International Journal of Geographic Information System
Vol. 4, No. 4, November 2017, pp. 1-24
Available Online at http://acascipub.com/Journals.php

Research article

Validating Causal Relationships of Chronic Hepatitis B Infections stratified by frequentism and urbanicity for Identifying Potential Endemic County Regions in Florida

S.Li,[a] T.Panaou [a], K. Waterson [a], and B.G Jacob [a]*

[a]*Department of Global Health, University of South Florida, Tampa, U.S.A.;*

*Corresponding Author Email: bjacob1@health.usf.edu

Abstract

The Hepatitis B virus (HBV) is the unknown silent killer that has infected over 2 billion people in the world, of which, over 300 million are chronic carriers (Franco et al. 2012). Highly infectious, HBV is transmitted through transfer of bodily fluids (including saliva, blood, and mucosal secretions) and can be spread between sexual partners and family members (primarily through vertical transmission from mother to child). In the state of Florida, the rate of acute Hepatitis B infections increased by 91% between 2011 and 2015, with many infected people unaware of their infection status (Florida State Health Profile, 2017). Chronic Hepatitis B infections can lead to cirrhosis of the liver, fibrosis, and primary liver cancer (hepatocellular carcinoma), as well increasing the rate of transmission. Currently, in the literature there are no geographic studies on chronic HBV employing linear or geospatial analyses. Many studies available focus solely on the use of traditional statistical methods (i.e., binary logistic regression) and fail to address the geographical factors (i.e. rurality or urbanicity, and land use) that may aggravate disease occurrence. This study compares and validates causal relationships of Chronic Hepatitis B cases with clinical and geographic explanators as independent variables in Proc Genmod

and Proc Reg in SAS, Further, density and Land Use Land Cover (LULC) mapping were used to describe demographic relationships in the empirical Chronic Hepatitis B empirical case data. New geographic relationships between prevalence statistics and multiple covariates were revealed. A linear regression model was determined to analyze chronic HBV case distribution. The geo-classified, LULC map revealed that urban residential areas had the highest proportion of cases. In the HBV model, a strong positively correlated association on the map when foreign born populations were compared with the number of cases and airport geolocations. Geolocations and urbanicity are key geographic identifiers when regressing
endemic chronic HBV cases.

Keywords: Hepatitis B, chronic HBV, land use land cover (LULC), SAS, GIS,

1. Introduction

The Hepatitis B virus (HBV) is the unknown silent killer that has infected over 2 billion people in the world, of which, over 300 million are chronic carriers (Franco et al. 2012). Approximately 600,000 deaths a year have been attributed to Hepatitis B infections worldwide with many cases in developing countries located in Southeast Asia and Sub-Saharan Africa (Ott et al. 2012). Highly infectious, HBV is transmitted through transfer of bodily fluids (including saliva, blood, and mucosal secretions) and can be spread between sexual partners and family members (primarily through vertical transmission from mother to child). The infection can occur as an acute or chronic infection.

In the United States (U.S.) alone, up to 350,000 cases a year have been identified, half of which are asymptomatic; and in the state of Florida, the rate of acute Hepatitis B infections increased by 91% between 2011 and 2015 (Florida State Health Profile 2017). Consequently, many infected people are unaware of their infection status and are at higher risks of developing Chronic Hepatitis B infections, which leads to cirrhosis of the liver and fibrosis, and primary liver cancer (hepatocellular carcinoma).Fortunately, Hepatitis B is a vaccine preventable disease and there are treatments and improvements in diagnostic procedure available. Ghumbre and Ghatol (2010) developed a rules-based intelligent system using hepatitis datasets and a Multilayer Perceptron structure and Back Propagation algorithm to diagnose and predict the severity level of a patient infected with HBV, and there are currently 7 FDA treatments approved for HBV treatment (Hepatitis B and C Treatments

2017) in the U.S. Even with advancements in technology and treatment, the number of HBV cases continues to rise.

Since the introduction of routine testing for Hepatitis B surface antigens (HBsAG) during pregnancy, and the universal birth dose of the Hepatitis B vaccine has increased detection and vaccination coverage and decreased the prevalence of childhood Hepatitis B in many countries. In the U.S., Koya et al. (2008) also showed a trend towards increased vaccine coverage in adults due people who were vaccinated as adolescents entering adulthood. However, a study by McMahon (2005) showed that vaccine efficacy starts to declines after 15 years, with 84% of those vaccinated retaining antibodies at that time. This leaves many adults who were vaccinated more than 15 years ago, as well as foreign born nationals and their families from high prevalence regions, at risk of contracting the virus.

On a Global scale, Asians have the highest rate of Chronic Hepatitis B infection and foreign born populations from Africa and Asia are up to 5 times more likely to be infected in the U.S. In a study from Walker et al (2016), African Americans have the highest rate of infection, and Asian Americans and Pacific Islanders were the second highest in the state of Florida. This difference in infection dynamic was observed in pregnant women, so there may be undetected cases from the general population that reflect the global statistics. In response to these statistics, there are several programs in the U.S. that target immigrant and migrant communities for HBV prevention. One program created by the University of Florida called the Hepatitis B Awareness and Service Linkage (HBASL) program aims to improve HBV infection status awareness and to increase treatment and counselling linkage for infected individuals in Jacksonville Florida (Stanford et al. 2016). Such programs would greatly benefit from geospatial analyses of disease distribution to better mobilize resources, especially during program expansion.

The innovation for this study comes from the comparison of statistical models that identify causal relationships and geospatial analyses to validate them. Currently, there are almost no geographical studies on chronic HBV using geospatial analyses. Many of the studies available now focus more on the use of traditional statistical methods to determine the efficacy of vaccination or a community program, and the use of global maps is limited to showing globalendemicity. While these studies have yielded much demographic information on at risk populations (i.e. age sex, ethnicity SEM), most fail to address localgeographical factors (i.e. rurality or urbanicity, and land use) that may aggravate demographic variables and in so doing affect transmission and prevalence. In addition,there have not been any geospatial studies on HBV, let alone in the state of Florida, and

there are no known studies at the county zip-code level which is what we employed for cartographic grid-stratification of the empirical infectious disease datasets.

Based on surveillance data from the Florida Charts website (Hepatitis B, Chronic 2017), this study focuses on the county with the highest counts of Chronic HBV cases, Miami-Dade (as shown in Figure 2.1.1). Our assumption was that remote sensing and geospatial analyses in GIS could validate results from forecast, vulnerability, epidemiological, statistical models generated in SAS, andcould determine the geographic landscape of HBV infections and refine target areas for guiding future HBV county-level prevention programs. GIS and remote sensing methods have been used in identifying geographical factors in other infectious diseases such as Onchoceriasis (river blindness) and Tuberculosis. A study by Jacob et al. (2013) used remote sensing and GIS to find breeding grounds for the onchocerca vector. *S.damnosum*, in Togo and was later able to validate the same model in Northern Uganda, demonstrating the universal applicability of geospatial analyses. In another study also by Jacob et al (2010), researchers used spatial autocorrelation to identify at risk populations for multidrug resistant Tuberculosis. These studies established applications of GIS to a variety of diseases. And though HBV is neither vector borne nor airborne, it is highly infectious, with multiplegeo-referenceable, bio geophysical, socio-demographic and geographic variables, therefore, the use of GIS would be appropriate for spatially describing county-level HBV predictors . The geographic investigation of an infectious disease ofteninvolves a search for spatial patterns of cases, followed by theassociation of other geographic or environmental phenomenathat coincide with those patterns(Esri 2017). Our assumption was that once an exploratory analyses have been performed, hypotheses can be tested as to how geographic space influences, and sometimes promotes HBV disease presence at the county zip code initially and then to the state and national level.

This study aims to geographically compare and validate causal relationships of Chronic Hepatitis B cases by comparing results generated from Proc GenmodPoissonian probability model and Proc Reg general linear modelin SAS, Density mappingand Land Use Land Cover (LULC) GIS mapping algorithms were employed to cartographically describe and compare the regression relationships identified in SAS.The objective of this study was to construct regression models in SASand then overlay the data onto geo-classified LULC maps in ArcGIS to target vulnerable HBV unbiased estimators.These methods may allow a researcher of analysts to sift meaningfully through county-level spatial data, identify "unusual" regression-related patterns, and formulate hypotheses to guide future HBV research.

Materials and Methods

2.1 Study Site;Florida is a state located in the southeastern region of the United States. It is bordered to the west by the Gulf of Mexico, to the north by Alabama and Georgia, to the east by the Atlantic Ocean, and to the south by the Straits of Florida and Cuba. Florida is the 3rd most populous,and the 8th most densely populated state of the United States (U.S. Census 2014). County level Chronic Hepatitis B infection counts were obtained for each Florida County through the Florida Charts databases (Florida Department of Health 2017). Based on county level data, the county with the highest count of chronic Hepatitis B cases was identified as Miami-Dade County, and selected for further analysis (Figure 2.1.1).

