Spring 5-4-2022

A Data Visualization Framework for ESDA: Understanding Pre-Diabetes and Diabetes Prevalence in Florida

Brindal Dhol

A Data Visualization Framework for ESDA: Understanding Pre-Diabetes and Diabetes Prevalence in Florida

by

Brindal Dhol

Under the Direction of Anu Bourgeois, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2022

ABSTRACT

Exploratory spatial data analysis (ESDA) is a technique for analyzing data from different geographic regions. To examine patterns, ESDA uses univariate and multivariate graphical approaches. Through a case study of diabetes and pre-diabetes prevalence in Florida, we built a novel data visualization framework for ESDA.

Diabetes is a rapidly increasing global disease that is a major global health concern with significant implications for healthcare spending. Information about the relationship between diabetes and geographical sociodemographic characteristics could assist public health programs better targeting those who are at risk. We show the regional prevalence of disease in Florida and its relationship to the geography of risk variables using our multivariate data visualization framework.

Our methodology can be applied to wide range of problems and domains that require complex analysis of disparate data to identify correlations. The method can be used to find patterns and clusters for any problem at any spatial scale.

INDEX WORDS: Multivariate choropleth maps, Clustering, GIS, Hotspot analysis, Spatial Epidemiology, Pre-Diabetes, Diabetes, Risk factors, ESDA

A Data Visualization Framework for ESDA: Understanding Pre-Diabetes and Diabetes Prevalence in Florida

by

Brindal Dhol

Committee Chair: Anu Bourgeois

Committee: Chetan Tiwari

Suhasini Ramisetty-Mikler

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2022

ACKNOWLEDGEMENTS

I would like to thank and acknowledge the support of Dr. Anu Bourgeois, Dr. Chetan Tiwari, Dr. Suhasini Ramisetty-Mikler, and the DICE research team members for guiding me in the right direction throughout the research process. Everyone's contribution to shaping this paper is invaluable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The rapid rise in diabetes prevalence, particularly among children and young adults, is a major worldwide health concern that is expected to worsen significantly in the following decades. Urbanization and economic growth are causing demographic and socio-cultural changes in the population composition and lifestyles, particularly in rising and developing economies. The latest evidence shows that diabetes continues to be a significant global health challenge and is likely to continue to grow substantially in the following decades, which would have significant implications for healthcare expenditures, particularly in developing countries [1, 2, 3].

Diabetes imposes an increasing economic burden on national health care systems worldwide [4]. In developing countries, most diabetics are between the ages of 45 and 65 years, but in developed countries, the majority are over 64 years [5]. According to demographic trends, the number of people with diabetes aged 64 and over in developing nations will exceed 82 million by 2030, while the number in affluent countries will exceed 48 million. The Middle East crescent, Sub-Saharan Africa, and India are expected to grow at the fastest rates. Nearly 170 million men and women in developing nations will be diagnosed with diabetes throughout their reproductive years in less than 30 years.

As per the Centers for Disease Control and Prevention (CDC) National Diabetes Statistics Report, 88 million American adults i.e., approximately 1 in 3 have prediabetes and more than 8 in 10 are unaware of their condition [6]. The normal range for hemoglobin A1c is between 4% and 5.6 % in those who do not have diabetes. Hemoglobin A1c readings of 5.7 % to 6.4 % indicate pre-diabetes, whereas levels of 6.5 % or more indicate an increased risk of developing diabetes [6]. As the seventh leading cause of death in the United States, diabetes will incur $327 billion in medical costs and lost work and wages. The patients with pre-diabetes do not exhibit any

symptoms. Consequently, millions of people are already pre-diabetic, with most of them being unaware of their condition. Although there is no cure for diabetes, the state of pre-diabetes can be reversed, and numerous studies have shown that the best method to prevent diabetes is to take precautions and seek to alleviate symptoms before they advance to diabetes. Pre-diabetes must be recognized early because it cannot be prevented from progressing to type 2 diabetes, and if left untreated, it can lead to diabetes in 4 to 10 years. Because pre-diabetes has no symptoms, a regular checkup is the only way to be diagnosed before it is too late. To efficiently help communities prevent and manage diabetes, health departments need to be able to target populations with high risk but low resources [7]. Information on the relationship between diabetes prevalence and built environment attributes could allow public health programs to better target populations at risk for diabetes [3,8].

Spatial data exploration is an application of exploratory data analysis that focuses directly on the characteristics of geographic data [9]. Spatial clustering analysis has grown prevalent in many research domains, with epidemiology being the most prominent application [10]. With aggregated data, approaches such as Moran's *I* of local spatial associations could widen the analytical range to investigate spatial dependency and spatial heterogeneity, which refers to the uneven distribution of different concentrations of each species within a given area. It is recognized that different visualization strategies must be evaluated for their usefulness in enabling and encouraging exploratory spatial data analysis [11]. Therefore, we created an exploratory spatial data analysis framework in this study that uses bivariate and multivariate maps for data comparison to allow data integration and lead public health practitioners through the process of finding the geographic prevalence of disease burdens in a community. To demonstrate how this paradigm can be used to

understand the geography of this disease, we used the prevalence of diabetes and pre-diabetes in Florida and compared it to the geography of the associated risk factors.

This article begins with a brief overview of the related work section and then the explanation about overall concept of spatial epidemiology and analysis and the process followed to create this framework, followed by descriptions of two methods for creating bivariate and multivariate maps and hotspot clusters for data comparison in this framework: GIS analysis and hotspot clustering. The article then goes on to give a quick overview of all the pertinent risk factors and the results achieved utilizing this framework in the form of multivariate maps. It finishes with a discussion section in which the complete framework-building process is outlined, as well as the need for this type of comparison framework and where else this framework can be used.

2 RELATED WORK

The study of the distribution and determinants of diseases in populations, particularly human populations, is known as epidemiology. The primary goal of epidemiology study is to determine the types and extent of illnesses that affect populations and the factors that influence disease outcomes. Epidemiologists examine the interactions that occur between the host, the agent, and the environment to determine what causes disease and how to avoid and control it. Spatial epidemiology has been mostly used for studying communicable diseases, but its application has currently expanded to the study of non-communicable diseases like diabetes [1]. Quantitative methods in spatial epidemiology may make it easier to create and implement control strategies by evaluating the net impact of geographical hotspots (areas with a disproportionately large burden of disease) and disease burden determinants at the ecological and individual levels [12]. The identification of these areas can uncover the locations of high-risk populations as well as reveal the factors that facilitate the persistence and spread of epidemics.

Understanding the geographic differences in the distribution of a given health condition and identifying high-prevalence areas is crucial to guiding control and prevention initiatives, according to the research 'Investigation of geographic disparities of pre-diabetes and diabetes in Florida.' The objective of this study was to investigate clusters of pre-diabetes and diabetes risk in Florida and identify significant predictors of the conditions [13]. The 2013 Florida Behavioral Risk Factor Surveillance System (BRFSS) data was used to evaluate county-level geographic disparities and determinants of prediabetes and diabetes prevalence. The findings of this study show that at the county level in Florida, there are spatial patterns of the high prevalence of pre-diabetes and diabetes. Three of the eight counties in the primary diabetes cluster, as well as the secondary cluster adjacent to it (Hendry County), are designated as geographic Health Professional Shortage Areas

(HPSAs) by the Health Resources and Services Administration (HRSA), implying that primary care services are in shortage.

Another study, titled 'Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009–2010,' looked at how spatial epidemiology and Geographic Information Systems be applied to the study of diabetes. They looked at a various spatial methodology for understanding the disease's spatial structure and identify potential geographical causes of diabetes's spatial distribution. Finally, they explored how spatial epidemiology might be used to create and implement diabetes prevention and treatment strategies that are regionally targeted.

Various environmental factors have been proposed to influence type 2 diabetes mellitus (T2DM). ‘Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review’ is the study that incorporates information from four databases on environmental factors of T2DM. It proposes a theoretical framework illustrating the link between environment and T2DM, and briefly discusses some methodological challenges and potential solutions, and opportunities for future research [14]. The most common environmental parameters investigated were walkability, air pollution, food and physical activity environment, and roadway proximity. Current evidence is limited in amount and study quality, making causal judgments impossible. However, the evidence suggests that environmental factors may play a role in Type 2 diabetes mellitus prevention, and it also provides a solid foundation for further research using higher-quality data and longitudinal studies involving policy-relevant environmental measures. This quest for more significant evidence is crucial to promoting health-oriented urban design and city planning.

The study 'Geospatial and geodemographic insights for diabetes in the United States examines the geodemographic correlates of Type 2 diabetes in the US. Further, with such a large

percentage of the U.S. population being diabetic, an interesting spatial pattern for the disease has emerged [15]. Specifically, we provide an exploratory geographical analysis of lifestyle groups and their relationship with diabetes using a countrywide database of age-adjusted, county-level estimates for diabetes prevalence. The findings imply that geodemographic data can be useful in detecting risky lifestyle environments and providing basic guidelines for identifying at-risk groups so that intervention efforts can be targeted more effectively.

Pre-diabetes is a significant risk factor for developing type 2 diabetes (T2D). The goal of the study, titled 'Factors associated with progression to pre-diabetes: a recurrent events analysis', was to look into factors linked to normal glucose maintenance and the prevention or delay of pre-diabetes. Despite frequent studies on lifestyle modification for pre-diabetes prevention, less information is available about the role of nutritional components [16]. In first-degree relatives, we found direct effects on macronutrient intake, such as fat, carbs, and protein. To examine these relationships in general populations, more research is needed.

