THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 14 Issue 4

RESEARCH ARTICLE

Using Geographically Weighted Random Forests to Analyze County-level Diabetes Prevalence in the USA

Edmund T. Ampofo [1], Erich Seamon [2,4], Benjamin J. Ridenhour [1,3]

[1] Bioinformatics and Computational Biology Program, University of Idaho, Moscow, Idaho, USA
[2] Department of Design and Environments, University of Idaho, Moscow, Idaho, USA
[3] Department of Mathematics and Statistical Science, University of Idaho, Moscow, Idaho, USA
[4] Department of Environmental Sciences, Baylor University, Waco, TX, USA.
*These authors contributed equally to this work.

OPEN ACCESS

PUBLISHED
30 April 2026

CITATION
Ampofo, ET., Seamon, E., et al., 2026. Using Geographically Weighted Random Forests to Analyze County-level Diabetes Prevalence in the USA. Medical Research Archives, [online] 14(4).

ISSN
2375-1924

ABSTRACT

Diabetes poses a major public health challenge in the United States, ranking among the top ten leading causes of death. Its prevalence is closely tied to factors such as obesity and lifestyle behaviors, yet it varies significantly across different geographic regions. Traditional regression models often fail to capture the entirety of the relationships between dependent and independent variables, especially when spatial heterogeneity is present. To understand county-level diabetes prevalence and its associated risk factors, researchers have employed spatial linear regression models, which have limitations, including the assumption of linear relationships and inadequate handling of multicollinearity. To address this, a geographically weighted random forest model (GW-RF), which combines random forests and locally weighted regressions via a spatially weighted matrix, is employed as an exploratory and predictive tool in this study. County-level diabetes prevalence data for the USA, along with twelve other independent variables, from 2010 to 2020 were divided into two time periods: pre- and post-National Health Information Survey updates (referred to as "historical" and "current" periods, respectively). These data were then used to explore the nature and pattern of county-level diabetes prevalence and to estimate the performance of GW-RF against other global and spatially weighted models. In this study, we found that all geographically weighted models outperformed their non-spatial counterparts across periods, indicating that spatial variation plays an important role in explaining county-level diabetes prevalence. Our results further indicate that the GW-RF model more effectively captures spatial heterogeneity and predicts diabetes prevalence than both global and local models. Compared to global ordinary least squares regression, global random forests, and geographically weighted ordinary least squares, our GW-RF model achieved higher R^2 values by 3.5%, 1.1%, and 0.6% (historic), and 2.3%, 0.5%, and 0.4% (current), as well as lower normalized root mean squared error values by 6.1%, 2%, and 1% (historic), and 0.8%, 0.3%, and 0.2% (current), respectively. We also found that, although models generally performed well, their performance dropped in the current period. This decline in model performance may be because the current period showed less spatial autocorrelation in diabetes prevalence (historical Moran's I: 0.559, $p < 0.001$; current Moran's I: 0.45, $p < 0.001$). This shift in the underlying spatial patterns of diabetes could reflect known changes in survey methodology or actual epidemiological changes, both of which warrant further investigation. The findings also suggest that the GW-RF model can support health professionals and policymakers in making accurate projections, detecting emerging hotspots, and guiding targeted prevention and control efforts.

Introduction

Diabetes is a significant health problem in the United States. There are two types of diabetes: type I, where the body produces no insulin, and type II, where the body does not properly use the insulin produced by the pancreas[1]. In the USA, type II diabetes is common and is associated with obesity and lifestyle choices[2-6]. Diabetes ranks as the eighth leading cause of death in the United States based on a staggering number of 103,294 death certificates in which diabetes was recorded as the underlying cause. In 2021, this increased to 399,401 certificates[1]. Diabetes not only leads to serious health complications and loss of human capital, but also adds a great cost burden to the country. Diabetes costs the US government a whopping $412.9 billion, including $306.6 billion in direct medical costs and $106.3 billion of indirect expenses[7]. Looking at how dangerous diabetes is to the health and economy, there is a need to study it extensively and provide fresh and updated information to the public and health professionals to help fight against it.

There is currently no cure for diabetes and the only medical assistance one can get is management[8]. As a result, it is very crucial to identify diabetes in its early stages to help prevent further complications[9]. According to the American Diabetes Association, about 38.4 million Americans have diabetes as of 2021. This includes 2 million people with type I diabetes with about 304,000 children and adolescents. The percentage of Americans age 65 and older remains high, at 29.2% or 16.5 million, this includes both diagnosed and undiagnosed. American Diabetes Association also states that about 1.2 million diabetes cases are recorded each year. What makes this scarier is that, as of 2021, an estimated 97.6 million adults 18 years or older who have prediabetes according to the US CDC[10], which is nearly 1 in 3. This gets more alarming as it is known from research that 81% of people who have this condition are unaware of their condition. Prediabetes is defined as blood glucose higher than the normal but lower than the known threshold to be diagnosed as diabetes.

Given that individuals with prediabetes are at a higher risk of developing type II diabetes, chronic kidney diseases, and cardiovascular disease[2], it is very crucial to identify these individuals and diagnose prediabetes before they add to the large pool of an existing diabetic population. The National Library of Medicine suggests that close surveillance of the prevalence of prediabetes is critical to projecting the future burden of diabetes and knowing the resources that will be required to combat diabetes[11].

Diabetes, just like any other disease, has known risk factors. A study done by Sulaskan et al., 2024[4], on the trends and disparities in diabetes prevalence in the United States from 2012 to 2022 provided very insightful information on the impact of some known or common diabetes risk factors. According to Sulaskan, age-standardized diabetes prevalence significantly increased by 18.6% between 2012 and 2022, there were also some disparities in the increase of diabetes for racial groups with non-Hispanic blacks accounting for 15.8%. Males had a higher prevalence than females suggesting that gender plays a role in diabetes prevalence and disparities. Also, physically inactive individuals and individuals with low income had higher prevalence compared to more active and high-income earners respectively. Obesity emerged as the most prevalent condition, showing a 19.23% increase over the period.

Diabetes in the USA is known to exhibit geographical disparities and according to Sulaskan et al., 2024[5] this has not changed. From 2012–2022, the South/Midwest experienced an out-sized increase from 9.2% to 12.8% which implies that although diabetes increased nationally during this period it affected some sociodemographic groups more than others. States like Arkansas, Kentucky, and Nebraska reported the highest increase, these three states add up to seven other states that also experienced high increases in diabetes that deserve a mention: Texas, Alabama, Minnesota, Illinois, West Virginia, Delaware and Massachusetts.

A geographically weighted random forest (GW-RF) is a tree-based non-parametric ensemble method[6]. It fits a local version of the traditional random forest algorithm to help explore spatial non-stationarity in the relationship between a dependent variable and a set of independent variables by taking into account only a set of neighboring observations[12]. A GW-RF has higher predictive performance over the traditional random forest by exploring spatial heterogeneity, without the need to consider multicollinearity and without the need to *a priori* examine independent variables[6,13]. Furthermore, a GW-RF gives improvements over geographically weighted regression (GWR) by relaxing the assumption of linearity between dependent and independent variables and is also not prone to overfitting due to its bootstrapping nature[12]

In 2019, the National Health Interview Survey (NHIS) underwent its first major questionnaire redesign since 1997, with the primary goal of improving data quality and relevance. This redesign also introduced new estimation methods for key variables. The CDC has advised caution when comparing data collected before and after these methodological changes. In response, we developed two separate models one for the pre-redesign period and another for the post-redesign period, to fairly evaluate the model's performance across both time periods. This approach also allows us to examine whether the revised methods yield significantly different insights into the associations between diabetes and its correlates.

As of 1 May 2025, literature database searches reveal that only one study has been conducted on utilizing geographically weighted random forest (GW-RF) methods to investigate geographic disparities in the relationship between county-level diabetes prevalence and its associated risk factors in the USA[6]. The study focused solely on health and socioeconomic risk factors. However, various studies[3,5,14] indicate that additional risk factors are crucial for understanding the geographic variability in diabetes and its associated risks. This research adds several novel risk factors such as race (Black, Hispanic, American Indian/Alaska Native and

white population percentages), percentage aged 65 and older, single parent household, percentage unemployed and commute time to provide a more comprehensive understanding of the variables that correlate with diabetes geographically.

This study aims to address four key objectives: (i) investigate the geographic associations between diabetes and its risk factors to support targeted prevention and intervention efforts, (ii) assess the predictive performance of a geographically weighted random forest (GW-RF) model in comparison to traditional and global models, (iii) examine how spatial scaling and parameter weighting influence the predictive accuracy of the GW-RF model, and (iv) compare the performance of the GW-RF model before and after the National Health Interview Survey (NHIS) revised its questionnaire to enhance data relevance and quality.

Materials and methods

AN OVERVIEW OF THE DATA

In this study we analyzed data from 3,080 counties across the 48 contiguous United States after excluding those with incomplete data for the relevant variables and time periods. Two primary data sources were utilized: the Topologically Integrated Geographic Encoding and Referencing (TIGER) database and the U.S. Centers for Disease Control and Prevention (CDC) Diabetes Surveillance System (https://www.cdc.gov/diabetes/data). Diabetes prevalence and its associated risk factors were obtained from the CDC's Diabetes Surveillance System, while geographic shapefiles were sourced from the TIGER database. The Diabetes Surveillance System uses data from multiple sources, including the Behavioral Risk Factor Surveillance System (BRFSS), the National Health Interview Survey (NHIS), and others, to provide comprehensive insights and trends of diabetes at the national, state, and county levels.