Miami-Dade County is geolocated in south-eastern point of Florida with more than half of it taken up by the Everglades National Park. As of 2009, approximately 62% of the county is occupied by Parks (conservational and recreational) and 5% is for agricultural use while the rest of the county is more urbanized; 8.6% is residential, 1.1% is commercial, 1% institutional, 6.9% is transportation, 3.2% is inland water sources, 1.4% is industrial, and 0.1 % is hotels and motels (Miami-Dade Gov 2010).

With so much of the county covered in wildlife, the urbanized are of the county is more compact and concentrated in one area which includes the City of Miami and its surrounding municipalities. The Miami metropolitan area is Florida's most populous urban area. It is also home to several large immigrant communities as well as a popular destination for many tourists making the area much more susceptible Hepatitis B outbreaks. Data about land use was obtained from a recent publication by the county for Land Use and development in the county (Miami-Dade Gov 2010). Geocoded data for health care facilities and residential areas in the county were obtained from the Miami-Dade County's self-service GIS website (Services M.-D.C.O. 2015).

2.2 Case Distribution: Zip code level, time series, demographic data for the Miami-Dade County was obtained from the U.S. census website (DADS 2010). Demographic data included age, sex, poverty level, race, ethnicity, and the number of Foreign Born Nationals (FBN).Since case count was only available at the county level, zip code level count data had to be inferred. Based on the county's total population and the population in each zip code, the number of cases in each zip code was calculated using the following equation:

$$Y_z = \left[\frac{P_x}{P_T}\right] * C$$

Where Y_z = the number of cases in each zip code, P_x = the population in the zip code, P_T = the total population in the county, and C= the total number of cases in the county. By taking the proportion of the county's population

in each zip code and multiplying it by the total number of cases in the county, it was possible to estimate the number of cases in each zip code area.

2.3 Environmental parameters: Environmental variables for each zip code in Miami-Dade were retrieved from the Miami-Dade GIS self-service page (Services M.-D.C.O. 2015). This included the number of HIV testing facilities, hospitals or other health centers, and airports in each zip code and their geocoded addresses. Other demographic data such as Foreign Born National population sizes, race and ethnicity, education level, age, sex, and poverty and unemployment levels were obtained from the most recent U.S. census data available (DADS 2010).

Univariate statistics were generated, and regression models were produced by employing zip code level (TIGER/Line with Selected Demographic and Economic Data 2015) data for initially summarizing the zip code level socio-demographic, epidemiological, covariate coefficients.

2.4 Regression Analysis: We generated a misspecification term for constructing explanative probability HBV, forecast vulnerability models in PROC REG and PROC GENMOD.Wecompared the outputs to find the model with the bestdiagnostic fit. Since count data was used for this study, a PROC GENMOD application to perform a Poisson regression and a Negative binomial regression (to rule out overdispersion) in SAS. For $Y\sim$ Poisson(λ) if we$E(Y)=\lambda$and $Var(Y)=\lambda Var(Y)=\lambda$ exist , the variance should at least be well-approximated by the mean (height 1967). Often in real infectious disease applications the variance for count data is larger than the meanas shown in previous studies by Jacob et al, and Griffith (2009, 2005). Fortunately, $Y\sim$negative binomial (rp) can render $E(Y)=rp/(1-p)$which is less than $Var(Y)=rp/(1-p)^2$so the negative binomial is sometimes used to account for "overdispersion" due to infectious disease sampled outliers.

Outliers can occur by chance in any county-level HBV distribution, which can indicate either measurement error or that the population has a heavy-tailed distribution. In probability theory, heavy-tailed distributions are probability distributions whose tails are not exponentially bounded: that is, they have heavier tails than the exponentialdistribution(Asmussen 2003).In the former case an HBV researcher or analyst may wish to discard them or use statistics that are robust to outliers, whilst in the latter case they may indicate that the county distribution has high skewness. In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean (Dean andIllowsky 2017) Hence a researcher or analyst should be very cautious in employing tools or intuitions that assume a normal HBV distribution. A frequent cause of outliers is a mixture of two distributions, which may be two

distinct HBV county geo-classified sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modeled by a mixture model.

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low(Bowley1901). However, a sample maximum and minimum HBV regressor value may not always be outliers as they may not be unusually far from other county, zip code, explicative observations.

As the zip-code level data was aggregated data, SAS was used to run a multiple linear regression model with PROC REG and a Pearson test for multi-collinearity was also run to identify any confounding variables. In statistics, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy (Belsley 1991*).* In this situation the coefficient estimates of a multiple regression HBV county-level zip code, forecast vulnerability model may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors (O'Hagan and McCabe1975).Hence, a multiple regression county HBV zip code model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual,sociodemographic or LULC explanatory predictor, or about which predictors are redundant with respect to others.

The models compared multiple demographic covariates such as age, sex, employment as a health care worker, and education level to the number of cases in each zip code, and identified significant variables as having a p-value below 0.05 or within a 95% confidence interval. Reliability of data was determined by looking at the Degrees of Freedom (DF) values for the Poisson regression and the R^2 values for the Multiple Linear Regression. Multiple data layers were created using different coded values for the various known zip code-levelgeo-referenceable, asymptotically, normalized, sociodemographic feature attributes. Due to missing data from the Census, fifteen zip-code areas were removed from the overall data analysis.

2.5. Spatial Analyses: ArcGIS was used to create a cartographic description of the demographic, environmental and case count HBV data. A choropleth map (Figure 2.1.1) of the case distribution in the state showed 4 counties with the highest counts of Hepatitis B cases and identified Miami-County as the county with the highest case count based on selected attributes (Figure 2.5.1). Graduated colors and shapes were used to

delineate areas with higher socioeconomic status (SES) (i.e. education, poverty level, unemployment rate, etc.) and areas with higher numbers of foreign born individuals from Africa and Asia compared to the case distribution in each county zip-code polygon. In an additional map, geocoded data of airports, free standing clinics, and HIV testing centers from the Miami-Dade GIS site (Services M.-D.C.O. 2015) were mapped out as points on the maps and case counts in each zip code which were represented employing the dot density option in GIS.

2.6. LULC Cover: ArcGIS was also used to create a LULC map of the county showing where the urban areas were concentrated.The map differentiated between the types of urban areas (Urban Commercial and Urban Residential) and rural and agricultural areas in the county.

We employed a digital elevation model to determine geomorphological, terrain-related (e.g., slope oriented coefficients) regressors for the county HBV, forecast, vulnerability model. Characteristics of drainage networks and drainage basin physiographic parameters have been used in hydrologic calculation and modeling flood and swamp water malaria mosquito abundance, in real time, using high resolution data (Jacob et al. 2008). Automated generation of drainage networks has become increasingly popular with the use of GIS and availability of DEMs (www.esri.com). These models account for topographic variability and their control over soil moisture heterogeneity and runoff within a shed. Patz et al. (1998) used a water balance ArcGIS model to hindcast weekly soil moisture levels in the Lake Victoria basin. These soil moisture levels were then associated with local human biting rates and entomologic inoculation rates. Jacob et al. (2010) evaluated environmental factors such as elevation range to determine human tuberculosis risk in San Juan de Lurangcho, a district in Lima, Peru. The model yielded several catchment hydrological variables including percent surface saturation, and total surface runoff for identification of potential productive endemic sites.

DEM's used in this study were Aster Global DEM's retrieved from the U.S. earth explorer database (Survey U.S.G.S. 2017). Asupervised classification was used to manuallygeo-classify different, geo-referenceable, LULC attribute, features in the zip code datasets(Esri, 2017). A map depicting zip-code contour lines and case count distribution was layered over the LULC map to identifyall geographic relationships.

3. Results

Based on data from the Florida Charts website, a map of case counts in each county was created (Figure2.1.1). This map showed that there were two counties with the highest number of cases in the state,

Broward County and Miami-Dade County. From the map attribute tables actual case count numbers, Miami Dade County was identified as the county with the highest number of cases. As such we selected thisfor the zip-code level HBV analysis.

In the statistical HBV probability distribution models, the Pearson Correlation (Table 3.1) revealed multicollinearity between several variables including poverty, income level, education (HS), insurance coverage (Private), and unemployment. Poverty had the most far reaching effect, correlating strongly with all the listed variables (R= 0.7 or above) and with p-values < 0.0001. It had a strong negative correlation with Private Insurance coverage (R=-0.877), education (R=-0.83790), and income level (R=0.84), and a strong positive correlation with Unemployment (R=0.72). There was also some negative correlation between poverty and the chances of being a HCW (R= -0.54) and some positive correlation between Poverty and Race (R= 0.5), both were statistically significant with p-values < 0.0001. Based on these results, Poverty, education, income level, insurance coverage, and unemployment were removed from further analyses, leaving 5 variables of interest: Race (black), employment as a HCW, sex (Male), Foreign Born Nationals, and age (under34).

