International Journal of Geographic Information System
Vol. 5, No. 2, March 2018, pp. 1-35, E-ISSN: 2576-375X
Available Online at http://acascipub.com/Journals.php

Research article

A linear least squares fitting technique, a nontrivial diagonal matrix and a zero-inner product for interpolating disproportionately weighted landscape regression estimates for identifying vulnerable populations to chlamydia in Miami-Dade County, Florida

Jordan Moberg[1], Toni Panaou[2], Benjamin G. Jacob[3]

[1]Department of Global Health, College of Public Health, University of South Florida, Tampa, Florida, 33612 United States of America

[2]Department of Integrative Biology, University of South Florida, Tampa, Florida, 33620 United States of America

[3] Department of Global Health, College of Public Health, University of South Florida, Tampa, Florida, United States of America, 33612

Contact: *Dr. Benjamin Jacob[2] bjacob1@health.usf.edu

Jordan Moberg is a graduate student in the MPH Global Health Practice Program in the Department of Global Health at the University of South Florida.

Dr. Toni Panaou is a postdoctoral research associate at the University of South Florida in the Department of Integrative Biology. She obtained her doctorate in Civil Engineering while working as a Teacher's Assistant for the Public Health Geographic Information Systems mapping class for Dr. Benjamin Jacob.

Dr. Jacob is an expert in the field of global health and has published a plethora of scholarly articles that scholastically contribute to current literature and learning. He is a research assistant professor at the University of South Florida.

International Journal of Geographic Information System
Vol. 5, No. 2, March 2018, pp. 1-35, E-ISSN: 2576-375X
Available Online at http://acascipub.com/Journals.php

Abstract

Chlamydia is the most highly reported infections disease in the United States (CDC 2015). The case count of chlamydia in Hillsborough County, Florida increased from a rate of 329.6 in 2006 to 569.4 in 2015, making it the fourth highest county for chlamydia in the state (Florida Health Charts 2017). Identification of regression covariates and spatially referenced locations are vital for predicting the causes associated with the rising trend of this disease. Unfortunately, frequentistic regression models are not available to identify populations vulnerable to chlamydia by the zip code level, or to make geospatial robust predictions about future hyperendemic geographic locations. In this analysis, cartographic and statistical algorithmic estimates were used to determine geographic locations of vulnerable populations to chlamydia at both county and zip code levels employing land use land cover (LULC) maps in Geographic Information Systems and regression analysis Statistical Analysis System covariates. Initially, a georeferenced map was constructed using the prevalence of chlamydia at the county level with data obtained from Esri (2014) and Florida Health Charts (2017). Next, the covariance's most likely to be associated with chlamydia at the highest Florida County prevalence cluster, i.e. Miami-Dade, with socio-demographics obtained from (http://zip-codes.com), were extrapolated. A LULC map was then created to associate specific geomorphological data feature attributes with the disease. The regression estimate model that was synthesized from the high-density chlamydia Miami-Dade County cluster (11,862 cases per 100,000 population) was iteratively optimally interpolated. This county cluster was then georeferenced to create a robust predictive map to identify where vulnerable chlamydia populations were geographically located in Hillsborough County. Within the highest Miami-Dade County cluster region, zip code 33012, Hispanic race (p-value=0.0007) and LULC (p-value=0.0262) covariates were found to be statistically significant at the 95% confidence interval level with an R-squared value of 0.9663. Conversely, average income, White, Black, Asian, American Indian, Hawaiian, Other, male, female, and median age were not statistically significant covariates associated with chlamydia at the 95% confidence interval level.

Keywords: Chlamydia; linear regression; geographic information systems; mapping; land use land cover; predictive covariates, Bayesian.

1: Introduction

Is it possible that the most commonly reported sexually transmitted infection usually persists and spreads silently through young populations without signs or symptoms? Chlamydia, on the list of nationally reportable diseases, was responsible for 1,526,658 cases of infection in 2015, with the south region leading the way for transmission in the United States (CDC 2015).

Chlamydia is a sexually transmitted bacterial infection that often appears symptomless and can spread undetected, causing a myriad of physical, psychological, and emotional problems (CDC 2004). If left untreated or undiagnosed, chlamydia infections can lead to cervical cancer, pelvic inflammatory disease, premature labor, tubal pathology, infertility, and silent transmission to other sexual partners (Klaveren et al. 2016). Therefore, screening to detect this mostly asymptomatic disease is vital for the treatment of infected individuals and prevention of further complications and transmission. However, due to the high costs and implausibility of screening all persons, specific at-risk populations must be targeted to help ensure all infected individuals are identified with a minimal number of infected cases missed (Klaveren et al. 2016).

To help identify groups at-risk for chlamydia infections, Klaveren et al. (2016) constructed a multivariate logistic regression hierarchical model with participant surveys and found that young age and Surinamese, Antillean, or Sub-Saharan African ethnicity were the strongest explanators of chlamydia infections, whilst neighborhood socioeconomic status was not a significant predictor. Chang, Pearson, and Owusu-Edusei (2017) conducted a hot spot analysis for the three most common sexually transmitted infections: syphilis, gonorrhea, and chlamydia, and then applied spatial logistic regression covariates to identify associations between hot spots and sociodemographic georeferenced information. This study found hot spots to exist in the south region of the United States. It identified that a 10-point increase in black non-Hispanic individuals was associated with a 44% increased chance of being a hot spot, while a 10-point increase in Hispanic individuals was associated with a 22% increased chance of being a hot spot (Chang, Pearson, and Owusu-Edusei 2017). While this research determined forecastable regression covariates associated with the prevalence of chlamydia, it did not reveal specific hyperendemic populations geographic locations (henceforth geolocations) employing a predictive probability georeferenced map with grid-stratifiable, land use land cover (LULC) regressors at the zip code level. This analysis builds on the finding that Florida is one of the southern-most states to be identified as a

major hot spot for sexually transmitted infections through examination of the county with the highest chlamydia prevalence at the zip code level using a LULC map (Chang, Pearson, and Owusu-Edusei 2017).

Although linear regression is the most basic type of regression employed for predictive analysis (Hosmer and Lemeshew 2000), these frequentistic models have not been utilized extensively in chlamydia endemicity research. In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data (e.g., county-level, georeferenced, empirical zip code, socio-demographic chlamydia related data) (Lindley 1987).

County-level, georeferenced, grid-stratified, chlamydia data may be represented employing a scatter plot, scatter diagram, or x-y plot. During analysis, a chlamydia researcher or sexually transmitted disease (STD) specialist may try to find the equation of a line that best fits the county data (i.e., regression line). From an algebraic perspective, points which are (x,y) pairs can be plotted on the Cartesian coordinate system (Kruskal and Tanur 1978). A straight line on the Cartesian coordinate system has the equation $y = mx + b$, where m is the slope of the line, and b is the y-intercept (Fox 1997).

Fortunately, the slope would always be the coefficient of the x term in a prognosticative, county-level, linear, chlamydia regression equation. Following this pattern, the slope of a regression line in the chlamydia model can be optimally quantitated employing various forms of the equation (e.g., $y = ax + b$, $y = a + bx$, etc). A researcher could determine the slope and the y-intercept by looking at the equation. In so doing, the county-level scatter diagrams may show a direct relationship between x (e.g., independent variable) and y (e.g., county-level chlamydia prevalence statistics). Hence, if an equation, $y = a + bx$ where b is the slope is employed to analyze an empirical dataset of county level, geosampled, grid-stratifiable, zip code georeferenceable polygons, then, a direct relationship may be determined to exist when the slope of the line (b) is positive. Conversely, an inverse relationship may exist when the slope of the line (b) is negative. When the slope (b) of the line equals zero, there may be no quantifiable relationships in the model.

As noted above, a plotted straight line employing an empirical dataset of georeferenceable, county-level, chlamydia covariates on the Cartesian coordinate system can have the equation $y = mx + b$. Presumably, then, a regression line will have the general form $y = a + bx + e$ where: a is the y-intercept, b is the slope of the line, and e is an error term. In practice, under ordinary circumstances, the value of the error term may not be given in a regression paradigm following the form of the equation $y = a + bx$. Although, according to Jacob et al. (2009) alternative forms such as $y + ax + b$ may also yield the same results in an infectious disease forecast vulnerability, uncertainty, probabilistic model.

From the study of correlation in a county-level, georeferenceable, forecast, vulnerability, endemic, chlamydia transmission-oriented model, it may be that when the slope of the regression line in the chlamydia model is positive the value of y may increase as the value of x increases (i.e. a positive correlation between a geosampled, georeferenecable, chlamydia covariate and county-level case distribution). Hence, when the slope of the regression line in the chlamydia model is negative the value of y would decrease as x increases in the model. The strength of these relationships may be given by the correlation coefficient (r), which can be calculated at a georeferenced, zip code polygon level within a county forecast, vulnerability, chlamydia linear regression model.

Most commonly, the conditional mean of Y given the value of X is assumed to be an affine function of X; less commonly, the median or some other quantile of the conditional distribution of y given X is expressed as a linear function of X (Lindley 1987). An affine function is a function composed of a linear function plus a constant (Sen and Srivastava 2011). Robustly calculating interval, forecast, county-level, georeferenceable, chlamydia linear graphs may require a straight regression line. This line may also be employed for constructing a general equation for an affine function in any one dimensional, county-level, vulnerability model (e.g., $y = ax + c$) where c is equal to the error distribution.

In probability theory, the conditional expectation, conditional expected value, or conditional mean of a random variable (e.g. county-level, geosampled regression, chlamydia, predictor variable) has an expected value given that a certain set of "conditions" is known to occur. In the case when the random variable is defined over a discrete probability space, the "conditions" are a partition of this probability space (Freedman 2009). With multiple random variables (e.g., a time series, empirical, georeferenced dataset of optimizable, diagnostic, clinical, field, or remote geosampled, county-level, chlamydia covariates), if one random variable mean is independent of all others both individually and collectively, then each conditional expectation would equal the random variable's (unconditional) expected value. This rule would hold in any time series, forecast, vulnerability, county-level, chlamydia regression model, but only if the variables are independent.

Depending on the nature of the conditioning in a county-level, forecast, vulnerability, time series, chlamydia model, the conditional expectation can be either a random variable itself or a fixed value. Two random explanatory, county-level, geosampled, georeferenecable, chlamydia-related, forecast variables are selected; if the expectation of the random variable, X, is expressed conditionally on another random variable, Y, without a particular value of Y being specified, then the expectation of X conditional on Y, in the vulnerability model may be denoted $E[X|Y]$. This expression would be a function of the geosampled random

variable, Y, and hence would itself be a random variable in the county-level model. Alternatively, if the expectation of X is expressed conditionally on the occurrence of a particular value of Y, in the chlamydia model denoted y, then the conditional expectation E[X|Y = y] would be a fixed value. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis (Hosmer and Lemeshew 2002).

