International Journal of Geographic Information System
Vol. 4, No. 3, October 2017, pp. 1-21
Available Online at http://acascipub.com/Journals.php

Research article

Displaying non-subtle randomness within a Durbin-Watson statistic for quantitating non-zero autocorrelation in grid- stratified, prostate cancer, zip code polygons geographically sampled in Hillsborough County, Florida

Mmadili N. Ilozumba[a], Toni Panaou[b], Fahad Mukhtar[a], Nnadozie Emechebe[a], Samuel Alao[c] and Benjamin G. Jacob[c]

[a]Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida,
Tampa FL, USA
[b]Department of Civil and Environmental Engineering, College of Engineering, University of South Florida,
Tampa FL, USA 33612
[c]Department of Global Health, College of Public Health, University of South Florida,
Tampa FL, USA 33612

*Corresponding author:
Benjamin Jacob, Department of Global Health, College of Public Health, University of South Florida,
12901 Bruce B. Downs Blvd., Tampa, 33612, USA, Tel: (813) 974-9784;
E-mail: bjacob1@health.usf.edu

ABSRACT

Although linear and non-linear regression models have been constructed to quantitate explanatory covariates associated with the prevalence of many chronic infections, they have been underutilized in prostate cancer research at the county level. The fundamental assumptions in the linear regression of vulnerability, prostate, cancer forecast, epidemiological models could be that the error terms εi have mean zero and constant variance and are uncorrelated whenst E (εi) = 0, Var (εi) = σ^2 , and E (εiεj) = 0. For purposes of testing hypotheses associated with county-level, prostate cancer inferential covariates, confidence intervals (CI) generated must be within the assumption of normality of error distribution; hence, εi should optimally be NID (0, σ^2). Some applications of this type of regression would involve multivariate regressors and response variables that have a natural sequential order (e.g., iterative interpolated, ArcGIS derived, prostate cancer, zip code, grid-stratified polygon renderings). Parsimoniously quantitating county-level, prostate cancer, vulnerability, regression, prognosticative model, diagnostic explanators may require that the assumption of uncorrelated or independent explicative errors be adjusted (e.g., from 95 percent CIto 90 percent CI) in order to fit non-violations of frequentistism (e.g., homoscedasticty of variance). Propagational, non-normalities in county-level prostate cancer, empirical datasets may exhibit serial correlation, that is, E ($\varepsilon i\varepsilon j$ 6= 0). Such error terms may be autocorrelated. This randomness may be ascertained by computing non-zero, latent, autocorrelation coefficients within a geographically weighted matrix for optimizing, county-level, polygonized, semi-parametric, zip code values at varying lags in AUTOREG. If random, such prostate cancer regression autocorrelations should be near zero for any and all lag separations. If non-random, then one or more of the residual autocorrelation coefficients in the zip code, polygon, geographically classified, regression model would be significantly non-zero. In this analysis, we initially constructed a multivariate linear regression of prostate cancer withna vulnerability model framework where $yt = \beta 0 + \beta 1xt + \varepsilon t$, and where $\varepsilon t = \rho\varepsilon t-1$ in PROC REG. In the model, yt and xtgeographically represented zip code,polygonized, observational covariates on the response and regressor variables which were revealed at t in AUTOREG. A first-order Durbin-Watson autocorrelation model was constructed where y was a first-order, vulnerability, county-level, prostate cancer affiliated process observed at homogeneously spaced levels, that is, $\varepsilon t = \rho\varepsilon t-1$, where εt was the error term at t, and NID (0, σ^2) represented a random,georefernceable, zip code polygonized, explanatory, predictor variable geosampled in Hillsborough County in Florida. p ($|\rho| < 1$) was the autocorrelation coefficient at the zip code level. We were able to robustly quantitate a linear estimate (median household income) at a 95% CI. We defined clustering propensities in the prostate cancer vulnerability explanatorialregressors. Our model illustrated positive autocorrelation. We were also able to identify a georeferenceable, hotspot location within the zip code polygon in ArcGIS. Situations where negative autocorrelation occurs were not encountered in the model. Regressively targeting grid-stratifiable, zip code, grid-stratified, regression polygon georeferenced geolocations within county areas can represent populations vulnerable to prostate cancer.

Keywords: Prostate cancer; prevalence rate; autocorrelation; zip code polygon; multiple linear regression.

1. Introduction

Prostate cancer is a chronic disease specific to only males. It involves the abnormal growth of cells in the prostate gland which is typically malignant (American Cancer Society 2017). In the United States (U.S.), prostate cancer is the first most common incident cancer and the second leading cause of cancer deaths among males (Centers for Disease Control and Prevention 2016). While there is a greater number of new cases of prostate cancer among White Americans, death rates from prostate cancer is higher among African Americans, compared to other races in the U.S. (National Cancer Institute 2017).

Identification of cancer clusters with the use of geographical analytical tools have been well established (Grubesic and Matisziw 2006, DeChello and Sheehan 2007, Meliker et al. 2009). Abe et al.(2006) employed the use of Spatial Scan Statistics (SaTScan) to identify high proportion of men in New Jersey who were diagnosed with late-stage prostate cancer. They found that Northeastern New Jersey had the highest cluster with its population comprising mostly of African Americans, Asians and Hispanics. In a recent study, Wagner et al.(2013) used Geographical Information System (GIS) to analyze and identify prostate cancer incidence hotspots in Southwestern Georgia. They also used a discrete Poisson count variab;e model to detect prostate cancer incidence rate in both African American and European American men in the north and northwest central Georgia. The application of GIS in analyzing health data has helped to inform health related interventions due to its benefit in mapping out geographical areas in need of such interventions and other strategies required for the control of diseases.

The number of new cases of prostate cancer in the U.S. varies from state to state. As a result, several research on prostate cancer have used Florida as a case study. Goovaerts et al. (2016) applied a binary, logistic regression analysis in a geographical context with the aim of analyzing the relationship between the proportion of late stage prostate cancer cases and potential demographic factors in Florida. Xiao et al. (2007)employed GIS to analyze racial disparities. They equally employed multilevel,multivariate, logistic regression to examine the impact of individual, county and census tract level factors on prostate cancer incidence in Florida. Additionally, Xiao et al. (2011) studied the disparities in late-stage prostate cancer in Florida in terms of race and geography. Although the aforementioned studies had some merits, they failed to cartographically geographically locate georeferenceable, vulnerable regions of total prostate cancer patients,grid-stratified by county levels.

Optimally, a regression approach for GIS modelling the relationship between a scalar dependent, geosampled, county-level, randomized, prostate cancer explanator Y and one or more time series, prognosticativevariables (independent variables),denoted, X may reveal vulnerable county-level, zip code,grid-stratifiable,georeferenceable populations in SAS. In a linear regression probability model the predictor function is linear in the parameters but not necessarily linear in the regression variables(Draper & Smith 1998). A geosampled empirical dataset of county-level, prostate cancer, vulnerability parameters may be unbiasedly estimated so that a measure of fit is optimally diagnosed in an ArcGIS cyberenvironment or inSAS employing geostatistical algorithms (e.g., eigenfunction, orthogonal, spatial filter, decomposition algorithm in AUTOREG) . For example, the equation for the $i$thgeoreferenced, zip code,polygonizedclinical observation (prostate-specific antigen (PSA) affiliated covariate) might be$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$where Y_i is the response variable, x_i is a county-level, prostate cancer, regression variable, β_0 and β_1 are unknown parameters, and ε_i is an error term. This model may be termed the simple linear regression (SLR) model, as it would be linear in β_0 and β_1 which maycontain only a single diagnostic, prostate cancer, grid-stratified, explicative regressor. The relationship may be optimally modelled using linear predictive functions, where the conditional mean of Y given the value of X may be assumed as an affine function of X. Affine functions in linearized represent vector-valued functions of the form $f(x_1, \dots, x_n) = A_1 x_1 + \dots + A_n x_n + b$, where coefficients can be scalars or dense or sparse matrices[Fox 199.]. These models although simplistic in nature may be fitted using a least square approach in SAS but they may be also fitted in other ways in ArcGIS (e.g., county model, zip code calculations based on either Euclidean or Manhattan distance measures from projected,hyperendemic,georeferenceablegeolocations) minimizing the lack of fit. In so doing, the least absolute deviations in a predictive, county-level, prostate cancer, SLR, stratified, risk-related,georeferenced polygon may be exemplified.

Least absolute deviations (LAD), also known as least absolute errors (LAE), least absolute value (LAV), least absolute residual (LAR), sum of absolute deviations, or the L_1 norm condition, is a statistical optimality criterion and the statistical optimization technique that relies on it [Fox 1997]. In statistics, an optimality criterion provides a measure of the fit of the data to a given hypothesis, to aid in model selection[Draper 1998]. A, vulnerability epidemiological, prostate cancer, county-level, epidemiological, regression diagnostic, forecast, model may be designated as the "best" of the candidate models if it’s renderings delineate the optimal value of an objective function measuring the degree of satisfaction of the criterion used to evaluate the alternative hypotheses. Similar to the popular least squares technique, the model may attempt to find a function which closely approximates an empirical dataset of geosampled, regressable prostate cancer covariates

The method of least squares is a standard approach in regression analysis to the approximate solution of overdetermined systems,[i.e., sets of equations in which there are more equations than unknowns]. "Least squares" incounty-level, prostate cancer, vulnerability, forecast, regression model would mean that the overall solution (e.g., statistically significant zip code hyperendemic foci, explanator) could minimize the sum of the squares of the model residuals made in the results of every single predictive equation.