BRFSS estimates the percentage of diabetes prevalence and its associated risk factors (obesity, physical inactivity) using Bayesian multilevel small area estimation methods[6,15,16]. Other variables such as the percentage of the population below poverty level were predicted with a regression model using a single year county-level observations from the American Community Survey as the dependent variable and administrative records and census data as the explanatory variables[17]. Sociodemographic variables (percentage Black, percentage white, percentage Hispanic, percentage American Indian and Alaska Native, percentage single parent household, percentage unemployed, percentage over age 65 (≥65) were obtained from the US Census Bureau's American Community Survey 5-Year Estimate.

CDC defines the prevalence of diabetes as the age adjusted estimated percentage of adults with diagnosed diabetes (Type I and II). Diabetes prevalence estimates do not include individuals below 18 as well as gestational diabetes. Obesity prevalence is defined as the age adjusted estimated percentage of adults with BMI $\geq$ 30. The variable "physical inactivity" refers to percentage of adults who have not participated in any physical activity or any form of exercise in the preceding 30 days. "Commute greater than 60 minutes" refers to the percentage of workers 16 years and older who drive with a commute time of ≥ 60 minutes long. "Single parenthood household" is defined as the percentage of household with single parent. The percentage Hispanic, Black, white, AIAN are individuals who self-identify as Hispanic, Black, white and AIAN respectively. The percentage unemployed are individuals who are jobless but actively seeking in the last four weeks. "Percentage below poverty" is the proportion of the population falling below the poverty-line as defined by the American Census Bureau and "food insecurity" is defined by CDC as the percentage of the population who did not have access to a reliable source of food during the past year. A summary of the information on the variables can be found in Table 1.

Table 1. *Source and description of model variables*

Variable Name	Description
Source: CDC Diabetes Surveillance System	
Percentage Obese	Age-adjusted percentage of adults with BMI ≥ 30
Percentage Diagnosed Diabetes	Age-adjusted percentage of adults with diabetes
Physically Inactive	Percentage of adults reporting no physical activity in the past 30 days
Percentage Aged ≥ 65	Percentage of individuals aged 65 years or older
Percentage Below Poverty	Percentage of the population below U.S. Census poverty income threshold
Single-Parent Households (SP)	Percentage of single-parent households
Percentage Black Population	Percentage of the population identifying as Black or African American
Percentage white Population	Percentage of the population identifying as white
Percentage Hispanic Population	Percentage of the population identifying as Hispanic
Percentage American Indian/Alaska Native(AIAN)	Percentage of the population identifying as American Indian or Alaska Native
Commute Time	Percentage of workers ≥ 16 years old who drive with a commute time of ≥ 60 minutes long
Percentage Unemployed	Percentage unemployed and actively seeking work (past 4 weeks)
Food insecurity	Percentage of the population who did not have access to a reliable source of food during the past year
Source: Topologically Integrated Geographic Encoding and Referencing (TIGER)	
GEOID	Geographic identifier (e.g., county FIPS code)
Geometry	Geospatial boundaries (multipolygon)

Counties lacking data for all study years were excluded from the analysis, while those with missing values for only a few years were imputed using mean values. To ensure clarity, consistency, and ease of reference, all maps are presented using the regional divisions established by the Department of Health and Human Services (HHS; Figure 1). Utilizing HHS regions aligns with the structure of national healthcare policy, providing a meaningful and informed approach to subdividing the United States.

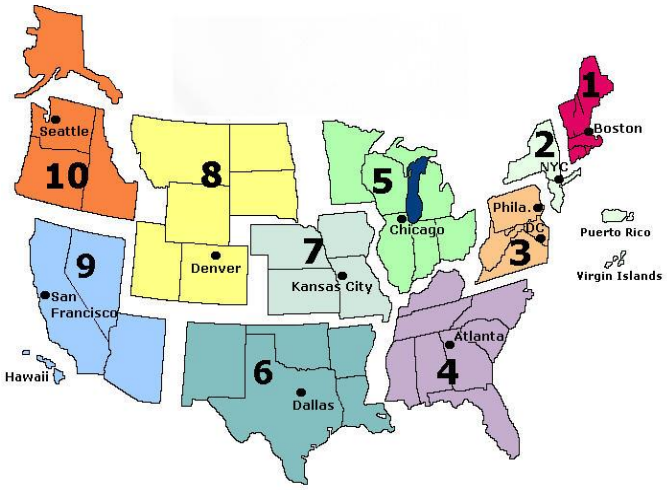


Figure 1. *Regional divisions of the United States based on U.S. Department of Health and Human Services (HHS) designations. Points on the map show where the HHS regional offices are located (U.S Department of Health and Human Services 2025).*

Non-Spatial Models

THE GLOBAL ORDINARY LEAST SQUARE REGRESSION (G-OLS)

The G-OLS regression model is a foundational technique in machine learning and serves as the basis for geographically weighted regression (GWR) models[18]. It estimates the relationship between a dependent variable and a set of independent variables. However, its effectiveness is limited when applied to spatial datasets, like those used in this study, due to several critical assumptions: linearity, constant error variance, normally distributed errors with zero mean, and no autocorrelation in errors. The general model is expressed as: $y = X\beta + \epsilon$ where y is the dependent variable, X is a matrix of independent variables, β is a column vector of coefficients and ϵ is a column vector of error terms.

Two inherent assumptions of OLS are particularly problematic for spatial data: (1) independence of observations is often violated due to spatial clustering[3], and (2) spatial stationarity of relationships is unlikely due to local context affecting both the magnitude and direction of variable relationships[3]. Due to these limitations, G-OLS was used only as an exploratory tool in this research: to identify important variables, assess multicollinearity, and justify the use of geographically weighted models. The *lm* function in R (version 4.5.0) was used to fit the OLS model, and the *vif* function was employed to assess multicollinearity.

Table 2. *Comparison of VIF values from the global ordinary least squares model for both historical and current periods. A VIF of 1 indicates no multicollinearity, 2–5 indicates moderate multicollinearity, 5–10 indicates high multicollinearity, and > 10 indicates serious multicollinearity.*

(a) G-OLS Historical Period

Variable	VIF
Below poverty	3.92
Obesity	3.63
Physical inactivity	3.92
SP	1.89
Food insecurity	3.49
Race: Black	27.72
Race: AIAN	6.32
Race: Hispanic	25.05
Race: white	48.83
Unemployed	2.33
Commute time	1.15
Age ≥ 65	1.41

(b) *G-OLS Current Period*

Variable	VIF
Below poverty	3.71
Obesity	2.30
Physical inactivity	2.51
SP	1.61
Food Insecurity	3.24
Race: Black	27.61
Race: AIAN	6.30
Race: Hispanic	24.87
Race: white	48.82
Unemployed	2.10
Commute Time	1.13
Age ≥ 65	1.42

The reason for the strong correlation among racial variables is that these variables are percentages and must sum to one for each county; including all the variables in the model creates a situation where one variable is by definition a linear combination of the others, leading to high multicollinearity, as seen in Table 3. However, the models used in this study are not negatively affected by multicollinearity and knowing the effects of these racial variables on county-level diabetes is important. We therefore retained these variables in our study without addressing multicollinearity.

RANDOM FOREST (RF) MODELS

Random forests are an ensemble non-parametric machine learning method that use hundreds and thousands of decision trees for regression or classification. In a random forest (RF) model, multiple decision trees are built using random vectors independently sampled from the same distribution, and their outputs are averaged to produce the final result[19]. RF methods improve on G-OLS as they are more resistant to issues produced by multicollinearity, learn non-linear relationships, and are less prone to overfitting due to the use of bootstrapping.

During training, the RF algorithm uses bootstrapped samples: data sampled with replacement, such that approximately two-thirds of the observations (referred to as "in-bag" samples) are used to form the training set, denoted as $D = \{1, \dots, d\}$. This dataset is then used to "grow" a specified number of trees. For this study, 200 trees were used, as increasing the number to the default 500 did not yield any noticeable improvement in model fit or performance. Therefore, 200 trees were chosen for their comparable accuracy and improved computational efficiency. While growing the trees, the RF model also randomly selects a set of features, $F = \{1, \dots f\}$, to include in each tree at each node to determine the split (decision rule used to divide data at each node). The number of features used can be user defined and in this study we specified 8 features. The remaining one-third of the data (counties) also known as the "out-of-bag" samples (OOB) are used to evaluate the model performance by using an internal cross-validation approach. The *ranger* function in the **ranger** package[20] was used to fit random forest models in R.

VARIABLE IMPORTANCE MEASUREMENT (VIM)

A key feature of RF models is the ability to assess the contribution of each variable to the prediction, commonly referred to as variable importance. Although various methods exist for measuring variable importance, this study utilized both the mean decrease in impurity (IncNodePurity) and permutation-based mean-squared error (MSE) reduction. Note that the permutation importance metric is also sometimes called the increase in mean-squared Error since if a variable is very important, we expect a large increase in the OOB error and vice versa when we compare the performance of the original model before and after. The permutation based approach is widely regarded as the most reliable[21,22]. However, this metric may introduce bias towards collinear features. The MSE reduction method estimates the variable importance using the MSE value from the OOB sample. The MSE of the OOB for each tree is computed by:

$$\text{MSE}_t = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \hat{y}_{i,t})^2$$

where $y_{i,t}$ for our model is the predicted diabetes percentage prevalence for tree t in the i^{th} county, y_i is the observed diabetes prevalence in the same county, and N_t is the total number of counties (in-bag sample) for tree t. The target variable j (e.g., physical inactivity) is randomly replaced and a new MSE for the new tree t is computed by:

$$\text{MSE}_t(j) = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \hat{y}_{i,t}(j))^2.$$

Finally, the variable importance measurement (value) for variable j of the RF is the average over MSE reduction of all n trees and it is obtained by:

$$\text{VIM}(j) = \frac{1}{n}\sum_{i=1}^{n}(\text{MSE}_t - \text{MSE}_t(j)).$$

The *ranger* function in the **ranger** package was used to obtain the viable importance as well. Impurity feature importance, also known as mean decrease impurity, measures how each feature contributes to reducing (Gini) impurity when building a RF model. It quantifies the

average reduction in impurity across all splits in the forest where a particular feature is used. Essentially, features that are more frequently used to create "purer" nodes in the decision trees are assigned higher impurity importance. However, impurity feature importance can be biased when predictor variables vary in scale of measurement and number of categories[23].

