The Positive Relationship of Walkability on Diabetes Prevalence in the Southern United States

A Preprint

Arkaprabho Bose
Undergraduate Program in Department of Computer Science
Texas A & M University
College Station, TX, 77843
abose0267@tamu.edu

Sebastian Oberg
Undergraduate Program in Department of Computer Science
Texas A & M University
College Station, TX, 77843

Abhinav Cheruvu
Undergraduate Program in Department of Mathematics
Texas A & M University
College Station, TX, 77843

April 28, 2024

Abstract

The diabetes epidemic in the United States presents a nuanced public health challenge, influenced by factors such as socioeconomic status and climate. While the impact of these factors is well-documented, the influence of walkability on diabetes prevalence has been underexplored. This study investigates how both socioeconomic and climate variables, alongside walkability, affect diabetes prevalence in the Southern U.S. Contrary to expectations, our findings indicate that higher walkability indexes correlate with an increase in diabetes prevalence. This effect persists even when controlling for high blood pressure and low physical activity, which indicates significant regional variance. Our findings show that the relationship between walkability and diabetes prevalence varies significantly by region, driven by distinct socioeconomic and environmental contexts. This variability highlights the need for urban planning as a public health strategy that is tailored to the specific regional characteristics to effectively address diabetes.

1 Introduction

Diabetes is a common chronic illness that is caused due to consistently high blood sugar levels, and can be prevented through sugar intake management, exercise and dieting. In a study done on 2016 and 2017 National Center for Health Statistics data, it was shown that among adults in the United States, there was a prevalence of 9.7% (Xu, et. al). This high prevalence can impact humans on a daily basis by directly impacting the quality of life both physically and mentally. Diabetes can affect organs all around the body such as the eyes, pancreas and kidneys. In addition to having direct impact on people, high prevalence of diabetes puts stress on the existing healthcare systems by forcing hospitals and doctors to put resources into solving issues that are preventable.

In recent years, there have been speculations that lifestyle changes, specifically walkability of a region can impact the prevalence of diabetes in that given region. The Environmental Protection Agency has developed a standardized scale on which regions can be ranked based on how walkable it is. The scale ranges from 1-20 with 1 being the least walkable and 20 being the most walkable. It takes into account various things such as intersection density, and proximity to transit (Glazier et al.). According to a temporal analysis study done in 2016, areas with highest walkability score, which is a value calculated had lower rates of diabetes prevalence (Creatore et. al). An area being walkable results in less reliance on cars, and forces the population to walk which is a form of exercise that is often overlooked and can have a meaningful impact on ones health.

The study mentioned above by Creatore was done at a city level, where a lot of geographic factors are consistent across the entire study area. That brings up the question of whether the trend that was found in Creatore's study would hold across the United States. Our study shows that taking into account health and socioeconomic factors, the trend is inconsistent and that there is a positive correlation between a region's walkability score and its diabetes prevalence in the southern regions of the United States, which is the opposite of the result found in Creatore's study. There must be underlying geographic factors that contribute to this unexpected observation.

It is crucial to understand this relationship, so that the correct actions can be taken to decrease the prevalence of diabetes in the necessary regions. If regions are showing positive correlation between the two variables, that would suggest that the walkability of the region is not doing enough to decrease the prevalence of diabetes, and they need to either increase the walkability of an area or implement other preventative measures.

2 Related Works

2.1 Exploring how location affects diabetes risk, focusing on two studies

Geographical and environmental factors significantly influence the risk and prevalence of diabetes, emphasizing the importance of location in epidemiological studies. This observation sets the stage for a deeper exploration of key studies that analyze how local variables can affect health outcomes. Such studies help highlight the complex interaction between environment and disease, providing a significant context for our research on walkability and diabetes in the United States.