We have proposed a novel method of data comparison and analysis in this study. We were unable to locate any other study that used bivariate and multivariate maps for data comparison in any GIS (Geographic Information Science)-related diabetes studies.

3 SPATIAL EPIDEMIOLOGY AND ANALYSIS

Quantitative tools in spatial epidemiology, as we discussed earlier, may make it easier to develop and implement control plans by assessing the net impact of geographical hotspots. Therefore, this is how we approached here about making this data visualization framework.

Phase 1: Starting with creating a list of essential aspects by reading various published papers.

Phase 2: Acquiring data for those aspects from standardized data sources like Florida Health Charts for county-wise diabetes and pre-diabetes data, BRFSS (Behavioral Risk Factor Surveillance System) to collect data for diabetes and pre-diabetes related lifestyle risk factors such as obesity, sedentary lifestyle, smoking, heavy alcohol drinking, IRS Statistics of Income Division to collect data for median household income, ACS (American Community Survey for population distribution based on gender, race, and ethnicity, and CDC (Centers for Disease Control and Prevention) for social vulnerability index of socioeconomic status. Most of the data is available on sites like SimplyAnalytics and PolicyMap which are data and mapping tool used by government agencies, universities, healthcare institutions, and nonprofits to solve location-based problems.

Phase 3: We processed and extracted the county-wise data in the CSV file from these sources.

Phase 4: Then integrated and analyzed the data to confirm its accuracy using Excel.

Phase 5: The data was then imported into QGIS (Quantum Geographic Information System) software and mapped with the Florida state's county-by-county shapefile which was obtained from the United States Census Bureau TIGER Geodatabase, followed by the creation of multivariate choropleth maps for data visualization. To see how one object compares to others nearby, we used the local Moran's *I* coefficient. It is used to identify diabetes and pre-diabetes clusters in GeoDa which is a software that conducts spatial data analysis, geovisualization, spatial autocorrelation, and spatial modeling.

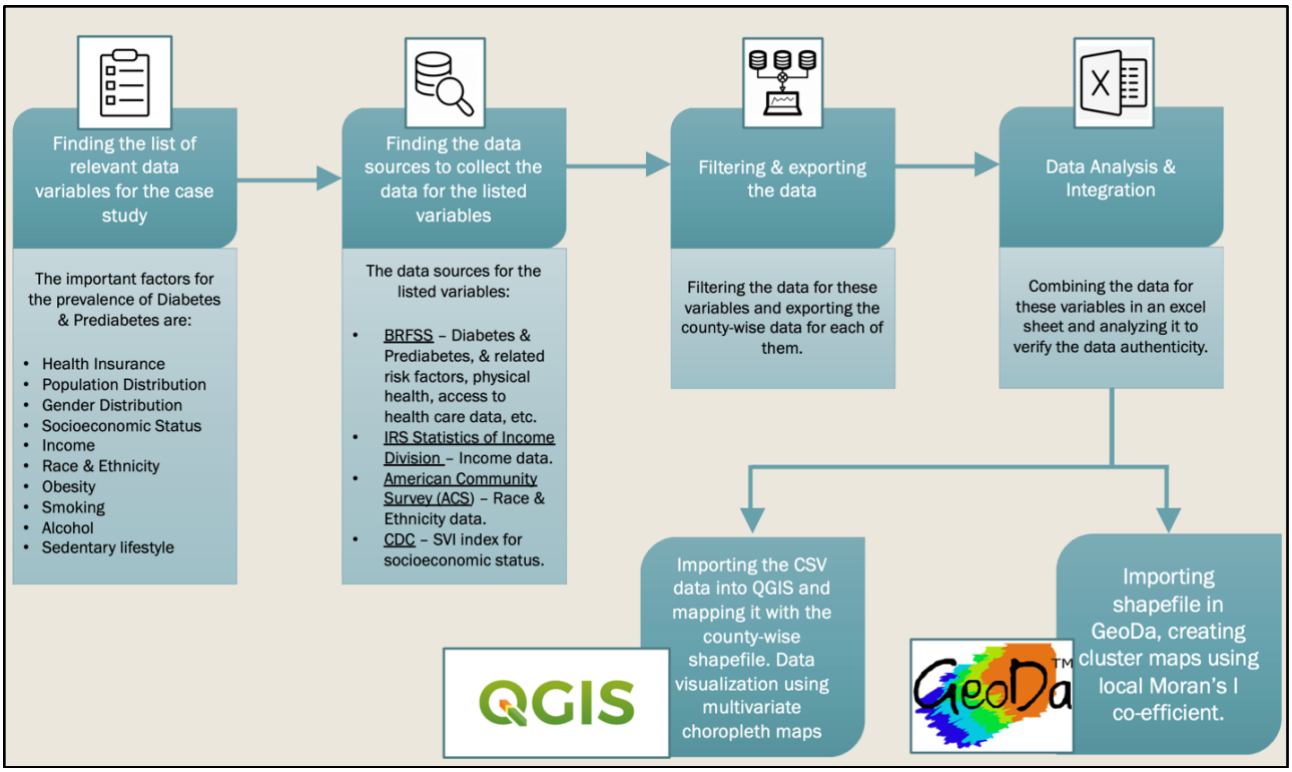

Figure 3.1 Process Workflow

The biggest challenge we ran into was that while diabetes data is widely available, pre-diabetes data is not. The only way to obtain this information is to make contact with CDC personnel from that state and request the data individually. Also, we used BRFSS data for diabetes and pre-diabetes, which has its own set of limitations because it is a cross-sectional, self-report survey that is susceptible to recall bias and social desirability bias, both of which can influence what events respondents recall or describe during the interview.

3.1 GIS Analysis

We used geographic information systems (GIS) software to analyze diabetes-related rates at the county level from web-based sources. Age-adjusted county-wise diabetes rates, population estimates, demographics, socioeconomic status (below poverty, unemployed, income, no high school education), health insurance, lifestyle risk factors, etc. were all included in the data.

We developed bivariate choropleth GIS maps highlighting locations with low, average, and high prevalence of diabetes and pre-diabetes.

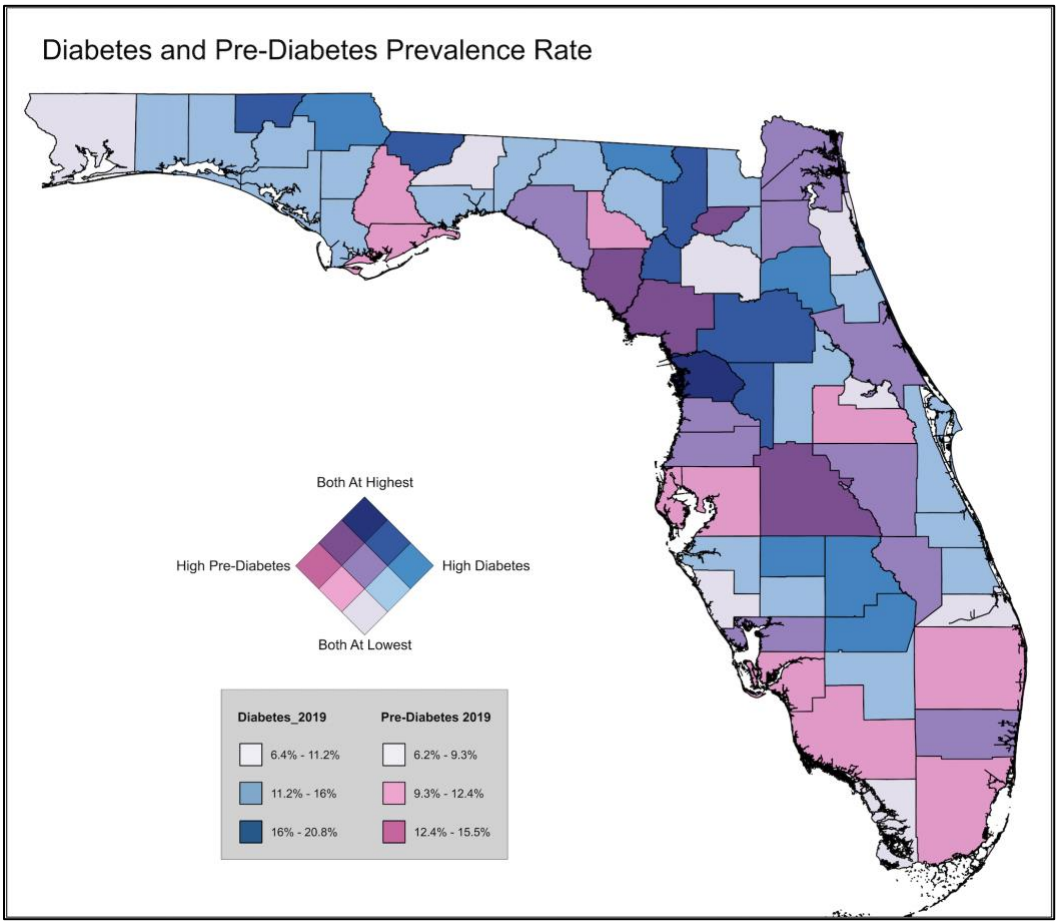

Figure 3.2 Diabetes and Pre-Diabetes Prevalence Rates (2019)

This research is based on diabetes and pre-diabetes statistics from 2019. Using QGIS (Quantum Geographic Information System) software, we multiplied the layers of diabetes and pre-diabetes to create this bivariate map. Bivariate choropleth maps integrate two datasets into a single map, allowing us to illustrate how much two variables exist in each enumeration unit relative to each other. This aids us in determining the difference in prevalence level in each county. According to the map's legend, the diabetes and pre-diabetes rates are divided into 3 equal ranges: low, average, and high. The gray color represents counties with the lowest prevalence of diabetes and

pre-diabetes, while the shades of pink indicate a higher degree of pre-diabetes than diabetes. Similarly, blue shades indicate a higher level of diabetes than pre-diabetes. The greater the rate variation, the darker the shade. The lighter hue of purple represents the average prevalence of diabetes and pre-diabetes. The darker purple tint indicates high pre-diabetes and average diabetes, while the darker blue shade indicates high diabetes and average pre-diabetes. The deepest hue of blue denotes the highest degree of both. Out of 67 counties, there are only eight counties with the lowest incidence of both, nine counties with a somewhat higher prevalence of prediabetes than diabetes, and 11 counties with an average prevalence of both conditions. That suggests that these counties are still at a point where proper care and intervention can help to improve the situation. The number of counties with extremely high rates of the disease is relatively small. Similarly, we have constructed multivariate maps for risk factors and other variables to assist us in comprehending the relationship between them and the disease's prevalence rate. For example, obesity and sedentary population rates are presented in concentric circles on top of the bivariate map of diabetes and pre-diabetes (see figure 4.1). Instead of comparing four different maps for diabetes, pre-diabetes, obesity, and physically inactive rates, we integrated all the variables into a single map, which allows for a more direct comparison.