The results for the Poisson HBV probability Model in Table 3.2,the forecast only showed little to no correlation between case counts and the demographic variables, Race (black), employment as a HCW, sex (Male), Foreign Born Nationals, and age (under34). Even though being a Foreign Born National had a significant p-value, its parameter estimates were 0 indicating there was no correlation to case counts. There was also a very small positive correlation for sex (with a parameter estimate of 0.0087) and a very small negative correlation for age (with a parameter estimate of -0.0074), however their p-values (0.6757 for sex, and 0.5 for age) were far too large for these correlations to be considered significant. The remaining variables, race and employment as a HCW showed significant p-values (0.0027 for HCWs and 0.0039 for race), and their parameter estimates showed very small positive correlations to case counts (with parameter estimates of 0.0003 for HCWs and 0.0056 for race). In addition, the parameter estimate for the Scale was 1.00 indicating that there was no over dispersion of data and a histogram of cases also showed an approximately normal distribution.

The results for the linear HBV model in Table 3.3 showed greater correlation between the variables and case count than in the Poisson model (Table 3.2). The parameter estimates were still low, 0.06386 for race, 0.00322 for HCWs, and 0.0000377 for Foreign Born Nationals, the p-values were highly significant (all <0.0001). For the variable of sex, there was a positive correlation (parameter estimate of 0.157), however the p-value was insignificant (p=0.1543). The linear model also included an R-squared selection step in which the model independent variables included Race, HCWs, Sex, and Foreign Born Nationals. This model had the best

fit with an R-squared value of 0.8604 and an adjusted R-squared of 0.8509 (Table 3.4). There was also a stepwise selection that identified the most significant variables in Table 3.5 and Table 3.6 as Race (black), HCWs, and being foreign born. The p-values for all the HBV variables were <0.0001 in both tables and the model R-squared in Table 3.6 were 0.7 or above for each explanatory variable.

A comparison of the results from the Poisson Model (Table 3.2) and Multiple Linear regression Model (Table 3.3) showed that the linear model produced more significant and more pronounced association as evidenced by the p-values and parameter estimates respectively. In addition, the Fit diagnostics plots (Figure 3.2 and Figure 3.3) for the linear regression model and the stepwise selection showed a more linear relationship and an approximately normally distributed data. However the residual plots for the HBV variables revealed some skewedness, especially for Race (Black), and there were few outliers.These observations were deduced fromCook's distance (1979) plots and the leverage plots in Figure 3.2 and Figure 3.3.

In statistics, Cook's *D* is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis(Cookand Weisberg 1982). In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate influential data points that are particularly worth checking for validity; or to indicate regions of the design space where it would be good to be able to obtain more data points (Cook et al. 1979).

Figure 3.5 shows a comparison of the relationship between each of the significant demographic variables and case count data variables using chloropleth and dot density mapping. Figure 3.4A shows the relationship between the percentage of Males (green) based on populationdistribution and case counts. Figure3.4B revealsthe number of HCW related variables (yellow) and HBV case counts. Figure 3.4C, identifies the number of Foreign Born Nationals (purple) and HBV case count, and Figure 3.4D illustrates the percentage of Black people (red) and case count. In all four maps, there is no clear relationship shown between the variables and case density, which correlates with the SAS results where the associations were very weak.

Figure 3.5 shows the association between the demographic variable Foreign Born National, landmarks (such asgeo-referenceable, airports, freestanding clinics, and HIV testing centers), and case count. There is more obvious relationship between Foreign Born Nationals and case count, and airport locations. Areas with locations had larger Foreign Born Nationals (shown in darker shades of purple) and higher case counts (shown with larger red circles).

Figure 3.6 is a LULC map showing the relationship between different type of Land Uses (rural/agricultural, Urban Residential and Urban Commercial areas) and case count. The map shows that the

areas with the highest number of cases were Urban Residential areas, while Rural/Agricultural areas had almost no cases.Geo-classfiiedurban commercial LULC areas had some cases but very few. The map shows that a large part of the county is urban residential which may account for why there is a larger proportion of cases in those areas.

4. Conclusions

Based on the results in Tables 3.2 and Tables 3.3, the linear regression method was the better method for analyzing Chronic HBV cases distribution at the county zip code level. The results were more significant, and the associations were larger, though still very small. The cartographic depictions of the relationships in Figure 3.4 also confirmed that the associations were very small. However, it is interesting to note that in the case of Foreign Born populations, when compared using graduated symbols revealed a stronger association with the number of cases and airport locations (Figure 3.5). This suggests that airport locations may be a key identifier of potential risk areas for chronic HBV cases. This is logical since airports are a major entry point for immigrants and tourists that may carry the virus.

In addition, based on the LULC map (Figure 3.6), Urban Residential areas had the highest proportion of cases, with many of the cases occurring in zip codes closer to the coastline. Urban Residential areas would be more likely to have denser populations and crowded living spaces, forcing people closer together and increasing HBV transmission. People who live in close proximity to one another spread diseases more quickly and easily (Jones 2008).

One of the major limitations of this study was the absence of zip-code level data. The number of cases in each zip-code had to be based on the zip-code population size, so the distribution of cases in this studywas more theoretical. Another limitation to this study was the incompleteness of census data which required the removal of fifteen zip-codes. Future studies can expand on these findings and improve the models by using geocoded case data to create a hot spot analysis of cases to confirm a causal relationship with airports. Among the most important exploratory methods for epidemiology and public health are methods for identifying space-time clusters or "hot spots" of disease (Jacob et al. 2010, Griffith 2005)

Openshaw's geographic analysis machine (GAM) was an early method that worked completely within a hybrid GIS. The GAM's many applications included an attempt to determine if spatial clusters of childhood leukemia were located near nuclear facilities in Britain (Marshall 1991).The GAM can works with county-level

zip code, grid stratafiable capture point data on HBV cases and searches at regular intervals for statistically significant clusters of disease prevalence. HBV maps displaying the geolocations of significant clusters may reveal the proximity of clusters to hypothesized environmental threats such as elementary school senior citizen facilities). Although Openshaw's work was widely criticized on statistical grounds, it opened the door for an active body of research on exploratory spatial analysis of disease. Some of the new methods that have been developed as outgrowths of Openshaw's approach have been published (Cressie 1992).

Exploratory methods are also valuable in searching for zones or districts of high HBV county-level prevalence. Because zip code areas may differ greatly in population size, prevalence rates may have different levels of variability and thus reliability. Researchers have long used probability mapping to show the statistical significance of prevalence rates (ChoynowskiandKaldor1987).However, probability mapping does not give a sense of the actual rates or the populations on which they are based. An alternative method may be to smooth rates towards a regional or local mean value using empirical Bayes methods. Although GIS and empirical Bayes methods have developed separately, there is much scope for interaction especially for generating multivariate, geo-referenecable, zip-codelevel, HBV forecast, vulnerability, regression maps. For example, GIS can be used to generate geographically based regional or local means to which actual HBV county rates are smoothed. These might be based on averaging rates for contiguous areas, or they might rely on more complex, multivariate, spatial clustering procedures that incorporate proximity as well as zip code, grid-stratified, population attributes. In so doing complete census data on housing and demographics may be employed cartographically and mathematically identify areas of increased risk based on the number and type of housing available inan HBV probability county-level, zip code model. This model can then be transposed onto other counties in Florida to assess the applicability of the model in multiple geolocations.

In conclusion linear regression model was determined to analyze chronic HBV case distribution at the county-level. Thegeo-classified, LULC map revealed that UrbanResidential areas had the highest proportion of cases. When foreign born populations were compared with the number of cases and airport geolocations, in the HBV model, a strong positively correlated association was revealed on the map. ArcGIS may cartographically determine covariates associated to HBV while regression models quantitate their statistical significance. In so doing HBV prevention programs may be optimally remotely and geographically targeted.

References

Asmussen, S. R., 2003. “Steady-State Properties of GI/G/1". Applied Probability and Queues.*Stochastic Modelling and Applied Probability*, (51), 266–301.

Belsley, D., 1991. Conditioning Diagnostics: Collinearity and Weak Data in Regression. New York: Wiley.

Bowley, A. L. 1901. Elements of Statistics, London:King & Son.

ChoynowskiM,and Kaldor J. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*,(43),671-681.

Cook, D.R. 1979. Influential Observations in Linear Regression. *Journal of the American Statistical Association*, **74** (365), 169–174.

Cook, D.R. Weisberg, S. 1982. Residuals and Influence in Regression. New York: Chapman & Hall

Cressie, N. 1992.Smoothing regional maps using empirical Bayes predictors.*Geogr Anal*,(24),75–95.

DADS (Data Access and Dissemination Systems), 2010.American FactFinder. Available at: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml [Accessed June 24, 2017].

Dean, S., Illowsky, B. 2017.Descriptive Statistics: Skewness and the Mean, Median, and Mode. Available at: https://www.saylor.org/site/wp-content/uploads/2011/06/MA121-1.3.3.pdf [Accessed August 21, 2017]

Esri. 2017. Classifying Imagery Using ArcGIS [online]. *Esri Accounts*. Available at: https://www.esri.com/training/ [Accessed June 24, 2017].