PARTIAL DEPENDENCY PROFILE

The partial dependence profile provides insight into the average marginal effect of a given variable on the predicted outcome in machine learning models. It also aids in uncovering the relationships learned between features and the dependent variable. For regression tasks, the partial dependence function is defined as:

$$\hat{f}_s(X_s) = E_{X_c}[\hat{f}(X_s, X_c)] = \int \hat{f}(X_s, X_c)\, dp(X_c)$$

where X_s represents feature(s) for which the partial dependence should be obtained and X_c are the other features in the machine learning model $\hat{f}$ and are treated as random variables and $p(X_c)$ is the marginal distribution of X_c. The partial dependency profile was extracted from the RF model using the *partial* function from the **pdp** package[24] and plotted. Partial dependency profiles should be interpreted with caution as they assume independence among variables[25].

Spatial Models

SPATIAL AUTOCORRELATION AND HOTSPOT ANALYSIS

Global Moran's I

Tobler's first law of geography states that everything is related to everything else, but near things are more related than distant ones[26]. This law is extremely important when dealing with spatial data. Moreover, the assumption that the relationship between variables in space is constant is flawed and highly unrealistic, especially when dealing with a large number of spatial data points[27]. It is crucial to use a model that not only examines the relationships between variables but also accounts for spatial heterogeneity. To justify the use of such spatially aware models, it is important to first demonstrate the presence of spatial autocorrelation in the dependent variable. If no spatial autocorrelation exists, the results of a global model would closely resemble those of a localized model. Additionally, we must show that non-spatial models fail to capture this spatial heterogeneity.

To assess spatial autocorrelation, a commonly used metric is Moran's I statistic[3,19]. This statistic ranges from -1 to 1, where values greater than 0 indicate a clustered pattern, values less than 0 suggest a dispersed pattern, and a value of 0 reflects a random spatial distribution. The Moran's I statistic is given by:

$$I = \frac{1}{W}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(x_i - \bar{x})(x_j - \bar{x})\right) S^{-2}$$

where n is total number of counties (3080), w_{ij} is the weight between i^{th} and j^{th} location (county), $W = \sum_i \sum_j w_{ij}$ (is the sum of all spatial weights), x_i and x_j are the values of the variable under consideration at the i^{th} and j^{th} location, respectively; $\bar{x}$ is the mean of the variable. The weighting scheme is a critical component of Moran's I statistic, as the results are highly sensitive to how spatial weights are defined.

To quantify the spatial relationships between counties, we first construct a neighborhood list using the *poly2nb* function from the **spdep** package. This was based on queen contiguity, where two regions are considered neighbors if they share at least one point along their boundaries. It is worth noting that other contiguity-based approaches exist, such as linear, bishop, and rook contiguity. However, the queen is recommended when dealing with irregular polygons (counties)[19]. The *nb2listw* function, also from **spdep** package, was then used to transform the neighborhood list into a spatial weights object. We applied the row-standardized weighting style, although alternative styles, such as binary weighting, equal weights, globally standardized weights are also available. We also used the global bivariate Moran's I to assess spatial autocorrelation between diabetes prevalence (y_i) and each of the risk factors (x_i). The global bivariate Moran's I is given by

$$I_B = \frac{\sum_i (\sum_j w_{ij}\, y_j \times x_i)}{\sum_i x_i^2}.$$

Local Indicators of Spatial Association (LISA)

The global Moran's I, as the name suggests, is a global statistic, and as such, it does not give insight into local spatial autocorrelation patterns. In order to focus on local patterns of association and to allow for local instabilities in overall spatial association, Anselin[27] suggested the local indicator of spatial association (LISA). According to Anselin, the LISA decomposes the global Moran's statistic in order to have a fair idea of the contribution of each individual observation (county). The Anselin local Moran's I for the i^{th} county is given by:

$$I_i = \left((x_i - \bar{x})\sum_j w_{ij}\,(x_j - \bar{x})\right) S^{-2}$$

where S^2 is the variance of diabetes prevalence. In this study LISA was used for hot spot analysis to identify cold and hot spots of diabetes in the USA.

Similarly, bivariate Moran's I quantifies the spatial association between the value of one variable at a given location, x_i, and the spatial lag of a second variable defined as the weighted average of that variable in neighboring locations, $\sum_j w_{ij}\, y_j$. The statistic is calculated as the product of x_i and the spatial lag of y_i, with both variables standardized to have zero mean and unit variance. The local bivariate Moran's I which is like its global counterpart is given by:

$$I_i^B = c x_i \sum_j w_{ij}\, y_j.$$

GEOGRAPHICALLY WEIGHTED REGRESSION MODEL (GW-OLS)

The geographically weighted ordinary least squares model (GW-OLS) is an effective method for achieving high accuracy when exploring the heterogeneous spatial relationships between a dependent variable and a set of independent variables in spatial data sets. This approach has gained significant popularity among researchers interested in spatial analysis due to its precision and

ability to uncover local variations[28,29]. GW-OLS is an extension of the ordinary least squares method (G-OLS) discussed previously, it allows regression coefficients to be estimated locally using a weighted least squares method by fitting a regression model for each location in the study area. The GW-OLS relaxes the assumptions that are violated when G-OLS is used on a spatial dataset. Specifically, unlike G-OLS which produces a single global model, GW-OLS produces numerous local models (one for each location).

The coefficients are estimated locally using the coordinates for each location. It should be noted that the GW-OLS model works best when the relationship between the dependent and independent variables is non-stationary[28]. The equation for the GW-OLS is given by:

$$y_i = \beta_o(u_i, v_i) + \sum_{k=1}^{q} \beta_k(u_i, v_i) X_{ik} + \epsilon_i, \quad i = 1,2,\dots,n$$

where (u_i, v_i) is the coordinates for the i^{th} location (county), $\beta_k(u_i, v_i)$ is the coefficients of the k^{th} independent variable at the ith location, y_i is the value of the dependent variable (percentage diabetes prevalence) at the i^{th} location, $x_i(k = 1,2,\dots,q)$ is the value of the k^{th} explanatory variable at the i^{th} location. The coefficients are estimated by:

$$\hat{\beta}(u_i, v_i) = (X^T W_i X)^{-1} X^T W_i y$$

and the variance is give by:

$$V\left(\hat{\beta}(u_i, v_i)\right) = \left(X^T W_i^{-1} X\right)^{-1}$$

where W_i is the geographical diagonal weight matrix for the ith location (county) is given by:

$$W_i = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{pmatrix}$$

The weight matrix is derived from a spatial kernel, defined as $W_i = f(d_i, h)$, where $f()$ represents the spatial kernel function, d_i is the distance vector at location i and its neighbors, and h is the bandwidth or decay parameter. There are two main approaches to spatial kernel functions: a fixed kernel, which assumes a constant bandwidth for all locations, and an adaptive kernel that adjusts the bandwidth size for each county. In this study, we utilized an adaptive kernel where bandwidth size was specified using a fixed neighbor count. Due to disparities in geographic size, the bandwidth (distance) required to encompass the centroid of, for example, 20 neighboring counties varies drastically from county to county. We specifically use a bisquare distance decay function with variable bandwidth h_i given by:

$$w_{ik} = \begin{cases} \left[1 - \left(\frac{d_{ik}}{h_i}\right)^2\right]^2 & \text{if } d_{ik} \le h_i \\ 0 & \text{if } d_{ik} > h_i \end{cases}$$

where w_{ik} is the weight assigned to observation j when estimating coefficients at location i.

The adaptive approach offers advantages over the fixed kernel; since the fixed method uses a constant bandwidth across all locations, it can lead to the "weak data" problem. This occurs when a location in the study area has few data points for model calibration, particularly in sparse areas, an issue often encountered when working with county-level spatial datasets. Once we selected the adaptive approach, we also needed a strategy to determine the bandwidth. The options included: (i) cross-validation, (ii) predefined, and (iii) AIC based. In this study, we chose to use cross-validation (CV), which minimizes squared errors by selecting the least squares cross-validation. This process generates a fixed number of neighbors, $h_\#$, to fit the model at each location. Although the AIC approach is effective, it is not recommended for large datasets. The equation for the cross-validation approach is given by:

$$CV(h_\#) = \sum_i [y_i - \hat{y}_{\neq i}(h_\#)]^2$$

where y_i is the observed value at location i, and $\hat{y}_{\neq i}(h_\#)$ is the predicted value at location i based on all data except y_i given some number of nearest neighbors $h_\#$. In other words, the number of rows of data with non-zero weights in each local regression is fixed at a value of $h_\#$, and the optimal value of $h_\#$ is chosen by the iteratively evaluating the CV score given in the previous equation.