3.2 Hotspot Analysis and Clustering

Hotspot analysis is a geographical analysis and mapping technique that detects of spatial phenomenon clustering. By grouping points of occurrence into polygons or converging points that are close to one another based on a computed distance, it employs vectors to find areas of statistically significant hot spots and cold spots in your data. The analysis clusters characteristics when similar high or low values are detected in a cluster. Moran's *I* and Geary's C are the statistics

that are designed to reject the null hypothesis of spatial randomness in favor of a clustering alternative. This type of clustering is a feature of the entire spatial pattern and does not indicate where the clusters are located. To address this problem, Anselin (1995) proposed the concept of a local indicator of spatial association (LISA) [17]. A LISA is thought to have two distinct features. First, it delivers a statistic for each site and a relevance rating. Second, it provides a proportional relationship between the total of local data and the global statistic that corresponds. The prevalence of type 2 diabetes is rising worldwide, demanding the identification of such high-risk groups and the development of appropriate primary care interventions.

With row-standardized weights, the total of all weights, $S_0 = \sum_i \sum_j w_{ij}$, matches the number of observations, n, as indicated in the GeoDa documentation. As a result, the Moran's I statistic can be summarized as follows:

$$I = \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

The total of the local statistics is proportional to the global Moran's *I*, or, conversely, the global Moran's *I* correspond to the average of the local statistics (for details, see Anselin 1995) [17]. For the Local Moran, assessing significance is not really beneficial in and of itself. When a significant indication is combined with the location of each observation in the Moran Scatterplot, a compelling interpretation is achievable. The combined data allows the significant sites to be classified as High-High and Low-Low spatial clusters and High-Low and Low-High spatial outliers [10]. It is vital to remember that the references to high and low are relative to the variable's mean and should not be considered absolutes. Using GeoDa software, the Local Moran's *I* correlation coefficient, spatial clusters, and possible 'hotspots' of diabetes and pre-diabetes risk were evaluated on the respective datasets from the year 2019.

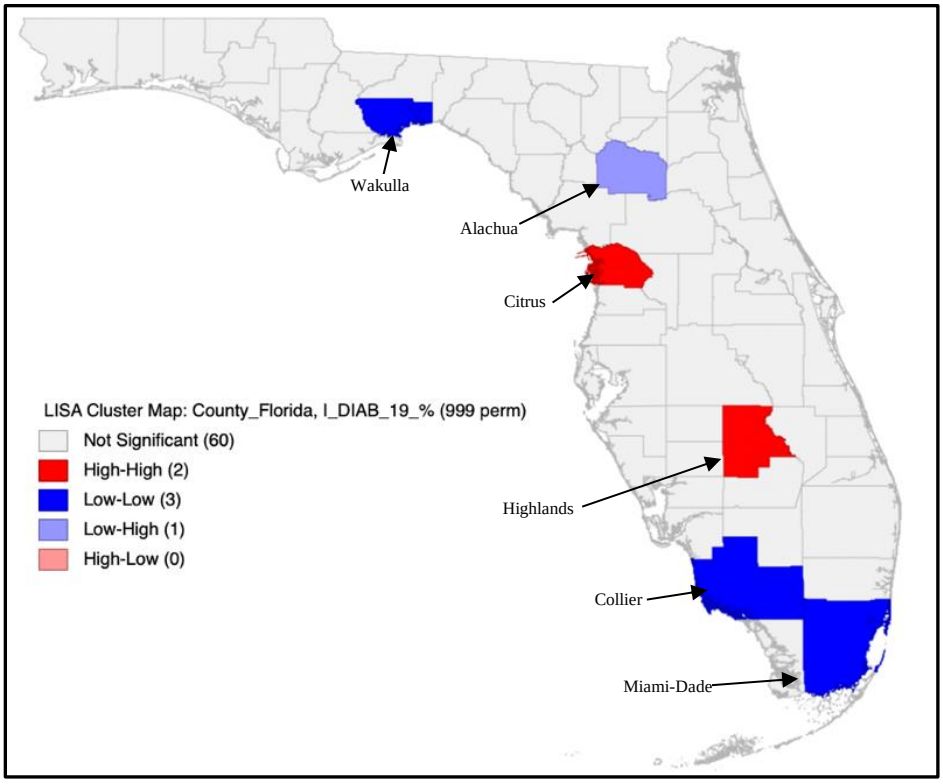

Figure 3.3 Diabetes Cluster

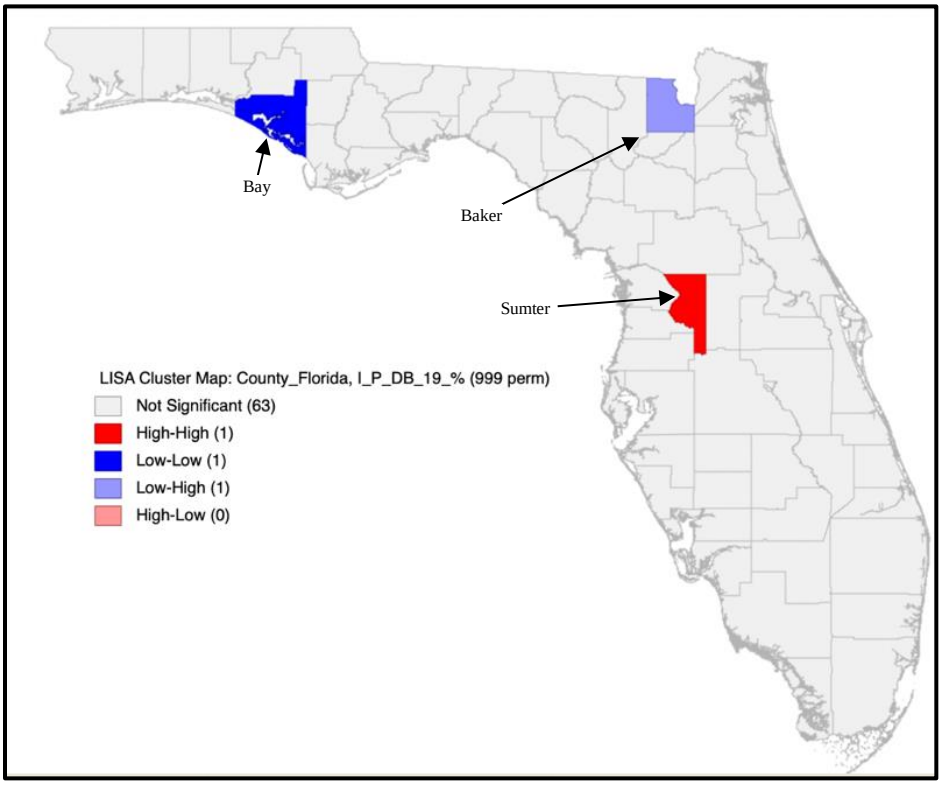

Figure 3.4 Pre-Diabetes Cluster

The map depicts four different cluster types: high-high, high-low, low-high, and low-low. For example, if a region is in the high-low cluster of diabetes, it signifies that the region's diabetes prevalence is higher, and the neighboring counties' diabetes prevalence is low.

Citrus county is in the high-high cluster in the diabetes cluster. As the bivariate map of diabetes and pre-diabetes in figure 3.2 indicates, Citrus is the only county in the state with the highest prevalence of diabetes and pre-diabetes. Sumter and Marion counties, which have the highest levels of diabetes, and Hernando County, which has the average level of diabetes, are its neighbors. Collier and Miami-Dade are in the low-low cluster, diabetes is lowest in those counties and pre-diabetes is also in the average range.

Sumter County is in the high-high prediabetes cluster, Citrus and Polk counties have a high prevalence of pre-diabetes, and Hernando and Pasco counties have an average prevalence of pre-diabetes. Bay County is part of the low-low cluster, including Gulf, Calhoun, Washington, and Walton counties, all of which have low pre-diabetes and average diabetes levels.

4 RESULTS OF MAPPING IMPORTANT FACTORS

Numerous elements contribute to the high prevalence of the disease. The most significant are lifestyle and environmental factors, which we can influence, whereas genetic risk factors are beyond our control. A few of these risk factors are discussed below.