2.2 Study on socio-economic impact in Northeastern Germany

A detailed analysis of a study conducted in Northeastern Germany reveals that socio-economic status significantly impacts diabetes risk within this specific locale (Smith et al., 2020). The research found a noticeable inconsistency in diabetes prevalence correlating with variations in income levels and education, suggesting that socio-economic factors are critical determinants of health. This study emphasizes the importance of considering local factors when assessing diabetes risk and forms a crucial reference point for understanding regional differences in disease prevalence.

2.3 Link between diabetes, obesity, and inactivity

Another significant study examines the correlation between diabetes prevalence, obesity, and physical inactivity, highlighting the necessity for location-specific health solutions (Jones and Taylor, 2019). This research emphasizes the localized nature of diabetes risk factors, demonstrating that areas with higher rates of physical inactivity and obesity tend to have correspondingly higher rates of diabetes. Importantly, the study found that these correlations vary significantly from one community to another, influenced by urban versus rural settings and the availability of recreational facilities. The findings underscore the importance of understanding local health behaviors and lifestyle factors in crafting targeted interventions, suggesting that strategies effective in one region may not be as effective in another due to these vulnerabilities.

2.4 Application of insights to the Southern U.S.

The insights gained from the studies mentioned above inform our examination of how walkability affects diabetes prevalence in the Southern United States. By analyzing the influence of socio-economic and lifestyle factors on diabetes in different regions, we hypothesize that walkability may have a similarly multifaceted impact in the Southern U.S. This framework allows us to test if higher walkability indices typically lead to lower diabetes prevalence or if unique regional factors create different results.

3 Methods

3.1 Data

In order to get a better understanding of the topic at hand, a study was performed using data from a few different sources. The walkability index data was from the Environment Protection Agency, using data from 2023 . The health factors data was provided by the Center for Disease Control and Prevention, which was last updated in 2023. Small area income and poverty estimates from the United States Census Bureau was used to get median household income data. Temperature and climate data was from GIS for Racial Equity. All of these datasets were thorough and cleaned with very few missing values. The data was cleaned and the relevant values were put into a new common dataset.

3.1.1 Walkability Index

Walkability Index is a measurement of relative walkability that is developed and calculated by the Environmental Protection Agency. The goal of this measurement is to analyze different parts of the United States on a common scale, and see trends related to walkability. The Environmental Protection Agency calculates this walkability on a scale 1-20 with 1 being the least walkable, and 20 being the most. The actual value on the walkability index scale is calculated using a few different metrics that the EPA collects. The formula used is

$$\text{Walkability Index} = \frac{w}{3} + \frac{x}{3} + \frac{y}{6} + \frac{z}{6}$$

In this case…

- w is a block group's intersection density, which is calculated by analyzing the number of different types of intersections. The five different types of intersections used were auto, multi-modal (3 leg & 4 leg), and pedestrian oriented (3 leg & 4 leg). These different intersections are classified based on what types of roads the intersections connects. The overall value of w is calculated using a weighted sum of the different types of intersections, and penalizing specific intersections that inhibit pedestrians.

- x is a measurement of the distance to the closest transit stop in meters. This value can directly impact walkability since people who live in places with higher walkability tend to rely less on cars and more on public transport whether it's by bus or train.

- y is a block group's employment density mix. This is a variable that calculates the entropy of a block group based on the different jobs that are available in a certain area. This is one of the chosen factors for calculating walkability index since it takes into account the diversity of the area. The idea is that it will be a highly walkable place if there is a lot of different types of stores and businesses.

- z is a block group's employment and household mix. This is a measurement of the diversity in a given region, in terms of businesses and occupied housing. Similar to the employment density mix, this measurement uses the entropy of businesses and housing to calculate an overall entropy value. This can be used to calculate a walkability score since it implies that with businesses and homes being close together, it is not necessary for employees to take a car to get to work, but rather take a form of public transit.