Florida Department of Health. 2017. FLHealthCHARTS. *FLHealthCHARTS*. Available at: http://www.flhealthcharts.com/charts/default.aspx [Accessed June 24, 2017].

Florida-State Health Profile, 2017. [online]. National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. Available from: https://www.cdc.gov/nchhstp/stateprofiles/pdf/florida_profile.pdf [Accessed 12 Jul 2017].

Franco, E. et al. 2012. Hepatitis B Epidemiology and prevention in developing countries.*World Journal of Hepatology*, (4)3, 74-80.

Ghumbre, S. and Ghatol, A., 2010.AN INTELLIGENT SYSTEM FOR HEPATITIS B DISEASE DIAGNOSIS.*International Journal of Computers and Applications*, 32 (4).

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, 1-14.

Hepatitis B and C Treatments, 2017.Fda.gov[online]. Available from: https://www.fda.gov/ForPatients/Illness/HepatitisBC/ucm408658.htm [Accessed 13 Jul 2017].

Hepatitis B, Chronic [online], 2017. Flhealthcharts.com. Available from: http://www.flhealthcharts.com/charts/LoadPage.aspx?l=~/OtherIndicators/NonVitalIndNoGrpCountsDataViewer.aspx?cid=8659 [Accessed 13 Jul 2017].

Jacob BG, Griffith DA, Novak RJ, 2008.Decomposing malaria mosquito aquatic habitat data into spatial autocorrelation eigenvectors in a SAS/GIS module.*Transactions in GIS*. (12): 341-364.

Jones, K.E. et al., 2008. Global Trends in Emerging Infectious Diseases. *Nature*. (451):7181. 990-993.

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

Jacob, B., et al., 2010. Accounting for autocorrelation in multi-drug resistant tuberculosis predictors using a set of parsimonious orthogonal eigenvectors aggregated in geographic space. *Geospatial health*, 4 (2), 201.

Jacob, B., et al. ., 2013.Validation of a Remote Sensing Model to Identify Simuliumdamnosums.l. Breeding Sites in Sub-Saharan Africa.*PLoS Neglected Tropical Diseases*, 7 (7).

Koya, D., Hill, E. and Darden, P., 2008. The Effect of Vaccinated Children on Increased Hepatitis B Immunization Among High-Risk Adults. *American Journal of Public Health*, 98 (5), 832-838.

Miami-Dade Gov. 2010. Land Use. Available at: https://www.miamidade.gov/greenprint/planning/library/milestone_one/land_use.pdf [Accessed June 24, 2017].

Marshall R. 1991.A review of methods for the statistical analysis of spatial patterns of disease.*J R Stat Soc* [Ser A].(154).421–441.

McMahon, B., et al., 2005. Antibody Levels and Protection after Hepatitis B Vaccination: Results of a 15-Year Follow-up. *Annals of Internal Medicine*, 142 (5), 333.

O'Hagan, J., McCabe, B. 1975. Tests for the Severity of Multicolinearity in Regression Analysis: A Comment. *Review of Economics and Statistics*. 57 (3): 368–370.

Ott, J., Stevens, G., Groeger, J. and Wiersma, S., 2012. Global epidemiology of hepatitis B virus infection: New estimates of age-specific HBsAgseroprevalence and endemicity. *Vaccine*, 30 (12), 2212-2219.

Services, M.-D.C.O. 2015. GIS Data. *Miami-Dade County - Information Technology - GIS Data*. Available at: http://www.miamidade.gov/technology/gis-data.asp [Accessed June 24, 2017].

U.S. Census Bureau, 2017.Florida Passes New York to Become Nation's Third Most Populous State [online].The United States Census Bureau. Available from: https://www.census.gov/newsroom/press-releases/2014/cb14-232.html [Accessed 20 Jul 2017].

Survey, U.S.G.S. 2017. U.S.G., EarthExplorer. *EarthExplorer*. Available at: https://earthexplorer.usgs.gov/ [Accessed June 24, 2017].

Stanford, J. et al., 2016. Community-engaged strategies to promote hepatitis B testing and linkage to care in immigrants of Florida.*Journal of Epidemiology and Global Health*, 6 (4), 277-284.

Walker, T. et al., 2016. Characteristics of Pregnant Women with Hepatitis B Virus Infection in 5 US Public Health Jurisdictions, 2008-2012. *Public Health Reports*, 131 (5), 685-694.

International Journal of Geographic Information System
Vol. 4, No. 4, November 2017, pp. 1-24
Available Online at http://acascipub.com/Journals.php

Appendix:

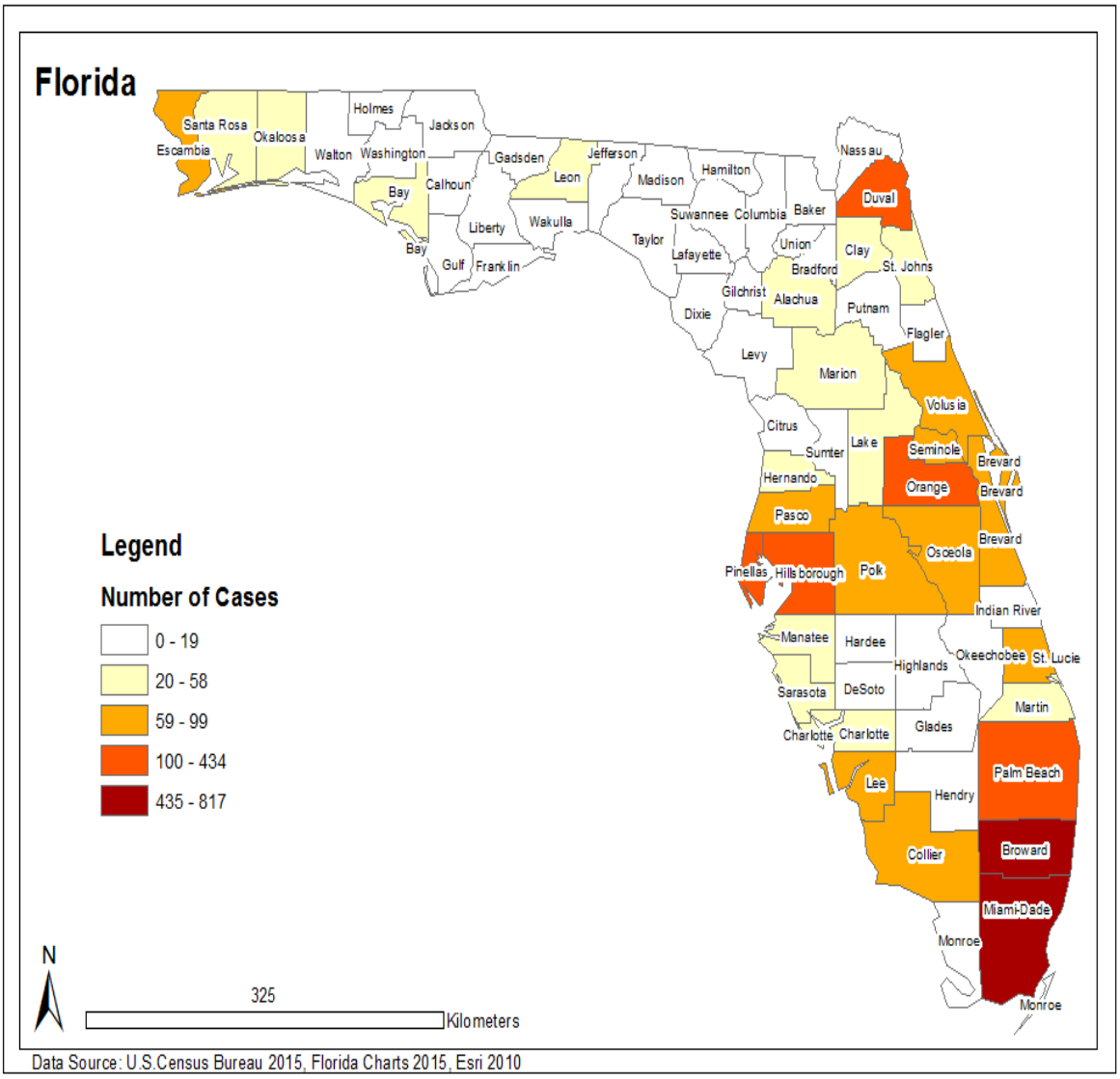
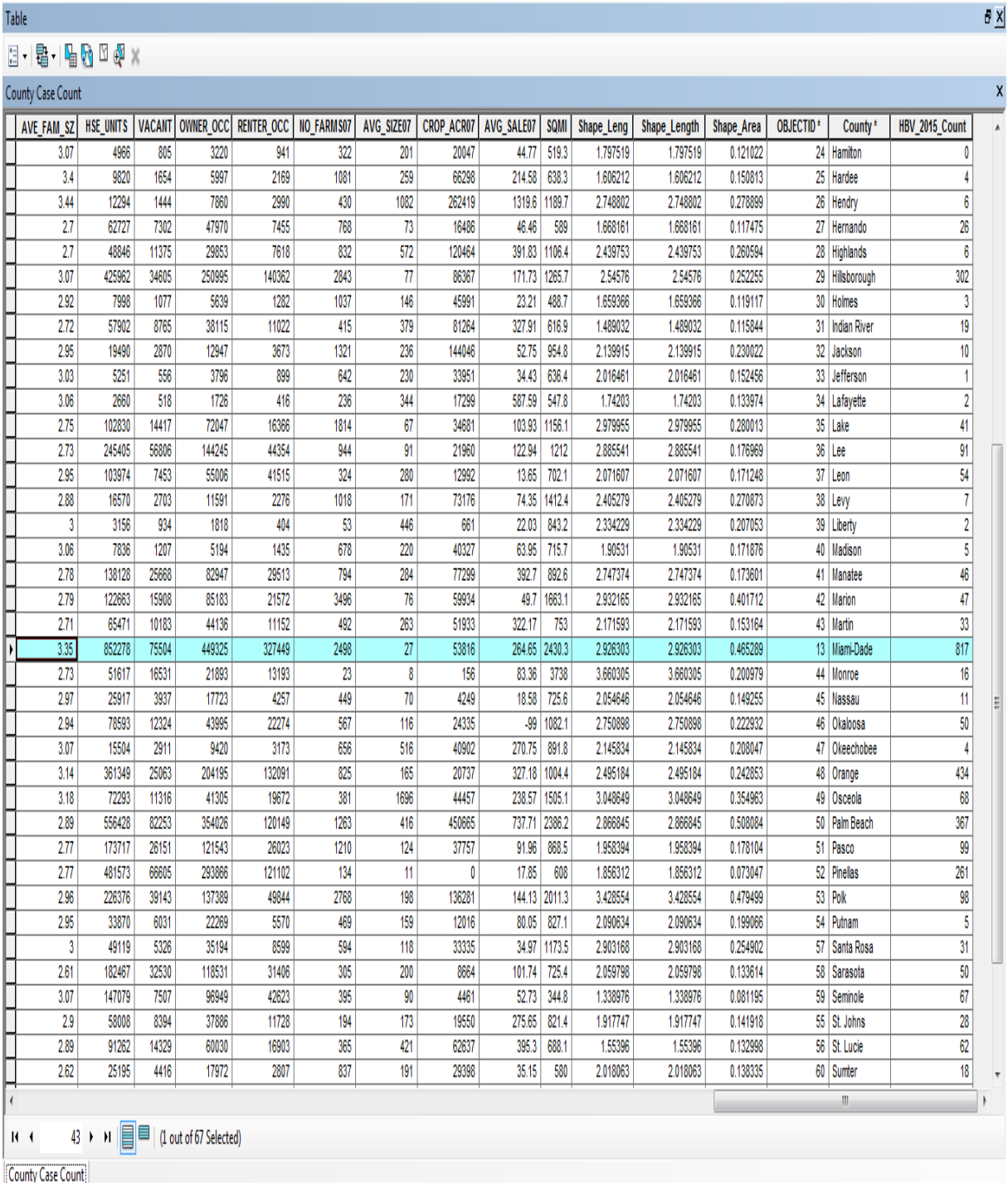