Statistical significance would play a pivotal role in statistical hypothesis testing of a county-level, zip code, grid-stratified, prostate cancer, regression model. It would be usable to determine whether the null hypothesis should be rejected or retained(e.g., vulnerable populations to prostate cancer reside in lower economic regions of a county)(< than 20,000/annual income), The null hypothesis would then be the default assumption that nothing happened or changed in the sociodemiographic.(, education, income, and occupation) specified, variables when regressed. For the null hypothesis to be rejected, an observed result has to be statistically significant, [i.e. the observed *p*-value is less than the pre-specified significance level (95% CI)].

To determine whether a result from a forecast vulnerable prostate cancer, regression model is statistically significant, a county-level epidemiologist or researcher may calculate a *p*-value, which may be the probability of observing a clinical effect given that the null hypothesis is true. Commonly. the null hypothesis is rejected if the *p*-value is less than a predetermined level, α. α is called the significance level, and is the probability of rejecting the null hypothesis given that it is true type I error[e.g., the incorrect rejection of a true null hypothesis (a "false positive"),]{ Fox 1997].

When α is set to 5% in a prostate cancer forecast, vulnerability, epidemiological model the conditional probability of a type I error, given that the null hypothesis is true, is 5%,and a statistically significant result is one where the observed p-value is less than 5%. Hence when drawing exploratory prostate cancer county-level, temporally dependent data from a regression model sample, the rejection region would comprise 5% of the sampling distribution. In statistics, a sampling distribution or finite-sample distribution is the probability distribution of a given statistic based on a random sample[Draper 1998].

Further, these 5% can be allocated to one side of the sampling distribution in the prostate cancer model as in a one-tailed test, or partitioned to both sides of the distribution as in a two-tailed test, with each tail (or rejection region) containing 2.5% of the distribution. One-tailed tests are used for asymmetric distributions that have a single tail, such as the chi-squared distribution, which are common in measuring goodness-of-fit, or for one side of a distribution that has two tails, such as the normal distribution, which is common in estimating location; this corresponds to specifying a direction.[Fox 1997]. Two-tailed tests are only applicable when there are two tails, such as in the normal distribution, and correspond to considering either direction significant. The use of a one-tailed test is dependent on whether the research question or alternative hypothesis specifies a direction such as whether a group of objects is *heavier* or the performance of students on an assessment is *better*.[

A two-tailed test may still be used in a vulnerability, prostate cancer, county-level, model to determine statistical significance but it may less powerful than a one-tailed test because the rejection region for a one-tailed test is concentrated on one end of the null distribution and is twice the size (5% vs. 2.5%) of each rejection region for a two-tailed test. As a result, the null hypothesis can be rejected with a less extreme result if a one-tailed test in a prostate cancer model is used. The one-tailed test is only more powerful than a two-tailed test if the specified direction of the alternative hypothesis is correct[Draper 1998]. If it is wrong, however, then the one-tailed test has no power.

A related topic in regression analysis, focuses more on questions of statistical inference such as how much uncertainty is present in a curve that is fit to data observed with random errors[Fox 1997]The best fit in the least-squares vulnerability county-level, prostate cancer, forecast model would optimally minimizethe sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a, endemic model estimator). When the problem has substantial uncertainties in the independent variable (the *x* variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares[Draper 1998] In statistics, errors-in-variables models or measurement error models are regression models that account for measurement errors in the independent variables[Fox 1997].

In the simple case of a set of (x,y) geographically and clinically geosampled, prostate cancer, county-level, grid-stratfied, regression prognosticators, the approximation function may be a simple "trend line" in two-dimensional Cartesian coordinates. The method mayminimize the sum of absolute errors (SAE) (the sum of the absolute values of the vertical "residuals" between georeferenced,geosampled,zip code, polygon points generated by the function that correspond toland cover points in the county empirical dataset). The least absolute deviations estimate also arises as the maximum likelihood estimate if the errors have a Laplace distribution. In probability theory and statistics, the Laplace distribution (i.e., the double exponential distribution) is a continuous probability distribution[Draper 1998]which may be defined in a vulnerability forecast, county-level, prostate cancer, regression model employing two exponential distributions with an additional georeferenced, zip-code polygon, geographic geolocation parameter) spliced together back-to-back.

Errors due to regression violations in a forecast, vulnerability, time series, county-level, prostate cancer models may be minimized by utilizing linear and non-liner techniques in SAS and ArcGIS. For example, The Spatial Statistics toolbox provides effective tools for quantifying spatial patterns in an empiricial dataset of clincial, field or remote geosampleddiagnsoticgeoreferenecable, covariates associated with county-level, hyperendemicgeolocations. Using the Hot Spot Analysis tool, for example, a county-level epidemiologist of public health official may answer questions such as: Are there places at the zip code where people are persistently dying due to underdiagnosed prostate cancer?Where are the hot spots for targeting prostate cancer interventions?Where do we find a higher than expected proportion of prostate cancer in a county? The Spatial Statistics toolbox contains statistical tools for analyzing spatial distributions, patterns, processes, and relationships of georeferenceable, prostate cancer geolocations. While there may be similarities between linear (traditional) and non-linear statistics in terms of concepts and objectives, spatial statistics are unique in that they were developed specifically for use with geographic data[www.esri.com]. Unlike traditional nonspatial statistical methods, spatial models incorporate space (proximity, area, connectivity, and/or other spatial relationships) directly into their mathematics.The tools in the Spatial Statistics toolboxhence may allow summarizing the salient characteristics of a spatial county distribution of time series geosampled prostate cancer variables (e.g., quantittaingthe mean center or overarching directional trend), identify statistically significant spatial georeferenceableclusters (e.g., hot spots/cold spots) or spatial outliers,(i.e, extreme observations) assess overall patterns of clustering or dispersion, group features based on attribute similarities,(e.g., autocorrelated diagnostic covarites) identify an appropriate scale of analysis, and explore spatial relationships

In this analysis, we constructed a linear and non-linear regression modelsto determine county-level socio-demographicexplanators associated with the prevalence of prostate cancer in Hillsborough County, Florida in PROC REG. We utilized cartographic data [land use land cover(LULC)geospatial analysis] to determine vulnerability geolocations of georeferenceablepopulations to prostate cancer at the zip code level at the study site in ArcGIS. GIS and remote sensing have often been used to describe and forecast chronic infectious diseases, such as tuberculosis (Jacob et al. 2013), West Nile virus (Griffith 2005), measles (Alao et al. 2016), just to mention a few. However, GIS and remote sensing have not been employed for delineating vulnerable, county-level,georeferenecable,LULC regions to prostate cancer.

Our objectives were:1)To construct a linear regression probability model to determine covariates at 95% confidence interval.2)To map areas that were vulnerable based on the regression model output,3)To utilize a first order autocorrelation analysis to quantitate clustering propensities in the empirical, county-level, social demographic and landscape dataset ;and,4)To conduct a hotspot analysis to identify and prioritize regions of vulnerability of prostate cancer in a zip code polygon in Hillsborough County, Florida.

2. Materials and Methods

2.1. *Sample Collection*

We obtained 61,139 prostate cancer cases of males diagnosed in Florida from 2010 to 2014 from the Florida Cancer Data System (FCDS). The FCDS is Florida statewide cancer registry lodged at the University of Miami. The FCDS was established in 1981 for the collection of cancer incidence data in Florida and contains approximately 3,400,000 cancer incidence records (Florida Cancer Data System 2017). The FCDS collects information on patients' demographics, characteristics of prostate tumor, information on tobacco use, primary payer of health insurance, residence and other information on cancer patients. Census tract level information which includes median age, population, percentage of high school graduates, median household income, individuals below poverty level and race (White, Black and Asian) were extracted from the U.S. Census Bureau (U.S. Census Bureau 2017) based on the county zip codes obtained from the Capitol Impact Government Gateway (Capitol Impact 2017). The zip codes with no demographic information were excluded from the study. A total of 48 zip codes were used in our study and they served as the unit of our analysis.

2.2. *Spatial Analysis*

ArcGIS 10.5 was used to perform GIS functions. Prostate cancer prevalence rates were tabulated at the county-level by superimposing point layers of prostate cancer prevalence rate onto a map of Florida counties (Xiao et al. 2007). The total combined number of prevalence rates of each county was affixed as a feature to each county record. Thematic map of high and low clusters of prostate cancer prevalence rates by counties (Figure 1) was generated as well as prostate cancer prevalence rate in Hillsborough County, as shown in Figure 2.

2.2.1. *Land use Land cover Classification (LULC)*

Using the Florida Department of Environmental Protection (FDEP) GIS data, four LULC classifications were generated for Hillsborough County. A LULC spatial analyses is beneficial for planning and monitoring a region for several purposes, such as determining changes over time. This tool is also powerful in providing important information for a large area (Esri 2017). The LULC for Hillsborough County included: water, agriculture, commercial and residential. These classes were defined as follows:

(1) Water: Bodies of water such as gulfs, oceans, seas, lakes and reservoir.
(2) Agriculture: Geographic areas dominated by the presence of thick vegetation, trees, nurseries, brush-lands and vineyards
(3) Commercial: Areas dominated by industrial activities and enterprises.
(4) Residential: High and low density areas with maintained housing.

The georeferenced Hillsborough County zip codes were then overlaid to determine the zip codes associated with the LULC (Figure 4).

2.2.2. *Hotspot Analysis*

Hotspots of prostate cancer prevalence at the zip code level in Hillsborough County was analyzed with the spatial statistics hotspot analysis tool referred to as Getis-OrdGi* algorithm in ArcGIS 10.5..The Getis-Ord local statistic was given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{\left[n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j} \right)^2 \right]}{n-1}}} \qquad (2.1)$$

wherex*j*was the attribute value set for a geosampled, county-level prostate cancer covariate feature,w*ij* was the spatial weight between*i* and *j;n*was equal to the total number of features and

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad \text{(2.2) and} \qquad S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \qquad \text{(2.3).}$$

The hotspot analysis (Figure 6) wa employed identify statistically significant spatial clusters of high and low prostate cancer prevalence rates in the study area. P-values and Z-scores were used to indicate the aggregate of spatial clustering in the geosampled datasets. A z-score (aka, a standard score) was employed to indicate how many standard deviations an element was from the mean in the prostate cancer model. A z-score can be calculated from the following formula. z = (X - μ) / σwhere z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation[Fox 1997].