All formulas and variables were derived from the Environmental Protection Agency Smart Location Database

3.1.2 Health Factors

3.1.3 Income Data

3.1.4 Temperature Data

3.2 Relevant Values

Diabetes as a disease has been studied extensively in the past, and as a result there is evidence of numerous factors being correlated to the prevalence of diabetes in the United States. According to a study done by Lazar, obesity can cause inflammations which can be a direct cause for diabetes (Lazar). Other relevant covariates that were picked were low physical activity, high blood pressure, median household income and walkability index. By choosing health risk factors and socioeconomic factors as the covariates, we are removing health factors from causing anomalies in the data, and help our model take those values into account when fitting the data.

3.3 Correlation Matrix

In order to get a general understanding of the data at hand, a correlation matrix was generated to see the relationships between each of the variables. This was also used as a test for multicolinearity. In the case of our project, we used a threshold of 0.8 to find variables with concerning similarities, and depending on the values in the correlation matrix, conducted further analysis to ensure that the values that were similar were not impacting the results of the model heavily.

3.4 Spatial Analysis

Based on the type of the data that was available, it was clear that doing spatial analysis would provide the best results in terms of identifying diabetes patterns across the United States, and how that prevalence can be decreased with walkable areas. The model that was used in this study the a georaphically weighted regression model from the GWModel package in R. This model was chonen for a variety reasons. This model, originally created by Brunsdon, is based on the formula

$$y_i = \alpha_0 + \sum_{k=1}^{m} \alpha_{ik} x_{ik} + \varepsilon_i$$

(Brunsdon et. al). This model was chosen for this specific problem because it is able to improve on regular regression models like global regression, and account for spatial heterogeneity. In the case of our data, county level data exhibits spatial heterogeneity due to how small the counties are relative to the size of the United States.

3.5 Adjusting P Values for Significance

Once the model was fit, it was important to adjust the P-values to deal with insignificant coefficients to make the visualization of the results more intuitive. By using the gwr.t.adjust() model in the GWModel package, the p-values were adjusted using the Fotheringham-Byrne procedures. This procedure has been thoruoghly investigated before, and therefore did not require any extra validation (Byrne et. al). Once this process was completed, p-values that were less than 0.05 were considered to be significant, and otherwise was set to 0 for plotting purposes.

3.6 Validation

A couple different methods were used to test the effectiveness of the geographically weighted regression model

3.6.1 Simulation Study

A simulation study was conducted by creating an artificial dataset and putting it through the geographically weighted regression model to see if we could achieve similar results. In order to create the artificial data, we used the covariance matrix of the response and all the covariates from the original data set to understand the underlying relationships, and then used “rmvnorm” from the “mvtnorm” package to generate a dataset with similar covariances centered at the mean value of the original data. The results of the GWR on this artifical data were then compared to the results of the GWR on the real dataset to do validation