GEOGRAPHICALLY WEIGHTED RANDOM FOREST (GW-RF)

The RF model is a global model and, as such, does not produce insight into addressing spatial heterogeneous relationships. The GW-RF is an extension of the RF model, but as a disaggregation consisting of multiple local RF models[30]. The fundamental idea is similar to the GW-OLS model where models are calibrated locally rather than globally. The GW-RF model integrates a spatial (geographical) weight matrix with the RF model to produce localized RF models. The main difference between the RF and GW-RF can be distinguished with some simplistic equations. For the RF model the model is given by:

$$y_i = ax_i + \epsilon$$

where y_i is the dependent variable of the RF model for the i^{th} location, ax_i is the non-linear prediction of the RF model for a set of x independent variables and ϵ is the random error term. A GW-RF model is given by:

$$y_i = a(u_i, v_i)x_i + \epsilon$$

where (u_i, v_i) is the coordinates of the i^{th} location (county), $a(u_i, v_i)x_i$ is the prediction of the RF model calibrated on location i. Before applying the GW-RF model, a spatial weight matrix for each observation unit in the study area must be established according to a predefined spatial weight rule, which can either be distance-based or edge-based. The spatial weight matrix (SWM) for the entire study area, consisting of p spatial observation units, is given by:

$$W = \begin{bmatrix} W(1) \\ W(2) \\ \vdots \\ W(i) \\ \vdots \\ W(p) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pp} \end{bmatrix} \quad i \in (1,2,\cdots,p)$$

where each $W(i)$ is the diagonal of the weight matrix shown in the previous section.

Parameter Settings

Optimum bandwidth selection for GW-RF is tedious since it is based on trial-and-error approaches[12]. Stephanos and Kalogirou (2022) propose that one can use the OOB accuracy of the GRF model as a selection method for the optimal bandwidth. You can test several bandwidths (predefined) with the GRF model and then select the bandwidth with the highest OOB accuracy (R^2 OOB) as the optimal bandwidth. The *grf.bw* function in the **SpatialML** package[31] can automate this process. However, the range to search for the optimal bandwidth is best predefined. For this research, we explore the effect of spatial scaling (bandwidth selection) on the predictive performance of the GW-RF model on county-level USA diabetes data. We ran the GW-RF model for multiple bandwidths and then used each of these models to make predictions on the completely new data set which acts as the OOB sample. Specifically, models trained on the historical data were tested on current data and vice versa. By doing this, we can address one of our objectives by assessing the effect of spatial scaling on model predictive performance and comparing the predictive performance of the best-performing historical data GW-RF model with the best-performing current data GW-RF model.

In this study, all the maps and results used to compare the GW-RF models with each other and with other models were based on a bandwidth of 275 (i.e., the 275 nearest neighbors), 200 trees (ntrees), and 8 variables sampled at each split. We selected a configuration of 200 trees and 8 variables while keeping other parameters at their default settings based on extensive experimental testing that demonstrated the best balance between computational efficiency and model performance. Although increasing the number of trees to 500 yielded a marginal improvement in both global and local performance (OOB R^2: 85.4% vs. 85.3%), it came at a significant cost: computation time more than doubled (12.4 minutes vs. 5.15 minutes), and memory usage nearly doubled (9.5 GB vs. 4 GB). Importantly, variable importance remained consistent across both models. Given the negligible performance gain, the 200-tree model was deemed more efficient and practical for our purposes. We evaluated models using bandwidths ranging from 50 to 500 nearest neighbors, increasing by 75 each time, which produces seven models with different adaptive bandwidths for each time frame. We chose this range because using a very small bandwidth can make the model unstable[30], while a very large one can lose important local details. Our goal was to capture local patterns in diabetes prevalence across the U.S. without making the model too sensitive or too broad, and this range gave us a good balance.

Before 2022, the initial version of the GW-RF model had a limitation: all data points within the chosen bandwidth were given equal weight by the weighting matrix. Stephanos and Kalogirou[12] addressed this issue by introducing a way to apply spatial weighting to the local observations. In their approach, the spatial weights matrix is passed to the *case.weights* parameter in the **ranger** implementation, which gives higher-weighted data points a greater chance of being selected during the decision tree bootstrapping process[12]. This enhancement was proposed to improve the accuracy and performance of the GW-RF model, so we incorporated it into our model development. Because the most recent study on county-level diabetes using GW-RF was published in 2021, to the best of our knowledge, we are the first to apply this updated method to U.S. county-level diabetes data.

Model Predictive Performance Metrics

One of our specific interests was to explore how the weighting parameter (α) affects prediction accuracy. To do this, we combined the predictions from both the local and global models using a weighted average controlled by α. The way the weight parameter does this can be expressed in simple terms by this equation:

$$\hat{y} = \alpha \times \hat{y}_{local} + (1 - \alpha) \times \hat{y}_{global}$$

The advantage of this approach is that it blends the low bias typically offered by local models with the low variance characteristic of global models. Stephanos and Kalogirou[30] suggest that this fusion leads to better predictive performance compared to relying solely on localized models. In our study, we tested four values of α: 0.25, 0.50, 0.75, and 1. These correspond to: 25% weight on the local model and 75% on the global, equal weighting for both, 75% local and 25% global, and finally 100% weight on the local model. All the predictions were done with the *predict.grf* function in the **SpatialML** package. After reviewing relevant research on GW-RF models, we chose three key metrics to evaluate the predictive performance of our models.

- The coefficient of determination which measures the fit of the model on test dataset. This is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2}.$$

- Normalized root mean-squared (NRMSE) measures the accuracy of the model prediction. This is given by:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}}{\max(y_i) - \min(y_i)}.$$

- Finally, we also used the mean absolute error (MAE), which also measures the models predictive accuracy and it is given by:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|.$$

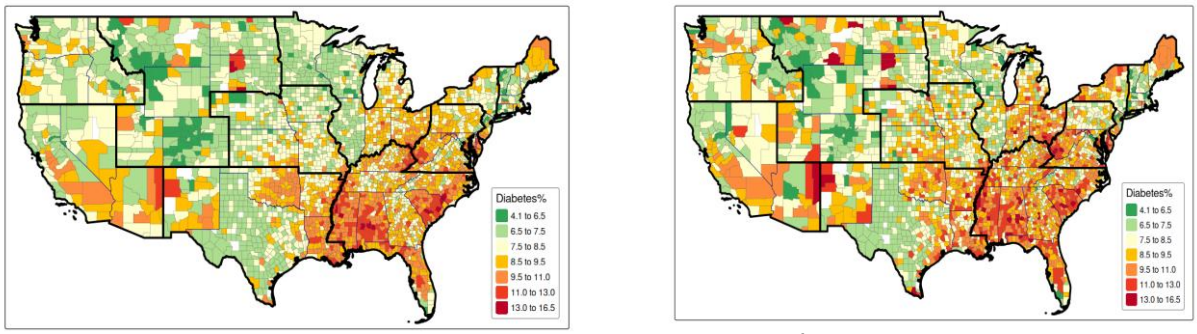
Results

In this section, we will present most of the model results side by side for easier identification and referencing. Also, note that the values of all variables are in percentages, so even though some values may appear to be small, they are significant.

EXPLORATORY DATA ANALYSIS

County-level diabetes prevalence in the USA ranged from 4.2% to 15% during the historical period (2010–2018), showing significant spatial clustering (Global Moran's $I = 0.559, p < 0.001$). In the more recent period (2019–2020), prevalence increased slightly, ranging from 4.2% to 16%, and continued to exhibit significant clustering (Moran's $I = 0.45, p < 0.001$).

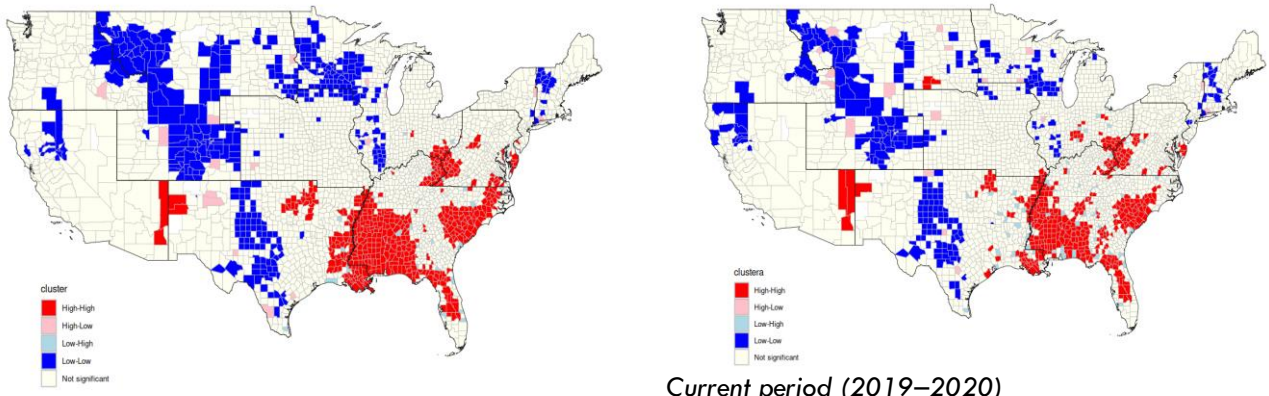
Figure 2 illustrates the spatial distribution of diabetes prevalence across the two periods. While many of the same counties had high diabetes prevalence in both periods, some counties experienced increases, and a number of counties with moderate prevalence in the historical period shifted to higher prevalence in the current period (Figure 2a compared to 2b). Overall, more counties displayed high diabetes prevalence in the current period compared to the historical. Additionally, clusters of high prevalence were observed in parts of the Philadelphia, Chicago, and Dallas regions (DHHS Regions 3, 5 and 6, respectively), with other regions showing scattered pockets of elevated prevalence. Note that counties shown as missing in the maps did have diabetes prevalence data but were excluded due to missing values for one or more risk factors analyzed in this study.



Historical period (2010–2018) *Current period (2019–2020)*

Figure 2. *Heat maps showing the average diabetes prevalence for the historical period (2010–2018) and the current period (2019–2020).*

The historical period Figure 3a shows high proportions of counties in the Atlanta Region (Region 4) exhibiting high clusters of diabetes (hot spots), with a few pockets of diabetes hot spots also present in the Philadelphia and Dallas Regions (Region 3 and 6). In contrast, a high proportion of counties in the Chicago and Denver Regions (Region 5 and 8) showed low clusters of diabetes (cold spots), with a few pockets of cold spots in the Dallas Region (Region 6), the San Francisco Region (Region 9), and Seattle Region (Region 10). Similar clusters were found in the current period (Figure 3b), but with fewer clusters that are less extreme and fewer hot spots, the majority of the counties were not significant.