4.1 Lifestyle Factors

Diabetes is becoming more prevalent worldwide, posing a significant socioeconomic and health concern. The number of people diagnosed with diabetes is expected to rise to 642 million by 2040. Around 90% of those with diabetes have type 2 diabetes mellitus (T2DM), which is characterized by high blood sugar levels. Here are a few key lifestyle factors that play a major role in the disease's prevalence [18, 19].

4.1.1 Obesity

Being an overweight (BMI > 24 and <30) or obese (BMI >= 30) increases the risk of high blood glucose (sugar), high blood pressure, and bad cholesterol, all of which increase the chance of life-threatening conditions like type 2 diabetes, heart disease, and stroke, according to the American Diabetes Association (ADA) [20].

4.1.2 Sedentary lifestyle

Sedentary behaviors, such as watching television for long periods or engaging in other activities that require sitting or lying down, can cause metabolic abnormalities and insulin resistance, leading to type 2 diabetes [21, 22].

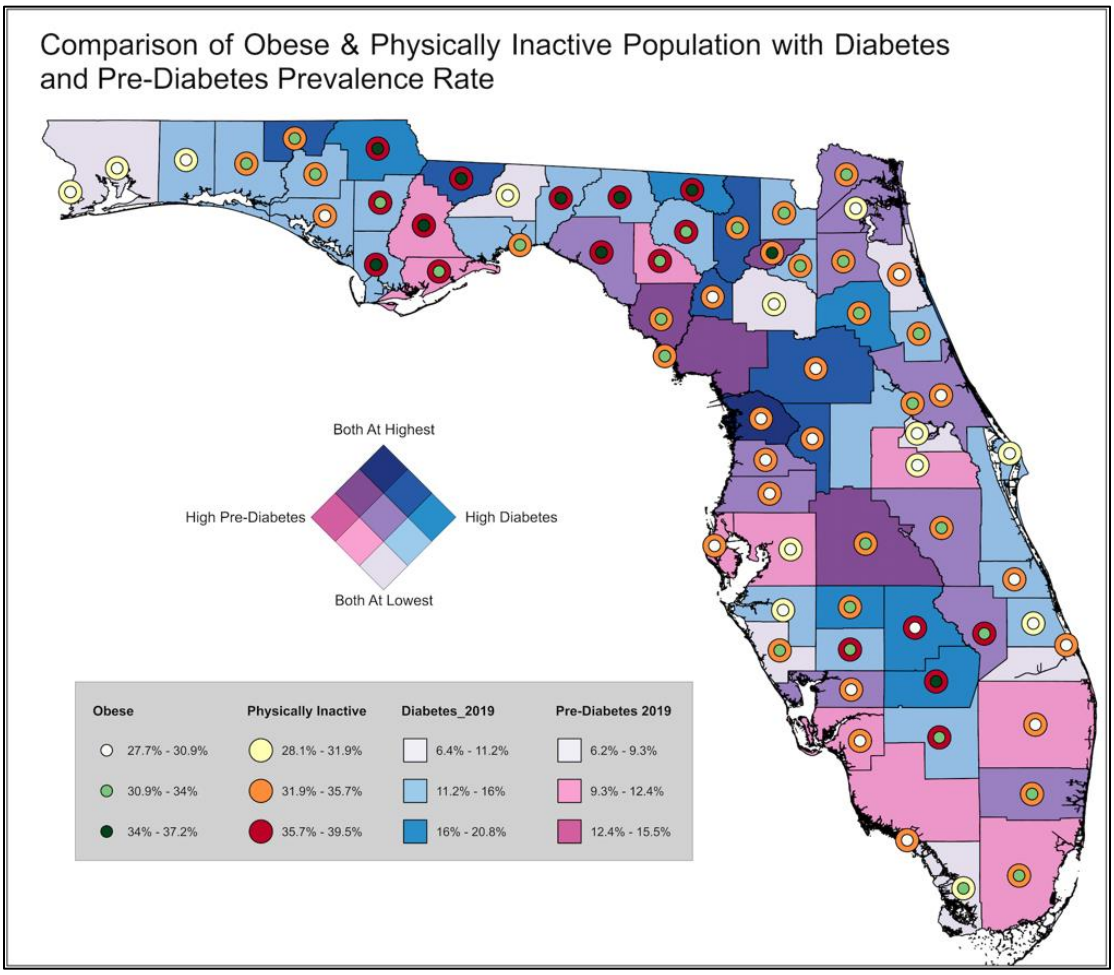

Figure 4.1 Comparison of Obese & Physically Inactive Population

Obesity rates in Florida range from 27.7% to 37.2 %, which we categorized into three equal quantiles: low, medium, and high. Ten counties are in the high range (34% to 37.2%), 28 are in the medium range (30.9% to 34%), and 29 are in the low range (27.7% to 30.9%) of the total 67 counties. This means that the bulk of the counties are in the low to medium risk zone, while nine of the ten counties in the high-risk range are in Florida's northern parts. The prevalence of diabetes and pre-diabetes, on the other hand, varies substantially among those 10 counties. Based on these results, it is safe to assume that it does not directly correspond to disease prevalence rates.

Sedentary behavior, defined as not engaging in any type of physical activity for the past 30 days, accounts for 28.1 % to 39.5 %. These rates are broken down into three equal quantiles: low, medium, and high. There are 17 counties in the high range (35.7% to 39.5%), 37 in the medium range (31.9% to 35.7%), and 13 in the low range (28.1% to 31.9%). For the most part, the rates of physical inactivity and obesity are similar among counties. If one is higher, the other must be as well, and vice versa.

Surprisingly, Citrus County with the highest prevalence of diabetes and pre-diabetes has the lowest obesity rate and a moderate level of a sedentary lifestyle.

4.1.3 Smoking and Alcohol Consumption

Some studies suggest that moderate alcohol consumption (one drink for women of all ages and men 65 and older) can reduce the incidence of diabetes; however, heavy drinking can be life-threatening since it can promote chronic pancreatic inflammation. It can hinder the body's capacity to secrete insulin, resulting in diabetes.

As for tobacco, smoking raises blood sugar levels, resulting in insulin resistance. Heavy smokers, defined as those who smoke more than 20 cigarettes per day, have a significantly increased chance of developing diabetes.

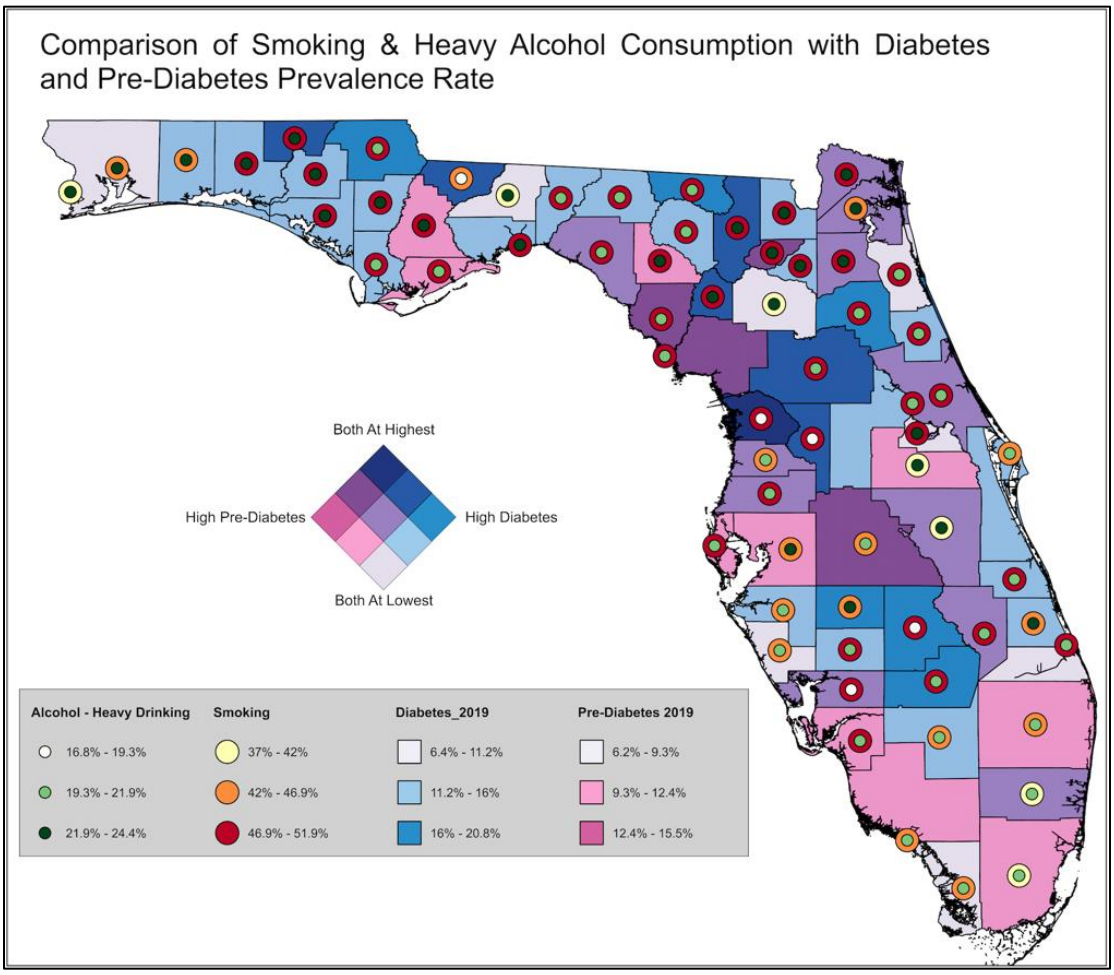

Figure 4.2 Comparison of Alcohol and Smoking Rates

The rate of excessive alcohol use ranges from 16.8 % to 24.4 %, as shown in the map below. There are 27 counties in the high range (21.9 % to 24.4 %), 35 counties in the medium range (19.3 % to 21.9 %), and just five counties in the low range (16.8 % to 19.3 %). However, all these five low-range counties have high diabetes rates, which is incongruent. The majority of the state's high-range counties are located in the north.