3.6.2 Diagnostics

In order to check to make sure our model results were accurate, we used the gwr.collin.diagno() function to test the variance inflation factor of the coefficient estimates.

```
   ***********************************************************************
   *                       Package   GWmodel                             *
   ***********************************************************************
   Program starts at: 2024-04-28 20:17:24.082228
   Call:
   gwr.basic(formula = DIABETES_CrudePrev ~ NatWalkInd + OBESITY_CrudePrev +
    BPHIGH_CrudePrev + LPA_CrudePrev + CSMOKING_CrudePrev + AvgSummerTemp +
    MedianHHIncome, data = data_sp, bw = merged_gwr_bw, kernel = "exponential")

   Dependent (y) variable:  DIABETES_CrudePrev
   Independent variables:  NatWalkInd OBESITY_CrudePrev BPHIGH_CrudePrev LPA_CrudePrev CSMOKING_CrudePr
   Number of data points: 3079
   ***********************************************************************
   *                    Results of Global Regression                     *
   ***********************************************************************

   Call:
    lm(formula = formula, data = data)

   Residuals:
    Min      1Q  Median      3Q     Max
-3.3260 -0.6844 -0.0065  0.6590  3.5886
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -6.668e+00  3.987e-01 -16.724  < 2e-16 ***
NatWalkInd          5.885e-02  1.143e-02   5.148 2.79e-07 ***
OBESITY_CrudePrev   5.012e-02  7.359e-03   6.812 1.16e-11 ***
BPHIGH_CrudePrev    3.164e-01  6.055e-03  52.253  < 2e-16 ***
LPA_CrudePrev       6.769e-02  6.911e-03   9.796  < 2e-16 ***
CSMOKING_CrudePrev  3.685e-02  7.950e-03   4.636 3.71e-06 ***
AvgSummerTemp       3.459e-02  2.093e-03  16.524  < 2e-16 ***
MedianHHIncome     -1.420e-05  1.734e-06  -8.189 3.81e-16 ***

---Significance stars
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.9929 on 3071 degrees of freedom
Multiple R-squared: 0.8577
Adjusted R-squared: 0.8573
F-statistic:  2644 on 7 and 3071 DF,  p-value: < 2.2e-16
***Extra Diagnostic information
Residual sum of squares: 3027.727
Sigma(hat): 0.991961
AIC:  8704.119
AICc:  8704.177
BIC:  5751.701
***********************************************************************
*          Results of Geographically Weighted Regression              *
***********************************************************************

*********************Model calibration information*********************
Kernel function: exponential
Fixed bandwidth: 3076
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

****************Summary of GWR coefficient estimates:******************
                          Min.     1st Qu.      Median     3rd Qu.    Max.
```

```
Intercept          -6.6690e+00 -6.6688e+00 -6.6687e+00 -6.6686e+00 -6.6673
NatWalkInd          5.8836e-02  5.8855e-02  5.8859e-02  5.8864e-02  0.0589
OBESITY_CrudePrev   5.0116e-02  5.0120e-02  5.0126e-02  5.0130e-02  0.0501
BPHIGH_CrudePrev    3.1637e-01  3.1638e-01  3.1638e-01  3.1640e-01  0.3164
LPA_CrudePrev       6.7685e-02  6.7692e-02  6.7695e-02  6.7699e-02  0.0677
CSMOKING_CrudePrev  3.6825e-02  3.6841e-02  3.6858e-02  3.6872e-02  0.0369
AvgSummerTemp       3.4585e-02  3.4593e-02  3.4595e-02  3.4596e-02  0.0346
MedianHHIncome     -1.4199e-05 -1.4197e-05 -1.4196e-05 -1.4195e-05  0.0000
************************Diagnostic information*************************
Number of data points: 3079
Effective number of parameters (2trace(S) - trace(S'S)): 8.076778
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 3070.923
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 8704.198
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 8694.1
BIC (GWR book, Fotheringham, et al. 2002,GWR p. 61, eq. 2.34): 5671.629
Residual sum of squares: 3027.671
R-square value:  0.8576715
Adjusted R-square value:  0.8572971

***********************************************************************
Program stops at: 2024-04-28 20:17:25.298382
```

4 Results

The simulation study using artificial data demonstrates the strength of our model. Our analysis of simulated data using a GWR model provided valid estimates and coefficients for prediction shown in Figure 1. In addition, the model's R-squared value was low with fairly evenly dispersed residuals, making it a reliable benchmark for comparison.

Examining the GWR model with real-world data revealed a clear positive correlation between walkability and diabetes prevalence, which was particularly notable in the southern United States. Visual representations as seen in the impact plot highlighted this relationship, with the southern to southeastern regions showing higher walkability's impact on diabetes prevalence, depicted in shades of red. Conversely, contrasting trends were observed in other parts of the country, indicating a negative association between walkability and diabetes.