The overall idea of regression for analyzing empirical, chlamydia data at the county level is to answer two basic questions: (1) Can a regressed empirical dataset of georeferenced, social demographic, predictor variables, forecast an outcome explanatory, vulnerability variable related to distribution (outcome prevalence and case distribution)? (2) Is the chlamydia model employing the prognosticators that account for variability of changes in a dependent variable associated with the disease? In so doing, research may be able to determine which regressors are significant predictors (statistical significance at 95% confidence interval) of the dependent variable (county-level prevalence). Further, a regression model may indicate the magnitude and sign of the beta estimates which may relate to the impact of sampled chlamydia estimators on the dependent variable. These chlamydia-based regression estimates may be employed to robustly quantitate the relationship between a series of independent, socio-demographic explanators. The simplest form of the equation may be defined by the form y=mx+b where y is equal to the estimated dependent score, c=constant, b=regression coefficients, and x is the independent variable.

Unfortunately, again, landscape heterogeneity has not been captured or regressed in a county level, prognosticative, chlamydia infection vulnerability model. Landscapes can be drivers of many different chronic infection processes, examples include: malaria (Jacob et al. 2009), tuberculosis (Jacob et al. 2013), and onchocerciasis (Jacob et al. 2015). By optimally delineating geoclassifiable LULCs and incorporating those statistics at the county level, a more robust model may be predicted revealing georeferenceable vulnerable populations to chlamydia. Further, by combining parameterizable, georeferenceable, and LULC time series covariates with socio-demographic geosampled covariates in a regression model, a hyperendemic capture point at the county-level may be revealed using a zip-code, polygon, grid-stratification.

In this analysis, a multivariate regression model for Miami-Dade was constructed, and multiple sociodemographic statistics that were retrieved from zip code shape files in GIS were employed where we optimally cartographically delineated a high-density prevalence cluster of chlamydia in Florida (Miami-Dade). We parsimoniously constructed the multiple regression and LULC models for the high density, geospatial,

georeferenced, cluster. In so doing, we synthesized heterogeneous landscape and sociodemographic variables, which were then iteratively quantitatively interpolated in Hillsborough County, Florida. This enabled us to generate a predictive map of vulnerability to chlamydia in this county. The objectives in this analysis were: (1) To construct a multivariate linear regression model using multiple sociodemographic variables; (2) to create a LULC model for a high-density county georeferenceable cluster; and (3) to interpolate landscape and regression estimates to geographically delineate areas of vulnerability in Hillsborough County.

2. Methodology

2.1 Study Site

Florida, also known as the "sunshine state" is home to the oldest city, St. Augustine, southern most city, Key West, and largest city by area, Jacksonville, in the United States of America (Alekinster 2015). It is bordered by Alabama and Georgia to the north, the Gulf of Mexico to the west, the Atlantic Ocean to the east, and Cuba to the south. Florida has 86,311square kilometers (km^2) of land and 19,516 km^2 of water, with a warm and humid tropical climate that is home to the Florida Everglades and the many reptiles that reside within it. Most of the state is at or below sea level with the highest point at only 105 meters above sea level (Alekinster 2015). The low elevation and peninsula shape make Florida vulnerable to natural disasters like hurricanes and flooding, while its exposure to the Atlantic Ocean, Gulf of Mexico, Caribbean Sea, and Florida straits on three out of its four sides enables foreigners to make land here, sometimes bringing with them, foreign diseases. The year-round warm climate makes Florida a popular state for retired persons to live and visit, with an average age of 50.7 years among residents (Zip-Codes.com 2017). In addition, Florida is home to an array of theme parks including: Disney World, Sea World, Islands of Adventure, Busch Gardens, Adventure Island, Epcot, and numerous others, which helped to bring in the states 112.8 million tourists in 2016 (Dineen 2017). With such an abundance of people entering and leaving the state every year, Florida has a higher exposure rate, and therefore higher risk factor than many other states for carrying and spreading diseases, including sexually transmitted infections.

According to Florida Health Charts (2017), the overall chlamydia case count for the state of Florida was 90,633 per 100,000 population in 2015. Of this total, Miami-Dade had 11,662 cases per 100,000 population, and thus ranked as the county with the highest prevalence of chlamydia in the state (Florida Health Charts 2017, Zip-Codes.com 2017). Two maps of the chlamydia prevalence in Florida by county are shown in Figure 1 and Figure 2. Miami-Dade is the most populous urban area in the state of Florida with approximately

2.5 million people, and an age demographic that is more than 10 years younger than the rest of the state (Miami-Dade: 38.2 years, Florida: 50.7 years). This county is also thriving in diversity with a population that is: 46.3% White, 40.8% Hispanic, 11.9% Black, and 0.01% Asian (Zip-Codes.com 2017). Due to its high prevalence of chlamydia, Miami-Dade was chosen as the county of study for this GIS analysis and interpolated regression analysis.

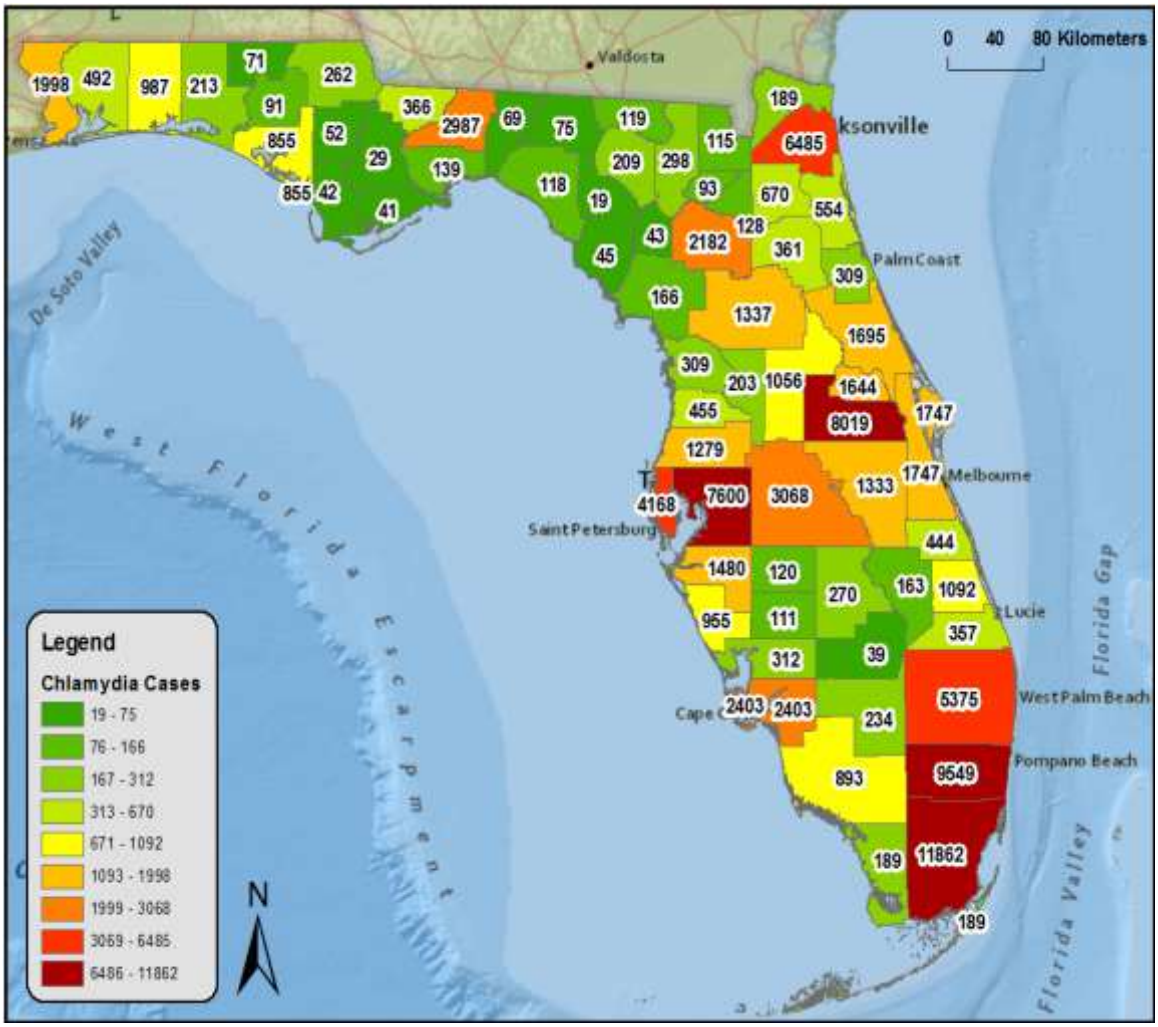

Figure 1. Chlamydia prevalence per 100,000 population in Florida Counties in 2015. Number of chlamydia cases is labelled in each county.