Historical period (2010–2018) *Current period (2019–2020)*

Figure 3. *Local indicators of spatial association (LISA) maps showing clusters of diabetes prevalence for both historical and current periods. High-high (red) and low-low (blue) are considered hot spots and cold spots, respectively. A classification of a county as high-high indicates, for example, that the county has a high prevalence of diabetes and is surrounded by other counties that also have high prevalence rates. Not significant: counties with statistically insignificant spatial patterns.*

GEOGRAPHICALLY WEIGHTED ORDINARY LEAST SQUARES (GW-OLS)

The global OLS (G-OLS) model applied to the historical data revealed a positive association between diabetes prevalence and most risk factors, except for poverty, Hispanic ethnicity, white race, and age 65 or older (Table 3). Although the G-OLS model explained 83% of the variation in diabetes prevalence, this was notably improved by the geographically weighted OLS (GW-OLS) model, which accounted for 88% of the variation. Additionally, the GW-OLS model had a significantly lower AIC value, indicating a better overall fit to the data. Beyond improved model performance, the GW-OLS also revealed spatial variation in both the direction and strength of associations between diabetes and the risk factors (Table 3).

(a) Diabetes and Obesity *(b)Diabetes and Unemployed* *(c)Diabetes and Physical Inactivity* *(d) Diabetes and Food Insecurity* *(e) Diabetes and Below Poverty* *(f) Diabetes and Black Population*

Figure 4. *Select bivariate local Moran's I (BLMI) cluster maps of diabetes and (a) obesity; (b) unemployed; (c) physical inactivity; (d) food insecurity; (e) below poverty; (f) Black population from the historical period.*

The current period GW-OLS model demonstrated similar improvements over the G-OLS model, namely, a lower AIC and a higher adjusted R^2. Most risk factors had a positive association with diabetes prevalence, except for single-parent households, poverty, Hispanic ethnicity, white race, and age 65 or older (Table 4). When comparing models across periods, the historical models generally outperformed those from the current period. This may be attributed to greater spatial heterogeneity in the historical data (as suggested by Moran's I values for diabetes prevalence in our exploratory data analyses) or other unknown factors.

Table 3. Summary results of the G-OLS and GW-OLS models using the historical data. AIC: Akaike's Information Criterion, NS: not statistically significant.

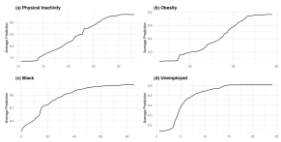
	G-OLS		GW-OLS				
Variable	**Estimate**	**Pr(>|t|)**	**Min**	**1st Qu**	**Median**	**3rd Qu**	**Max**
Intercept	2.21	<0.001	-8.09	-0.23	1.92	3.66	16.89
Obesity	0.1	<0.001	0.01	0.09	0.12	0.15	0.24
Physical inactivity	0.15	<0.001	-0.01	0.08	0.11	0.15	0.23
Unemployed	0.06	<0.001	-0.05	0.005	0.02	0.05	0.15
Food insecurity	0.05	<0.001	-0.07	0.003	0.03	0.05	0.20
Below poverty	-0.002	0.4763^{NS}	-0.08	-0.01	-0.003	0.008	0.07
Commute time	0.007	<0.001	-0.04	-0.01	-0.002	0.005	0.03
Age≥ 65	-0.01	<0.001	-0.06	-0.01	-0.002	0.008	0.07
SP	0.001	0.9708^{NS}	-0.17	-0.03	0.01	0.07	0.29
Race: AIAN	0.01	0.0066	-0.48	0.01	0.04	0.06	0.51
Race: Black	0.02	<0.001	-0.14	0.01	0.04	0.06	0.26
Race: Hispanic	-0.006	0.0848^{NS}	-0.18	-0.01	0.06	0.03	0.11
Race: White	-0.007	0.0542^{NS}	-0.17	-0.01	0.004	0.02	0.09
AIC		5254.688					3752.262
Adjusted R^2		0.83					0.88

GEOGRAPHICALLY WEIGHTED RANDOM FOREST REGRESSION (GW-RF)

The pseudo R^2 of the values of the historical period GW-RF models provide insight into the local fit and ranged from 0.26 to 0.90 with a mean value of 0.74 and a standard error of 0.22. The R^2 in the random forest is pseudo because the RF is non-parametric and ensemble-based method and does not model linear relationship or minimize squared-errors in the same way as a traditional linear regression model and hence the R^2 is merely analogous. The GW-RF model showed relatively high pseudo R^2 values as compared to the global RF model (Table 5). The local GW-RF model was more robust (pseudo $R^2 > 0.6$) in 2779 counties, which is about 90% of the counties under study, with 28.1% of these counties having pseudo $R^2 > 0.8$ (Table 9a). These counties were concentrated in the Dallas and Atlanta Regions, with a few pockets of such counties in the Denver and San Francisco Regions (Figure 5b). The GW-RF model also showed lower MSE values as compared to the global RF model (Table 5). The mean decreased Gini-score and the permutation-based feature importance both ranked physical inactivity as the number one most important variable, followed by obesity, Black race, and percentage unemployed across all counties under study globally (Figure 6a).

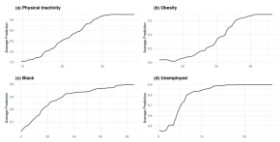
Table 4. *Summary results of the G-OLS and GW-OLS models using the current data. AIC: Akaike's Information Criterion, NS: not statistically significant.*

	G-OLS		GW-OLS				
Variable	**Estimate**	**Pr(>|t|)**	**Min**	**1st Qu**	**Median**	**3rd Qu**	**Max**
Intercept	3.95	<0.001	-7.16	1.66	3.83	6.19	22.17
Obesity	0.11	<0.001	0.02	0.09	0.11	0.13	0.19
Physical inactivity	0.15	<0.001	-0.02	0.08	0.12	0.15	0.22
Unemployed	0.04	<0.001	-0.08	-0.007	0.03	0.07	0.18
Food insecurity	0.06	<0.001	-0.10	0.01	0.03	0.06	0.25
Below poverty	-0.002	0.441^{NS}	-0.08	-0.022	-0.01	0.002	0.1
Commute time	0.004	0.275^{NS}	-0.09	-0.02	-0.004	0.006	0.04
Age ≥ 65	-0.01	0.816^{NS}	-0.07	-0.009	0.005	0.01	0.1
SP	-0.01	0.689^{NS}	-0.25	-0.02	0.03	0.08	0.28
Race: AIAN	0.0009	0.881^{NS}	-0.68	-0.01	0.02	0.07	2.13
Race: Black	0.005	0.368^{NS}	-0.19	-0.03	0.01	0.04	0.17
Race: Hispanic	-0.02	<0.001	-0.19	-0.06	-0.02	0.013	0.12
Race: White	-0.03	<0.001	-0.23	-0.05	-0.02	0.005	0.10
AIC		7883.82					7187.163
Adjusted R^2		0.71					0.76



(a) Historical period partial dependency profile

(b) Historical period local R^2



(c) Current period partial dependency profile

(d) Current period local R^2

Figure 5. *Partial dependency profiles of the top four important features for both periods of the global RF model are shown in panels (a) and (c). Spatial variation in local R^2 values from the GW-RF model for both historical and current periods are shown in panels (b) and (d).*

Additionally, the partial dependence plot (Figure 5a) reveal that the RF model captured distinct non-linear relationships between diabetes prevalence and the four most important predictors: obesity, physical inactivity, percentage of the Black population, and unemployment rate. Specifically, the relationships are logistic-like for obesity, mostly linear for physical inactivity and non-linear for both the Black population percentage and unemployment rate. The plots also indicate that the influence of these variables generally increases across most of their ranges but tends to plateau beyond certain thresholds when the effects of other variables are held constant.

Locally higher values of the mean decreased Gini-score (> 150) for physical inactivity were observed in a large number of counties in the San Francisco, Dallas, Atlanta, Philadelphia, New York, and Boston Regions, with a handful of such counties in Denver, Kansas, and Chicago Regions which implies increased importance of physical inactivity to diabetes prevalence in such counties (Figure 6b). Physical inactivity was the most influential variable to diabetes in 55.0% of the counties and ranked as the second most influential in about 41.1% of counties (Table 6). Obesity was the second most influential value, showing higher mean decrease Gini index (> 150) in a large number of counties in Seattle, San Francisco, Denver, Dallas, and Atlanta Regions, which implies the increased importance of obesity to diabetes prevalence in these counties (Figure 6b). Obesity ranked as the top most influential variable to diabetes in about 39.7% counties and second most influential in about 50.5% of the counties (Table 6). The percentage of the population that is white ranked the most influential in 10.2% of the counties, and all these counties were located in the Denver Region, while other variables ranked as the most influential variable in less than 10% of the counties (Table 6, Figure 6b).