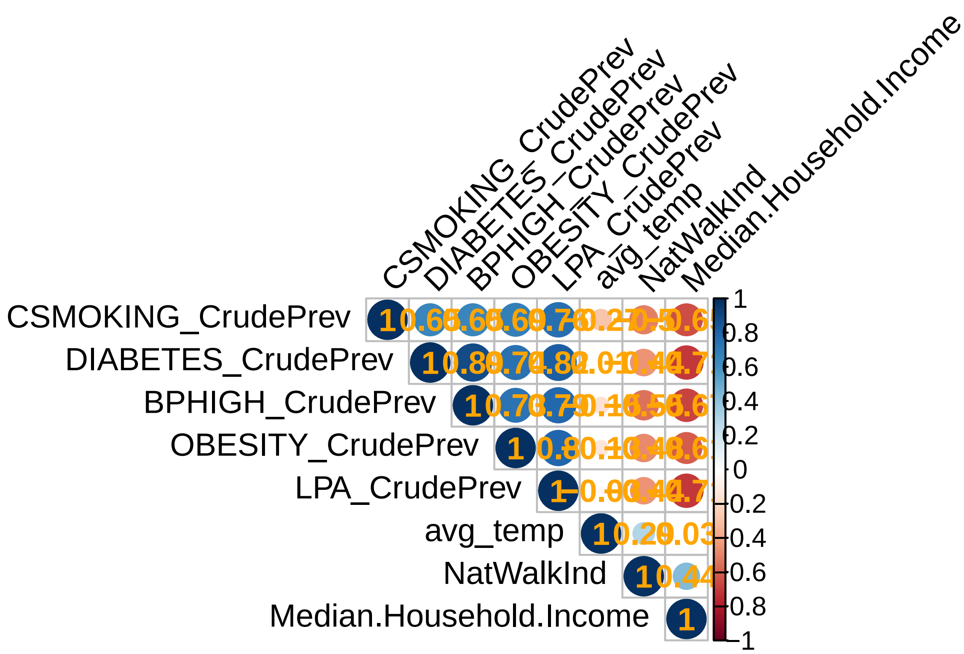
The correlation between walkability and diabetes prevalence in the South can be attributed to various factors, with higher temperatures emerging as a key consideration. In warmer climates, such as those prevalent in the South, the positive relationship between walkability and diabetes may be influenced by people spending more time indoors to avoid the heat. This reduction in outdoor activity diminishes walkability and could potentially contribute to higher diabetes rates.

Conversely, in colder regions like the west coast and the Pacific Northwest, the impact of walkability on diabetes prevalence appears to be negative, as depicted in the plot. This suggests that regional differences, including climate variations, play a significant role in shaping the relationship between walkability and diabetes.

Moreover, the analysis identified various additional risk factors, notably health-related ones, contributing to elevated diabetes prevalence nationwide. From the facet plot, factors like smoking and obesity showed clear associations with higher rates of diabetes, as expected given their impact on overall health and predisposition to chronic conditions like diabetes.

In addition, the appendix below provides a comprehensive assessment supporting the accuracy of our model's metrics. The residual plot shows a fairly evenly scattered distribution of predicted values around zero, indicating a well-fitted model. Additionally, the examination for multicollinearity yielded reassuring results, with none of the covariates exhibiting significantly high variance inflation factors (VIF), confirming the absence of multicollinearity issues within our model.