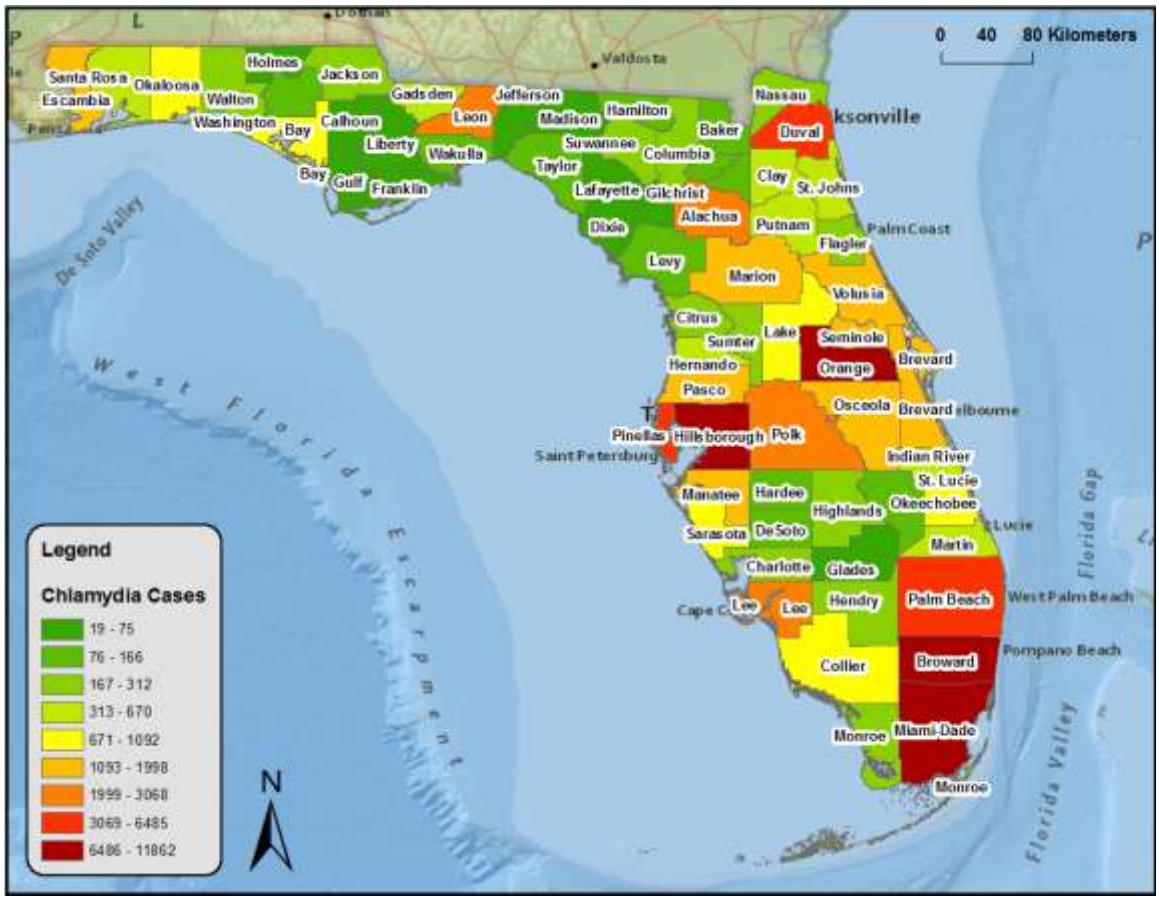

Figure 2. Chlamydia prevalence per 100,000 population in Florida Counties in 2015 with county names labelled.

2.2 *Data*

2.2.1 *Data Collection*

Chlamydia prevalence data was collected from the Florida Department of Health and Bureau of Communicable Diseases under the label "chlamydia" from (http://www.flhealthcharts.com/charts/OtherIndicators/NonVitalSTDDataViewer.aspx?cid=0143). The map of Florida Counties was obtained from the United States Census Bureau-Tiger/Line files at (http://www.esri.com) using the search term "Florida Counties." Zip codes and sociodemographic data by zip code were obtained by searching "Miami-Dade" from (https://www.zip-codes.com/county/fl-miami-dade.asp). The map of Miami-Dade was obtained using information from the census at (https://www.census.gov/geo/maps-data/data/tiger-geodatabases.html) under the title: CensusGeodatabase_MiamiDadeCounty.

2.2.2 *Zip Code Data*

83 Miami-Dade zip codes were selected as the primary focus for this study (Zip-Codes.com 2017). From this list of 83 zip codes, five were removed for the SAS analysis to prevent data skewing due to insufficient information at these specific locations. Table 1 shows a list of the zip codes used in this study while Table 2 provides a list of zip codes not used. By employing the zip code level data, we were able to generate univariate statistics and regression models for the zip code level and socio-demographic covariates.

Table 1. 83 Miami-Dade zip codes analyzed and used in ArcMap. Note: zip codes with (*) were removed for SAS analysis (Zip-Codes.com 2017).

83 Zip Codes						
33010	33125	33137	33154	33170	33183	33160
33012	33126	33138	33155	33172	33184	33122
33013	33127	33139	33156	33173	33185	33192*
33014	33128	33140	33157	33174	33186	33149
33015	33129	33141	33158	33175	33187	33031
33016	33130	33142	33161	33176	33189	33032
33018	33131	33143	33162	33177	33190	33033
33030	33132	33144	33165	33178	33193	33039*
33056	33133	33145	33166	33179	33194	33035
33101*	33134	33146	33167	33180	33196	33034
33109	33135	33147	33168	33181	33055	43*
33122*	33136	33150	33169	33182	33054	

Table 2. Five zip codes removed from SAS analysis. 78 out of 83 zip codes were analyzed using SAS (Zip-Codes.com 2017).

Zip Code	**Reason Removed**
33101	Population = 0, insufficient data.
33122	Population = 287, insufficient data.
33192	Population = 0, insufficient data.
33039	Homestead Air Reserve Base, insufficient data.
43	Everglades, population = 0, insufficient data.

2.3 *Covariates Selected for Regression Analysis*

2.3.1 *Chlamydia Cases per specific zip code*

According to the Florida Health Charts (2017) website, the total number of cases of chlamydia in Miami-Dade is 11,862, the highest rate in the state. The total Miami-Dade case count was proportionally scaled down according to the size of the population within each zip code for a more accurate representation of

individual case counts of chlamydia by geographic location. This calculation yielded 12,143 total cases of chlamydia in Miami-Dade, which is a slightly higher number of chlamydia cases than reported on Florida Health Charts. The variance is most likely due to differences in rounding and reporting. The calculation that was used is illustrated below.

Population proportion = Zip code population___
Miami-Dade population

Chlamydia cases at specific zip code = chlamydia prevalence * population proportion

Example:

Miami-Dade population: 2,651,195

Miami-Dade chlamydia prevalence: 11,862

Zip code 33010 population: 44,267

Population proportion = Zip code population___
Miami-Dade population

0.016697 population proportion = 44,267___
2,651,195

Chlamydia cases at zip code = chlamydia prevalence * population proportion

198.059801 chlamydia cases at 33010 = 11,862 * 0.016697

= 198 chlamydia cases at zip code 33010

2.3.2 *Land Use Land Cover Map*

A LULC map utilizes satellite and aerial imagery to create and classify pixilated surface areas of various landscapes to determine the land cover of varying geographical locations based on groups of clustering (Esri 2017). A LULC map was thus chosen to determine if varying land cover and usage, specifically relating to: urban residential, urban commercial, pastureland/parkland, or water are covariates associated with chlamydia prevalence at the county-level.

The LULC shape file with satellite imagery of Florida was obtained from (FL_Landuse/STATEWIDE_LANDUSE). The Miami-Dade land cover area was then clipped into its own layer for further analysis. The layer was re-categorized into four LULC areas particular to this study. These labels were then coded for as numbers in excel to export into SAS for regression analysis of covariates associated with chlamydia prevalence: 1=urban commercial, 2=urban residential, 3=pastureland/parkland, 4=water. This classification breakdown of LULC features and codes assigned in SAS is also shown in Table 3.

Table 3. LULC re-categorization with codes assigned for SAS analysis and colors used.

New Classification of Categories	Urban Commercial	Urban Residential	Pastureland/Parkland	Water
# Code in SAS	1	2	3	4
Map Colors				
Original Categories	Commercial and Services	Low Density	Cropland and Pastureland	Bays and Estuaries
	Communication	Medium Density	Disturbed Lands	Lakes
	Feeding Operations	High Density	Extractive	Oceans Seas and Gulfs
	Industrial		Herbaceous	Reservoirs
	Institutional		Mixed Rangeland	Streams and Waterways
	Recreational		Non-Vegetated	Beaches other than swimming
	Transportation		Nurseries and Vineyards	
	Utilities		Open Land	
			Other Open Lands <Rural>	
			Sand Other Than Beaches	
			Salt Flats	
			Shrub and Bushland	
			Specialty Farms	
			Tree Crops	
			Tree Plantations	
			Upland Coniferous Forests	
			Upland Hardwood Forests	
			Upland Mized Forests	
			Vegetated Non-Forested Wetlands	
			Wetland Coniferous Forests	
			Wetland Forested Mixed	

By polygonizing Miami-Dade counties based on georeferenceable land cover, a visual representation of land cover by category was created. A zip code map was then layered on top of the LULC map to identify land cover by zip code as shown in Figure 3 and Figure 4.

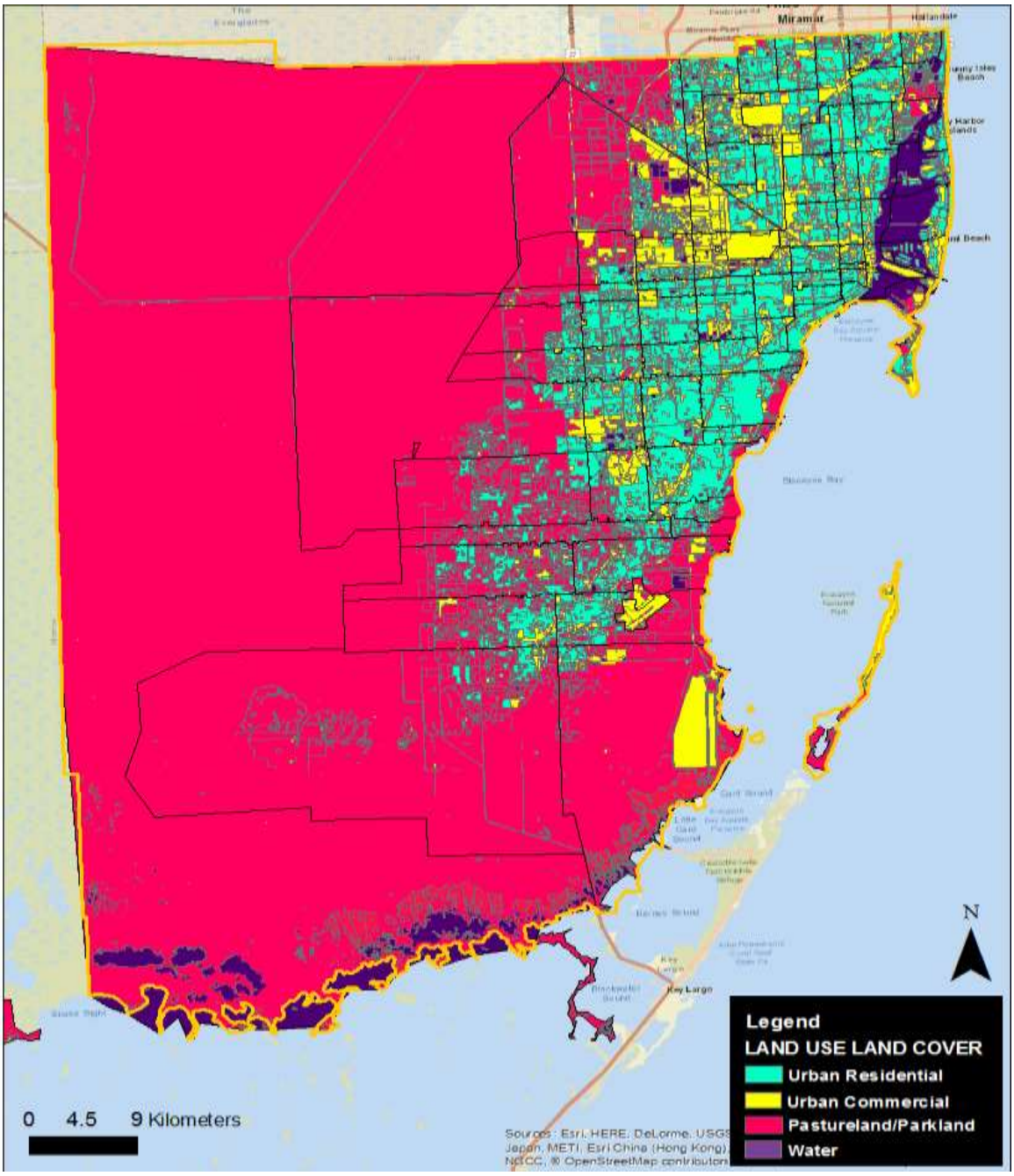

Figure 3. Miami-Dade land use land cover map with zip code regions outlined in black, 2015.