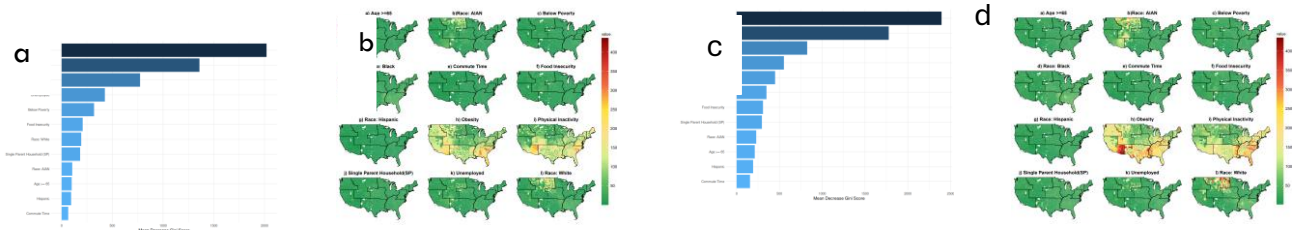


Figure 6. *Global variable importance based on Gini impurity index for both historical and current period are shown in panels (a) and (c). Spatial distribution of local feature importance based on the Gini impurity index for both historical and current periods are shown in panels (b) and (d).*

The GW-RF model for the current period also showed relatively higher pseudo R^2 values compared to the global RF model, with values ranging from 0.12 to 0.84, and MSE values ranging from 0.33 to 1.27 also relatively lower than those of the global RF model (Table 7). The GW-RF model had good fits (pseudo $R^2 > 0.6$) for 1,333 counties, representing approximately 43.28% of all counties (Table 9b). These counties were concentrated in the Denver, Dallas, and Atlanta regions, with more robust fits (pseudo $R^2 > 0.8$) observed particularly in counties within the Dallas region (Figure 5d).

The top four most influential variables globally, as ranked by both the mean decreased Gini Score and the permutation-based feature importance, were the same as in the historical period (Figure 6c). The partial dependence plot (Figure 5c) also shows that the RF model learned a similar relationship between diabetes prevalence and these risk factors, consistent with the historical period. Physical inactivity exhibited higher importance (mean decrease Gini index value > 150) to diabetes prevalence in counties located in the same regions as in the historical period, but with more intensity (evidenced by the presence of more red areas in Figure 6d). For the current period, physical inactivity was the most important risk factor for diabetes prevalence in approximately 39.6% of the counties studied and ranked as the second most influential risk factor in about 50.1% of counties (Table 8). The influence of obesity in the current period was also somewhat similar and dominant in the same areas as in the historical period, but with increased intensity (more red areas in Figure 6d). Obesity was ranked as the most important variable for diabetes prevalence locally in about 56.0% of counties and ranked as the second most important in 39.4% of the counties studied (Figure 6d).

Table 5. *Summary results of global random forest (RF) and geographically weighted random forest regression (GW-RF) models. Both models were trained on the historical period data (i.e., mean data (2010-2018)).*

	RF (Global)	GW-RF(Local)			
Variable	**IncNodePurity**	**Min**	**Max**	**Mean**	**Std**
Obesity	1112.99	4.53	423.46	107.20	76.46
Physical inactivity	2828.36	3.93	417.47	110.28	68.90
Unemployed	350.61	1.63	130.76	12.37	11.88
Food insecurity	150.41	1.62	38.11	8.18	4.88
Below poverty	172.22	1.69	42.00	7.65	4.82
Commute time	57.35	1.28	15.23	4.83	2.17
Age ≥ 65	77.44	1.53	22.43	6.00	2.54
SP	70.28	1.63	76.72	11.59	8.52
Race: AIAN	88.13	1.72	158.67	12.41	20.97
Race: Black	720.23	2.39	90.91	17.55	16.90
Race: Hispanic	69.48	0.79	27.57	4.99	3.30
Race: White	153.95	0.66	287.51	22.21	34.45
MSE	0.28	0.13	0.57	0.31	0.14
R^2	0.85	0.26	0.90	0.74	0.22

Table 6. *The proportion of counties with local risk factors (the risk factors with the 1st, 2nd, 3rd and 4th value of local variable importance) on the county level diabetes prevalence for the historical period.*

Variable	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**
Obesity	39.7	50.5	6.2	3.6
Physical inactivity	55.0	41.1	2.0	1.9
Unemployed	6.4	12.5	31.4	49.8
Food insecurity	0.0	2.3	31.2	66.5
Below poverty	0.0	0.0	44.9	55.1
Commute time	0.0	0.0	0.0	100.0
Age ≥ 65	0.0	0.0	32.6	67.4
SP	5.4	6.2	31.4	57.1
AIAN	1.5	19.2	43.3	36.0
Race: Black	0.1	0.7	65.9	33.3
Race: Hispanic	0.0	0.0	88.0	12.0
Race: White	15.6	5.8	37.2	41.5

THE EFFECT OF GEOGRAPHICAL SCALE AND WEIGHT PARAMETER ON THE PREDICTIVE PERFORMANCE OF THE GW-RF MODEL

A key aspect of the GW-RF model in capturing spatial associations between variables is the choice of geographical scale, while the weight parameter plays a crucial role in optimizing predictive accuracy. To evaluate their effects, we compared the R^2, MAE, and NRMSE across GW-RF models using different values of $h_{\#}$ (nearest neighbors) and weight parameters. We assessed model performance by applying historical models to current data and vice versa. In total, we tested seven different bandwidths, starting from 50 and increasing by increments of 75. As shown in Figure 7 and Figure 8, there is a clear pattern in the performance of the models, regardless of the period or the model used. In both periods, our model performs best with weight parameters of 0.25 and 0.50, which aligns with previous

research[30] suggesting that moderate or light weighting of the local model may yield better predictions compared to heavier weighting at certain bandwidths.

MODEL PREDICTIVE PERFORMANCES

The performance of all models was evaluated using tenfold cross-validation, and the results are presented in Table 10. The GW-RF model achieves the best performance across periods, with the highest R^2, the lowest MAE, and the lowest NRMSE. This demonstrates that incorporating both spatial weighting and non-linearity through the use of random forest enhances predictive accuracy, regardless of how minimal it may seem. Although the GW-RF outperforms all the other models, the performance improvement was not massive. We argue that with proper hyperparameter tuning, the global and locally weighted random forest may perform even better. In the current period, we observe the same pattern in model performance. However, the performance of all models turned out to be reduced in the current period compared to the historical period.

Table 7. *Summary results of global random forest (RF) and geographically weighted random forest regression (GW-RF) models. Both models were trained on the current period data (i.e., mean data (2019-2020)).*

	RF (Global)	GW-RF (Local)			
Variable	**IncNodePurity**	**Min**	**Max**	**Mean**	**Std**
Obesity	1329.28	13.89	443.43	156.80	85.01
Physical inactivity	3496.48	11.90	430.27	147.35	90.33
Unemployed	451.40	2.91	208.20	21.72	16.07
Food insecurity	254.55	2.15	69.52	16.76	9.67
Below poverty	299.23	3.14	66.49	16.37	7.70
Commute time	143.18	2.17	45.46	13.47	5.56
Age ≥ 65	179.80	2.89	46.00	14.97	5.23
SP	169.61	3.98	83.28	23.02	11.92
Race: AIAN	208.08	3.60	307.33	26.58	38.00
Race: Black	759.72	3.33	102.43	27.39	18.57
Race: Hispanic	161.16	1.81	44.85	13.20	5.44
Race: White	371.50	2.71	415.43	36.61	53.31
MSE	0.72	0.33	1.27	0.82	0.19
R^2	0.72	0.12	0.84	0.57	0.13

Discussion

We conducted a spatial analysis to explore the spatial heterogeneity of the relationship between diabetes prevalence and its associated risk factors at the county level in the U.S. using a geographically weighted random forest (GW-RF) model, along with traditional non-spatial and spatial weighted models. As part of our motivation to model diabetes within the U.S., the NHIS updated its method of data collection and estimation of the variables in the diabetes dataset in 2019, following a previous update in 1999. The Centers for Disease Control and Prevention (CDC) advises researchers to exercise caution when dealing with data from before and after these updates, as any potential differences could be due to the updates rather than actual changes. To address this and assess how our models perform on data sets from before and after the updates, we divided the data set, spanning from 2010 to 2020, into two periods: before the updates (historical period) and after the updates (current period). The historical period data set was summarized as the mean of data from 2010 to 2018, while the current period data set was summarized as the mean of data from 2019 to 2020; these summaries were then used for modeling diabetes prevalence in each period. A handful of studies have been conducted on diabetes prevalence and its correlates, most of which employ geographically weighted regression (GW-OLS)[32-35]. Although this model effectively assesses the heterogeneous spatial relationship between diabetes prevalence and its associated risk factors, it has limitations.

Table 8. *The proportion of counties with local risk factors (the risk factors with the 1st, 2nd, 3rd and 4th value of local variable importance) on the county level diabetes prevalence for the current period.*

Variable	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**
Obesity	56.0	39.4	2.0	2.7
Physical inactivity	39.6	50.1	6.9	3.4
Unemployed	0.0	1.9	61.6	36.5
Food insecurity	0.0	0.3	33.1	66.7
Below poverty	0.0	4.0	18.7	77.3
Commute time	0.0	0.0	36.0	64.0
Age ≥ 65	0.0	0.0	23.2	76.8
SP	0.4	10.3	41.6	47.7
Race: AIAN	6.8	19.7	49.3	24.2
Race: Black	0.0	3.1	62.1	34.7
Race: Hispanic	0.0	0.0	31.3	68.7
Race: White	14.4	8.7	29.8	47.2

The GW-RF model is a non-parametric machine learning model designed to address the limitations of the GW-OLS model. To the best of our knowledge, the GW-RF model has been used to study diabetes prevalence with its risk factors in the USA in only one paper[6]. Although this study effectively explored spatial heterogeneity in diabetes prevalence, it included only a few variables such as obesity, physical inactivity, accesses to exercise, food environment index, poverty, and education, highlighting the need for more comprehensive research. It also used data from 2013 to 2017, collected before the NHIS updated its method of data collection and estimation.