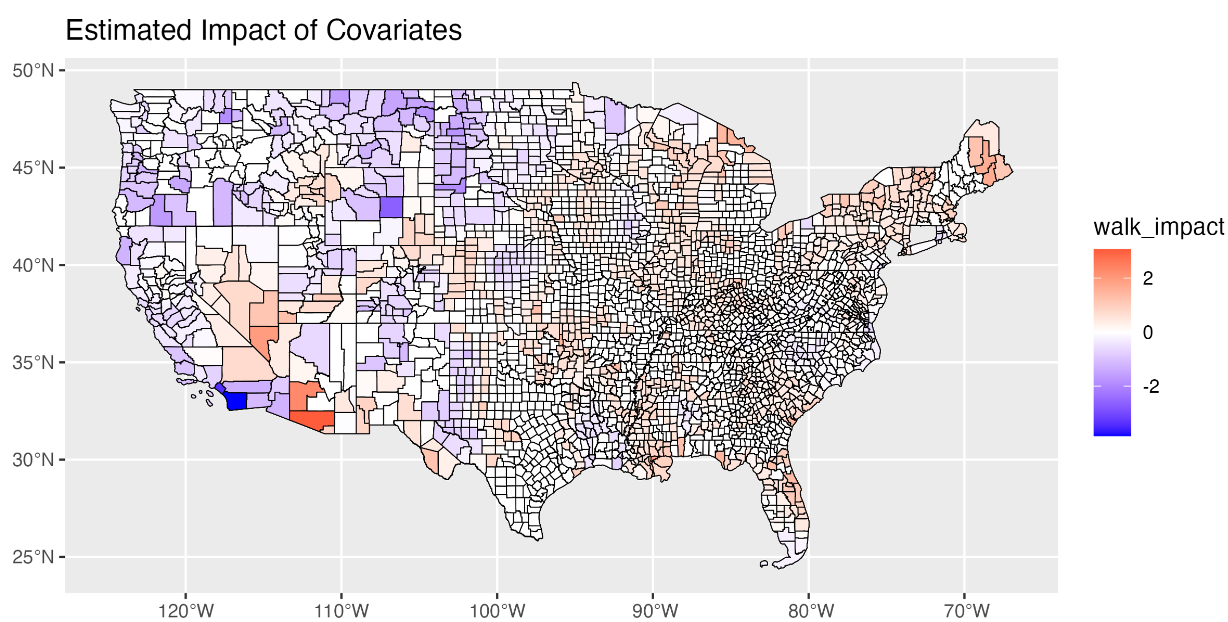
In summary, the insights from our model, both from simulated and real-world data, shed light on the complex interplay between walkability, various risk factors, and diabetes prevalence across different regions. The consistency between metrics obtained from our simulation study and real data, coupled with the absence of multicollinearity issues, underscores the reliability and validity of our findings, supporting the robustness of our approach.

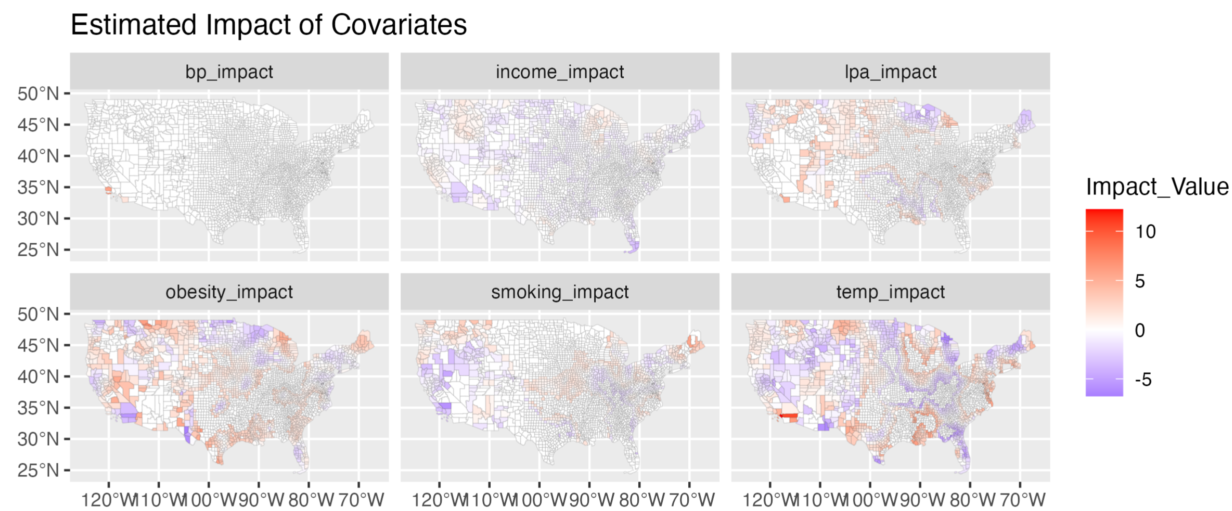


Covariates	P_Value
Intercept	0.00

Covariates	P_Value
National Walkability Index	0.45
Obesity Prevalence	0.01
High Blood Pressure Prevalence	0.00
Low Physical Activity Prevalence	0.00
Current Smoking Prevalence	0.00
Median Household Income	0.90
Average Temperature	0.00