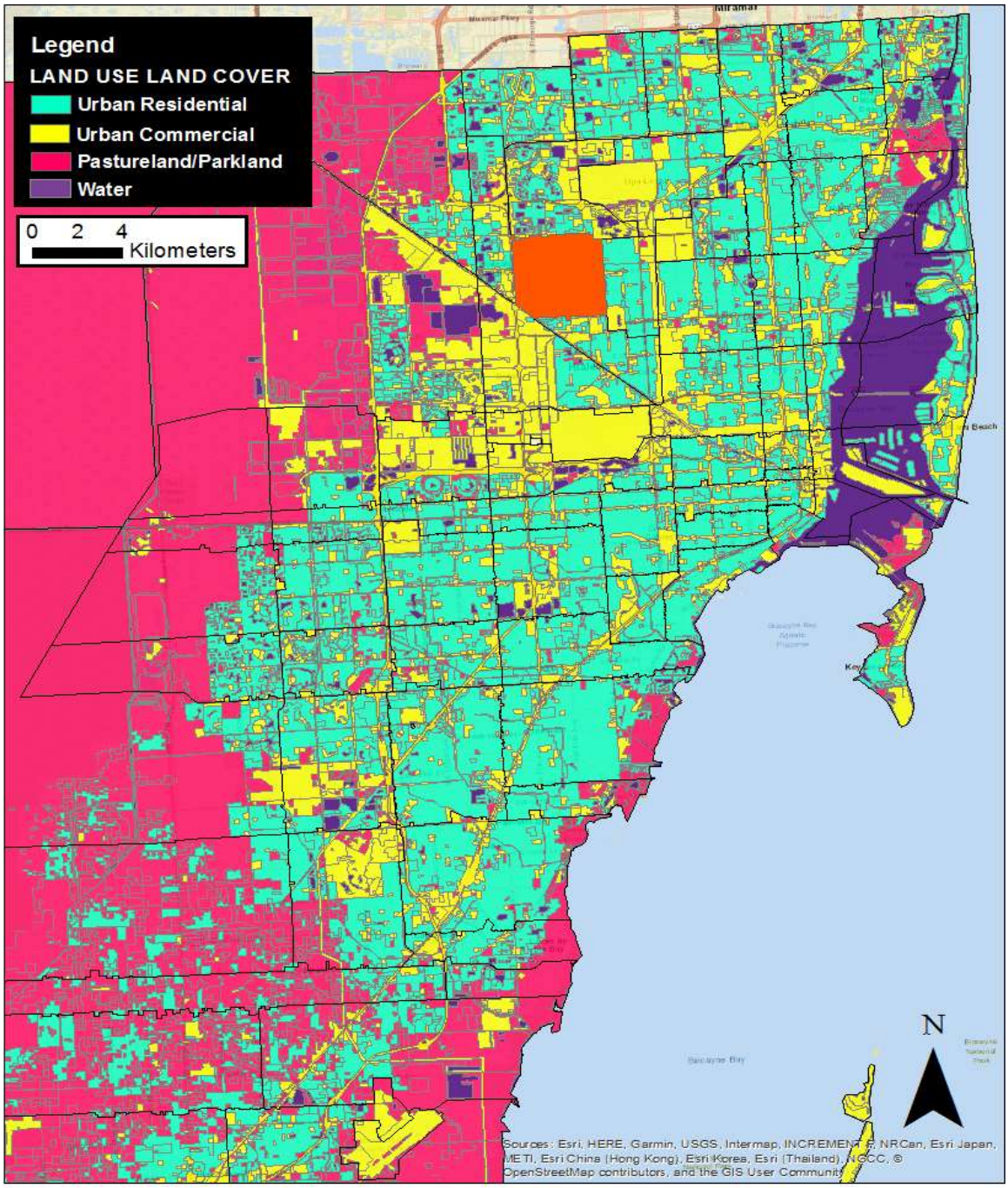

Figure 4. Zoomed in view of densely populated areas by land cover features in Miami-Dade with zip code 33012 highlighted in the color orange.

2.4 Regression Analysis

A simple linear regression model was used to identify the relationship between the dependent variable, chlamydia prevalence, and the independent variables: average income per household, White population, Black population, Hispanic population, Asian population, American Indian population, Hawaiian population, Other population, female population, male population, median age, and LULC. Linear regression is used in statistics to

model the relationship between a single dependent variable, y, and one or more descriptive independent variables, x (Hosmer and Lemeshow 2000). LULC was a categorical variable while all other explanatory variables were continuous.

The estimates of the unknown parameters in the epidemiological, chlamydia-model were obtained from the linear least squares regression, which employed the optimal estimates from a broad class of possible parameter estimates under the usual assumptions used for process modeling. Linear least squares regression is a method of solving statistical mathematical problems. It uses the least squares algorithmic technique to increase accuracy of solution approximations, corresponding with a particular problem's complexity (Freedman 2005). The linear least squares fitting technique is the simplest and most commonly applied form of linear regression and provides a solution to the problem of finding the best fitting straight line through a set of points (e.g., county-level georeferenced LULC and sociodemographic variables) (Draper and Smith 1998).

The condition for R^2 to be a minimum in the epidemiological, chlamydia county-level, forecast model was that $\frac{\partial(R^2)}{\partial a_i} = 0$ for $i = 1, ..., n$. For the linear fit, $f(a, b) = a + b\,x$, so $R^2(a, b) \equiv \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2$.

$\frac{\partial(R^2)}{\partial a} = -2\sum_{i=1}^{n} [y_i - (a + b\,x_i)] = 0$ and $\frac{\partial(R^2)}{\partial b} = -2\sum_{i=1}^{n} [y_i - (a + b\,x_i)]\,x_i = 0.$ This lead to the equations

$n\,a + b\sum_{i=1}^{n} x_i \sum_{=i=1}^{n} y_i$ and $a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \sum_{=i=1}^{n} x_i\,y_i.$

In matrix form, the county-level model revealed $\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix}$, so $\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,y_i \end{bmatrix}$. The 2×2 matrix inverse was

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\,y_i \\ n\sum_{i=1}^{n} x_i\,y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

so $a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\,y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{(\sum_{i=1}^{n} x_i\,y_i) - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}\,\bar{y}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}\sum_{i=1}^{n} x_i\,y_i \over \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2$ whilst $b = \frac{n\sum_{i=1}^{n} x_i\,y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$. The inverse of a square matrix $\mathbf{A}$, sometimes called a reciprocal matrix, is a

matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix (Courant and Hilbert 1989) used the notation $\check{\mathbf{A}}$ to denote the inverse matrix.

According to Wolfram Research, Inc (2018), the identity matrix was then defined for the chlamydia county-level model such that $\mathbf{I}(\mathbf{X}) \equiv \mathbf{X}$ for all vectors $\mathbf{X}$. An identity matrix is denoted $\mathbf{1}$, $\mathbf{I}$, $\mathbf{E}$ (the latter being an abbreviation for the German term "Einheitsmatrix" (Courant and Hilbert 1989), or occasionally $\mathbb{I}$, with a subscript sometimes used to indicate the dimension of the matrix. Identity matrices are also known as unit matrices (Akivis and Goldberg 1972).

The $n \times n$ identity matrix was given explicitly by $I_{ij} = \delta_{ij}$ for $i, j = 1, 2, \ldots, n$, where δ_{ij} I= was the Kronecker delta in the forecast, vulnerability, chlamydia, county-level, epidemiological model which was written explicitly as, $\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$. The simplest interpretation of the Kronecker delta in the model was the discrete version of the delta function defined by $\delta_{ij} \equiv \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j. \end{cases}$. The delta function is a generalized function that can be defined as the limit of a class of delta sequences (Lipschutz and Lipson 2009). Delta sequence is a sequence of strongly peaked functions for which $\lim_{n\to\infty} \int_{-\infty}^{\infty} \delta_n(x) f(x)\, dx = f(0)$ so that the limit was $n \to \infty$ (Weisstein 2017).

The $n \times n$ identity matrix was implemented in SAS. The "Square root of identity" matrices was defined for $\mathbf{I}_n$ by solving $\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$. For $n = 2$, the most general form of the resulting square root matrix in the forecast, vulnerability chlamydia, epidemiological, county-level model residuals revealed $\mathbf{I}_2^{1/2} = \begin{bmatrix} \pm d & \frac{1-d^2}{c} \\ c & \mp d \end{bmatrix}, \begin{bmatrix} \pm d & c \\ \frac{1-d^2}{c} & \mp d \end{bmatrix}$ rendering $\begin{bmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{bmatrix}, \begin{bmatrix} \pm 1 & 0 \\ c & \mp 1 \end{bmatrix}, \begin{bmatrix} \pm 1 & c \\ 0 & \mp 1 \end{bmatrix}$ as limiting cases. These were rewritten in a simpler form in our model by defining the sums of squares when

$$\mathrm{SS}_{x\,x} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\,\bar{x}^2,\ \mathrm{SS}_{y\,y} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{n} y_i^2\right) - n\,\bar{y}^2 \text{ and } \mathrm{SS}_{x\,y} = \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \left(\sum_{i=1}^{n} x_i\,y_i\right) - n\,\bar{x}\,\bar{y},$$

The Kronecker delta was implemented in SAS [*i*, *j*], where the contour integral in the chlamydia county-level model was quantitated parsimoniously by $\delta_{m\,n} = \frac{1}{2\,\pi\,i}\oint_{\gamma} z^{m-n-1}\,dz,$ where γ was a contour corresponding to the unit circle. In the model m and n were finite, discrete, integer values. Watson (1996) uses the notation $\int^{(a+)} f(z)\,dz$ to denote the contour integral of $f(z)$ with contour encircling the point a once in a counter clockwise direction.