2.3. *Statistical Analyses*

The prevalence rate of prostate cancer was calculated by dividing the existing number of prostate cancer cases in each county by the total number of population in each county per 10,000 population Our multiple linear regression model involved the use of prostate cancer prevalence rate in Hillsborough County across the different zip codes available in the county as our dependent variable. To introduce heterogeneity in our dependent variable (Y), it was necessary to calculate the proportion of prostate cancer prevalence across each zip code. This was obtained by dividing the population at each zip code by the total population in Hillsborough County. The result was then multiplied by the Hillsborough County's prostate cancer prevalence rate of 30 per 10,000 population.

2.3.1. *Pearson's Correlation Coefficient*

Pearson's r correlation coefficient was employed to determine whether there exists any linear relationship between the variables in the county-level prostate cancer model. The PROC CORR procedure in SAS was utilized to achieve this aim. A sample Pearson's correlation co-efficient (r) falls between the values of +1 and -1, where a positive value indicates a positive correlation while a negative value indicates a negative correlation[ ].

2.3.2.*Multiple Linear Regression*

A linear regression with statistical significancewas determined by a p-value less than 0.05 to ascertain whether the proportions of prostate cancer in each grid-stratified, georferenced, zip code could be predicted by certain independent variables. The linear regression model assumeda random sample between Yi, (proportion of prostate cancer prevalence rate), the regress and regressorsXi1, ...Xip (age, education, income, race, poverty, landcover). A disturbance term εi, helped capture the influence of all covariates geosampled on Yi other thanXi1, ...Xip. The random error term, ε, in a regression analysis istypically assumed to be normally distributed with meanzero and variance S^2(Jacob et al. 2010). Theregression analyses were performed using PROC REG. P-values less than 0.05 was considered statistically significant. The multiple linear regression model was described as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i,\ i = 1, \ldots, n. \quad (2.4)$$

It was important to distinguish the model in terms ofrandom variables and the observed values of the random variables. Thus, we determined;$p + 1$ parameters $\beta_0, \ldots, \beta_p.$ (2.5). In order to estimate the sampled parameters, it was useful to use the matrix notation:) $Y = X\beta + \varepsilon,$ (2.6) where Y was a column that included proportion of prostate cancer prevalence rates of, $Y_1, \ldots, Y_n,$ whichincluded the unobserved stochastic components $\varepsilon_1, \ldots, \varepsilon_n$and the matrix X as in Jacob et al. (2010).

The weighted matrix was the observed sampelde county-level prostate cancer, grid-stratified, explanatory, parameter values of the regressorsexpressed as

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}. \quad (2.7)$$

In the forecast, vulnerability, county-level, prostate cancer, epidemiological model, X included a column that did notvary across the geosampledregressors which we employed to represent the intercept term$\beta_0.$ The presence of multicollinearity was examined and the predictor variables with a variance inflation factor of >10 were excluded from the model. Diagnostics for the presence of any assumption violations of a multiple linear regression model were carried out. The PLOT statement with residuals plotted against the predictors checked for non-linearity and non-constancy of error variance whilst a normal quantile-quantile (Q-Q plot) was plotted to check for the normality assumption. In chronic infectious statistics, a Q–Q plot may be employed as a probability plot, for comparing two probability distributions by plotting their quantiles against each other (Jacob et al. 2013, Griffith 2005). In statistics and the theory of probability, quantiles are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities, or dividing the observations in a sample in the same way.(Fox 1997).

Our assumption was that a county-level,geosampled, zip code, grid- stratified, prostate cancer, explanative, georeferenecable, regressor point (x, y) on the plot would correspond to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Hence, the regression line in model we assumed would reveal a parametric curve with the parameter which would be the number of the interval for the quantile. If the two prostate cancer probability distributions being compared were similar we could then assumethat, the points in the Q–Q plot would approximately lie on the line $y = x$. If the distributions were linearly related, the points in the Q–Q plot we assumed would then approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots in ArcGIS can be used as a graphical means of estimating parameters in a location-scale family of distributions(Anselin 1995).

2.3.3. *Durbin-Watson First-Order Autocorrelation Test*

The Durbin-Watson test was employed to evaluate the presence of a first-order autocorrelation in our model. We applied PROC AUTOREG for Durbin-Watson statistics to detect inherent residual error coefficients in the regression analysis for the county-level geosampled, prostate cancer data. The DWPROB option in PROC AUTOREG allowed for the ascertainment of significance level within the selected predictor variables. The DW statistics tests the null hypothesis:$H_o: \varphi_1=0$ against $H_1: \varphi_1 >0$. Likewise, the residuals associated with the observation at time t wasprovided by e_t. Then, the DurbinWatson test was given by the equation below

$$d = \frac{\sum_{t=2}^{T}(e_t - e_t)^2}{\sum_{t=1}^{T} e_t^2}$$

(2.8) where T was the number of geo-referenced prostate cancer prevalence rate in each grid-stratifiable, georeferenceable, zip code as in Alao et al. (2016).

The value of d in the DW test for the prostate cancer prevalence modelwas approximately 2 (1-r), where r was the sample autocorrelation ofthe residual. Similarly, the distribution of d falls within the values 0 and4 and is symmetrical around 2 while d=2 shows no serial correlationin errors (Chen 2016). When a value of d<1, exists in a first order autocorrelation model this would indicate that the errors within the modelare serially correlated to one another, and there is possible violation of the assumption of error independence in a given regression model (Griffth 2003). Durbin Watsond values <2 indicates a positive serial correlation and d>2 signifiessuccessive error terms are much dissimilar from oneanother i.e. negatively correlated.

3. Results

Figure 1 shows the counties with the highest and lowest prostate cancer prevalence in Florida. Sumter County has the highest prostate cancer prevalence rate of 93 per 10,000 population compared to other counties in Florida. On the other hand, Lafayette County and Glade County had an equal prostate cancer prevalence rates of 9 per 10,000 population compared to the rest of the counties in Florida.

Figure 2 depicts the prostate cancer prevalence rate of our study site. Hillsborough County has a prostate cancer prevalence rate of 30 per 10,000 population. In addition, there were other counties in Florida, which had a prevalence rate of 30 per 10,000 population, similar to Hillsborough County. These counties are Duval County, Osceola County, Desoto County, Broward County, Miami-Dade County and Monroe County.

Figure 4 illustrates the result of our LULC classification. Eligible zip codes in our study site were mapped to identify their individual georeferenceable, geolocations in the LULC polygons. Based on the LULC, majority of the zip codes were located in the residential areas, followed by commercial area, agricultural area and water.

Figure 5 shows the results of the Pearson r correlation algorithm. This procedure is useful in determining both the magnitude and the direction of the association amongst the variables in our model. Hence, the output from the prostate cancer model indicates that income (r = -0.51113, p-value= 0.0002) is inversely associated with prostate cancer prevalence. In other words, as income decreases, the prevalence of prostate cancer becomes higher. Variables associated to povertyshowed a positive association with prostate cancer prevalence although not statistically significant (r = 0.17434, p = 0.2360). This means that as poverty increases at the epidemiological study site the prostate cancer prevalence also increases. Additionally, the output shows the relationship between income and race. White race (r = 0.43108, p = 0.0022 has a statistically significant positive association with income while Black race (r = -0.13534, p = 0.3591) and Hispanic race (r = -0.24287, p = 0.0962) are negatively associated with income although the relationship is not statistically significant. An inverse association was found between poverty and White race (r = -0.80115, p < 0.0001). On the other hand, Black race (r = 0.77179, p < 0.0001) and Hispanic race (r = 0.45202, p = 0.0013) had a significantly positive association with poverty.

Figure 6 depicts the output from a multivariate, linear regression model. The result indicates that income (p-value =0.0033) is significantly associated with prostate cancer prevalence. Poverty (p-value = 0.0633) had an association with prostate cancer prevalence, although it was not statistically significant at an alpha level of 0.05. Other independent variables in our model were not significantly associated with prostate cancer prevalence. Specifically, the LULC variable showed no association with the prostate cancer risk.

Figure 7 shows a map that illustrates the result of our multiple linear regression. Based on our findings in the regression model, income was significantly associated with the prevalence of prostate cancer at the zip code level. Thus, zip codes with lower income are more likely to have a greater risk of prostate cancer. Zip codes 33603, 33604, 33605, 33612, 33613, 33614, 33619, 33563, 33592 have lowest median income within the range of $25,341-$38,132. Likewise, they have an increased likelihood of having a greater prostate cancer prevalence rate compared to other zip codes in Hillsborough County. On the contrary, zip codes 33547, 33548, 33556, 33558, 33569, 33602, 33606, 33624, 33626, 33629, 33647, 33572 are more likely to have a lower prostate cancer prevalence since they have the highest median income within the range of $61,512 - $100,817.

Figure 8 shows the output from our first-order Durbin-Watson test of autocorrelation. The PROC AUTOREG procedure generated a positive auto correlation (1.5577), in other words, a positive spatial clustering among the predictor variables and the prostate cancer prevalence rate outcome variable among Hillsborough County zip codes in our model.

Figure 9 illustrates the result of our hotspot analysis. Our analysis targeted the highest region (hotspot) of prostate cancer vulnerability. Zip code 33610 is the hotspot, surrounded by zip codes 33510, 33603, 33612, 33613, 33614, 33617, 33637, which were located in the hotspot region as shown in the figure. On the other hand, we found cold spots in zip codes 33572, 33621 and 33616 while the remaining zip codes were neither hotspots nor cold spots of prostate cancer vulnerability.