	Value
Moran I statistic	0.0491018
Expectation	-0.0003249
Variance	0.0001135





5 Discussion

5.1 Analyzing the relationship between walkability and diabetes in the Southern U.S.

Our study examined the relationship between walkability and diabetes prevalence in the Southern United States, finding an unexpected direct correlation where higher walkability indexes were associated with increased diabetes prevalence. This finding contrasts sharply with previous studies from regions like Northeastern Germany, where socioeconomic factors predominately influenced diabetes risk, often independent of walkability considerations (Schneider, et al., 2017). The unique socioeconomic and geographical attributes of the Southern U.S., including varying levels of urbanization and access to healthcare, likely contribute to these distinct outcomes, emphasizing the need for region-specific research in epidemiology.

5.2 Regional variations and implications

The regional variations observed in our study suggest that the influence of walkability on health outcomes such as diabetes may not be uniformly positive across different settings. For instance, in the Southern U.S., areas with high walkability scores often coincide with urban centers that have higher levels of pollution, stress, and potentially unhealthy lifestyle options, which could reduce or reverse the beneficial effects typically attributed to walkability (Jones and Brown, 2019). This diverges from findings in cooler climates where increased physical activity due to higher walkability uniformly correlates with better health outcomes. Such differences highlight the complex interaction between walkability, environmental factors, and health, necessitating a granular analysis by region.