Table 9. *The local R^2 of the GW-RF for both periods. The left table (a) shows results from the historical period model, while the right table (b) represents the results from the current period model. Categories are based on rounded R^2 values.*

(a) Local R^2 distribution, historical

Category	Percentage of counties
≤0.3	0.1
(0.3, 0.5]	5.0
(0.5, 0.6]	4.7
(0.6, 0.8]	62.2
>0.8	28.1

(b) Local R^2 distribution, current

Category	Percentage of counties
≤0.3	6
(0.3, 0.5]	15.6
(0.5, 0.6]	35.2
(0.6, 0.8]	39.4
>0.8	3.4

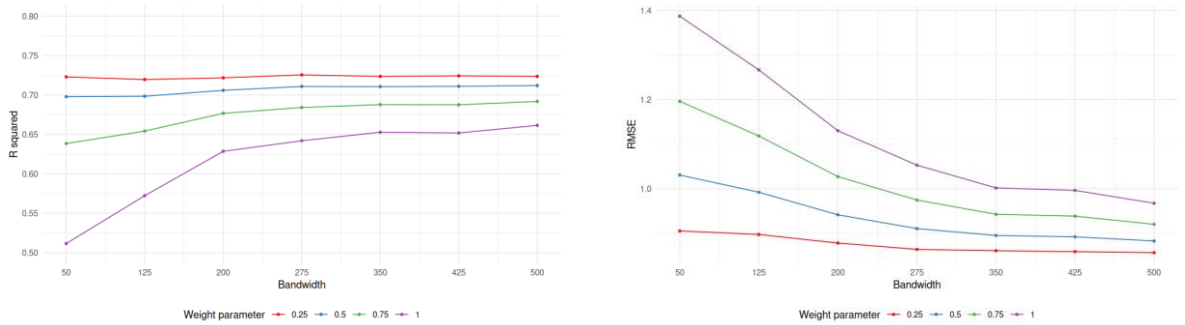


Figure 7. *R^2 and RMSE of the GW-RF model trained on historical data and tested on current data with incrementing bandwidth (number of nearest neighbors included in the local models). A step size of 75 was used across a range of bandwidth values from 50–500.*

In this study, we address these shortcomings and more. The risk factors considered are obesity, physical inactivity, age ≥ 65, commute time, food insecurity, unemployment, living below the poverty line, single-parent households, and race (Black, white, Hispanic, and American Indian/Alaskan Native). As expected, the GW-RF model explained higher variability in diabetes prevalence compared to traditional global models (OLS, RF) and the GW-OLS model for both periods. Our study ranked obesity and physical inactivity as the two most influential variables in diabetes prevalence in the U.S., both locally and globally, aligning with several studies in this area[3,6,36]. Although the national trend in diabetes prevalence, as reported by the CDC, indicates a decline from 2019–2020, our study showed that the average diabetes prevalence in the diabetes belt region (Atlanta Region) remains high in this period. This suggests that while diabetes prevalence may be falling nationally, this is not necessarily the case locally. Our study also indicated that the importance of the two most influential risk factors intensified during the current period. Counties with higher diabetes prevalence than the national average, and even higher prevalence in the current period, were concentrated in the Atlanta Region, with a few pockets in other regions.

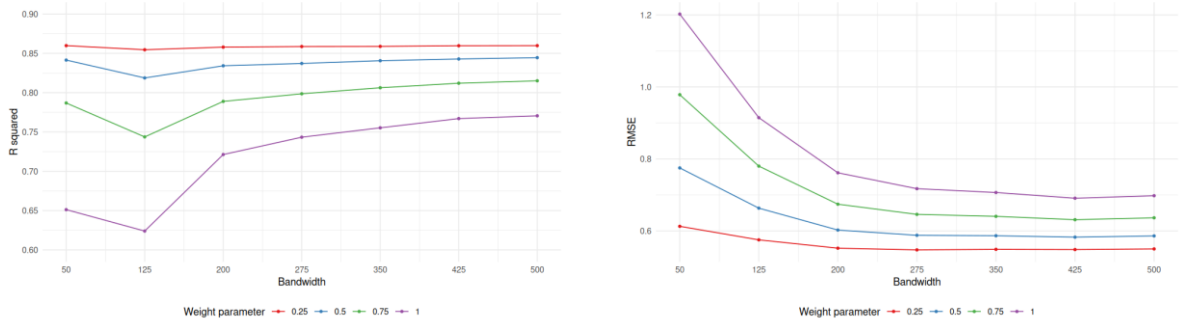


Figure 8. *R^2 and RMSE of GW-RF model trained on current data and tested on historical data with incrementing bandwidth (number of nearest neighbors included in the local models). A step size of 75 was used across a range of bandwidth values from 50–500.*

Table 10. *Model performance metrics for the historical and current periods across four regression models: G-OLS, RF, GW-OLS, and GW-RF.*

Historical Period	G-OLS	RF	GW-OLS	GW-RF
R^2	0.832	0.856	0.861	0.867
MAE	0.439	0.398	0.388	0.378
NRMSE	0.068	0.063	0.062	0.060
Current Period				
R^2	0.706	0.724	0.725	0.729
MAE	0.671	0.642	0.643	0.634
NRMSE	0.100	0.097	0.087	0.087

The CDC has identified 15 states in the U.S. with historically high diabetes prevalence, termed the diabetes belt. This belt includes 644 counties in parts of Alabama, Florida, Texas, Georgia, Louisiana, Kentucky, North Carolina, South Carolina, Mississippi, Ohio, Pennsylvania, Tennessee, Arkansas, Virginia, and West Virginia. Eight of these states are in the Atlanta Region, which we found to be a hot spot for diabetes across periods; the remaining states are in the Philadelphia and Dallas Regions. We also found pockets of counties outside the diabetes belt with high diabetes prevalence and increased importance of obesity and physical inactivity. While a few of such counties existed in the historical period, more appeared in the current period, possibly due to the 2019 updates rather than reflecting real-life occurrences.

Counties below the poverty line also showed high diabetes prevalence, as did counties with a high percentage of the Black population, particularly in the Atlanta and Dallas Regions. Counties with high levels of food insecurity also experienced high diabetes prevalence, this was not a surprise as food-insecure populations are more likely to have limited access to healthy foods and depend on high-calorie foods, contributing to obesity and diabetes[6,37,38]. High unemployment rates were also associated with high diabetes prevalence, possibly due to inactivity and limited access to healthy food or gym memberships. The percentage of Hispanic individuals showed more regional variation compared to other variables, as did the percentages of American Indian and Alaska Native population. Unexpectedly, age $\geq$ 65 was not a significant risk factor for county-level diabetes, both globally and locally, despite higher diabetes prevalence in this age group according to the National Diabetes Association and the CDC[1]. The global RF models for both periods ranked age $\geq$ 65 as the tenth most important variable for diabetes globally, and it did not rank as the most important variable in any of the counties studied.

Figure 4 shows selected bivariate local Moran's I cluster maps of diabetes with its risk factors for the historical period. The red color (high-high) corresponds to significant clusters of high diabetes prevalence with a high prevalence of obesity (Figure 4a), unemployment (Figure 4b), physical inactivity (Figure 4c), food insecurity (Figure 4d), poverty (Figure 4e) and Black population percentage (Figure 4f). These counties were mostly concentrated in Atlanta Region and Dallas Region with few pockets of such counties in Philadelphia Region. The blue color (low-low) corresponds to significant clusters of low diabetes prevalence with a low prevalence of the risk factors. Spatial anti-correlation was observed in some counties across nearly all regions for the variables studied (see Figure 4). Counties with high diabetes prevalence surrounded by neighbors with low prevalence of risk factors may reflect a lag in the health benefits of recent behavioral or policy improvements. Conversely, counties with low diabetes prevalence but surrounded by high-risk neighbors may be spatial outliers, possibly due to stronger healthcare infrastructure, more favorable demographic characteristics, or other protective factors. However, if high prevalence of risk factors persists in neighboring counties, these currently low-diabetes counties may become increasingly vulnerable in future years.

Our models demonstrated consistent improvements in accounting for local and global diabetes variability (R^2) and predictive accuracy (NRMSE, MAE) when transitioning from G-OLS to RF global models. These improvements were even more pronounced when comparing the GW models to the global models. However, the performance of all models declined (by comparison) during the current period. This could be attributed to several factors, such as the reduced spatial autocorrelation observed in the current period compared to the historical period. It might also be due to changes in the patterns of diabetes prevalence or possibly a shift in data quality, which warrants further investigation. For model performance in relation to geographical scale (bandwidths), we observed that the model performed better at higher bandwidths. However, beyond a bandwidth of 275 (Figure 7, Figure 8), further increases did not significantly enhance performance. In other words, using larger bandwidths (i.e., including more nearest neighbors) tends to shift the model toward a more global approach, but this does not necessarily translate into better predictive performance for the GW-RF model.

With regard to the predictive performance of our models, we observed that the differences in performance when moving from G-OLS and GW-OLS to RF and GW-RF were not as substantial as reported in other studies[6,12]. In those studies, the GW-RF model often performed exceptionally well, while other models performed poorly. In contrast, all models in our study performed well, and we did not observe a dramatic improvement in the performance of the RF and GW-RF models as expected. One possible explanation for this could be that more than half of the variables used in our study exhibited a linear relationship with diabetes—something we observed during the data exploration stage. Additionally, although we detected spatial autocorrelation in both time periods, it was not particularly strong, as indicated by the Moran's I values. This is a valid reason why OLS and GW-OLS

may perform comparably to RF and GW-RF models: when the underlying data relationships are predominantly linear and spatial heterogeneity is moderate, simpler models can perform just as well.

Furthermore, if interactions between features are limited and the dataset is relatively clean or low in noise, the added complexity of RF and GW-RF may not offer a significant advantage. This points to a potential shift in data quality, which we have alluded to throughout this study. Another valid reason could be the hyperparameter tuning of the models, as the performance of RF and GW-RF is highly dependent on their hyperparameters. This warrants further investigation in future studies.