In three-space, the Kronecker delta in the forecast, vulnerability, chlamydia, county-level model satisfied the identities: $\delta_{i\,i} = 3$, $\delta_{i\,j}\,\epsilon_{i\,j\,k} = 0$, $\epsilon_{i\,p\,q}\,\epsilon_{j\,p\,q} = 2\,\delta_{i\,j}$ and $\epsilon_{i\,j\,k}\,\epsilon_{p\,q\,k} = \delta_{i\,p}\,\delta_{j\,q} - \delta_{i\,q}\,\delta_{j\,p}$, when the Einstein summation was implicitly assumed, $i, j = 1$, 2, 3, and $\epsilon_{i\,j\,k}$ was the permutation symbol. Einstein summation is a notational convention for simplifying expressions including summations of vectors, matrices, and general tensor with the permutation symbol (Freedman 2005). In mathematics, particularly in linear algebra, tensor analysis, and differential geometry, the permutation symbol represents a collection of numbers; defined from the natural numbers 1, 2, …, n, for some positive integer n (Fox 1997).

Technically, the Kronecker delta in the chlamydia, epidemiological, county-level model was a tensor defined by the relationship $\delta_l^k \frac{\partial x_i'}{\partial x_k}\frac{\partial x_l}{\partial x_j'} = \frac{\partial x_i'}{\partial x_k}\frac{\partial x_k}{\partial x_j'} = \frac{\partial x_i'}{\partial x_j'}.$ Since, by definition, the model sociodemographic was gridded, LULC explanative coordinates x_i and x_j were independent for $i \neq j$, $\frac{\partial x_i'}{\partial x_j'} = \delta'^{i}_{j}$, so $\delta'^{i}_{j} = \frac{\partial x_i'}{\partial x_k}\frac{\partial x_l}{\partial x_j'}\delta_l^k$, and δ_j^i was really a mixed second-rank tensor. The model output satisfied $\delta_{a\,b}{}^{j\,k} = \epsilon_{a\,b\,i}\,\epsilon^{j\,k\,i} = \delta_a^j\,\delta_b^k - \delta_a^k\,\delta_b^j$, $\delta_{a\,b\,j\,k} = g_{a\,j}\,g_{b\,k} - g_{a\,k}\,g_{b\,j}$ and $\epsilon_{a\,i\,j}\,\epsilon^{b\,i\,j} = \delta_{a\,i}{}^{b\,i}$.

In this research, we also wrote $\sigma_x^2 = \frac{\mathrm{SS}_{x\,x}}{n}$, $\sigma_y^2 = \frac{\mathrm{SS}_{y\,y}}{n}$ and $\mathrm{cov}(x, y) = \frac{\mathrm{SS}_{x\,y}}{n}$ for robustifying the geosampled, chlamydia parameter estimators. In so doing, we found that cov (x, y) was the covariance, and, σ_x^2 and σ_y^2 were the variances. Covariance provides a measure of the strength of the

correlation between two or more sets of random variates (Freedman 2005). The covariance for the geosampled, chlamydia variates X and Y, each with sample size N, was definable by the expectation $\operatorname{cov}(X, Y) = \langle (X - \mu_X)(Y - \mu_Y) \rangle = \langle X\,Y \rangle - \mu_X\,\mu_y$ where $\mu_x = \langle X \rangle$ and $\mu_y = \langle Y \rangle$ were the respective means, which was written explicitly as $\operatorname{cov}(X, Y) = \sum_{n=1}^{N} \frac{(x_n - \bar{x})(y_n - \bar{y})}{N}$.

For quantitating uncorrelated variates, forecast, vulnerability, chlamydia-related, county-level model, unbiased estimators $\operatorname{cov}(X, Y) = \langle X\,Y \rangle - \mu_X\,\mu_Y = \langle X \rangle \langle Y \rangle - \mu_X\,\mu_Y = 0$, were employed, which revealed that the covariance was zero. If the geosampled county-level variables were correlated in the same way, then their covariance would have been nonzero. In fact, if $\operatorname{cov}(X, Y) > 0$, in a model, then Y tends to increase as X increases, and, if $\operatorname{cov}(X, Y) < 0$, then Y tends to decrease as X increases (Jacob et al. 2014). Note that whilst statistically independent interpolative signature explanators are always uncorrelated, the converse is not necessarily true in a regression infectious disease model (Griffith 2005, Jacob et al. 2011).

For un-biasing the county-level, chlamydia, epidemiological, model estimators we used the $\sigma_{X\,Y} = \operatorname{cov}(X, Y)$, which then provided a consistent way of denoting the variance as $\sigma_{X\,X} = \sigma_X^2$, where σ_X was the standard deviation. The derived $\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X\,\sigma_Y} = \frac{\sigma_{X\,Y}}{\sqrt{\sigma_{X\,X}\,\sigma_{Y\,Y}}}$, was the statistical correlation of X and Y in the model. The covariance is especially useful when looking at the variance of the sum of random variates, since $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y)$. (Freedman 2005). The covariance in the chlamydia model was symmetric by definition, since $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$.

Hence, given n forecast, vulnerability, chlamydia, county-level, epidemiological, uncorrelated, random variates, the model estimators (denoted X_1, ..., X_n,) the covariance $\sigma_{i\,j} \equiv \operatorname{cov}(X_i, X_j)$ of X_i and X_j was robustly definable by $\operatorname{cov}(X_i, X_j) = \langle (X_i - \mu_i)(X_j - \mu_j) \rangle = \langle X_i\,X_j \rangle - \mu_i\,\mu_j$, when $\mu_i = \langle X_i \rangle$ and $\mu_j = \langle X_j \rangle$ were the means of X_i and X_j, respectively. The matrix $(V_{i\,j})$ of the quantities $V_{i\,j} = \operatorname{cov}(X_i, X_j)$ is called the covariance matrix (Fox 1997). The covariance in the chlamydia, county-level, epidemiological model obeyed the identities $\operatorname{cov}(X + Z, Y) = \langle (X + Z)\,Y \rangle - \langle X + Z \rangle \langle Y \rangle = \langle X\,Y \rangle + \langle Z\,Y \rangle - (\langle X \rangle + \langle Z \rangle) \langle Y \rangle = \langle X\,Y \rangle - \langle X \rangle \langle Y \rangle + \langle Z\,Y \rangle - \langle Z \rangle \langle Y \rangle = \operatorname{cov}(X, Y) + \operatorname{cov}(Z, Y)$. By induction, it therefore followed that

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} \operatorname{cov}(X_i, Y) \operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \operatorname{cov}\left(X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \operatorname{cov}\left(\sum_{j=1}^{m} Y_j, X_i\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{cov}(Y_j, X_i) \sum_{=i=1}^{n}\sum_{j=1}^{m} \operatorname{cov}(X_i, Y_j).$$

Note that the quantities $\sum_{i=1}^{n} x_i\, y_i$ and $\sum_{i=1}^{n} x_i^2$ in the chlamydia model were interpreted as the dot products employing $\sum_{i=1}^{n} x_i^2 = \mathbf{x}\cdot\mathbf{x}$ and $\sum_{i=1}^{n} x_i\, y_i = \mathbf{x}\cdot\mathbf{y}$. In mathematics, the dot product or scalar product is an algebraic operation that takes two equal-length sequences of numbers (usually coordinate vectors) and returns a single number (Jain, Ahuja, and Ahmad 1995). In Euclidean geometry, the dot product of the Cartesian coordinates of vectors is widely used and often called an inner product (Lipschutz and Lipson 2009). In linear algebra, an inner product space is a vector space with an additional structure called an inner product as well (Arfken and Weber 2000).

The additional structure associated with each vector is derived from the chlamydia, county-level, epidemiological model (i.e., the inner product of the vectors) in regression space. The quantitation of the inner products allowed the rigorous introduction of intuitive geometrical notions, such as: the length of a vector, or the angle between the vectors in the model. They also provide the means for defining orthogonality between vectors (i.e., zero inner product) in the model. Inner product spaces generalize Euclidean spaces (in which the inner product is the dot product, also known as the scalar product) to vector spaces of any (possibly infinite) dimension (Borisenko and Taparov 1968).

In terms of the sums of squares, the regression coefficient b was given by $b = \dfrac{\operatorname{cov}(x, y)}{\sigma_x^2} = \dfrac{\mathrm{ss}_{xy}}{\mathrm{ss}_{xx}}$, and a was given in terms of b using (◇) as $a = \bar{y} - b\,\bar{x}$. In so doing, the overall quality of the fit was then parameterized in terms of a quantity known as the correlation coefficient, defined by $r^2 = \dfrac{\mathrm{ss}_{xy}^2}{\mathrm{ss}_{xx}\,\mathrm{ss}_{yy}}$, in the epidemiological, chlamydia, county-level model, which then rendered the proportion of ss_{yy} which was accounted for by the regression model residuals.

Let $\hat{y}_i$ be the vertical coordinate of the best-fit line with x-coordinate x_i, so $\hat{y}_i \equiv a + b\,x_i$, then the error between the actual vertical point y_i and the fitted point is given by [illegible] (Freedman 2005). We

defined s^2 as an epidemiological, county-level, chlamydia model, estimator for the variance in e_i, whenst $s^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-2}$. In so doing, s was given by $s = \sqrt{\frac{\mathrm{ss}_{yy} - b\,\mathrm{ss}_{xy}}{n-2}} = \sqrt{\frac{\mathrm{ss}_{yy} - \frac{\mathrm{ss}_{xy}^2}{\mathrm{ss}_{xx}}}{n-2}}$. The standard errors for a and b in the model were $\mathrm{SE}(a) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{ss}_{xx}}}$ whilst $\mathrm{SE}(b) = \frac{s}{\sqrt{\mathrm{ss}_{xx}}}$.

2.5 *Regression Estimate*

We generated a predictive model from the Miami-Dade regression analysis to find associated covariates at the zip code level in Hillsborough County to identify populations vulnerable to Chlamydia. We then used a map from Google Earth to create an image for cartographically displaying population vulnerabilities to chlamydia.

3. Results

Twelve covariates were tested for association with chlamydia prevalence in 78 of Miami-Dade's zip codes using a linear regression model in SAS and LULC mapping in GIS's ArcMap 10.4 as shown in Table 4. Hispanic race (p value = 0.0007) and LULC (p-value=0.0262) were found to be statistically significant at the 95% confidence interval as shown in Table 7. Conversely, average income (p-value=0.9341), White race (p-value=0.9164), Black race (p-value=0.9790), Asian race (p-value=0.9602), American Indian race, which was labelled AM_IND within SAS software (p-value=0.3923), Hawaiian race (p-value=0.2361), Other race (p-value=0.04291), male (p-value=0.9483), female (p-value=0.7711), and median age (p-value=0.2136) were not statistically significant at the 95% confidence interval as shown in Table 7. The regression analysis had an r-squared value of 0.966, which indicates an excellent fit to the model as shown in Table 6. Together, this indicates that Hispanic race and the environment are both instrumental to the outcome of this disease. Therefore, clustering of Hispanic populations and specific land cover features must be identified to determine possible hot spots of the disease, as well as areas of correlation. Thus, Miami-Dade was further evaluated at the zip code level due to its nearly 41% Hispanic population located in densely populated urban residential areas with commercial shorelines to the east and everglades to the west.