4. Discussion

Socio-demographic factors have a role to play in disease occurrence. Our analysis found that income had a negative association with the prevalence of prostate cancer, such that lower income level is more likely to increase the prevalence of prostate cancer. Income has been delineated as an important independent variable in prostate cancer research (Howard et al. 2000, Lund Nilsen et al. 2000, Du et al. 2006, Talcott et al. 2007,Kilpeläinen et al. 2016). Studies by (Lund Nilsen et al. 2000, Kilpeläinen et al. 2016) showed that high income increases the incidence rate of prostate cancer. On the other hand, lower income increases the mortality rate from prostate cancer as shown in studies by (Howard et al. 2000, Du et al. 2006). The effect of income on prostate cancer prevalence from our finding indicates that grid-stratified, zip codes with lower income are more likely to lack adequate funding and access to prostate cancer treatment. Thus, the number of existing cases of prostate cancer would tend to be high. As a ripple effect, this would increase the death rate of low income prostate cancer patients.

Environmental and genetic factors contribute to the development of prostate cancer. It is important to understand the interactions between and within factors that encourage prostate cancer risk (Klassen & Platz 2006). Hence, we included a landscape variable in our analysis to demonstrate the possible impact of geographical location on prostate cancer prevalence although we found no association between them in our analysis. However, Sharp et al. (2014) revealed that the relative risk of prostate cancer was more prominent in rural areas compared to urban areas.

There were some limitations present in our study. We evaluated sociodemographic factors at the census/ county level and not at the individual level. Hence, the issue of fallacy comes into play as we could not make inferences on the individuals in our study area based on our findings. Additionally, in our analysis, we estimated the proportionality assignment of prostate cancer cases in each zip code. In other words, we used the prevalence rate of prostate cancer in Hillsborough County to estimate the prostate cancer prevalence rate by each zip code since we did not have information on the prostate cancer prevalence rate by zip codes.

The Durbin-Watson test, can decide if autocorrelation correction is needed in a county level, empirical geosampled, dataset of stratified, zip code polygonized, prostate cancer data points. However, generalized Durbin-Watson tests should not be used to decide on the autoregressive order in these datasets. The higher-order tests will assume the absence of lower-order autocorrelation in the dataset. Also, Durbin-Watson tests may not be valid when a lagged dependent, geosampled, stratified, prostate cancer, specified, explanatory variable is employed in a county-level, grid-stratified, zip code, SAS, regression model. In this case, the Durbin *h* test or Durbin *t* test may be used to test for first-order autocorrelation.

For the Durbin *h* test, a prostate cancer researcher could specify the name of the lagged, county-level, explicative dependent variable (e.g., case distribution,) in the LAGDEP= option. For the Durbin *t* test, specify the LAGDEP option without giving the name of the lagged dependent variable(SAS Institute Inc 2017). For example, the following statements can add a county-level, grid- stratifiable, geosampled, zip code prostate cancer related time series variable YLAG to the dataset and henceforth regress Y on YLAG instead of TIME:

```
data b;
set a;
ylag = lag1( y );
run;

procautoreg data=b;
county zip code prostate cancer  model y = ylag / lagdep=ylag;
run;
```

If the ordinary Durbin-Watson test indicates no first-order autocorrelation in a dataset of county, zip code, grid-stratified, regression model renderings, a prostate cancer researcher may use the second-order test to check for second-order autocorrelation. The capabilities for visualizing, rapid data retrieval, and manipulation in ArcGIS have created the need for applying new techniques of exploratory zip code, grid-stratified, prostate cancer, time series, diagnostic analysis that focus on the "spatial" aspects of the data. Outlining a new general class of local indicators of spatial association (LISA) may allow for the decomposition of global indicators, such as Moran's *I*, into the contribution of each geosampledprostate cancer, explanative georeferenecable, observation. In statistics, Moran's *I* is a measure of spatial autocorrelation which may be characterized by a correlation in a signal among nearby locations in space(Moran 1950).

The LISA statistics may serve two purposes for optimizing prostate cancer empirical geosampled, county-level, optimizabledatasets. On one hand, they may be interpreted as indicators of local pockets of nonstationarity, or hot spots, similar to the Gi and G*i statistics of Getis and Ord (1992) at a county, prostate cancer, grid-stratified, zip code specification level. On the other hand, they may be used to assess the influence of individual geolocations in a georeferencible zip code polygon based on the magnitude of the global statistic for identification of outliers, as in Anselin's Moran scatterplot(Anselin 1993). In statistics, an outlier is an observation point that is distant from other observations(Grubbs 1969). An initial evaluation of the properties of a LISA statistic may be carried out for the local Moran, which may be applicable for robustly quantitating spatial patterns in a grid-stratifiable, georeferenceable, prostate cancer, zip code polygon employing number of Monte Carlo simulations. In chronic infectious, predictive, probabilistic models, Monte Carlo simulation performs risk analysis by building models of possible results by substituting a range of values—a probability distribution—for any factor that has inherent uncertainty. It then calculates results repeatedly, each time using a different set of random values from the probability functions (see Jacob et al. 2013, Griffith 2005).

If a test for negative autocorrelation is desired to quantitate non-linear, stratified, county-level, prostrate, cancer, zip code, gridded, ArcGIS polygon, a prostate cancer researcher could use the statistic 4−d. Then the decision rules for H0: ρ = 0 versus H1: ρ < 0 may not be the same as those used in testing for positive autocorrelation. It may be also possible to conduct a two-side test (H0 : ρ = 0 versus H1 : ρ = 0) in an AUTOREG procedure by employing both one-side tests simultaneously in a county-level, zip code, prostate cancer, georeferenced, polygon model. If this is conducted parsimoniously, the two-side procedure could quantitate Type I error 2α, whenst α is the Type I error is used for each one-side test. Accurate CI estimators for ascertaining stratifiable, zip code level, county, diagnostic, prostate cancer regression gridded proportions, rates, and their differences may be also described in MATLAB programs. The programs may search for CIs and utilize an integration of the Bayesian posterior with diffuse priors which may measure CI at 99% for optimally quantitating time series, prostate cancer, zip code grid-stratified, georeferenceable, county-level regressors.

In generalizable, Bayesian hierarchical, prognosticative, county-level, prostate cancerprobability models, a prior probability distribution (i.e., the prior**)**, of an uncertain vulnerability related quantity would be the optimal distribution that would express beliefs about this quantity before some evidence is taken into account (see Jacob et al. 2013, Griffith 2005). For example, the prior in a county-level, prostate cancer, regression, research, time series, diagnostic model could be the probability distribution representing the relative proportions of vulnerability, grid-stratifiable county-geolocationfor determining prevalence in a specific zip code polygon in ArcGIS. The unknown quantity may be a parameter of the model or a latent variable rather than an observable variable. Bayes' theorem calculates the renormalized pointwise product of the prior and the likelihood function, to produce the posterior probability distribution, which is the conditional distribution of the uncertain quantity given the data (Gelman 2005). The CI estimators may find one or two-sided intervals in the model. For two-sided intervals, either minimal-length, balanced-tail probabilities, or balanced-width may be henceforth selected for optimization of biased, geosampled, county-level, prostate cancer, regression, zip code polygon, model predictors.

5. Conclusion

Our findings reveal that a socio-demographic covariate, in this case, income is an optimal explanatory predictor of prostate cancer prevalence at the zip code level within a county. We further identified areas of high prostate cancer vulnerability at the county-level.This information can help target prostate cancer intervention strategies in communities and help allocate resources to areas highly in need of them. Further studies on the linear and spatial association between prostate cancer prevalence rate and potential socidemiographic and LULC predictors may optimally forecast time series georefereneceable, grid-stratifiable vulnerable, hyperendemic, county-level, zip code regions.

References

Abe, T., Martin, I.B. & Roche, L.M., 2006. Clusters of census tracts with high proportions of men with distant-stage prostate cancer incidence in New Jersey, 1995 to 1999. *American Journal of Preventive Medicine*, 30(2 SUPPL.), pp.S60-6.

Alao, S., Mati, K. & Jacob, B., 2016. Journal of Remote Sensing & GIS Differentiating Non-Homoscedasticity and Geospatially Extreme Outliers for Urban and Rural Landscape Dataset Using Pearson â€^{TM} s Product Moment Correlation Coefficients for Quantitating Clustering Tendencies in Non- Vaccinate. , 5(4).

American Cancer Society, 2017. What Is Prostate Cancer? [online] Available from: https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html [Accessed June 10, 2017].

Anselin, L., 1993. Discrete space autoregressive models. In *Environmental Modeling with GIS*. pp. 454–469.

Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), pp.93–115.

Capitol Impact, 2017. CI Gateway Zip Code List [online]. Available from: http://www.ciclt.net/sn/clt/capitolimpact/gw_ziplist.aspx?ClientCode=capitolimpact&State=fl&StName=Florida&StFIPS=&FIPS=12115 [Accessed June 18, 2017].

Centers for Disease Control and Prevention, 2016. Cancer Statistics - Men [online]. Available from: https://www.cdc.gov/cancer/dcpc/data/men.htm [Accessed June 11, 2017].

Centers for Disease Control and Prevention, 2012. Principles of Epidemiology | Lesson 3 - Section 2 [online]. Available from: https://www.cdc.gov/ophss/csels/dsepd/ss1978/lesson3/section2.html [Accessed June 18, 2017].

Chen, Y., 2016. Spatial autocorrelation approaches to testing residuals from least squares regression. *PLoS ONE*, 11(1), pp.1–19.

DeChello, L.M. & Sheehan, T.J., 2007. Spatial analysis of colorectal cancer incidence and proportion of late-stage in Massachusetts residents: 1995–1998. *International Journal of Health Geographics*[online], 6(1), p.20. Available from: http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-6-20.