Figure 2.1.1.shows a chloropleth map of case count distributions in Florida counties.



Table

County Case Count

AVE_FAM_SZ	HSE_UNITS	VACANT	OWNER_OCC	RENTER_OCC	NO_FARMS07	AVG_SIZE07	CROP_ACR07	AVG_SALE07	SQMI	Shape_Leng	Shape_Length	Shape_Area	OBJECTID *	County *	HBV_2015_Count
3.07	4966	805	3220	941	322	201	20047	44.77	519.3	1.797519	1.797519	0.121022	24	Hamilton	0
3.4	9820	1654	5997	2169	1081	259	66298	214.58	638.3	1.606212	1.606212	0.150813	25	Hardee	4
3.44	12294	1444	7860	2990	430	1082	262419	1319.6	1189.7	2.748802	2.748802	0.278899	26	Hendry	6
2.7	62727	7302	47970	7455	768	73	16486	46.46	589	1.668161	1.668161	0.117475	27	Hernando	26
2.7	48846	11375	29853	7618	832	572	120464	391.83	1106.4	2.439753	2.439753	0.260594	28	Highlands	6
3.07	425962	34605	250995	140362	2843	77	86367	171.73	1265.7	2.54576	2.54576	0.252255	29	Hillsborough	302
2.92	7998	1077	5639	1282	1037	146	45991	23.21	488.7	1.659366	1.659366	0.119117	30	Holmes	3
2.72	57902	8765	38115	11022	415	379	81264	327.91	616.9	1.489032	1.489032	0.115844	31	Indian River	19
2.95	19490	2870	12947	3673	1321	236	144046	52.75	954.8	2.139915	2.139915	0.230022	32	Jackson	10
3.03	5251	556	3796	899	642	230	33951	34.43	636.4	2.016461	2.016461	0.152456	33	Jefferson	1
3.06	2660	518	1726	416	236	344	17299	587.59	547.8	1.74203	1.74203	0.133974	34	Lafayette	2
2.75	102830	14417	72047	16366	1814	67	34681	103.93	1156.1	2.979955	2.979955	0.280013	35	Lake	41
2.73	245405	56806	144245	44354	944	91	21960	122.94	1212	2.885541	2.885541	0.176969	36	Lee	91
2.95	103974	7453	55006	41515	324	280	12992	13.65	702.1	2.071607	2.071607	0.171248	37	Leon	54
2.88	16570	2703	11591	2276	1018	171	73176	74.35	1412.4	2.405279	2.405279	0.270873	38	Levy	7
3	3156	934	1818	404	53	446	661	22.03	843.2	2.334229	2.334229	0.207053	39	Liberty	2
3.06	7836	1207	5194	1435	678	220	40327	63.95	715.7	1.90531	1.90531	0.171876	40	Madison	5
2.78	138128	25668	82947	29513	794	284	77299	392.7	892.6	2.747374	2.747374	0.173601	41	Manatee	46
2.79	122663	15908	85183	21572	3496	76	59934	49.7	1663.1	2.932165	2.932165	0.401712	42	Marion	47
2.71	65471	10183	44136	11152	492	263	51933	322.17	753	2.171593	2.171593	0.153164	43	Martin	33
3.35	852278	75504	449325	327449	2498	27	53816	264.65	2430.3	2.926303	2.926303	0.465289	13	Miami-Dade	817
2.73	51617	16531	21893	13193	23	8	156	83.36	3738	3.660305	3.660305	0.200979	44	Monroe	16
2.97	25917	3937	17723	4257	449	70	4249	18.58	725.6	2.054646	2.054646	0.149255	45	Nassau	11
2.94	78593	12324	43995	22274	567	116	24335	-99	1082.1	2.750898	2.750898	0.222932	46	Okaloosa	50
3.07	15504	2911	9420	3173	656	516	40902	270.75	891.8	2.145834	2.145834	0.208047	47	Okeechobee	4
3.14	361349	25063	204195	132091	825	165	20737	327.18	1004.4	2.495184	2.495184	0.242853	48	Orange	434
3.18	72293	11316	41305	19672	381	1696	44457	238.57	1505.1	3.048649	3.048649	0.354963	49	Osceola	68
2.89	556428	82253	354026	120149	1263	416	450665	737.71	2386.2	2.866845	2.866845	0.508084	50	Palm Beach	367
2.77	173717	26151	121543	26023	1210	124	37757	91.96	868.5	1.958394	1.958394	0.178104	51	Pasco	99
2.77	481573	66805	293866	121102	134	11	0	17.85	608	1.856312	1.856312	0.073047	52	Pinellas	261
2.96	226376	39143	137389	49844	2768	198	136281	144.13	2011.3	3.428554	3.428554	0.479499	53	Polk	98
2.95	33870	6031	22269	5570	469	159	12016	80.05	827.1	2.090634	2.090634	0.199066	54	Putnam	5
3	49119	5326	35194	8599	594	118	33335	34.97	1173.5	2.903168	2.903168	0.254902	57	Santa Rosa	31
2.61	182467	32530	118531	31406	305	200	8664	101.74	725.4	2.059798	2.059798	0.133614	58	Sarasota	50
3.07	147079	7507	96949	42623	395	90	4461	52.73	344.8	1.338976	1.338976	0.081195	59	Seminole	67
2.9	58008	8394	37886	11728	194	173	19550	275.65	821.4	1.917747	1.917747	0.141918	55	St. Johns	28
2.89	91262	14329	60030	16903	365	421	62637	395.3	688.1	1.55396	1.55396	0.132998	56	St. Lucie	62
2.62	25195	4416	17972	2807	837	191	29398	35.15	580	2.018063	2.018063	0.138335	60	Sumter	18

43 (1 out of 67 Selected)

County Case Count

Figure 2.5.1 shows the screenshot of the attributes table from which the county with the highest number of cases was identified.