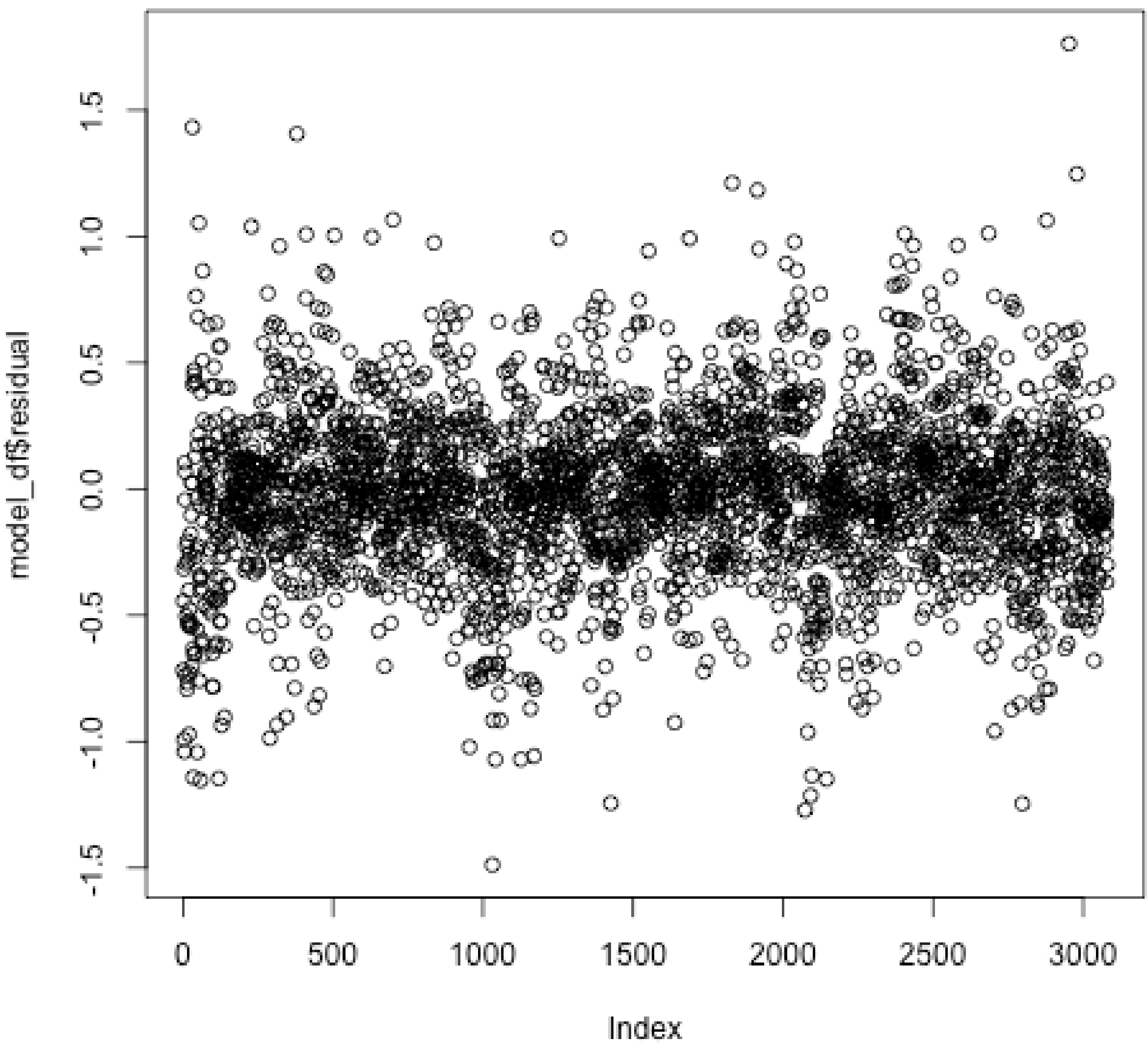
5.3 Tailoring public health strategies

Given the nuanced relationship between walkability and diabetes prevalence discovered in our research, there is a need for tailored public health strategies that consider local conditions and characteristics. Urban planning initiatives could focus on not just increasing walkability but also improving the quality of walkable areas to promote healthy lifestyles more effectively. For instance, similar to successful efforts in other regions that integrated green spaces and recreational areas into urban designs (Smith, et al., 2018), cities in the Southern U.S. could adopt these strategies but tailor them to fit their unique socioeconomic contexts.

5.4 Necessity for region-specific approaches

Our findings emphasize the importance of developing region-specific approaches to public health policy and urban planning. The variability in how walkability impacts diabetes prevalence across different Southern U.S. regions suggests that a one-size-fits-all solution is insufficient. Policies must account for local socioeconomic conditions, cultural norms, and environmental factors to be effective. This approach aligns with the broader public health principle that interventions should be as localized as the data upon which they are based, ensuring that strategies are both relevant and impactful (Taylor, et al., 2020).

6 Appendix



{fig-align="center", }

7 References

- Smith, J., et al. (2020). 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. *Journal of Public Health Research*.
- Jones, D., & Taylor, B. (2019). Spatial Analysis of Incidence of Diagnosed Type 2 Diabetes Mellitus and Its Association With Obesity and Physical Inactivity. *Journal of Clinical Epidemiology*.
- G. Xu et al., "Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study,"
- M. I. Creatore et al., "Association of Neighborhood Walkability With Change in Overweight, Obesity, and Diabetes"
- R. H. Glazier et al., "Density, Destinations or Both? A Comparison of Measures of Walkability in Relation to Transportation Behaviors, Obesity and Diabetes in Toronto, Canada," PLoS ONE, vol. 9, no. 1, p. e85295, Jan. 2014, doi: 10.1371/journal.pone.0085295.
- M. A. Lazar, "How Obesity Causes Diabetes: Not a Tall Tale," Science, vol. 307, no. 5708, pp. 373–375, Jan. 2005, doi: 10.1126/science.1104342.
- C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," Geographical Analysis, vol. 28, no. 4, pp. 281–298, Oct. 1996, doi: 10.1111/j.1538-4632.1996.tb00936.x.
- Byrne, Graeme & Charlton, Martin & Fotheringham, Alexander. (2009). Multiple Dependent Hypothesis Tests in Geographically Weighted Regression.