The smoking prevalence in the population ranges from 37 % to 51.9 %, and 44 of the 67 counties are in the upper quartile (46.9% to 51.9%), which is cause for concern. There are 16 counties in the moderate range (42% to 46.9%) and seven in the low range

(37% to 42%). Citrus is the only county with the highest prevalence of diabetes and pre-diabetes. Alcohol consumption is low, obesity is low, a sedentary lifestyle is medium, and only smoking level is high in that county, which is extremely contradicting.

Let us compare these lifestyle risk factors to the hotspot and cold-spot cluster counties identified by the Local Moran's *I* coefficient. The table below shows the outcome of combining the results of identifying clusters and the level of prevalence of each lifestyle risk factor. Diabetes cluster counties are represented by the first six records in blue, whereas the last three records in pink represent pre-diabetes cluster counties. As we saw in detail, the risk factor variables are divided into three categories: low, average, and high. Let us take a look at the prevalence of these risk factors in these cluster counties and see how well they correlate.

Table 4.1 Comparison of all risk factors for cluster counties

County Name	Cluster Type	Diabetes Level	Pre-Diabetes Level	Obesity Level	Physical Inactivity Level	Smoking Level	Alcohol Level
Alachua	Low-High	Low	Low	Low	Low	Low	High
Citrus	High-High	High	High	Low	Average	High	Low
Collier	Low-Low	Low	Average	Low	Average	Average	Average
Highlands	High-High	High	Low	Low	High	High	Low
Miami-Dade	Low-Low	Low	Average	Average	Average	Low	Average
Wakulla	Low-Low	Average	Low	Average	Average	High	High
Bay	Low-Low	Average	Low	Low	Average	High	High
Baker	Low-High	Average	Low	Average	Average	High	High
Sumter	High-High	High	Average	Low	Average	High	Low

All the counties in the pre-diabetes clusters have a high smoking rate and an average sedentary lifestyle, as can be seen in the table. However, the rate of excessive drinking of alcohol is high for a county with a low rate of pre-diabetes. The levels of diabetes, pre-diabetes, physical inactivity, smoking, and alcohol are similar in Bay and Baker counties, which are in the low-low and low-high clusters, respectively, except obesity, which is low in Bay and medium in Baker. Even though both counties have low rates of diabetes and pre-diabetes, they have high rates of alcohol and smoking. It is possible that alcohol and smoking do not directly correlate with pre-diabetes. All the risk factors for diabetes cluster counties are at varying levels, and there is no predictable pattern. Although, obesity is the risk factor that is regarded as the most important and yet none of these cluster counties have a high prevalence of obesity. This may also be the case of a data discrepancy.

4.2 Other Impacting Factors

The distribution of age and gender, race and ethnicity, the population with health insurance, the average median income level, and other essential elements are important to consider.

4.2.1 Age/ Gender

As we already know, diabetes causes the body's cells to become less receptive to insulin, which is responsible for controlling blood sugar levels. As a result, it can harm the heart, kidneys, nerves, and eyes. It can impact men and women differently because it has extensive effects throughout the body. There were more men than women among those diagnosed with diabetes, and the majority of those diagnosed were middle-aged or older.

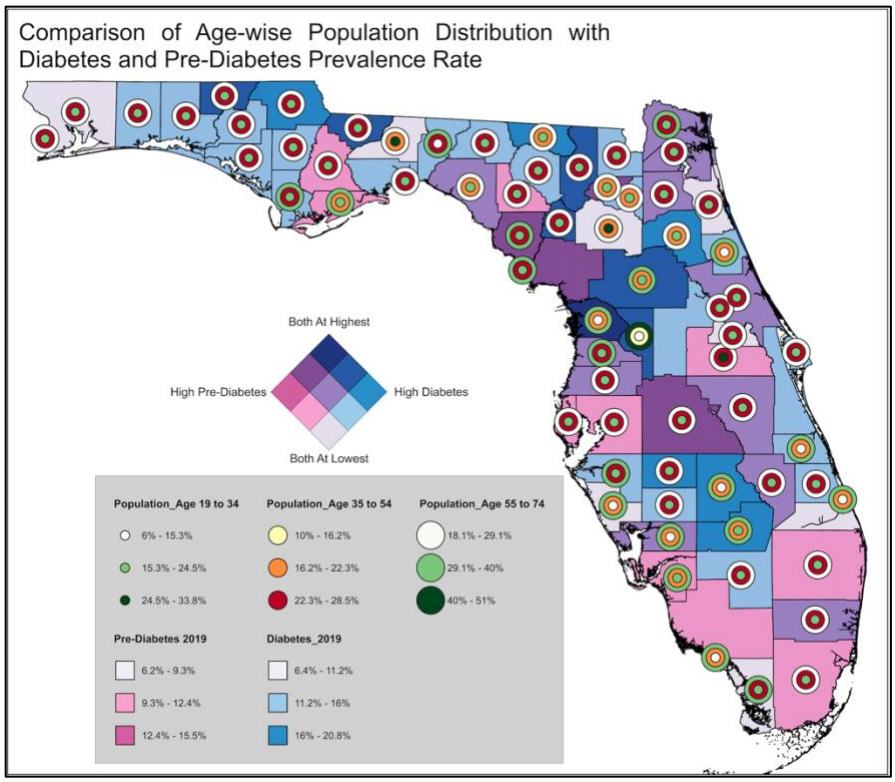

Figure 4.3 Age-wise Population Distribution

The population distribution chart for three age groups is shown above: 19 to 34 years old, 35 to 54 years old, and 55 to 74 years old. The highest level of population group 35 to 54 years is scattered throughout the state, as seen on the map. Sumter County is the only county in the state having 51% of the population between the ages of 55 and 74, and it has the highest rate of diabetes and the average rate of pre-diabetes. The population distribution is identical in 35 of the 67 counties: the lowest population is between 55 and 74 years old, the maximum population is between 35 and 54 years old, and the average population is between 19 and 34 years old. Diabetes and pre-diabetes prevalence differ significantly across these 35 counties.

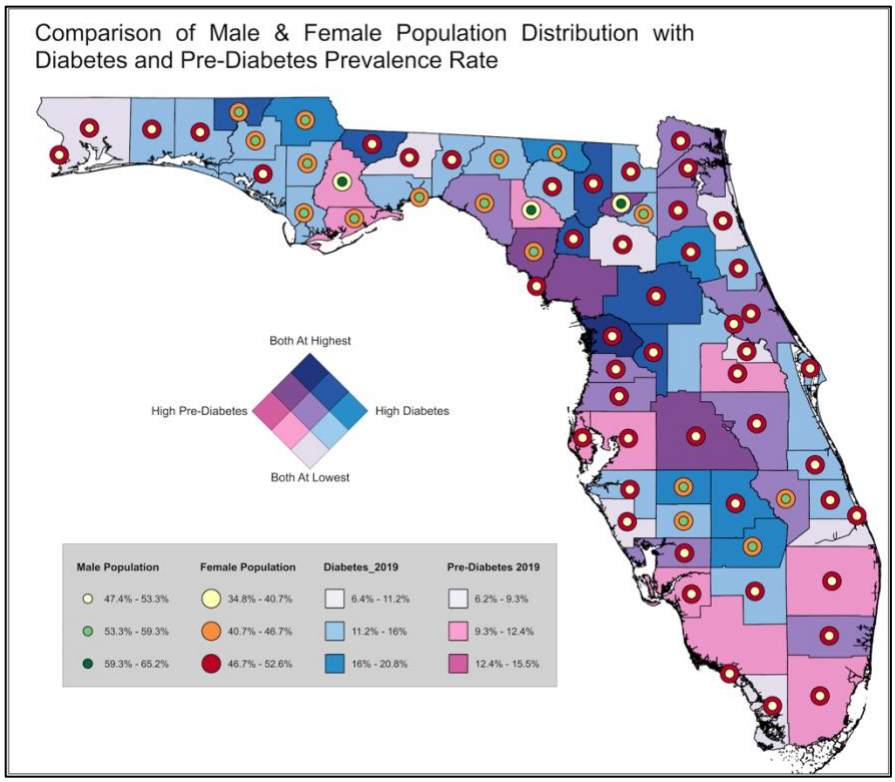

Figure 4.4 Gender-wise Population Distribution

As can be seen in the graph above, the gender distribution is nearly equal in most counties. Only three counties had male populations of more than 60% to 65%, and all those counties have more pre-diabetes than diabetes, but we cannot attribute this just to gender distribution.

4.2.2 Race/ Ethnicity

The age-adjusted prevalence of diabetes showed, the non-Hispanic whites (4.1%) having the lowest prevalence and Asian Indians having the highest, according to a study to evaluate the risk of diabetes by race and ethnicity (11.1%) [23]. In multivariable models, being born outside of the United States was associated with an increased risk of diabetes in black, Asian Indian, Filipina, Pacific Islander, Chinese, Mexican, and non-Hispanic white women [23]. Let us examine the distribution of races and ethnicities in Florida.