Table 3.1 shows the results of a Pearson Correlation performed on all demographic variables. HCW represents Health Care Workers, and HS represents High School level education. The top value in each cell represents the correlation value and the bottom value represents the p-value.

Pearson Correlation Coefficients, N = 64 Prob > |r| under H0: Rho=0										
	Private	**HCW**	**HS**	**Unemployment**	**Male**	**Age**	**Poverty**	**Foreign**	**Black**	**Income**
Private	1.00000	0.47333 <.0001	0.93088 <.0001	-0.68684 <.0001	-0.00374 0.9766	-0.09709 0.4453	-0.87700 <.0001	-0.24662 0.0495	-0.39639 0.0012	0.93055 <.0001
HCW	0.47333 <.0001	1.00000	0.52533 <.0001	-0.41365 0.0007	-0.26882 0.0317	-0.14862 0.2412	-0.54196 <.0001	0.40399 0.0009	-0.24017 0.0559	0.45608 0.0002
HS	0.93088 <.0001	0.52533 <.0001	1.00000	-0.67207 <.0001	-0.15469 0.2223	-0.10702 0.4000	-0.83790 <.0001	-0.16925 0.1812	-0.29798 0.0168	0.82359 <.0001
Unemploymen t	-0.68684 <.0001	-0.41365 0.0007	-0.67207 <.0001	1.00000	-0.02959 0.8165	0.14866 0.2410	0.72813 <.0001	-0.19653 0.1196	0.70224 <.0001	-0.60816 <.0001
Male	-0.00374 0.9766	-0.26882 0.0317	-0.15469 0.2223	-0.02959 0.8165	1.00000	0.17450 0.1679	0.15570 0.2192	-0.32417 0.0090	-0.15809 0.2122	0.02812 0.8255
Under 34	-0.09709 0.4453	-0.14862 0.2412	-0.10702 0.4000	0.14866 0.2410	0.17450 0.1679	1.00000	0.21119 0.0939	-0.30168 0.0154	0.38095 0.0019	-0.01301 0.9187
Poverty	-0.87700 <.0001	-0.54196 <.0001	-0.83790 <.0001	0.72813 <.0001	0.15570 0.2192	0.21119 0.0939	1.00000	-0.01312 0.9181	0.50374 <.0001	-0.84299 <.0001
Foreign	-0.24662 0.0495	0.40399 0.0009	-0.16925 0.1812	-0.19653 0.1196	-0.32417 0.0090	-0.30168 0.0154	-0.01312 0.9181	1.00000	-0.38282 0.0018	-0.19232 0.1279
Black	-0.39639 0.0012	-0.24017 0.0559	-0.29798 0.0168	0.70224 <.0001	-0.15809 0.2122	0.38095 0.0019	0.50374 <.0001	-0.38282 0.0018	1.00000	-0.40790 0.0008
Income	0.93055 <.0001	0.45608 0.0002	0.82359 <.0001	-0.60816 <.0001	0.02812 0.8255	-0.01301 0.9187	-0.84299 <.0001	-0.19232 0.1279	-0.40790 0.0008	1.00000

Table 3.2shows the results of the Poisson Model

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**P-value**
Intercept	1	1.3230	1.0452	-0.7255	3.3716	1.60	0.2056
Black	1	0.0056	0.0019	0.0018	0.0094	8.35	0.0039
HCW	1	0.0003	0.0001	0.0001	0.0004	8.99	0.0027
Foreign	1	0.0000	0.0000	0.0000	0.0000	48.39	<.0001
Male	1	0.0087	0.0207	-0.0319	0.0493	0.18	0.6757
Under34	1	-0.0074	0.0112	-0.0293	0.0145	0.44	0.5075
Scale	0	1.0000	0.0000	1.0000	1.0000		

Table 3.3shows the results of the Multiple Linear Regression

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**P-value**	**Variance Inflation**
Intercept	Intercept	**1**	-7.39717	5.69612	-1.30	0.1991	0
Black	Black	**1**	0.06386	0.01024	6.24	<.0001	1.34396
HCW	HCW	**1**	0.00322	0.00054627	5.90	<.0001	1.25977
Foreign	Foreign	**1**	0.00037738	0.00002600	14.51	<.0001	1.51853
Male	Male	**1**	0.15711	0.10886	1.44	0.1543	1.30160

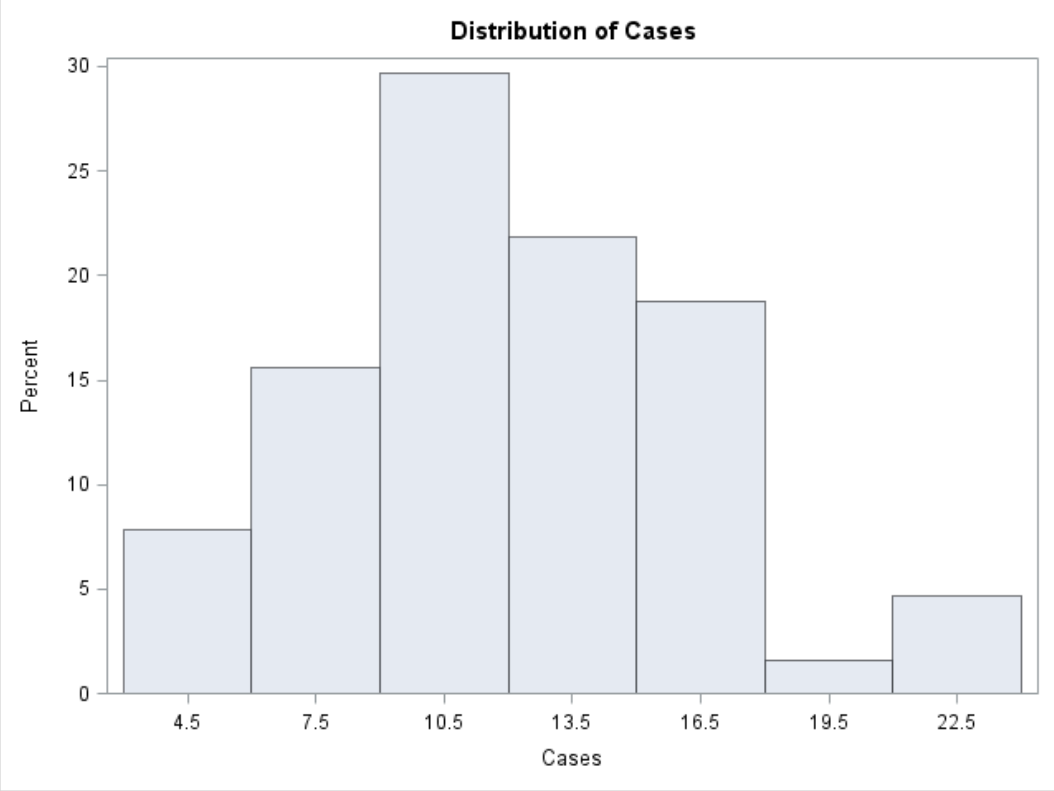


Figure 3.1shows a histogram of case distribution in the study, demonstrating an approximately normal distribution.