Draper, N.R. & Smith, H., 1998. Applied Regression Analysis. *Technometrics*, 47(3), p.706.

Du, X.L. et al., 2006. Racial disparity and socioeconomic status in association with survival in older men with local/regional stage prostate carcinoma: Findings from a community-based cohort. *Cancer*, 106(6), pp.1276–1285.

Esri, 2017. Classifying Imagery to Create a Thematic Raster [online]. Available from: http://trainingbeta.esri.com/courses/57d0800584b087dd46817ceb-11612//Content/player.html?endpoint=http%3A%2F%2Ftrainingbeta.esri.com%2FEngine%2FTCAPI%2F&auth=Basic OjRmYzFjOWQyLTU1MjctNGY1ZS1hMzNiLTUxMmY2NjYyNmUxOA%3D%3D&actor=%7B%22objectType%22%3A%22A [Accessed June 28, 2017].

Florida Cancer Data System, 2017. Home-Data requests [online]. Available from: https://fcds.med.miami.edu/inc/welcome.shtml [Accessed June 18, 2017].

Fox, J., 1997. *Applied Regression Analysis, Linear Models, and Related Methods,*pp. 597

Gelman, A., 2005. Analysis of variance — why it is more important than ever [online]. *The Annals of Statistics*, 33(1), pp.1–53. Available from: http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1112967698/.

Getis, A. & Ord, J.K., 1992. The Analysis of Spatial Associatiion by Use of Distance Statistics. *Geographical Analysis*, 24(3), pp.189–206.

Goovaerts, P. et al., 2016. Geographically-Weighted Regression Analysis of Percentage of late-stage prostate cancer diagnosis in Florida. *Applied Geography*, 62, pp.191–200.

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*[online], 4, p.18. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1215506&tool=pmcentrez&rendertype=abstract.

Grubbs, F.E., 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1), pp.1–21.

Grubesic, T.H. & Matisziw, T.C., 2006. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International journal of health geographics*[online], 5, p.58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17166283%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1762013.

Howard, G. et al., 2000. Race, socioeconomic status, and cause-specific mortality. *Annals of epidemiology*, 10(4), pp.214–23.

Jacob, B. et al., 2014. Pseudo R2 Probablity 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 Re. *American Journal of Applied Mathematics and Statistics*, 2(5), pp.252–301. Available at: http://pubs.sciepub.com/ajams/2/5/1/index.html.

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, pp.122–143.

Jacob, B.G. et al., 2010. Developing GIS-based eastern equine encephalitis vector-host models in Tuskegee, Alabama. *International journal of health geographics*, 9, p.12.

Kilpeläinen, T.P. et al., 2016. Prostate Cancer and Socioeconomic Status in the Finnish Randomized Study of Screening for Prostate Cancer. *American Journal of Epidemiology*[online], 184(10), pp.720–731. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kww084.

Klassen, A.C. & Platz, E.A., 2006. What can geography tell us about prostate cancer? *American Journal of Preventive Medicine*, 30(2 SUPPL.).

Lund Nilsen, T.I., Johnsen, R. & Vatten, L.J., 2000. Socio-economic and lifestyle factors associated with the risk of prostate cancer. *British journal of cancer*[online], 82(7), pp.1358–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10755415%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2374496.

Meliker, J.R. et al., 2009. Spatial cluster analysis of early stage breast cancer: a method for public health practice using cancer registry data. *Cancer causes & control : CCC*[online], 20(7), pp.1061–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19219634.

Moran, P.A.P., 1950. Notes on Continuous Stochastic Phenomena. *Biometrika*[online], 37(1), pp.17–23. Available from: http://www.jstor.org/stable/2332142.

National Cancer Institute, 2017. Cancer of the Prostate - Cancer Stat Facts [online]. Available from: https://seer.cancer.gov/statfacts/html/prost.html [Accessed June 11, 2017].

SAS Institute Inc, 2017. SAS Product Documentation [online]. Available from: http://support.sas.com/documentation/ [Accessed August 1, 2017].

Sharp, L. et al., 2014. Risk of several cancers is higher in urban areas after adjusting for socioeconomic status. Results from a two-country population-based study of 18 common cancers. *Journal of Urban Health*, 91(3), pp.510–525.

Talcott, J.A. et al., 2007. Hidden barriers between knowledge and behavior: The North Carolina prostate cancer screening and treatment experience. *Cancer*, 109(8), pp.1599–1606.

United States Census Bureau, 2017. American FactFinder - Community Facts [online]. Available from: https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml [Accessed June 18, 2017].

Wagner, S.E. et al., 2013. Prostate cancer incidence and tumor severity in Georgia: Descriptive epidemiology, racial disparity, and geographic trends. *Cancer Causes and Control*, 24(1), pp.153–166.

Xiao, H. et al., 2007. Analysis of prostate cancer incidence using geographic information system and multilevel modeling. *Journal of the National Medical Association*[online], 99(3), pp.218–25. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2569639&tool=pmcentrez&rendertype=abstract.

Xiao, H., Tan, F. & Goovaerts, P., 2011. Racial and Geographic Disparities in Late-Stage Prostate Cancer Diagnosis in Florida. *Journal of Health Care for the Poor and Underserved*[online], 22(4A), pp.187–199. Available from: http://muse.jhu.edu/content/crossref/journals/journal_of_health_care_for_the_poor_and_underserved/v022/22.4A.xiao.html.

Appendix 1.

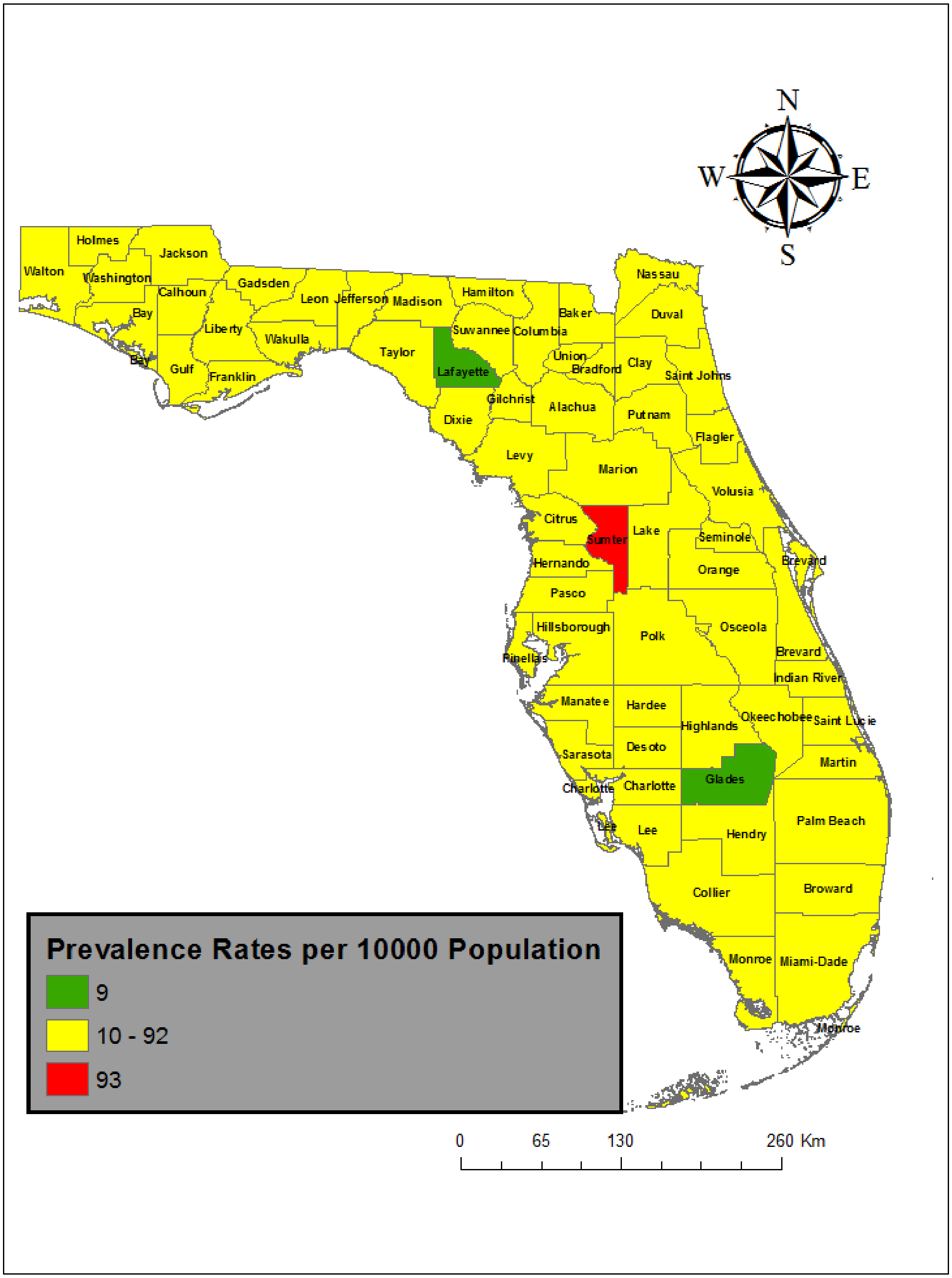


Figure 1. High and low clusters of prostate cancer prevalence rates in Florida.