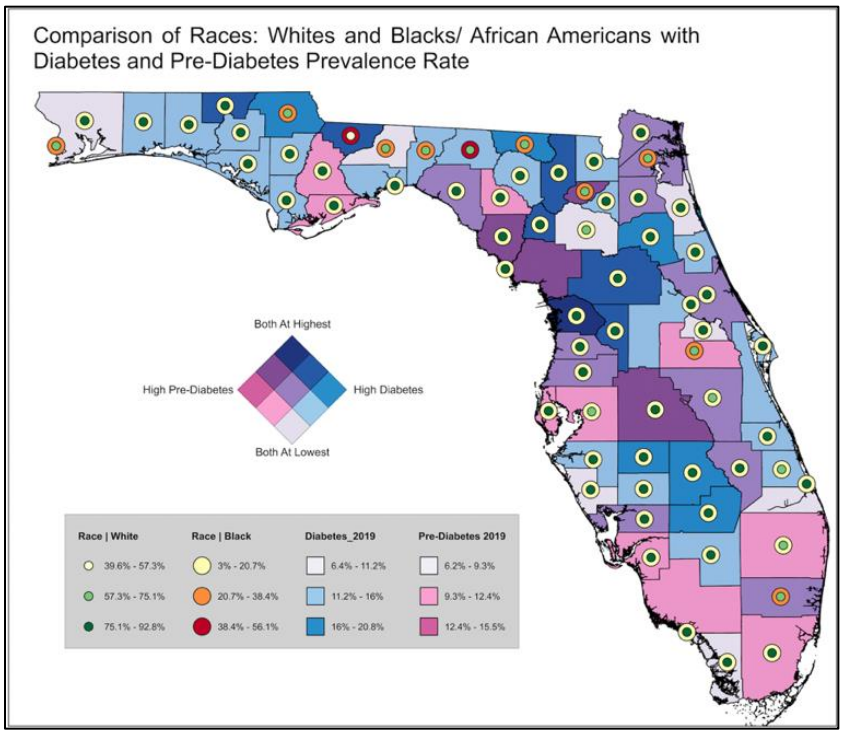

Figure 4.5 Racial Distribution (White and African Americans)

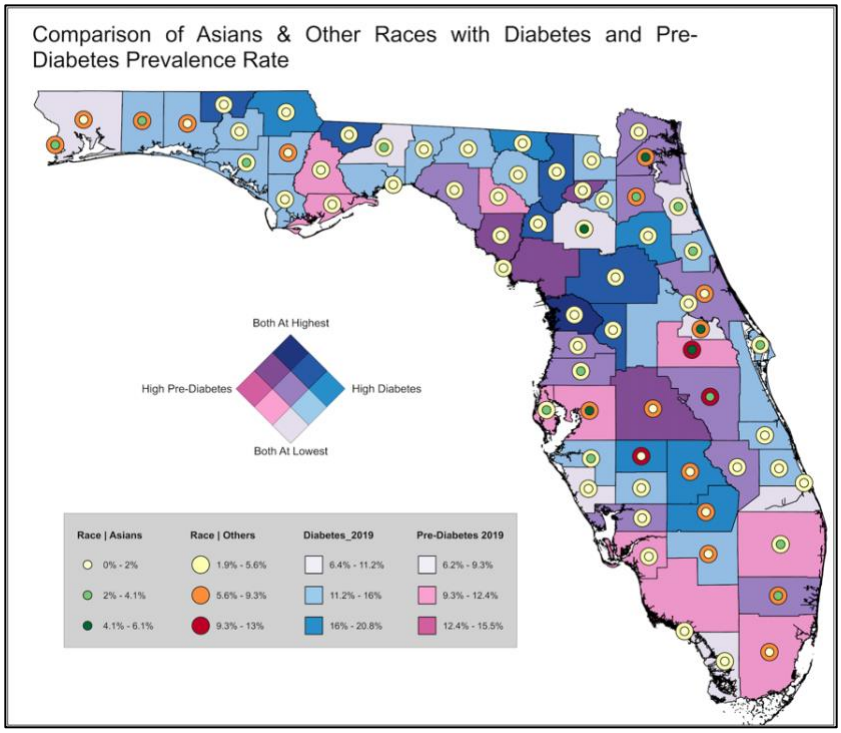

Figure 4.6 Racial Distribution (Asians and Other Races)

According to the maps, 51 of the 67 counties have a more than 75 % white population. In these 51 counties, the black population is less than 20%, Asians less than 2%, and other races less than 5.6 %. Only two counties have black populations ranging from 38 % to 56 %. The Asian and other races population in Florida is less than 19%.

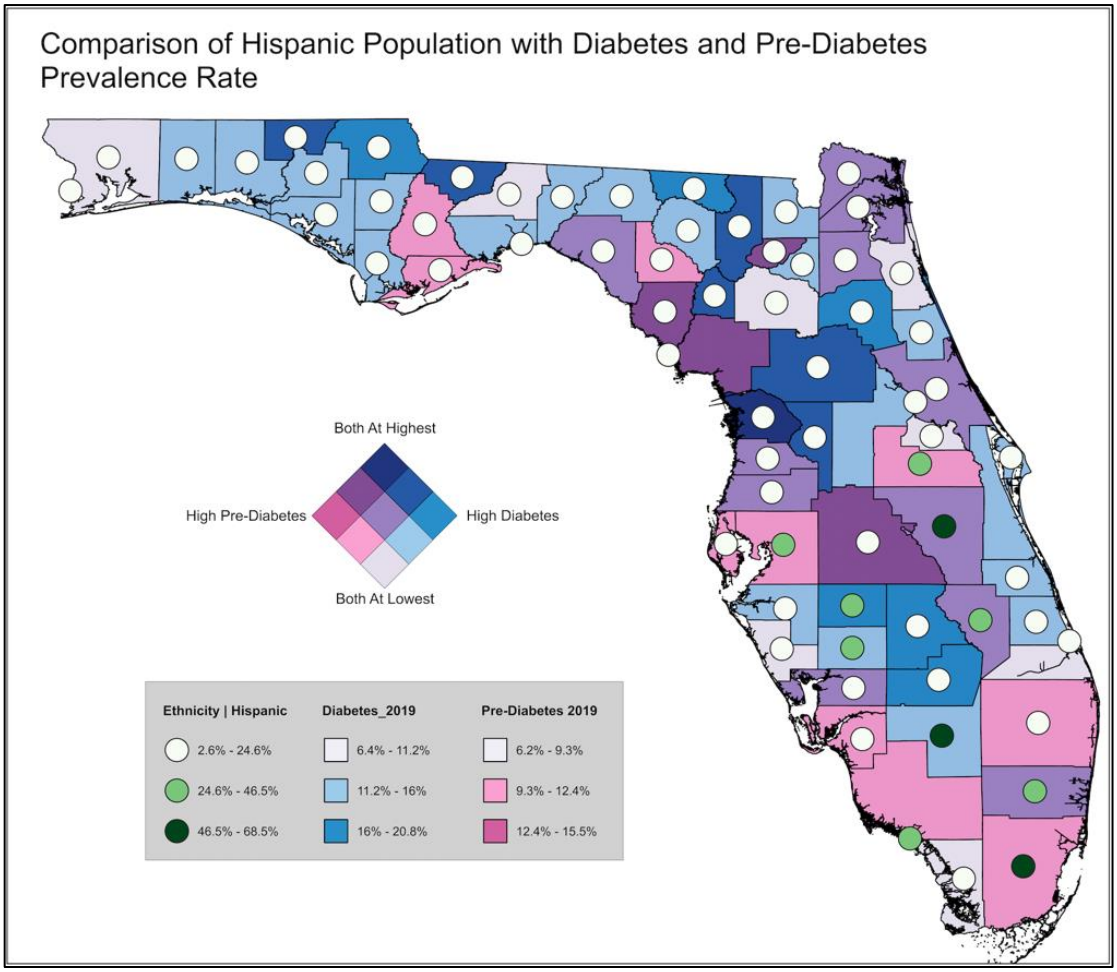

Figure 4.7 Hispanic Population Distribution

The majority of counties have a Hispanic population of less than 25%. However, three counties have a Hispanic population of 46.5% to 68.5% and all three counties have a different prevalence level of diabetes and pre-diabetes.

4.2.3 Health Insurance

Diabetes is one of the diseases that has the most significant impact on the health of Americans, as evidenced by virtually every national metric. Among the more than 200 illnesses studied by the BCBS Health Index, diabetes ranks third in terms of national health impact on quality of life and third in terms of per member cost for the commercially insured population. The impact of diabetes continues to grow, with the largest growth occurring in the 18-34 age group, which also happens to be the age group with the highest obesity rates. Obesity is a major contributor to the onset of diabetes [24].

Table 4.2 Age-wise health insured population

Under 6 Years	**6 to 18 Years**	**19 to 25 Years**	**26 to 34 Years**	**35 to 44 Years**
94.24%	91.55%	74.96%	75.02%	78.80%
45 to 54 Years	**55 to 64 Years**	**65 to 74 Years**	**Above 75 Year**	**Overall Average %**
80.93%	86.76%	98.89%	99.44%	91.51%

The rate of health insured population falls dramatically between the age groups of 6 to 18 years and 19 to 25 years, as seen in the table. Following that, it gradually increases. After the age of 65, the health-insurance population reaches 98 % to 99 % courtesy of Medicare and Medicaid [25].