Table 4. SAS linear regression output. (The REG Procedure; Model: MODEL1; Dependent Variable: CASES_ZIP)

Number of Observations Read	80
Number of Observations Used	78
Number of Observations with Missing Values	2

Table 5. Analysis of Variance

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	12	449023	37419	155.09	<.0001
Error	65	15683	241.27240		
Corrected Total	77	464706			

Table 6. Mean and R-Squared Values

Root MSE	15.53295	**R-Square**	0.9663
Dependent Mean	155.67654	**Adj R-Sq**	0.9600
Coeff Var	9.97770		

Table 7. Parameter Estimates

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	13.77362	19.45987	0.71	0.4816
AVG_INCOME	AVG_INCOME	**1**	-0.00000685	0.00008255	-0.08	0.9341
WHITE	WHITE	**1**	0.00234	0.02222	0.11	0.9164
BLACK	BLACK	**1**	0.00058440	0.02211	0.03	0.9790
HISPANIC	HISPANIC	**1**	-0.00227	0.00063619	-3.56	0.0007
ASIAN	ASIAN	**1**	0.00084162	0.01682	0.05	0.9602
AM_IND	AM_IND	**1**	-0.05537	0.06430	-0.86	0.3923
HAWAIIAN	HAWAIIAN	**1**	-0.06565	0.05490	-1.20	0.2361
OTHER	OTHER	**1**	0.01407	0.01768	0.80	0.4291
MALE	MALE	**1**	0.00146	0.02239	0.07	0.9483
FEMALE	FEMALE	**1**	0.00651	0.02227	0.29	0.7711
MED_AGE	MED_AGE	**1**	-0.64574	0.51414	-1.26	0.2136
LULC	LULC	**1**	6.02223	2.64659	2.28	0.0262

4. Discussion

The regression variables showed that Hispanics were the most important statistic in Miami-Dade for association with chlamydia case prevalence. Hispanics may be at greater risk for infection of chlamydia due to their lack of appointments and check-ups with medical examiners (Marks, Garcia, and Solis 1990). This irregularity in appointments can be a major indicator of a problem since chlamydia is typically an asymptomatic disease, which can easily spread throughout communities' undetected if routine screenings and management are not properly upheld.

The least squares method for fitting the geosampled, vulnerability, forecast, chlamydia prevalence prognosticators for the unknown parameters in the county-level model were estimated by minimizing the sum of the squared deviations between the data and the model. The minimization process reduced the overdetermined system of equations formed by the geosampled chlamydia data to a sensible system of p, (where p was the number of parameters in the functional part of the county model). This new system of equations was then solved to obtain the statistically significant parameter estimators.

The SAS output in Table 7 indicated that LULC is a significant covariate associated with chlamydia prevalence. We thus constructed a LULC model to determine which land cover feature was most prevalent in the zip code with the highest case rate of chlamydia in Miami-Dade. Zip code 33012 had the highest prevalence of chlamydia, 320 cases per 100,000 population, in Miami-Dade County (Florida Health Charts 2017). Therefore, the land cover features of zip code 33012 were identified to determine the highest in this area: urban commercial, urban residential, water, or parkland/pastureland. To do this, the total square meters (m2) of all features was added together, and then subsequently divided to determine percent breakdowns by land feature type. The calculation breakdown of the different land use features is shown in Table 8. Urban residential was determined to be the primary land cover in zip code 33012 with 70% of the land covered by this feature. The second highest land cover feature was urban commercial which covered 26% of the land. Water made up only 4% of the land cover and parkland/pastureland made up 0% of the land cover in this area as shown in Table 8 and Figure 5. This indicates that a land cover that is primarily composed of urban residential followed by urban commercial and little to no water or pastureland/parkland may be significantly associated with the prevalence of chlamydia.

Table 8. Miami-Dade zip code 33012 LULC shape areas with percent breakdown by category.

LULC Categories	Urban Commercial	Urban Residential	Pastureland/ Parkland	Water
Areas Calculated for zip code 33012 (m^2)	276179.80	5742.02	0	42623.04
	36344.76	169782.08		8259.21
	24439.63	33086.15		88522.74
	47427.78	10096.83		171872.57
	47579.82	15530.17		75555.25
	30303.19	11109.14		62780.33
	48850.97	176312.02		91022.90
	27172.24	230857.41		65250.79
	3758.67	203221.93		27858.67
	4403.50	120081.76		
	3398.80	3735379.09		
	44088.48	82810.23		
	112593.32	102036.90		
	9165.36	1631.83		
	1675.07	10467.15		
	67620.03	1684608.53		
	55577.01	67117.51		
	30158.72	261029.14		
	30878.10	36590.91		
	6617.87	37897.26		
	22608.37	266051.25		
	24992.22	2332323.33		
	21593.88	117064.33		
	1587.50	54476.58		
	17844.73	140324.92		
	80165.86	145838.73		
	56570.65	94128.37		
	51458.96	205705.18		
	82161.79	146465.23		
	4325.31	259681.06		
	13702.17	56586.87		
	110032.39	83243.52		
	13193.66	8092.30		
	29280.48			
	51959.52			
	101580.05			
	36682.05			
	91209.22			
	151962.72			
	39696.15			
	29285.81			
	160205.07			
	150312.42			
	129765.16			
	61318.25			
	175971.21			
	38363.51			
	249724.34			

Table 8 Continued…

LULC Categories	**Urban Commercial**	**Urban Residential**	**Pastureland/ Parkland**	**Water**
Areas Calculated for zip code 33012 (m^2)	77575.048			
	123989.22			
	133209.47			
	182411.52			
	31117.66			
	26080.60			
	26345.90			
	23391.16			
	283093.75			
	134624.28			
	39185.54			
LULC total sq. meters by category	**3,986,810.71**	**10,905,369.74**	**-**	**633745.50**
LULC % breakdown by category	**26%**	**70%**	**0%**	**4%**
LULC total sq. meters	**15,525,925.96**			

Figure 5. LULC map of zip code 33012, which had the highest prevalence of chlamydia in Miami-Dade (320 cases per 100,000 population) in 2015.

It is possible that the dense population of commercial businesses next to highly populated living quarters breeds disease between the community enterprises and neighborhoods. Of particular importance are

some of the establishments within zip code 33012, which may contribute to this being the highest area for chlamydia within Miami-Dade. Establishments include: an Immigration and Travel Center, Obamacare Center, Florida Patrol Investigators, Southern Winds Hospital, Larkin Community Hospital, Westland Hialeah Senior High School, Florida National University, Hialeah-Miami Dade College, Ramada Hialeah/Miami Airport, Westland Mall, Hialeah Mercado, Epworth Village Retirement Community, and a multitude of car dealerships, body shops, apartments, shopping enterprises, and ethnically diverse restaurants and service companies. The number of universities may indicate a younger population that is more sexually active, high density neighborhoods may allow closer daily contact of individuals, the Immigration Center indicates that the area may be home to foreigners with various sexual and medical practices, and the retirement community may also be a place of higher sexual activity and therefore transmission of sexual infections. Future studies may further examine some of these specific features within the area to determine any possible statistical connections to high chlamydia rates. This LULC analysis serves as a proxy for other counties to create predictive maps for determining where chlamydia rates may be higher based on geographic location and topographical land cover features.

Landscapes have been noted as drivers of many chronic infectious disease processes (Patz et al. 2000). By utilizing an LULC estimate, our assumption was that we would be able to localize a vulnerable region to chlamydia in Hillsborough County, which has the fourth highest rate of chlamydia in the state of Florida at 7,600 cases per 100,000 population. Combination of the regression and LULC estimates enabled the specific predictive mapping of vulnerable regions to chlamydia in Hillsborough County. Such a predictive map will enable more precise resource allocation in Hillsborough County for targeting of these vulnerable georeferenceable, geographic locations, which may be associated with high rates of chlamydia infection. In so doing, the rates of transmission of chlamydia in this county would likely be definitively lowered. Additionally, predictive chronic infectious disease maps are crucial for helping to implement control strategies at the county level (Griffith 2005). The LULC map created for Hillsborough County is shown in Figure 6.