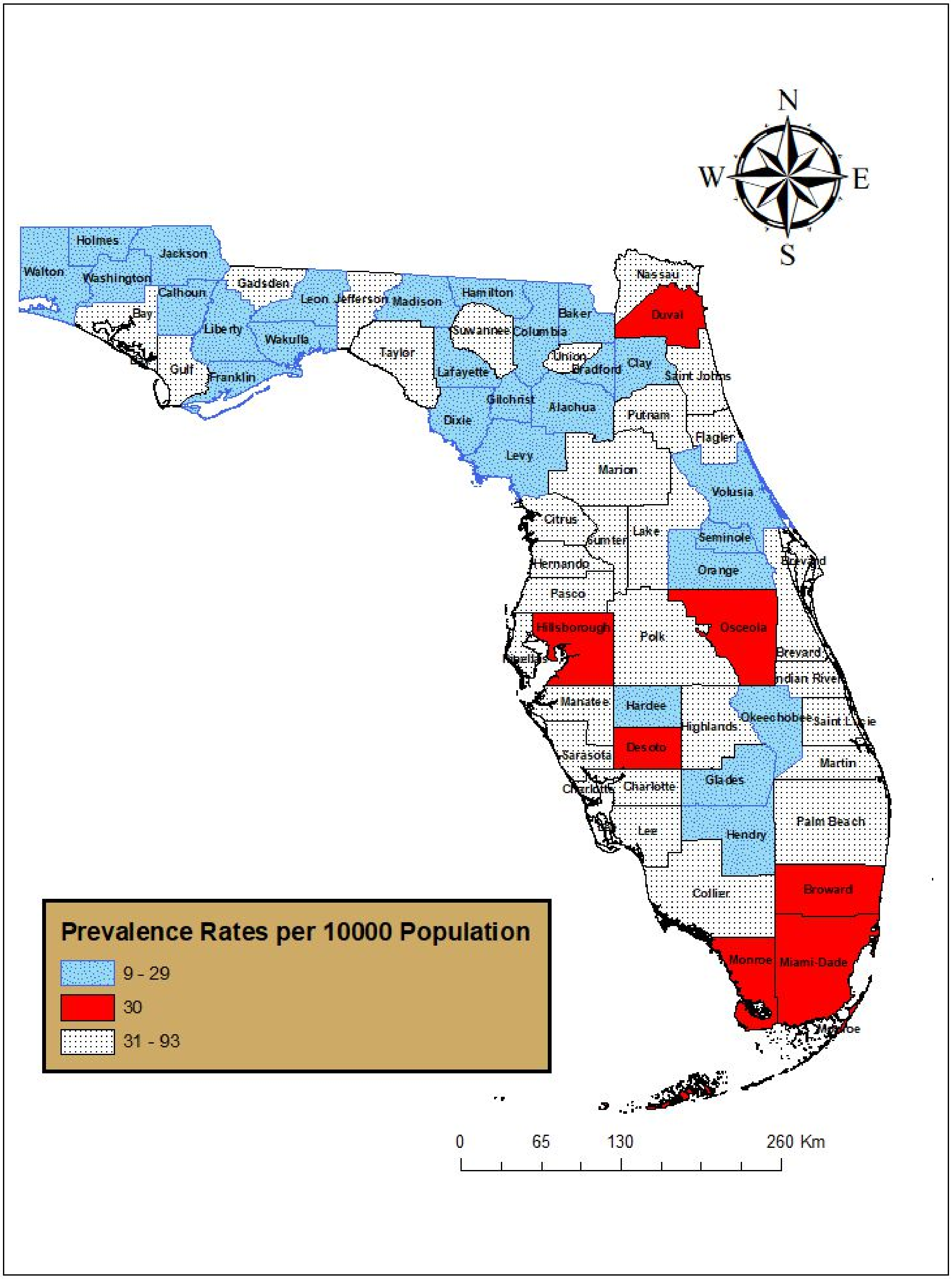


Figure 2. Counties in Florida and their prostate cancer prevalence rates, highlighting those with 30 per 10000 population