Table 3.4 shows the table with results from the R-squared selection

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	MSE	SSE	Variables in Model
4	**0.8509**	0.8604	4.4263	73.1030	2.90761	171.54875	Black HCW Foreign Male
5	**0.8495**	0.8614	6.0000	74.6343	2.93616	170.29713	Black HCW Foreign Male Under34
3	**0.8482**	0.8555	4.4889	73.3234	2.96008	177.60487	Black HCW Foreign
4	**0.8460**	0.8558	6.3716	75.1992	3.00442	177.26054	Black HCW Foreign Under34
2	**0.7706**	0.7779	34.9550	98.8214	4.47427	272.93057	Black Foreign
3	**0.7671**	0.7782	36.8371	100.7401	4.54307	272.58423	Black Foreign Under34
3	**0.7669**	0.7780	36.8953	100.7803	4.54592	272.75534	Black Foreign Male
4	**0.7634**	0.7784	38.7345	102.6694	4.61497	272.28299	Black Foreign Male Under34
3	**0.7610**	0.7724	39.2454	102.3793	4.66093	279.65562	HCW Foreign Under34
2	**0.7594**	0.7671	39.4893	101.8695	4.69252	286.24387	HCW Foreign
4	**0.7590**	0.7743	40.4776	103.8612	4.70171	277.40109	HCW Foreign Male Under34
3	**0.7568**	0.7684	40.9242	103.4975	4.74308	284.58484	HCW Foreign Male
3	**0.7015**	0.7157	62.9813	116.6199	5.82247	349.34796	Foreign Male Under34
2	**0.7010**	0.7105	63.1666	115.7847	5.83220	355.76433	Foreign Male
1	**0.6999**	0.7047	63.5875	115.0507	5.85278	362.87224	Foreign
2	**0.6995**	0.7090	63.7903	116.1132	5.86222	357.59539	Foreign Under34
2	**0.3407**	0.3617	209.1567	166.3875	12.85925	784.41413	HCW Male
3	**0.3365**	0.3681	208.4556	167.7371	12.94139	776.48314	HCW Male Under34
3	**0.3302**	0.3621	210.9882	168.3471	13.06532	783.91932	Black HCW Male
4	**0.3289**	0.3715	209.0476	169.3954	13.09066	772.34901	Black HCW Male Under34
3	**0.3138**	0.3465	217.5000	169.8893	13.38398	803.03903	Black HCW Under34
2	**0.3119**	0.3337	220.8544	169.1301	13.42230	818.76046	HCW Under34
1	**0.3112**	0.3221	223.7015	168.2330	13.43536	832.99236	HCW
2	**0.3045**	0.3266	223.8320	169.8099	13.56563	827.50318	Black HCW
2	**0.1077**	0.1360	303.5962	185.7598	17.40497	1061.70339	Male Under34
2	**0.1063**	0.1346	304.1776	185.8626	17.43296	1063.41051	Black Male
1	**0.1042**	0.1184	308.9738	185.0524	17.47363	1083.36500	Male
3	**0.0994**	0.1423	302.9875	187.2964	17.56740	1054.04376	Black Male Under34
1	**0.0209**	0.0364	343.2909	190.7441	19.09880	1184.12559	Under34
2	**0.0048**	0.0364	345.2902	192.7440	19.41186	1184.12337	Black Under34
1	**-.0109**	0.0051	356.3874	192.7894	19.71902	1222.57893	Black

Table 3.5 shows the significant variables found in the stepwise selection

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.76399	0.69041	3.62459	1.22	0.2729
Black	0.05876	0.00970	108.63900	36.70	<.0001
HCW	0.00306	0.00053960	95.32570	32.20	<.0001
Foreign	0.00036426	0.00002458	649.89832	219.55	<.0001

Table 3.6shows the summary of results from the stepwise selection

Summary of Stepwise Selection								
Step	**Variable Entered**	**Label**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Foreign	Foreign	1	0.7047	0.7047	63.5875	147.96	<.0001
2	Black	Black	2	0.0732	0.7779	34.9550	20.10	<.0001
3	HCW	HCW	3	0.0776	0.8555	4.4889	32.20	<.0001

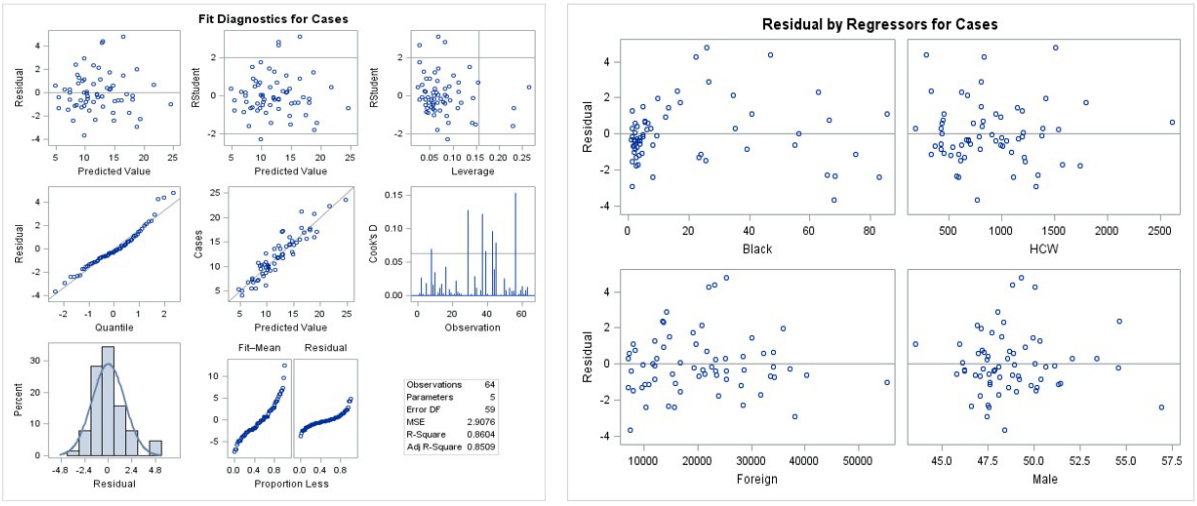


Figure 3.2shows the fit diagnostics plots for the linear regression model

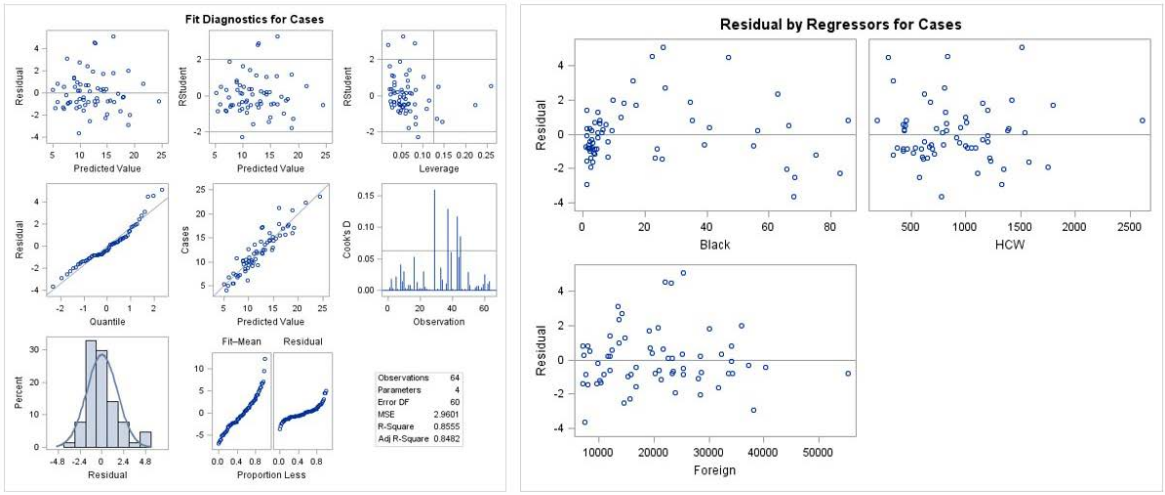


Figure 3.3shows the fit diagnostics plots for the stepwise selection of the linear regression model