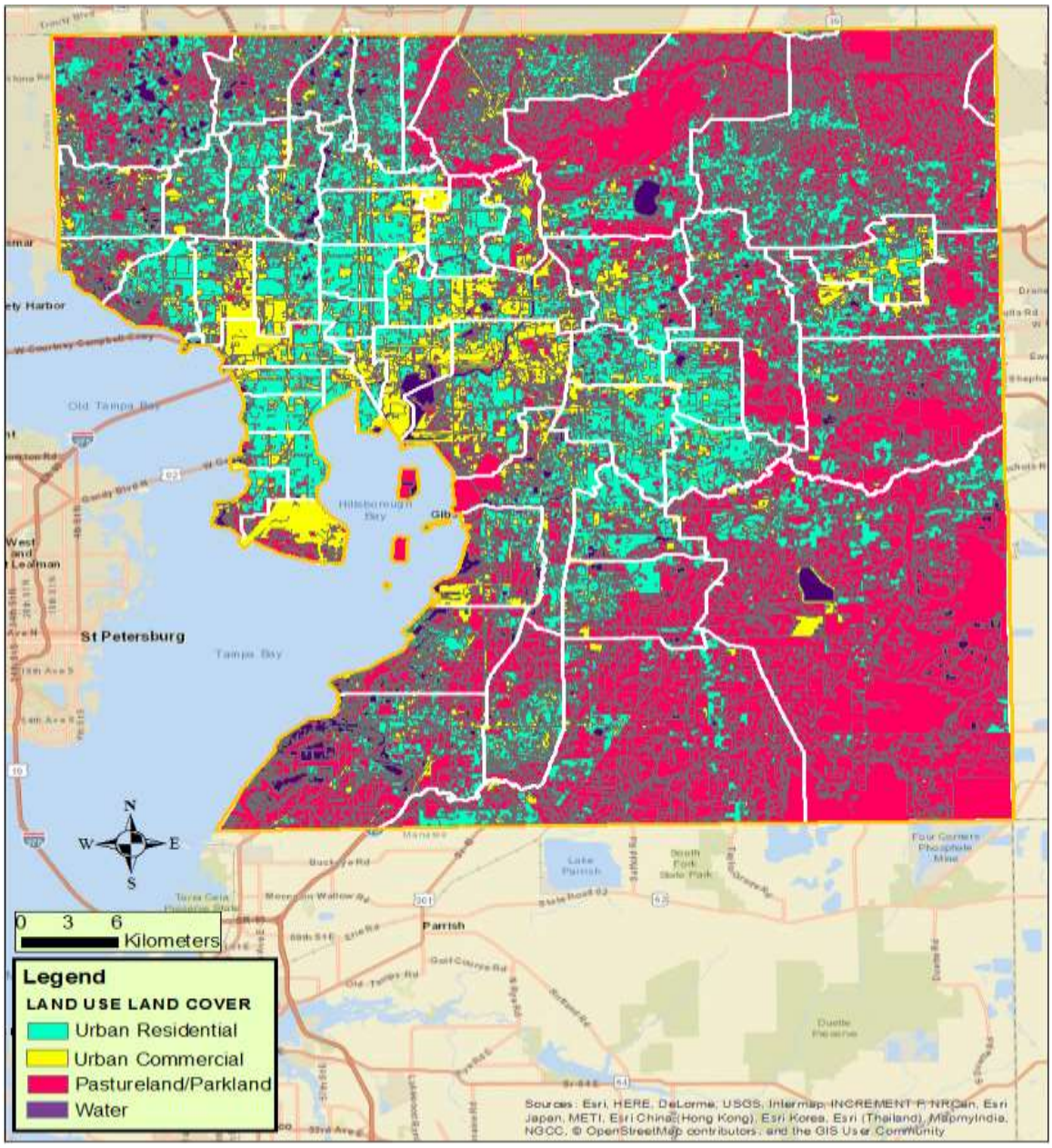

Figure 6. Hillsborough County LULC map with zip code regions outlined in white, 2015.

Although the findings in this paper were able to determine vulnerable regions of potential transmission sites of chlamydia in Hillsborough County, the research warrants further investigation.

Autocorrelation models may reveal clustering tendencies within georeferenced county-level, zip code, polygons. This would enable a researcher to optimally determine "hot spots" throughout a county level, zip code polygon model. Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional surface, introducing a deviation from the independent observations assumption of classical statistics (Griffith 2009). Spatial autocorrelation exists

because real world phenomena (chlamydia county-level case distribution) are typified by orderliness, pattern, and systematic concentration, rather than randomness. Tobler's First Law of geography encapsulates this situation: "everything is related to everything else, but near things are more related than distant things," (Tobler 1970). To this maximum should be added the qualifier when county-level chlamydia forecast vulnerability modelling: "but not necessarily through the same mechanisms," (Griffith 2009). In other words, spatial autocorrelation in a county-level, regression, forecast, vulnerability chlamydia model, would mean a dependency exists between values of a geosampled, heterogeneous landscape covariate or a georeferenceable, sociodemographic explanator in a neighboring or proximal county geolocation. A systematic pattern in values of a grid-stratifiable, predictor, explanatory, georeferenceable, zip code variable may occur across the county geolocations on a vulnerability, time series map due to underlying common factors (e.g., median income). Spatial autocorrelation has many interprets: a nuisance parameter, self-correlation, map pattern, a diagnostic tool, a missing variables surrogate, redundant information, a spatial process mechanism, a spatial spill over, and the outcome of area unit demarcation in a chronic infectious forecast vulnerability model (Griffith 2005, Jacob et al. 2005).

Interpreting spatial autocorrelation as an explanative, topological, chlamydia-related, georeferenecable, GIS map pattern may aid in emphasizing conspicuous trends, gradients, swaths, or mosaics across a time series, county-level, forecast, vulnerability, grid-stratified map. Consider a constant, in a vulnerable, county-level, diffuse chlamydia population model which has a degenerative case (i.e., a constant with no variance) of perfect positive spatial autocorrelation (e.g., aggregation of similar race attributes) in a GIS-derived, georeferenced, geospatial cluster. Once the value of a constant is known at a single county geolocation (e.g., zip code polygon), it may be known at all geolocations via an iterative quantitative interpolation in GIS.

Next, a chlamydia researcher may consider a geosampled sociodemographic or geoclassifiable explanatory, diagnostic, prognosticative, georeferenceable, LULC, variable in GIS that portrays a north-south (or east-west) linear trend across a vulnerability map. If this variable has a mean of zero, then it would be geometrically orthogonal to and uncorrelated with the constant in any GIS constructed, county-level, chlamydia model. These north-south and east-west oriented, explicative, linear trend variables may also be orthogonal and uncorrelated. A geosampled explicative, temporally dependent, chlamydia-oriented, predictor variable with mean zero whose values' magnitudes form a 3-dimensional symmetric GIS mound in a georeferenced zip code polygon of a grid-stratified, county map may constitute yet another mutually orthogonal and uncorrelated cartographic pattern. These three variables may display maximum levels of positive spatial autocorrelation in

GIS whenst geographic variance is present in an empirical, geosampled, geoclassified, landscape or socio-demographic, explanatory, county level, georeferenced, forecast, vulnerability, chlamydia-related dataset, which may be described as a global geographic pattern in GIS. Alternating sequences of small mounds and basins with either an east-west or a north-south county-level orientation in GIS may portray weak positive spatial autocorrelation, and constitute local chlamydia risk map patterns. This fragmentation may continue through randomness (i.e., zero spatial autocorrelation) to arrangements of increasingly alternating LULC or socio-demographic values (i.e., single value mounds and basins), which may portray increasing negative spatial autocorrelation in the chlamydia data. Most substantive chronic infectious disease variables have geographic distributions that can be described in GIS by linear combinations of some subset of these mutually orthogonal and uncorrelated data that may vary in size mound-basin map patterns at the county-level (Griffith 2005, Jacob et al. 2005). Further, inconspicuous error may be teased out (spatial heteroscedasticity pseudo-replicated data) in the county-level chlamydia model.

An iterative Bayesian model may also reveal evidential probabilities of regression covariates in GIS associated with the prevalence of chlamydia infection at the county level. Bayesian spatial modelling refers to the application of Bayesian methodology to spatial model data (non-linear county-level chlamydia covariates), such as spatial autoregressive models and conditional autoregressive models (Congdon 2001). The underlying Bayesian spatial modelling concept is known as Bayes' theorem. This theorem establishes both the distribution of data and the unknown coefficient estimates (LeSage and Pace 2009). Simulations within hierarchical, generalizable, non-frequentistic inferential frameworks in GIS may reveal temporally dependent, explanatory, parameterize-able, temporal, regressors associated with a dependent variable as this paradigm utilizes a subjective maximum likelihood estimation (Jacob et al. 2013). This GIS estimate may allow a Bayesian approach to specify a hierarchical, diagnostic, forecast, vulnerability, chlamydia, county-level model with a prior distribution over grid-stratifiable, zip code, polygonized hyperparameters whilst specifying the prior distribution of the weights relative to these hyperparameters. These residuals may be connected to county-level, geosampled, georeferenced, sociodemographic, or landscape geoclassified chlamydia datasets via an observation model; for example, in a regression context, the value of the dependent variable may be corrupted by Gaussian noise. Given an observed county-level geoclassified, demographic or landscape, grid-stratified empirical robustifiable dataset, a posterior distribution over the chlamydia-related regression weights and hyper-parameters (rather than just a point estimate) may be induced. However, for a neural network, county-level, time series, georeferenceable, chlamydia model, this posterior cannot be obtained analytically in GIS; computational

methods may have to include approximations in SAS (MacKay 1992) employing the evaluation of integrals using Monte Carlo methods (Neal 1996). SAS provides convenient tools for applying non-frequentistic regression methods, including built-in capabilities in the GENMOD, FMM, LIFEREG, and PHREG procedures (called the built-in Bayesian procedures), and a general Bayesian modelling tool in the MCMC procedure (https://support.sas.com/). In the Bayesian approach to neural networks, a prior on an empirical regressed dataset of grid-stratified, county-level forecast, vulnerability, georeferenceable, epidemiological, chlamydia model weights of a network would induce a prior over functions in SAS, which then may be mapped in GIS.

An alternative method for putting a prior over functions may be to employ a Gaussian process (GP) prior over functions in a county-level, georeferenceable, forecast vulnerability, GIS, chlamydia, regression model. This idea has been employed for a long time in the spatial statistics community under the name of "kriging" although it seems to have been largely ignored as a general-purpose regression method for chlamydia, county-level, predictive mapping. In statistics, kriging or Gaussian process regression is a method of interpolation for which the interpolated values are modelled by a Gaussian process governed by prior covariance's, as opposed to a piecewise-polynomial spline chosen to optimize smoothness of the fitted values (Cressie 1993). Kriging assumes that the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface (Griffith 2005). The Kriging tool in GIS will fit a mathematical function to a specified number of geosampled-stratified, chlamydia, data points, or all points within a specified grid radius, to determine the output value for each county-level, geolocation. Kriging is a multistep process; it includes exploratory statistical analysis of the data, variogram modeling, creating the surface, and (optionally) exploring a variance surface (www.esri.com).

The inverse distance weighted and Spline interpolation tools are referred to as deterministic interpolation methods because they are directly based on the surrounding measured values, or on specified mathematical formulas that determine the smoothness of the resulting surface. A second family of interpolation methods consists of geostatistical methods, such as kriging, which are based on statistical models that include autocorrelation—that is, the statistical relationships among the measured points. Because of this, geostatistical techniques in SAS and GIS not only have the capability of producing a prediction surface at the county-level regression dataset, but also provide some measure of the certainty or accuracy of the predictions.

Non-subjective probabilities play a central role in many chronic infection, regression-related decisions (Jacob et al. 2009, Griffith 2005) and may act as an immediate confounder of inferences about county-level, chlamydia, grid-stratifiable, mapping data unless controlled for. Several procedures to recover non-subjective

probabilities have been proposed in the literature, but in order to recover the correct latent probabilities in a county-level, forecast, vulnerability, real-time, cartographic, zip code, polygon, grid-stratified, chlamydia case distribution, or prevalence statistic, a researcher or STD specialist must either construct elicitation mechanisms that control for risk aversion (e.g., bootstrapping), or construct elicitation mechanisms (latent auto covariate matrices) which would undertake "calibrating adjustments" to elicited reports rendered from these models in GIS. To illustrate how the joint estimation of risk attitudes and non-subjective probabilities can provide the calibration adjustments for a county-level, prognosticative chlamydia, regression-oriented, time series, a GIS approach may be explored whereby the empirical data from a controlled county experiment is used initially to develop a landscape or sociodemographic signature. This would allow a STD specialist to make inferences about the latent non-subjective probability in the chlamydia model, under virtually any well-specified uncertainty specification (e.g., heteroscedasticity of variance) for optimally diagnosing subjective risk of chlamydia infection on vulnerable georeferenced populations, whilst still employing relatively simple elicitation mechanisms.