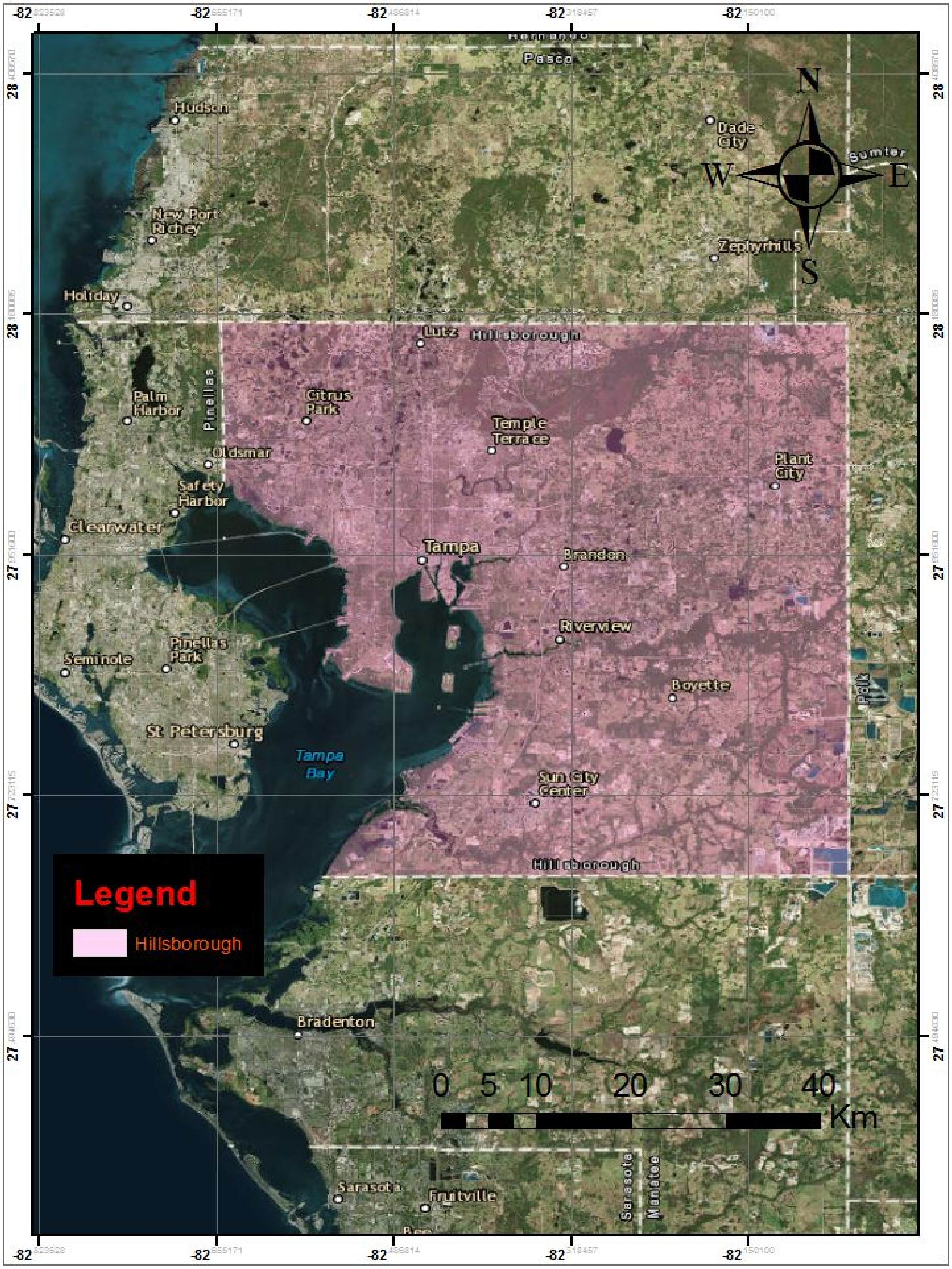


Figure 3. Cities in Hillsborough County, Fl

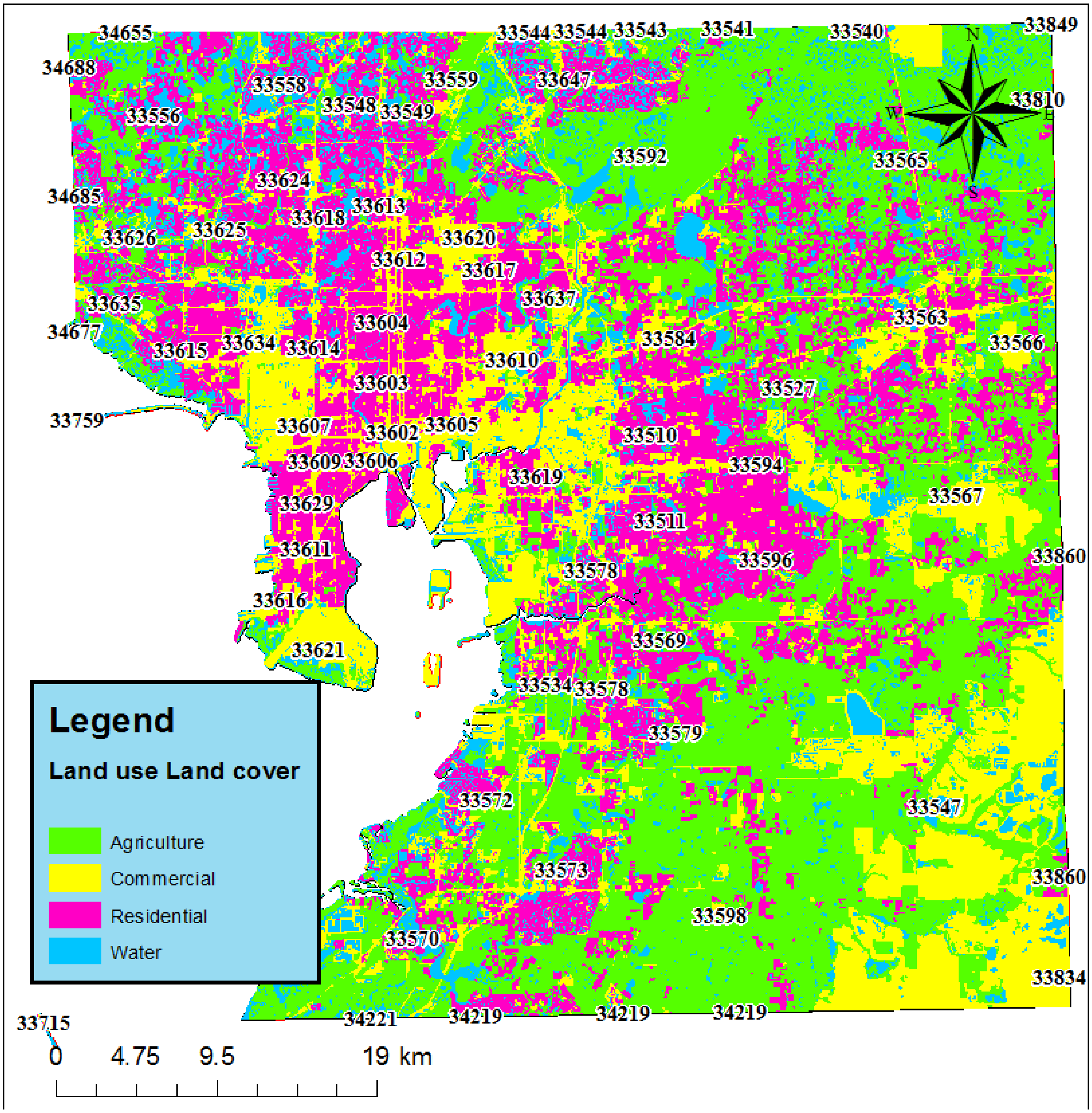
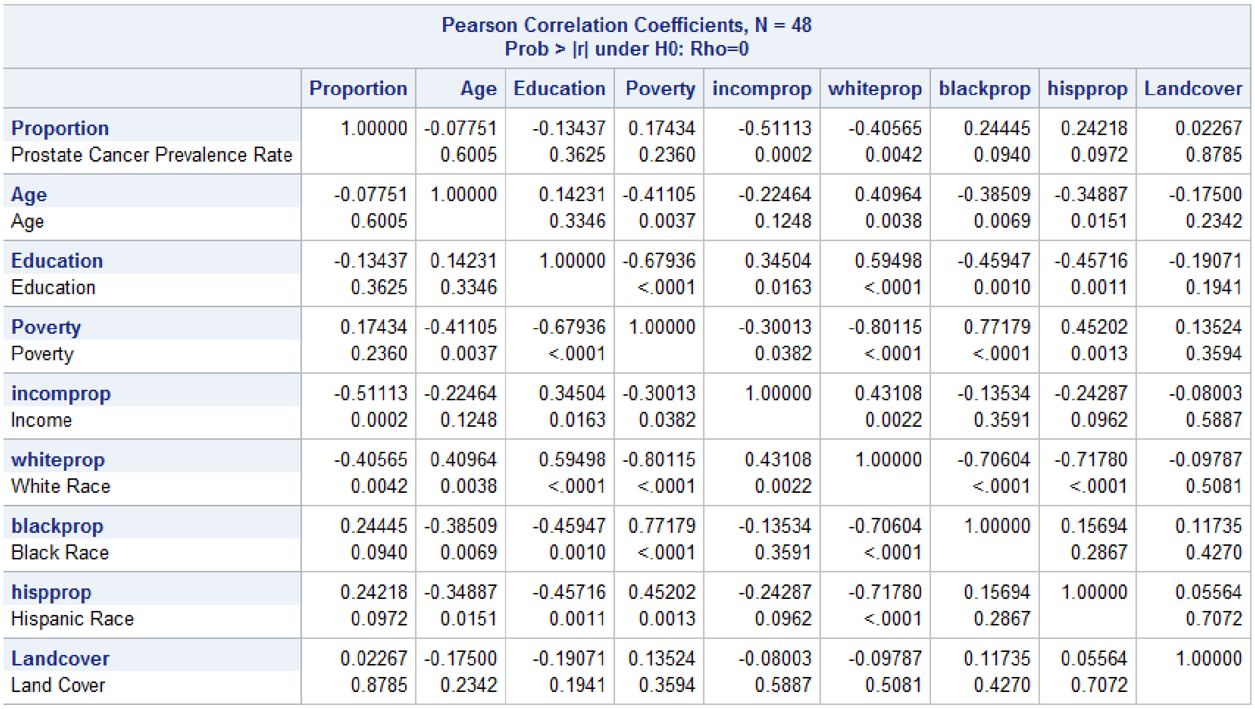


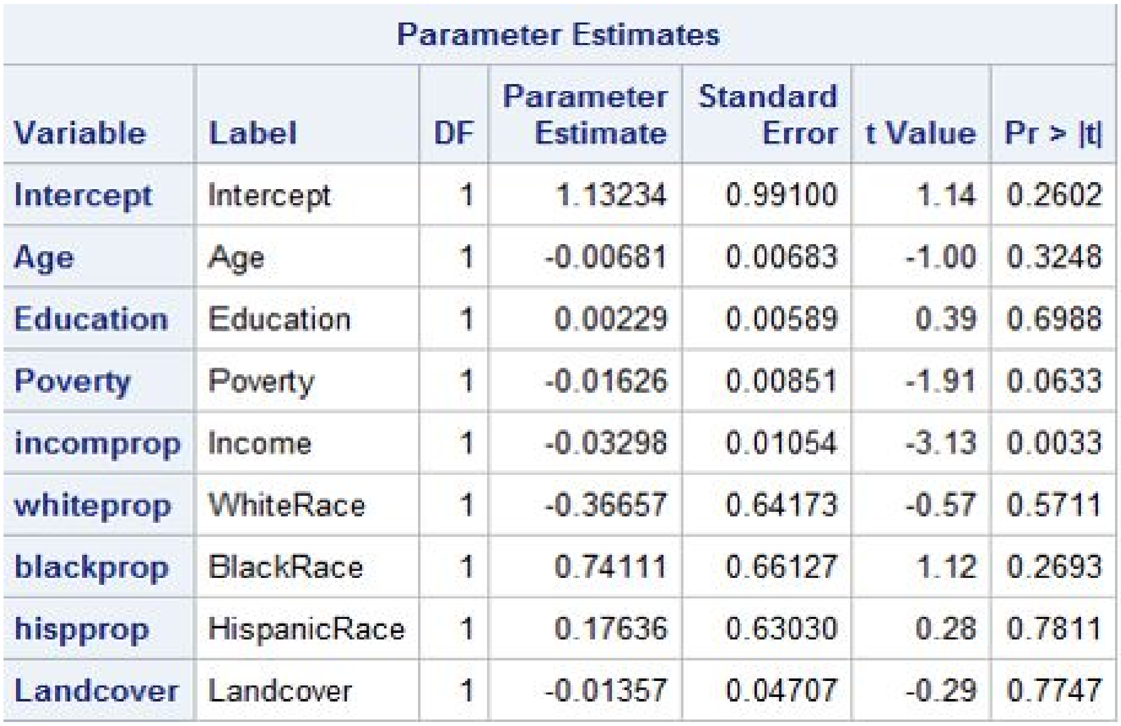
Figure 4. Land use land cover classification of zip codes in Hillsborough County, Fl.



Pearson Correlation Coefficients, N = 48
Prob > |r| under H0: Rho=0

	Proportion	Age	Education	Poverty	incomprop	whiteprop	blackprop	hispprop	Landcover
Proportion Prostate Cancer Prevalence Rate	1.00000	-0.07751 0.6005	-0.13437 0.3625	0.17434 0.2360	-0.51113 0.0002	-0.40565 0.0042	0.24445 0.0940	0.24218 0.0972	0.02267 0.8785
Age Age	-0.07751 0.6005	1.00000	0.14231 0.3346	-0.41105 0.0037	-0.22464 0.1248	0.40964 0.0038	-0.38509 0.0069	-0.34887 0.0151	-0.17500 0.2342
Education Education	-0.13437 0.3625	0.14231 0.3346	1.00000	-0.67936 <.0001	0.34504 0.0163	0.59498 <.0001	-0.45947 0.0010	-0.45716 0.0011	-0.19071 0.1941
Poverty Poverty	0.17434 0.2360	-0.41105 0.0037	-0.67936 <.0001	1.00000	-0.30013 0.0382	-0.80115 <.0001	0.77179 <.0001	0.45202 0.0013	0.13524 0.3594
incomprop Income	-0.51113 0.0002	-0.22464 0.1248	0.34504 0.0163	-0.30013 0.0382	1.00000	0.43108 0.0022	-0.13534 0.3591	-0.24287 0.0962	-0.08003 0.5887
whiteprop White Race	-0.40565 0.0042	0.40964 0.0038	0.59498 <.0001	-0.80115 <.0001	0.43108 0.0022	1.00000	-0.70604 <.0001	-0.71780 <.0001	-0.09787 0.5081
blackprop Black Race	0.24445 0.0940	-0.38509 0.0069	-0.45947 0.0010	0.77179 <.0001	-0.13534 0.3591	-0.70604 <.0001	1.00000	0.15694 0.2867	0.11735 0.4270
hispprop Hispanic Race	0.24218 0.0972	-0.34887 0.0151	-0.45716 0.0011	0.45202 0.0013	-0.24287 0.0962	-0.71780 <.0001	0.15694 0.2867	1.00000	0.05564 0.7072
Landcover Land Cover	0.02267 0.8785	-0.17500 0.2342	-0.19071 0.1941	0.13524 0.3594	-0.08003 0.5887	-0.09787 0.5081	0.11735 0.4270	0.05564 0.7072	1.00000

Figure 5. Pearson correlation analysis amongst prostate cancer prevalence rate, age, education, poverty, income, race and land cover.