4.2.4 Income

Florida's median family income is expected to be 57,435 dollars in 2020. This is somewhat lower compared to the previous year ($58,368). The median household income is a standard metric for determining the affluence of a region. Many statisticians believe

that the median income is a better predictor than the average household income since it is unaffected by extremely high or low values.

Anyone with a higher income will almost certainly have better health insurance coverage. On that basis, health insurance coverage should be proportional to the median income level in every given county. For this assumption, let us look at the map presented below.

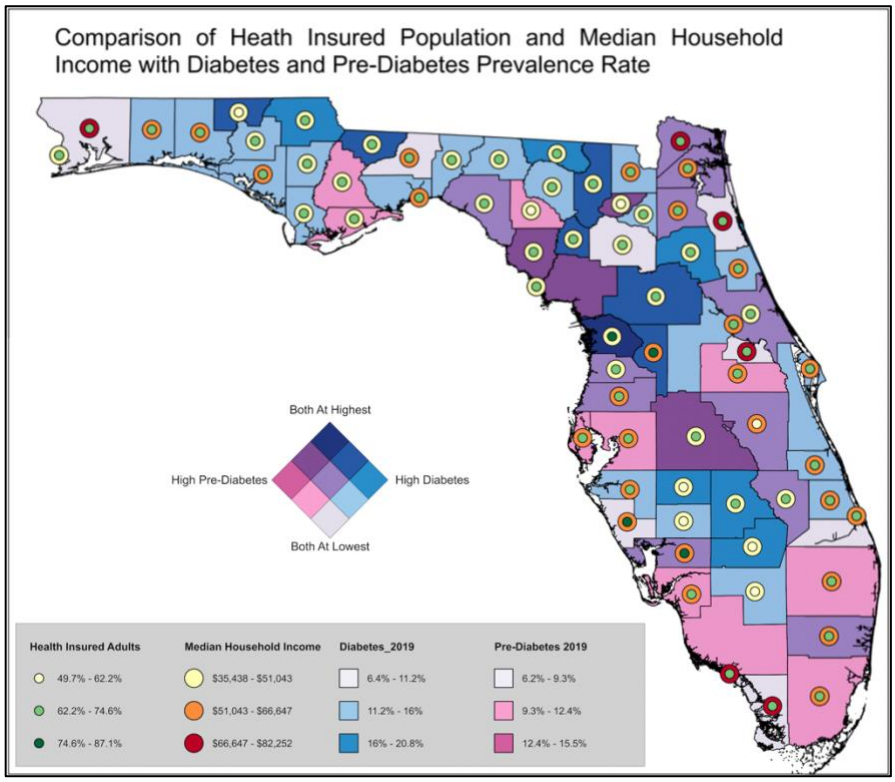

Figure 4.8 Health Insurance & Median Household Income

The correlation between health insurance and median household income in any county is at the same level or slightly above or below it. For example, if the county's median income is low, health insurance coverage is likely to be low or average. There are no counties where both variables are at the opposite level except Citrus. As an example, suppose you have a high median income but low health insurance coverage. Citrus is the

only county in the state with a low median income and a high rate of health insurance coverage, yet it has the highest rates of diabetes and pre-diabetes which is quite surprising.

4.2.5 Social Vulnerability Index (SVI) for Socioeconomic Status

The social vulnerability index for socioeconomic status includes the percentage of population below poverty, unemployed, income, and no high school diploma. It is divided into four categories of vulnerability levels: very low, low, medium, and high.

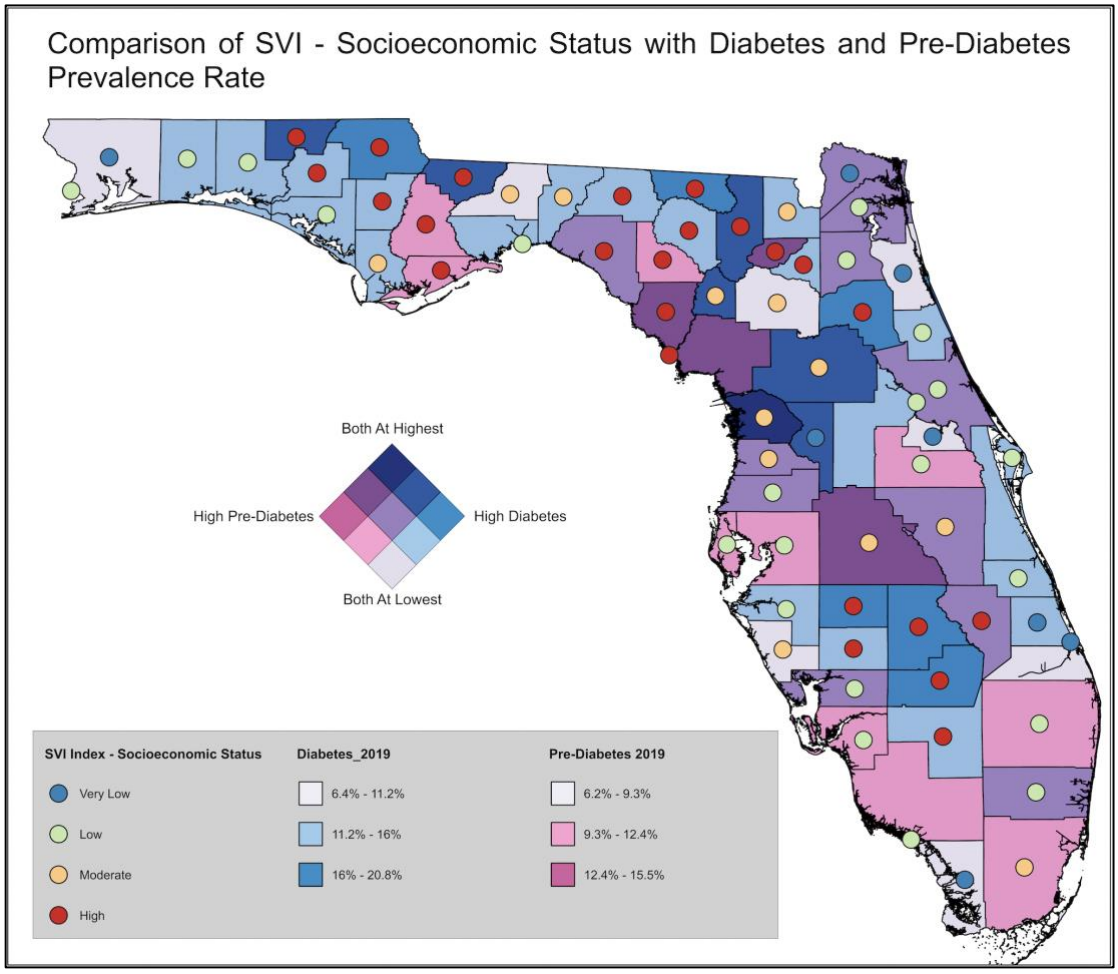

Figure 4.9 Socioeconomic Vulnerability Status

The majority of the high vulnerability counties are in the northern parts of Florida, as seen on the map. The rest of the high vulnerability counties are clustered together in the southern part which has a higher diabetes prevalence. Diabetes and pre-diabetes are very low in five of the eight counties in the very low vulnerability category. All of the counties in the low vulnerability category (total of 22 counties) have average pre-diabetes (light pink shade – 6 counties), average diabetes (light blue shade – 10 counties), or the average of both (light purple shade – 6 counties), and they are all on the state's boundaries. There are 13 counties in the moderate vulnerability range and the prevalence level of the disease is different for all of them. For most regions, there is a significant correlation between disease prevalence and socioeconomic status [26].

5 DISCUSSION

This paradigm can be applied to a variety of case studies involving the comparison of different variables to derive conclusions. This is one method for comparing many variables in a single map, and it can be applied to any case study that requires more than one univariate maps. Comparing variables in a single map is easier than comparing them in multiple modalities. We can combine up to 10 components in concentric circles to acquire a deeper understanding of a limited area. We had roughly 14 variables to analyze in our case study to determine if any of them corresponded to the disease prevalence rate and comparing the single maps of each variable made drawing any conclusions very challenging. This method of comparison simplified the research process, and we found a few discernible patterns. Also, creating a multivariate map entail considering color combinations that allow everyone, even colorblind people, to see the analyzed data. There is a plethora of software and tools available for undertaking spatial data analysis. However, this framework was built entirely with open-source software, such as QGIS and GeoDa. The data utilized in this study is also publicly available.

According to the information gleaned using this framework, most counties have low and average obesity rates, while nine out of ten counties in the high category are in the northern parts of Florida. The rates of physical inactivity and obesity are comparable across most counties. When one is higher, the other must be as well, and vice versa. When it comes to alcohol, there are 27 counties in the high range, 35 counties in the medium range, and only five counties in the low range. However, diabetes rates are high in all five low-range counties, which is incompatible. The prevalence of smoking in the population spans from 37% to 51.9 %, with 44 of the 67 counties falling into the upper quartile (46.9% to 51.9 %), which is worrisome. In 35 of the 67 counties, the population distribution is the same: the lowest population is 55 to 74 years old; the highest

population is 35 to 54 years old, and the average population is 19 to 34 years old. In most counties, the gender distribution is roughly equal. Only three counties have men more than 60% to 65%, and all those counties have more pre-diabetes than diabetes, but we cannot pin this only on gender. More than 75% of the people in 51 of the 67 counties are white. The black population in these 51 counties is less than 20%, Asians less than 2%, and other races less than 5.6 %. Most counties have less than 25% Hispanic residents. The association between health insurance and median household income is the same in every county, or slightly higher or lower. Citrus is the state's only county with a low median income and a high proportion of health insurance coverage, yet it has the highest rates of diabetes and pre-diabetes, which is surprising. The SVI for socioeconomic status and rates of disease prevalence rate for most counties are highly correlated.