Other forms of estimation for a regression, county-level chlamydia, signature model can be constructed by picking a different class of interpolates in GIS. For instance, rational interpolation of a regression iterative interpolative chlamydia estimate may be conducted by rational functions using Padé approximant, and trigonometric interpolation which may be enabled by trigonometric polynomials using Fourier series. In mathematics, a Fourier series is a way to represent a function as the sum of simple sine waves. More formally, the series may decompose any periodic function or periodic signal generated from a georeferenced, county-level, chlamydia, grid-stratified, zip code polygon into the sum of a (possibly infinite) set of simple oscillating functions, namely sines and cosines (or, equivalently, complex exponentials). The discrete-time Fourier transform may be a periodic function in a county-level, georeferenced, chlamydia polygon which may be defined in terms of a Fourier series in GIS.

5. Conclusion

GIS and SAS regression models are effective tools for revealing and generating regression models and LULC maps, which can geographically locate high prevalence areas of chlamydia case distribution rates in order to optimally forecast vulnerability to chlamydia in other counties based on the model created. The regression analysis indicates that urban residential is the most important LULC feature, followed closely by urban commercial as an important covariate associated with chlamydia prevalence. However, this analysis

provides only a broad scope for the role that LULC plays in the onset and prevalence of chlamydia. The linear regression identifies LULC as a significant covariate with preliminary findings that land covered 96% by urban residential and urban commercial is a possible predictor of where exactly chlamydia may be occurring. However, this analysis does not specifically explain which land cover features within the specified area may be significant, such as: buildings, enterprises, businesses, universities, or housing styles (apartments, homes, and townhouses). Thus, a more in-depth analysis of specific zip code breakdown mixtures employing geoclassified LULC attribute features are needed to further understand why this occurrence is significant. This analysis provides the baseline for future research to be conducted to determine the exact impact of varying landscapes on sexually transmitted infections like chlamydia. This is important to public health because rates of sexually transmitted infections are on the rise, and improved targeting for prevention and treatment programs are essential for curbing this upward trend.

References

Akivis, M.A. and Goldberg, V.V., 1972. An introduction to linear algebra and tensors. New York: Dover, 71.

Alekinster, M.M., 2015. 11 Interesting facts you may not know about Florida. *Interactive Education Concepts Inc.* [online]. Available from: https://www.myimprov.com/11-facts-about-florida/ [Accessed 19 June 2017].

Arfken, G.B. and Weber, H.J., 2000. *Mathematical methods for physicists.* 5th ed. Boston, Massachusetts: Academic Press, 14-15.

Borisenko, A.I. and Taparov, I.E., 1968. *Vector and tensor analysis with applications*. New York: Dover, 14.

Centers for Disease Control and Prevention, 2004. Trends in reportable sexually transmitted diseases in the United States, 2005. *CDC* [online]. Available from: https://www.cdc.gov/std/stats04/trends2004.htm [Accessed 3 June 2017].

Centers for Disease Control and Prevention, 2015. Sexually transmitted disease surveillance. *CDC* [online]. Available from: https://www.cdc.gov/std/stats15/STD-Surveillance-2015-print.pdf [Accessed 20 July 2017].

Chang, B.A., Pearson, W.S., and Owusu-Edusei, K., 2017. Correlates of county-level nonviral sexually transmitted infection hot spots in the US: application of hot spot analysis and spatial logistic regression. *Annals of Epidemiology,* 27 (4), 231-237.

Congdon, P., 2001. *Bayesian statistical modelling.* New York: John Wiley & Sons, Inc.

Courant, R. and Hilbert, D. 1989. *Methods of mathematical physics.* Vol. 1. New York: Wiley, 7.

Cressie, N., 1993. *Statistics for spatial data.* Revised ed. New York: John Wiley & Sons, Inc.

Dineen, C., 2017. Florida welcomed nearly 113 million tourists in 2016. *Orlando Sentinel* [online]. Available from: www.orlandosentinel.com/travel/os-bz-visit-florida-tourism-2016-story.html [Accessed 27 June 2017].

Draper, N.R. and Smith, H., 1998. *Applied Regression Analysis*. 3rd ed. New York: Wiley-Interscience.

Earth Explorer, 2016. Miami-Dade Data Se LE07_L1TP_015042_20160507_20160902_01_T1 [online]. Available from: https://earthexplorer.usgs.gov/ [Accessed 25 June 2017].

Esri, 2014. FL_Counties. *United States Census Bureau-Tiger/Line Files* [online]. Available from: http://www.esri.com [Accessed 8 June 2017].

Esri, 2017. Classifying imagery using ArcGIS. *Esri Training* [online]. Available from: https://www.esri.com/training/catalog/57d0800584b087dd46817ceb/classifying-imagery-using-arcgis/ [Accessed 20 June 2017]

Florida Health Charts, 2017. Chlamydia cases. *Florida Department of Health, Bureau of Communicable Diseases* [online]. Available from: http://www.flhealthcharts.com/charts/OtherIndicators/NonVitalSTDDataViewer.aspx?cid=0145 [Accessed 18 June 2017].

Florida Geographic Data Library, 2012. Zip code areas (five-digit) in Florida. *University of South Florida GeoPlan Center* [online]. Available from: http://www.fgdl.org/metadata/fgdl_html/zipbnd_2012.htm [Accessed 27 June 2017].

Fox, J., 1997. *Applied regression analysis, linear models and related methods*. California: Sage Publications, Inc.

Freedman, D., 2005. *Statistical models: Theory and practice.* New York: Cambridge University Press.

Freedman, D., 2009. *Statistical models: Theory and practice.* 2nd ed. New York: Cambridge University Press.

Griffith, D.A., 2005. A comparison of six analytical disease mapping techniques as applied to West Nile Virus in the coterminous United States. *International Journal of Health Geographics*, 4 (18).

Griffith, D.A., 2009. Spatial Autocorrelation. *Elsevier Inc.* [online]. Available from: https://booksite.elsevier.com/brochures/hugy/SampleContent/Spatial-Autocorrelation.pdf [Accessed 21 August 2017].

Hosmer, D.W. and Lemeshow, S., 2000. *Applied Logistic Regression*. New Jersey: John Wiley & Sons.

Jacob, B.G., et al., 2005. Evaluation of environmental data for identification of Anopheles (Diptera: Culicidae) aquatic larval habitats in Kisumu and Malindi, Kenya. *Journal of Medical Entomology,* 42 (5), 751-755.

Jacob B.G., et al., 2009. A heteroskedastic error covariance matrix estimator using a first-order conditional autoregressive Markov simulation for deriving asympotical efficient estimates from ecological sampled Anopheles arabiensis from ecological sampled Anopheles arabiensis aquatic habitat covariates. *Malaria Journal*, 8 (1), 216-225.

Jacob, B.G., et al., 2011. A cartographic analysis using spatial filter logistic model specifications for implementing mosquito control in Kenya. *Urban Geography*, 32 (2), 263-300.

Jacob, B.G., et al., 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. *Journal of Public Health and Epidemiology*, 5 (3), 122-143.

Jacob, B.G., et al., 2014. Denoising a model employing automated bandwidth selection procedures and pre-whitened Euclidean-based quadratic surrogates in PROC ARIMA for optimizing asymptotic expansions and simulations of onchocerciasis endemic transmission zones in Burkina Faso. *Journal of Public Health and Epidemiology,* 6 (11), 347-389.

Jain, P.K., Ahuja, O.P., and Ahmad, K. 1995. 5.1 Definitions and basic properties of inner product spaces and Hilbert spaces. *Functional analysis.* 2nd ed. New Delhi: New Age International, 203.

Kay, J., 2015. Busy waters off Florida: Illegal immigration surges from across Caribbean. *The Associated Press* [online]. Available from: www.tbo.com/news/florida/busy-waters-off-florida-illegal-immigration-surges-from-across-caribbean-20150104/ [Accessed 27 June 2017].

Klaveren, D.V., et al., 2016. Prediction of Chlamydia trachomatis infection to facilitate selective screening on population and individual level: a cross sectional study of population-based screening programme. *Sexually Transmitted Infections,* 92 (6), 433-440.

Kruskal, W.H. and Tanur, J.M., 1978. Linear hypotheses. *International Encyclopedia of Statistics.* New York: Free Press, 1.

LeSage, J. and Pace, R.K., 2009. *Introduction to spatial econometrics.* Boca Raton: Chapman & Hall/CRC, 5, 123-154.

Lindley, D.V., 1987. *Regression and correlation analysis.* 4th ed. New Palgrave: A Dictionary of Economics, 120–123.

Lipschutz, S. and Lipson, M., 2009. *Schaum's outlines linear algebra.* 4th ed. New York: McGraw Hill.

MacKay, D.J.C., 1992. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4 (3), 448-472.

Marks, G., Garcia, M., and Solis, J.M., 1990. III. Health Risk Behaviors of Hispanics in the United States: Findings from HHANES, 1982-84. *American Journal of Public Health*, 80, 20-26.

Neal, R. M., 1996. *Bayesian learning for neural networks*. New York: Springer Science+Business Media, LLC.

Patz, J.A., et al., 2000. Effects of environmental change on emerging parasitic diseases. *International Journal for Parasitology,* 2000 (30), 1395–1405.

Sen, A. and Srivastava, M., 2011. Regression analysis - theory, methods, and applications, *Springer-Verlag* [online], 4. Available from: https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Regression_analysis.html [Accessed 20 July 2017].

Tobler W., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-24

US Census TIGER Geodatabases, 2016. CensusGeodatabase_MiamiDadeCounty. *GISCIENCE 2015 TIGER Geodatabases* [online]. Available from: https://www.census.gov/geo/maps-data/data/tiger-geodatabases.html [Accessed 8 June 2017].

Watson, G.N., 1996. *A treatise on the theory of Bessel functions.* 2nd ed. London: Cambridge University Press.

Weisstein, E.W., 2017. Kronecker Delta. *Math world – a Wolfram web resource* [online]. Available from: http://mathworld.wolfram.com/KroneckerDelta.html [Accessed 3 September 2017].

Wolfram Research, Inc., 2018. *Identity matrix* [online]. Available from: http://mathworld.wolfram.com/IdentityMatrix.html [Accessed 12 June 2017].

Zip-Codes.com, 2017. MIAMI-DADE County, FL ZIP Codes. *Datasheer, L.L.C.* [online]. Available from: https://www.zip-codes.com/county/fl-miami-dade.asp [Accessed 11 June 2017].