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.13234	0.99100	1.14	0.2602
Age	Age	1	-0.00681	0.00683	-1.00	0.3248
Education	Education	1	0.00229	0.00589	0.39	0.6988
Poverty	Poverty	1	-0.01626	0.00851	-1.91	0.0633
incomprop	Income	1	-0.03298	0.01054	-3.13	0.0033
whiteprop	WhiteRace	1	-0.36657	0.64173	-0.57	0.5711
blackprop	BlackRace	1	0.74111	0.66127	1.12	0.2693
hispprop	HispanicRace	1	0.17636	0.63030	0.28	0.7811
Landcover	Landcover	1	-0.01357	0.04707	-0.29	0.7747

Figure 6. Multiple linear regression of prostate cancer prevalence rate with age, education, income, poverty, race, land cover as regressors.

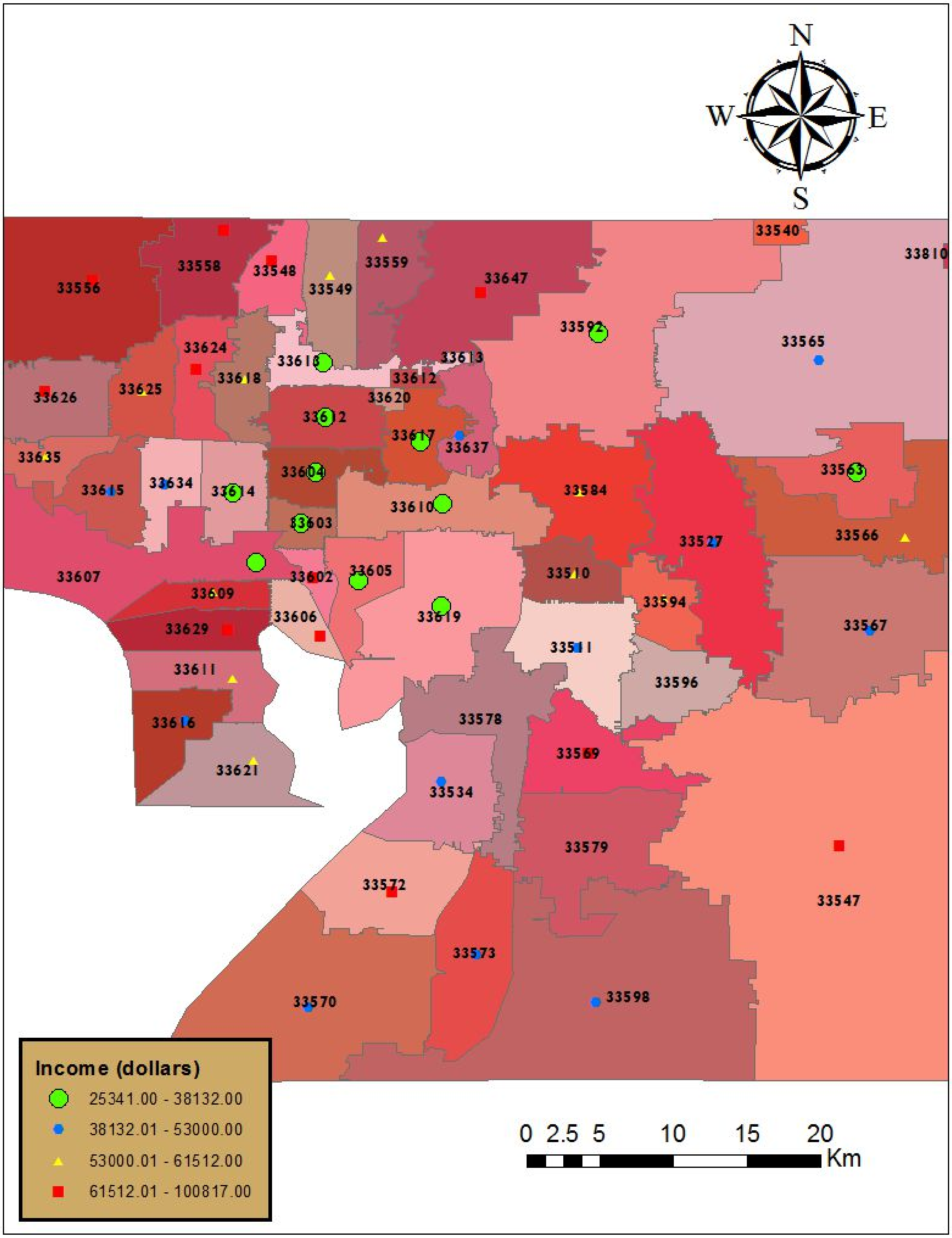
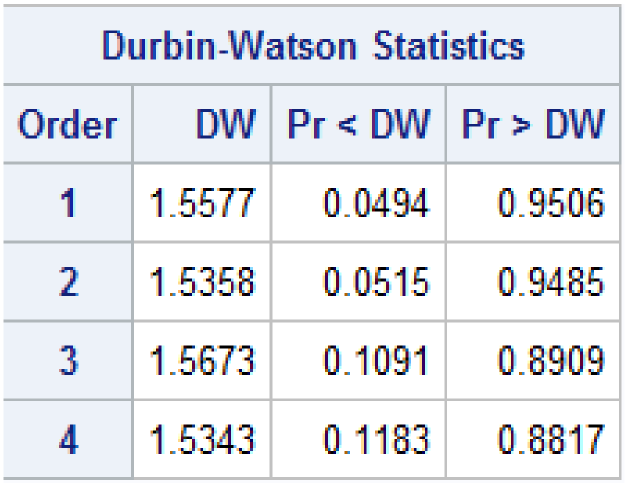


Figure 7. Income distribution of Hillsborough County by zip code.



Durbin-Watson Statistics

Order	DW	Pr < DW	Pr > DW
1	1.5577	0.0494	0.9506
2	1.5358	0.0515	0.9485
3	1.5673	0.1091	0.8909
4	1.5343	0.1183	0.8817

Figure 8. Durbin-Watson test for spatial clustering.

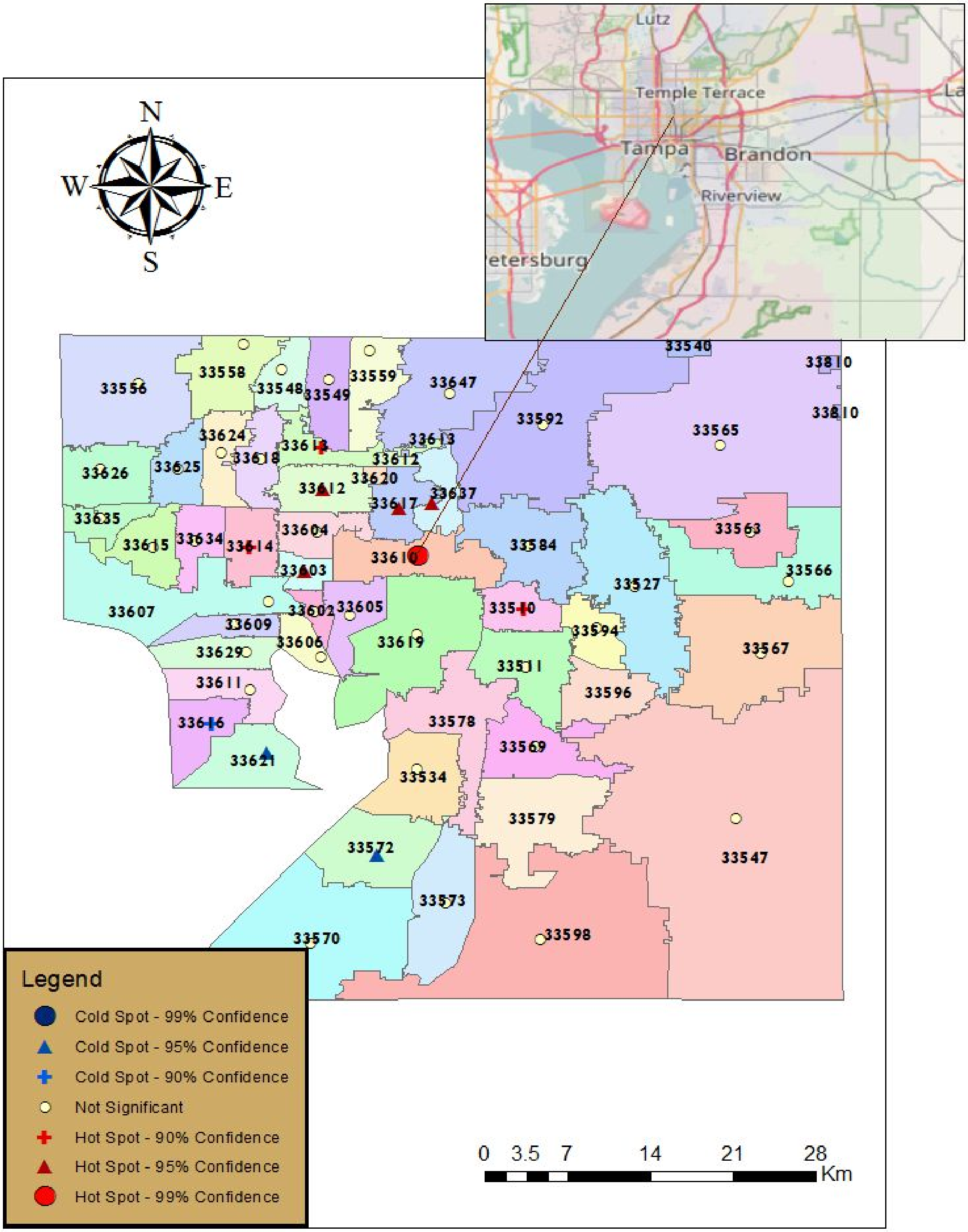


Figure 9. Hotspot Analysis of prostate cancer prevalence rate in Hillsborough County at zip code level