory variables and spatial weight matrices, albeit concentrating on the class of k-nearest neighbour spatial weight matrices. From a technical point of view, employing numerical integration techniques from these models may help obtain posterior model probabilities for endemic, TB-related specifications with different k-nearest spatial weight matrices which may then be usable to obtain Bayesian HIV modelled averages estimates. The computational costs of this procedure, makes it an impractical choice for a large set of alternative, spatial weight, TB-related, time series matrices.

A methodology may improve on Lesage and Fischer (2008) by adopting Bayesian information criterion (BIC) posterior model, grid-stratfied, TB weights to overcome such computational costs. In doing so, a county-level epidemiological model may allow for the consideration of a wide range of weight matrices for unbiasedly quantitating georeferenceable, zip-code, forecast vulnerability, model estimators as potential spatial structures which may underly the spillovers in the sample related data.

Many Bayesian approaches for analyzing spatial disease patterns focus on mapping spatially smoothed diseases rates (Clayton and Kaldor 1987). Mapping parameters in a spatial Bayesian probabilistic regression matrix may produce stable estimates for the cell-specific TB disease rates at the county-level, by shrinkage to the overall rate or by averaging over neighbouring, HIV, grid-stratified cells. A Bayesian description can be provided employing a simple exact uncertainty spatially-dependent cluster-based detection algorithm in an

ArcGIS cyberenvironment for optimally quantitating large spatial regions in a highly TB infected, county study area.

4. Conclusion

In this research, determining the difference in significance of various heterogeneous socio-demographic factors and HIV on TB, a multiple, linear, regression model using, zip-code level cases with time series, endemic TB zip-code level, socio-demographic, clinical diagnostics regressors is created. The model does find a strong association between Tuberculosis and HIV case, but that is expected as re-emergence of tuberculosis has been linked to HIV emergence. However, once HIV is taken out and only socio-demographic factors are regress all of the regressors show a significance association. This shows the operational power of HIV variable. The study uses county level data of number cases to calculate zip-code level cases based on the population. This is one of the biggest fallacies of the study as it uses an assumed data and not real data. Even for the risk prediction, assumed data of number of HIV cases at zip-code level in Hillsborough County is used. In addition, there could be multiple land use in various zip-codes and assuming it as one land use based on highest percentage may introduce a bias in the result. With HIV cases showing such high association with TB, a strategy to look only for HIV cases/clusters for TB therapy/surveillance will be as effective as surveillance of entire population and have better cost-benefit analysis. Hence, the model is extrapolated to look for HIV cases and classifying zip codes into risk zones for the surveillance, screening and effective use of resources for the control of Tuberculosis. HIV cases have shown a significant propensity for socio-demographic estimators like African American population, Hispanic population, population density and median income. Thus people belonging to these socio-demographic strata may be at increase risk of HIV and thus an increase risk of developing Tuberculosis. The model does however suggest a need for a further study based on the real data to quantitative land usage heterogeneity in determining clusters of tuberculosis. To determine the exact location of clusters of HIV in various zip-codes for prediction of future cases of TB, a study using spatial autoregressive and Bayesian probability model can be done in future. TB pattern can be better predicted if their spatio-temporal structure is understood. Measures of spatial association like Local Moran's I can be used in future to find if there is a critical inflection point when the disease changes its character or not. (Hardisty and Klippel2010). Currently, county wise data for migration population is not available, but with the developed world showing most of TB cases in migrant population (Haar et al. 2007, Hattori et al. 2016 and Sotgiu et al. 2017) that sociodemographic covariate could be included in future studies.

References:

1. Besag, J. and Newell J, 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistics Society A*; 154: 143-155.
2. Bojorquez, I. and Barnes et al., 2013. Multidrug-resistant tuberculosis among patients in Baja California, Mexico, and Hispanic patients in California. *American Journal of Public Health*, *103*(7), 1301–1305.
3. Centers for Disease Control and Prevention, 2008. *AIDS-Defining conditions* [Online]. Available fromhttps://www.cdc.gov/mmwr/preview/mmwrhtml/rr5710a2.htm [Accessed 27 June 2017].
4. Centers for Disease Control and Prevention, 2016a. *How TB spreads* [Online]. Available from https://www.cdc.gov/tb/topic/basics/howtbspreads.htm [Accessed 27 June 2017].
5. Centers for Disease Control and Prevention, 2016b. *CDC | TB | Fact Sheets | Multidrug-Resistant Tuberculosis (MDR TB)* [online]. Available from https://www.cdc.gov/tb/publications/factsheets/drtb/mdrtb.htm [Accessed 11 June,2017]
6. Centers for Disease Control and Prevention, 2017. *CDC | TB | Data and Statistics* [online]. Available from https://www.cdc.gov/tb/statistics/default.htm [Accessed 11 June 2017]
7. Clayton, D. and Kaldor, J., 1987. Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *International Biometric Society*, 43(3), 671-681.
8. Draper, N. and H. Smith, 1981. *Applied regression analysis: Series in probability and mathematical statistics*. Wiley.
9. Fernandez, C., Ley, E. and Steel, M., 2001. Model uncertainity in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5), 563-576.
10. Florida Department of Environmental Protection Geospatial open data, 2017. *Florida Land Usage, 2017[online]*. http://geodata.dep.state.fl.us/datasets/statewide-land-use-land-cover [Accessed 13 June 2017]
11. Florida Department of Health, 2017a. *FLHealthCHARTS Data Viewer* [online]. Available from http://www.flhealthcharts.com/charts/OtherIndicators/NonVitalIndNoGrpDataViewer.aspx?cid=0148 [Accessed 28 May 2017]
12. Florida Department of Health, 2017b. *FLHealthCHARTS Data Viewer* [online].Available from: http://www.flhealthcharts.com/charts/OtherIndicators/NonVitalHIVAIDSViewer.aspx?cid=0471 [Accessed 28 May 2017].
13. Gandhi, N. R et al., 2006. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *The Lancet*, *368*(9547), 1575–1580.
14. Haar, C. H. et al.,2007. Tuberculosis Drug Resistance and HIV Infection, the Netherlands. *Emerging Infectious Diseases*, *13*(5), 776–778.
15. Hardisty, F. and Klippel, A., 2010. Analysing spatio-temporal autocorrelation with LISTA-Viz. *International Journal of Geographical Information Science*, *24*(10), 1515–1526.
16. Hastie, T.J. and Tibshirani, R.J., 1990. *Generalized Additive Models*. Boca Raton: Chapman and Hall.
17. Hattori, T. et al., 2016. Nationwide HIV-, MDR-TB Survey in Japan and Collaborative Study in the Philippines. *International Journal of Mycobacteriology*, *5*, S18–S19.

18. Jacob, B. G. et al., 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.
19. Jacob, B. G. et al., 2014. Pseudo R^2Probablity 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. *American Journal of Applied Mathematics and Statistics*, *2*(5), 252–301.
20. Lee, S. et al., 2016. Is Multi-Drug Resistant Tuberculosis More Prevalent in HIV-Infected Patients in Korea? *Yonsei Medical Journal*, *57*(6), 1508–1510.
21. Lesage, J.P. and Fischer, M., 2008. Spatial Growth Regressions: Model Specification, Estimation and Interpretation. *Spatial Economic Analysis*, 3(3), 275-304.
22. LeSage, J.P. and Parent, O., 2007. Bayesian Model Averaging for Spatial Econometric Models. *Geographical Analysis*, 39(3), 241-267.
23. Madigan, D., York, J. and Allard, D, 1995. Bayesian Graphical Models for Discrete Data. *International Statistical Institute*, 63(2), 215-232.
24. Moonan, P. K. et al., 2013. Transmission of multidrug-resistant tuberculosis in the USA: a cross-sectional study. *The Lancet Infectious Diseases*, *13*(9), 777–784.
25. Rifat, M. et al., 2014. Development of Multidrug Resistant Tuberculosis in Bangladesh: A Case-Control Study on Risk Factors. *PLoS ONE*, *9*(8), e105214.
26. Sahebi, L. et al., 2016. Epidemiology and patterns of drug resistance among tuberculosis patients in Northwestern Iran. *Indian Journal of Medical Microbiology*, *34*(3), 362.
27. Sethi, S. et al, 2013. Prevalence of multidrug resistance in Mycobacterium tuberculosis isolates from HIV seropositive and seronegative patients with pulmonary tuberculosis in north India. *BMC Infectious Diseases*, *13*, 137.
28. Shringarpure, K. S. et al., 2015. Loss-To-Follow-Up on Multidrug Resistant Tuberculosis Treatment in Gujarat, India: The WHEN and WHO of It. *PLOS ONE*, *10*(7), e0132543.
29. Sotgiu, G. et al., 2017. Breaking the barriers: Migrants and tuberculosis. *PresseMedicale (Paris, France: 1983)*, *46*(2 Pt 2), e5–e11.
30. Streicher, E. M. et al., 2012. Emergence and treatment of multidrug resistant (MDR) and extensively drug-resistant (XDR) tuberculosis in South Africa. *Infection, Genetics and Evolution*, *12*(4), 686–694.
31. Suchindran, S., Brouwer, E. S. andRie, A. V., 2009. Is HIV Infection a Risk Factor for Multi-Drug Resistant Tuberculosis? A Systematic Review. *PLOS ONE*, *4*(5), e5561.
32. U. S. Census Bureau, 2010. *American FactFinder - Results* [online]. Available from:https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF[Accessed 18 June 2017].
33. *Zip Codes*. Hillsborough County Geoweb [online]. Hillsborough County. Available from http://gis-hillsborough.opendata.arcgis.com/datasets/d356e19e0fb34524b54d189fafb0d675_0 [Accessed 9 July 2017]