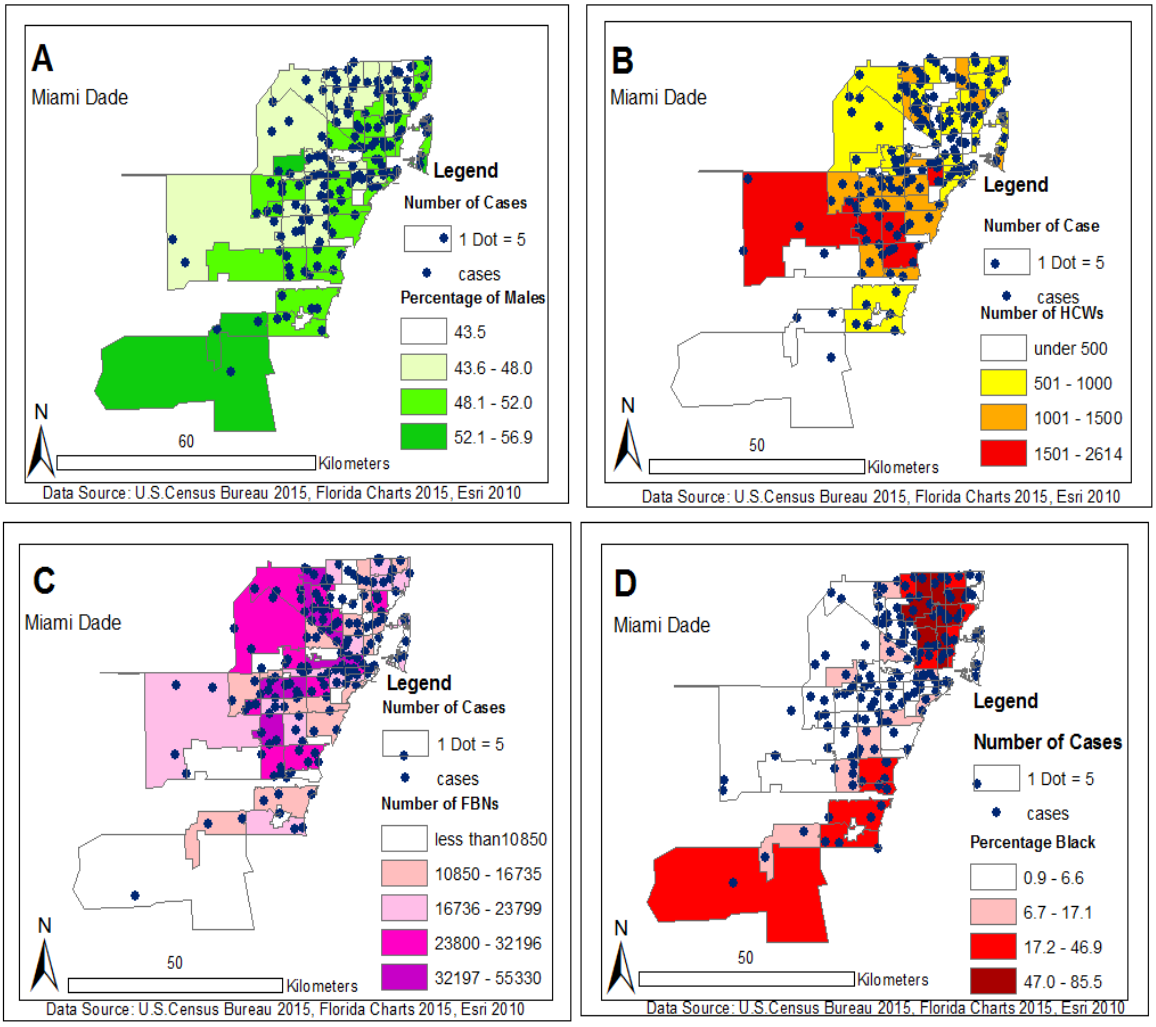


Figure 3.4shows the relationship between case count and demographic variables: A) Case count and sex (males), B) case count and number of HCWs (Health Care Workers), C) case count and FBNs (Foreign Born Nationals), and D) case count and percentage of black people. Each map is color coded and the cases are represented as dot densities, with each dot representing 5 cases.

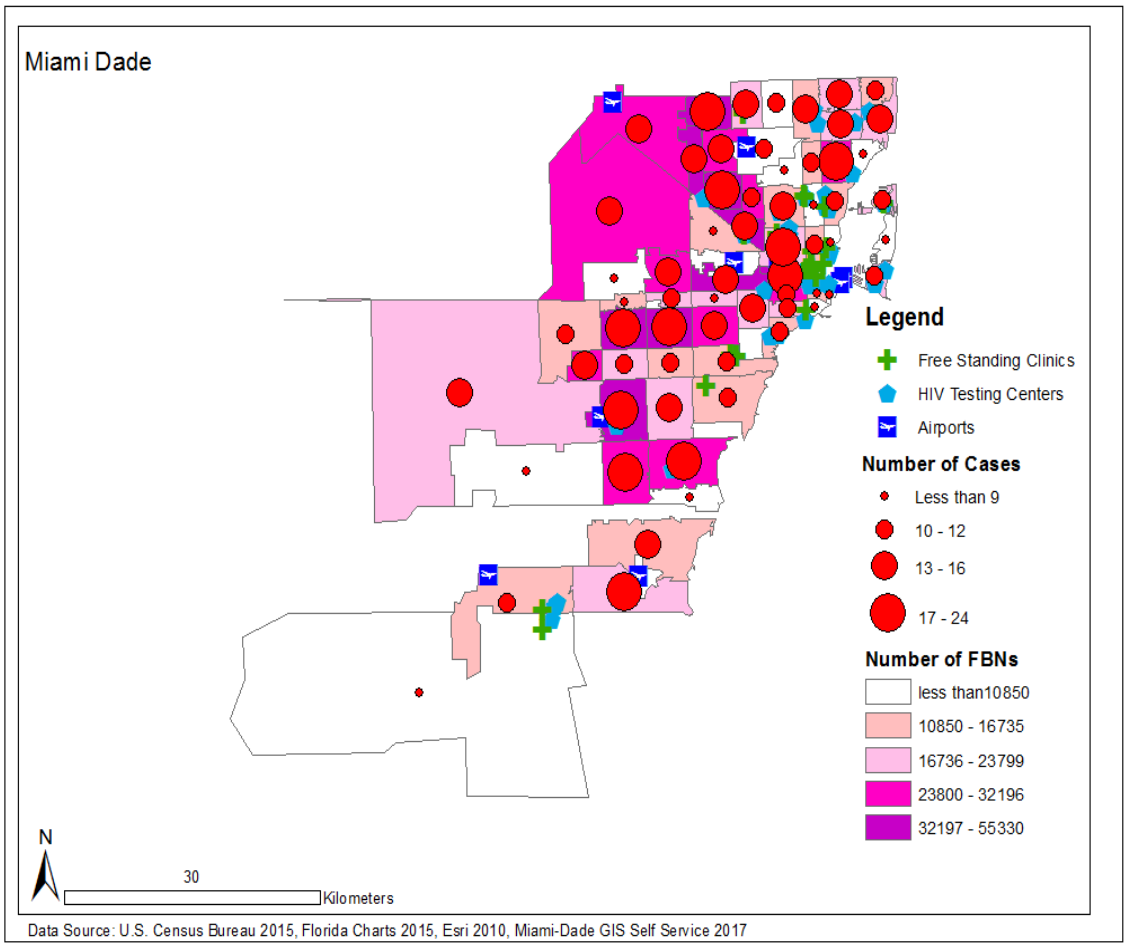


Figure 3.5is a chloropleth map showing the relationship between the number of Chronic Hepatitis B cases (as graduated circles), the Number of Foreign Born Nationals (FBN) (shades of Purple), and landmarks such as freestanding clinics (green crosses), HIV testing centers (blue pentagons), and airports (blue airport symbols).

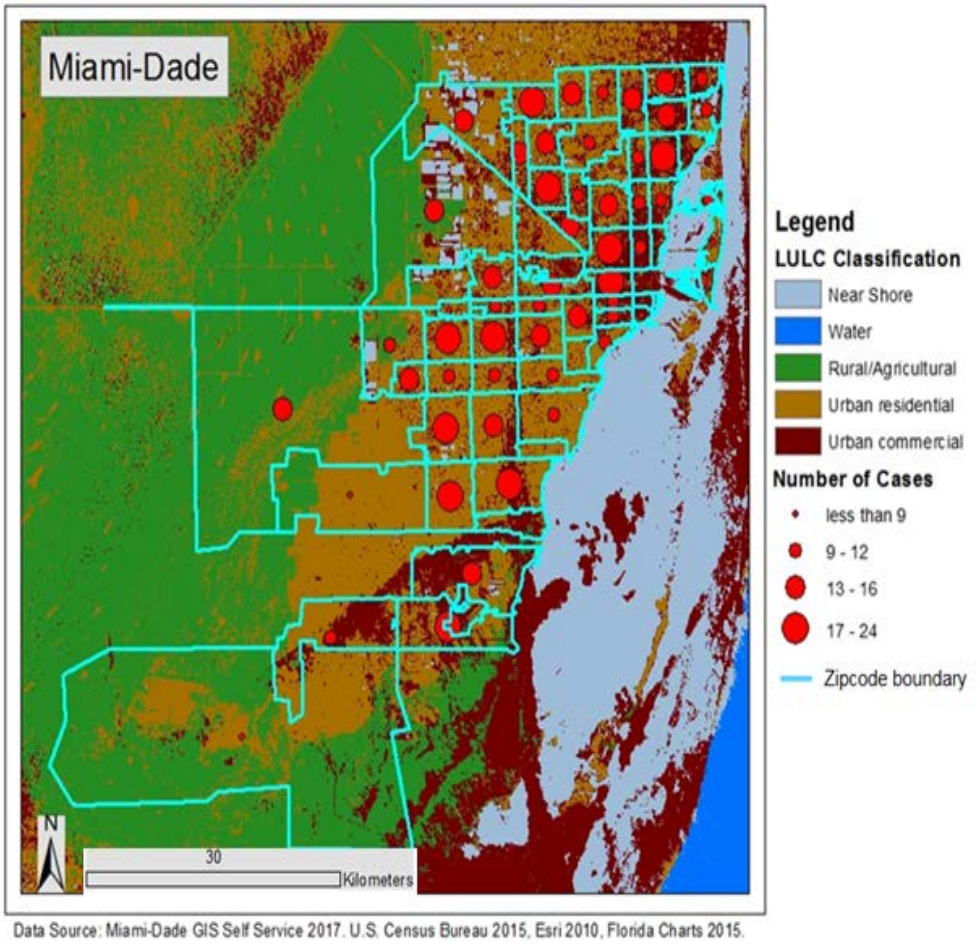


Figure 3.6shows the relationship between the number of cases and LULC in Miami-Dade County. The zip-code boundaries are highlighted in light blue, and case counts are depicted as graduated circles. The county was classified into 5 areas, Water (blue), Near Shore (grey) which represents the water near the shoreline, Rural/Agricultural (green), Urban Residential (light brown), and Urban Commercial (dark brown).