The analytical approach presented certain challenges. After collecting all the data, the biggest challenge we had was figuring out how to use it effectively so that we could get the best results. We began by making univariate maps for each variable, but this did not lead to the discovery of any patterns. We knew we needed to compare each variable's data for each county to determine whether there was a pattern that corresponded to disease prevalence, but the challenge was how to accomplish it. Many considerations arose, like which approach to pursue, which tools to use, how much time to spend on analysis, and so on.

We studied the counties in only one state, Florida, to determine whether there were any clusters. We can do the same process for an entire country and then compare different countries to identify patterns and clusters on a global scale using this framework. It can be used to assess other ailments and its factors or different levels of educational attainment at the county level. It can also be used for case studies such as crime mapping and analysis, and we can use zip-code level data for different types of crimes to compare which places have high crime rates for specific crimes.

6 CONCLUSION

Diabetes is quickly becoming one of the most pressing public health issues of our day. One of the most important public health breakthroughs in the study of diabetes is the demonstration that it can be prevented by implementing effective interventions that target the factors that enhance diabetes risk. Many adults are pre-diabetic yet are not identified until it is too late. It could be due to a lack of awareness or a disregard for regular checkups, for example. It is critical to identify this vulnerable demographic, which is already pre-diabetic and at high risk. Identifying the places where the danger is concentrated will assist us in better allocating resources. The value of spatial analysis could be fundamental in this task by identifying hotspots of diabetes across high-risk communities [27]. Different counties may require different levels of attention and resources. While some counties have high smoking and alcohol consumption rates, others have high obesity and sedentary behavior rates. It is critical to pay attention to the type of risk these counties pose and to focus on allocating resources appropriately.

In this study, we developed a novel approach to data comparison and analysis. We could not find any additional GIS (Geographic Information Science)-related diabetes studies that used bivariate and multivariate maps for data comparison. This framework aids in the identification of some vulnerable regions. To battle this disease, we cannot have the same approach or remedy for every county. Each region is unique, and each should have its strategy for dealing with them. Perhaps, depending on the results of this modal, we can create a new modal that splits all these counties into distinct groups and distributes resources according to these groups. This approach can be used to obtain conclusions from a number of case studies involving the comparison of distinct variables. We can perform the same process for an entire country and even compare countries to find patterns and clusters on a global scale using this approach. It may also be utilized

for case studies such as crime mapping and analysis or used at the county level to assess various degrees of educational attainment.

We can obtain data on pre-diabetes and diabetes from the Centers for Disease Control and Prevention (CDC) for all states in the future as an extension of this research and utilize this framework to verify prevalence at the national level. After that, we can move on to a global level comparison. We can also gather and analyze data from different years to see how the prevalence rate has changed over time, and therefore generate a time-series map. Because the process involves a lot of manual work and analysis, we can work on automating at least part of the procedures, if not the entire framework, to improve its efficiency.

REFERENCES

1. Cuadros, D. F., Li, J., Musuka, G., & Awad, S. F. (2021). Spatial epidemiology of diabetes: Methods and insights. World journal of diabetes, 12(7), 1042–1056. https://doi.org/10.4239/wjd.v12.i7.1042

2. Arun Nanditha, Ronald C.W. Ma, Ambady Ramachandran, Chamukuttan Snehalatha, Juliana C.N. Chan, Kee Seng Chia, Jonathan E. Shaw, Paul Z. Zimmet; Diabetes in Asia and the Pacific: Implications for the Global Epidemic. Diabetes Care 1 March 2016; 39 (3): 472–485. https://doi.org/10.2337/dc15-1536

3. Ruta, L. M., Magliano, D. J., LeMesurier, R., Taylor, H. R., Zimmet, P. Z., Shaw, J. E.; Prevalence of diabetic retinopathy in Type 2 diabetes in developing and developed countries. Diabet. Med. 30, 387– 398 (2013). https://doi.org/10.1111/dme.12119

4. Ping Zhang, Xinzhi Zhang, Jonathan Brown, Dorte Vistisen, Richard Sicree, Jonathan Shaw, Gregory Nichols, Global healthcare expenditure on diabetes for 2010 and 2030, Diabetes Research and Clinical Practice, Volume 87, Issue 3, 2010, Pages 293-301, ISSN 0168-8227, https://doi.org/10.1016/j.diabres.2010.01.026. (https://www.sciencedirect.com/science/article/pii/S0168822710000495)

5. Hilary King, Ronald E Aubert, William H Herman; Global Burden of Diabetes, 1995–2025: Prevalence, numerical estimates, and projections. Diabetes Care 1 September 1998; 21 (9): 1414–1431. https://doi.org/10.2337/diacare.21.9.1414

6. CDC - The Facts, Stats, and Impacts of Diabetes. https://www.cdc.gov/diabetes/data/statistics-report/index.html

7. Curtis, A. B., Kothari, C., Paul, R., & Connors, E. (2013). Using GIS and Secondary Data to Target Diabetes-Related Public Health Efforts. Public Health Reports, 128(3), 212–220. https://doi.org/10.1177/003335491312800311

8. Hipp, J. A., & Chalise, N. (2015). Spatial analysis and correlates of county-level diabetes prevalence, 2009-2010. Preventing chronic disease, 12, E08. https://doi.org/10.5888/pcd12.140404

9. Haining, R., Wise, S. and Ma, J. (1998), Exploratory Spatial Data Analysis. Journal of the Royal Statistical Society: Series D (The Statistician), 47: 457-469. https://doi.org/10.1111/1467-9884.00147

10. Chan, T. C., & King, C. C. (2010). Surveillance and Epidemiology of Infectious Diseases using Spatial and Temporal Lustering Methods. Infectious Disease Informatics and Biosurveillance: Research, Systems and Case Studies, 27, 207–234. https://doi.org/10.1007/978-1-4419-6892-0_10

11. Wise, S., Haining, R., & Signoretta, P. (1999). Scientific Visualisation and the Exploratory Analysis of Area Data. Environment and Planning A: Economy and Space, 31(10), 1825–1838. https://doi.org/10.1068/a311825

12. Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. Environmental health perspectives, 112(9), 998–1006. https://doi.org/10.1289/ehp.6735

13. Lord, J., Roberson, S. & Odoi, A. Investigation of geographic disparities of pre-diabetes and diabetes in Florida. BMC Public Health 20, 1226 (2020). https://doi.org/10.1186/s12889-020-09311-2

14. Dendup, T., Feng, X., Clingan, S., & Astell-Burt, T. (2018). Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. International journal of environmental research and public health, 15(1), 78. https://doi.org/10.3390/ijerph15010078

15. Tony H. Grubesic, Jennifer A. Miller, Alan T. Murray, Geospatial and geodemographic insights for diabetes in the United States, Applied Geography, Volume 55, 2014, Pages 117-126, ISSN 0143-6228, https://doi.org/10.1016/j.apgeog.2014.08.017.

16. Mansourian, M., Yazdani, A., Faghihimani, E. et al. Factors associated with progression to pre-diabetes: a recurrent events analysis. Eat Weight Disord 25, 135–141 (2020). https://doi.org/10.1007/s40519-018-0529-7

17. Anselin, L. (1995), Local Indicators of Spatial Association—LISA. Geographical Analysis, 27: 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

18. Zimmet, P.Z. Diabetes and its drivers: the largest epidemic in human history? Clin Diabetes Endocrinol 3, 1 (2017). https://doi.org/10.1186/s40842-016-0039-3

19. Paul Z Zimmet, Dianna J Magliano, William H Herman, Jonathan E Shaw, Diabetes: a 21st century challenge, The Lancet Diabetes & Endocrinology,Volume 2, Issue 1, 2014, Pages 56-64, ISSN 2213-8587, https://doi.org/10.1016/S2213-8587(13)70112-8. (https://www.sciencedirect.com/science/article/pii/S2213858713701128)

20. American Diabetes Association. https://www.diabetes.org/diabetes

21. Hu. (2003). Sedentary lifestyle and risk of obesity and type 2 diabetes. Lipids., 38(2), 103–108. https://doi.org/10.1007/s11745-003-1038-4

22. (2013). Diabetes Mellitus in Developing Countries and Case Series. In (Ed.), Diabetes Mellitus - Insights and Perspectives. IntechOpen. https://doi.org/10.5772/50658

23. Hedderson, M. M., Darbinian, J. A., & Ferrara, A. (2010). Disparities in the risk of gestational diabetes by race-ethnicity and country of birth. Paediatric and perinatal epidemiology, 24(5), 441–448. https://doi.org/10.1111/j.1365-3016.2010.01140.x

24. BCBS Health Index - https://www.bcbs.com/the-health-of-america/reports/diabetes-and-the-commercially-insured-us-population

25. Harris MI, Cowie CC, Eastman R. Health-insurance coverage for adults with diabetes in the U.S. population. Diabetes Care. 1994 Jun;17(6):585-91. doi: 10.2337/diacare.17.6.585. PMID: 8082529.

26. Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Staines, A., & Elliott, P. (2000). Statistical issues in the analysis of disease mapping data. Statistics in medicine, 19(17-18), 2493–2519. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2493::aid-sim584>3.0.co;2-d

27. Grubesic, Tony & Murray, Alan. (2001). Detecting Hot Spots Using Cluster Analysis and GIS.