There are limitations to GW models, the data set, and our analyses. The county-level prevalence data from CDC for variables such as diabetes, obesity and physical inactivity are model-based estimates from the BRFSS telephone survey, which has inherent limitations such as recall bias and social desirability bias[6,39]. Additionally, the diabetes prevalence data used in this study exclude individuals with undiagnosed diabetes[36], potentially affecting the results if counties significantly vary in the proportion of undiagnosed cases. Another limitation of our study is the exclusion of approximately 28 counties due to missing data for certain variables during the study years. We also imputed missing values for other counties, which may not accurately represent real-world conditions. These decisions could introduce bias, disrupt spatial patterns, and lead to underestimated uncertainty in our results.

Geographically weighted (GW) models also have inherent limitations. Unlike global models that rely on the entire dataset, GW models estimate local regression coefficients or variable importance using data from nearby locations, such as adjacent counties. In our study, we used an adaptive kernel approach to select the optimal number of neighboring counties, allowing for variation in county size. However, this method results in an undefined and spatially variable distance of influence. As a result, the number of neighbors or the bandwidth is shaped by the spatial configuration of surrounding counties as specified by the kernel function. This can lead to potential spillover effects or spatial autocorrelation in the residuals[6]. The GW model is also limited by the edge effect, meaning counties located on the edges of the U.S. (i.e., coastal regions and borders with Canada and Mexico) do not have the 360-degree influence of counties in the nation's interior[3]. Our findings also have some limitations, especially regarding the current period. The local R^2 accounted for a good amount of variability in diabetes prevalence; however, for a large number of counties in the Kansas and Chicago Regions, the explained variability was less than 0.3 (pseudo $R^2 < 0.3$). The low explanatory power in some regions could indicate that some factors associated with county-level diabetes prevalence in these geographic areas are missing from our model.

Conclusion

While this study is not the first of its kind, its primary strength lies in building upon the work of Quinones et al.[6] by incorporating additional demographic and socioeconomic variables. It is also the first to utilize data from 2019 following updates to the NHIS survey which offers a more current perspective on the relationship between diabetes prevalence and its correlates. The inclusion of this updated data set allows for a meaningful comparison between time periods and sheds light on potential changes in data quality, which may merit further investigation. This study also ranks the importance of diabetes correlates at both the local and global levels. Such insights are valuable to health researchers and practitioners by identifying the specific drivers of diabetes prevalence at the county level. This, in turn, can support the development of more targeted, context-specific public health interventions and policies particularly those with national scope but local implementation.

Furthermore, the findings provide a fresh perspective and perhaps a robust update of the spatial correlation between diabetes prevalence and its associated risk factors after the NHIS update. The study also demonstrates the effectiveness of a geographically weighted random forest (GW-RF) model as an exploratory and predictive tool. This modeling approach can help health professionals make accurate projections, identify emerging hotspots, and implement appropriate prevention and control strategies.

Finally, we recommend that future research continue to explore data from 2019 onward to better understand whether the observed shifts in diabetes prevalence and its spatial patterns reflect real-world changes or are the result of methodological updates in data collection and variable estimation.

Acknowledgments

We would like to thank Dr. Fan Yi for their helpful comments and guidance in improving this research.

References

1. American Diabetes Association. Statistics about diabetes. American Diabetes Association. 2024. Accessed March 13, 2026. https://diabetes.org/about-diabetes/statistics/about-diabetes
2. Andes LJ, Cheng YJ, Rolka DB, Gregg EW, Imperatore G. Prevalence of prediabetes among adolescents and young adults in the United States, 2005-2016. JAMA Pediatr. 2020;174:e194498. doi:10.1001/jamapediatrics.2019.4498
3. Hipp JA, Chalise N. Spatial analysis and correlates of county-level diabetes prevalence, 2009-2010. Prev Chronic Dis. 2015;12:E08.
4. Neupane S, Florkowski WJ, Dhakal C. Trends and disparities in diabetes prevalence in the United States from 2012 to 2022. Am J Prev Med. 2024.
5. Neupane S, Florkowski WJ, Dhakal U, Dhakal C. Regional disparities in type 2 diabetes prevalence and associated risk factors in the United States. Diabetes Obes Metab. 2024;26:4776-4782.
6. Quiñones S, Goyal A, Ahmed ZU. Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. Sci Rep. 2021;11:6955.
7. Parker ED, Lin J, Mahoney T, et al. Economic costs of diabetes in the US in 2022. Diabetes Care. 2024;47:26-43.
8. Bottaro A. Type 2 diabetes cure. Verywell Health. 2022. Accessed March 13, 2026. https://www.verywellhealth.com/type-2-diabetes-cure-6823636
9. Joshi RD, Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. Int J Environ Res Public Health. 2021;18:7346.
10. CDC Diabetes Surveillance System. CDC Diabetes Atlas. Centers for Disease Control and Prevention. 2024. Accessed March 13, 2026. https://gis.cdc.gov/grasp/diabetes/diabetesatlas.html
11. Abraham TM, Fox CS. Implications of rising prediabetes prevalence. Diabetes Care. 2013;36:2139.
12. Georganos S, Kalogirou S. A forest of forests: a spatially weighted and computationally efficient formulation of geographical random forests. ISPRS Int J Geo-Inf. 2022;11:471.
13. Luo Y, Yan J, McClure S. Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. Environ Sci Pollut Res. 2021;28:6587-6599.
14. Li X, Staudt A, Chien LC. Identifying counties vulnerable to diabetes from obesity prevalence in the United States: a spatiotemporal analysis. Geospat Health. 2016;11(1).
15. Cadwell BL, Thompson TJ, Boyle JP, Barker LE. Bayesian small area estimates of diabetes prevalence by US county, 2005. J Data Sci. 2010;8:171-188.
16. Barker LE, Thompson TJ, Kirtland KA, et al. Bayesian small area estimates of diabetes incidence by United States county, 2009. J Data Sci. 2013;11:269.
17. Bell WR, Basel WW, Maples JJ. An overview of the US Census Bureau's small area income and poverty estimates program. In: Analysis of Poverty Data by Small Area Estimation. 2016:349-378.
18. Khan SN, Li D, Maimaitijiang M. A geographically weighted random forest approach to predict corn yield in the US corn belt. Remote Sens (Basel). 2022;14:2843.
19. Seamon E, Ridenhour BJ, Miller CR, Johnson-Leung J. Spatial modeling of sociodemographic risk for COVID-19 mortality. Preprint. medRxiv. Posted 2024.
20. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77:1-17. doi:10.18637/jss.v077.i01
21. Luo Y, Yan J, McClure SC, Li F. Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model. Environ Sci Pollut Res. 2022:1-13.
22. Santos F, Graw V, Bonilla S. A geographically weighted random forest approach for evaluate forest change drivers in the northern ecuadorian amazon. PLoS One. 2019;14:e0226224.
23. R-Bloggers. Be aware of bias in RF variable importance metrics. R-Bloggers. Published June 2018. Accessed March 13, 2026. https://www.r-bloggers.com/2018/06/be-aware-of-bias-in-rf-variable-importance-metrics/
24. Greenwell BM. pdp: an R package for constructing partial dependence plots. R J. 2017;9:421-436. doi:10.32614/RJ-2017-016
25. Molnar C. Partial dependence plot (PDP). Interpretable Machine Learning. 2024. Accessed March 13, 2026. https://christophm.github.io/interpretable-ml-book/pdp.html
26. Tobler WR. A computer movie simulating urban growth in the Detroit region. Econ Geogr. 1970;46:234-240.
27. Anselin L. Local indicators of spatial association—LISA. Geogr Anal. 1995;27:93-115.
28. Arabameri A, Pradhan B, Rezaei K. Gully erosion zonation mapping using integrated geographically weighted regression with certainty factor and random forest models in GIS. J Environ Manage. 2019;232:928-942.
29. Chalkias C, Kalogirou S, Ferentinou M. Landslide susceptibility, Peloponnese Peninsula in south Greece. J Maps. 2014;10:211-222.
30. Georganos S, Grippa T, Niang Gadiaga A, et al. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 2021;36:121-136.
31. Kalogirou S, Georganos S. SpatialML: spatial machine learning. R package. 2024. Accessed March 13, 2026. https://CRAN.R-project.org/package=SpatialML
32. Siordia C, Saenz J, Tom SE. An introduction to macro-level spatial nonstationarity: a geographically weighted regression analysis of diabetes and poverty. Hum Geogr. 2012;6:5.
33. Lord J, Roberson S, et al. A retrospective investigation of spatial clusters and determinants of diabetes prevalence: scan statistics and geographically

weighted regression modeling approaches. PeerJ. 2023;11:e15107.
34. Kauhl B, Schweikart J, Krafft T, Keste A, Moskwyn M. Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in northeastern Germany using kernel density estimation and geographically weighted regression. Int J Health Geogr. 2016;15:1-12.
35. Sharma A. Exploratory spatial analysis of food insecurity and diabetes: an application of multiscale geographically weighted regression. Ann GIS. 2023;29:485-498.
36. Geiss LS, Kirtland K, Lin J, et al. Changes in diagnosed diabetes, obesity, and physical inactivity prevalence in US counties, 2004-2012. PLoS One. 2017;12:e0173428.
37. Adams EJ, Grummer-Strawn L, Chavez G. Food insecurity is associated with increased risk of obesity in California women. J Nutr. 2003;133:1070-1074.
38. Weigel MM, Armijos RX, Hall YP, Ramirez Y, Orozco R. The household food insecurity and health outcomes of US-Mexico border migrant and seasonal farmworkers. J Immigr Minor Health. 2007;9:157-169.
39. Barker LE, Kirtland KA, Gregg EW, Geiss LS, Thompson TJ. Geographic distribution of diagnosed diabetes in the US: a diabetes belt. Am J Prev Med. 2011;40:434-439.

