Obesity prevalence estimation in the 500 US cities at the neighborhood level

Ryan Zhenqi Zhou (zhenqizh), Prathamesh Ravindra Varhadpande (pvarhadp), and Nikhil Gokhale (gokhale7)

1. Introduction

Obesity is a serious public health problem. In the USA, nearly 42.4% of the U.S. adult population is considered overweight or obese (Hales, 2020). The estimated annual medical cost of obesity ranges from $147 billion to nearly $210 billion per year (Cawley and Meyerhoefer, 2012; Finkelstein et al., 2009). Obesity can lead to various physical and mental consequences affecting people across their lifespan, including heart disease, type 2 diabetes, sleep apnea, and depressive disorder (Akil and Ahmad, 2011; Robert and Reither, 2004). Given its substantial costs, reducing obesity is a critical task for public health policymakers.

Existing research has shown a strong association between obesity and an individual's socioeconomic and demographic status. These factors include race/ethnicity, gender, age, education, occupation, and income (Ball et al., 2002; Cui et al., 2021; Puciato and Rozpara, 2020; Shaikh et al., 2015; Wang and Beydoun, 2007). The methods used for collecting individual-level data were typically surveys and interviews. For example, Shaikh et al. (2015) used data from the 2010 US National Health Interview Survey with 23434 records to study the socio-demographic of obesity in the United States. Puciato and Rozpara (2020) studied the demographic and socioeconomic factors associated with obesity involving 4315 respondents (2206 women and 2109 men) aged 18-64 years from Wroclaw. The sample selection was random and purposive by using a multilevel stratification. Cui et al. (2021) presented an estimation of obesity levels based on socio-demographic and other factors. The dataset they used is based on the life routine and body condition of different volunteers, which was collected using a web platform with a survey. These surveys and interviews provide highly valuable individual-level insights about obesity and its associated risk factors. Meanwhile, they cost considerable financial and labor resources, especially when they are conducted for a large geographic area.

If we "zoom out" from individuals to their neighborhoods, measurements about the general neighborhood-level socioeconomic and demographic status of the residents may help us identify those neighborhoods that are more likely to have obesity issues (Fitzpatrick et al., 2018). This has an important meaning since obesity prevention programs could then put more focus on these neighborhoods, e.g., by distributing more educational materials promoting healthy lifestyles or further investigating the underlying factors that may have caused obesity in these neighborhoods.

In this project, we try to fill this gap: estimating the obesity prevalence at the neighborhood level. The objective of this study is twofold. First, we collect socioeconomic and demographic data at the neighborhood level related to obesity following existing literature. Second, we investigate the research question: *how well the socioeconomic and demographic variables can estimate the obesity prevalence, and which factors are associated with the obesity prevalence*. We experiment with three different statistical and machine learning (ML) models, including one spatially explicit model, to examine the effectiveness of the derived measurements for obesity estimation.

2. Methods

2.1. Study area and data source

This study focuses on the 500 largest cities in the US, such as New York City (NYC), Los Angeles (LA), and Buffalo (BUF). The time period of our study is the year 2021, and the geographic unit of analysis is census tract which is roughly comparable to neighborhoods. We choose this study area, this time period and this geographic unit largely due to data availability: the obesity data used in this study is from the

PLACES project of the Centers for Disease Control and Prevention (CDC), whose latest data contains 500 largest cities in the US in the year 2021 and the smallest geographic unit is census tract (Centers for Disease Control and Prevention, 2021). The socioeconomic and demographic variables are from the American Community Survey (ACS) of the US Census Bureau.

2.2. Data collection

The outcome variable that we focus on in this study is neighborhood-level obesity prevalence rate. We obtained the census tract-level obesity prevalence among adults (age >= 18) data from the CDC PLACES Project, and the obesity rates are recorded in percentages (e.g., a value of 26.6 indicates the obesity rate for that census tract is 26.6%).

In this study, we aim to understand how well the socioeconomic and demographic variables can estimate the obesity prevalence, and which factors are associated with the obesity prevalence. Learning from the literature, we select variables in five categories: (1) race and ethnicity, (2) gender, marital status, and age, (3) education, (4) economic status, and (5) housing condition. Table 1 presents the detailed notations and descriptions of these variables. Note that in the age category, we do not include % age <18 because the obesity data from CDC do not include age below 18 years old. We obtained data for these variables from the American Community Survey (ACS) of the US Census Bureau.

Table 1. Notations and descriptions of the five categories of neighborhood-level variables.

Variable notations	**Descriptions**
(1) Race and ethnicity	
% White	Percentage of population in White
% Black	Percentage of population in Black or African American
% Ame Indi and AK Native	Percentage of population in American Indian and Alaska Native
% Asian	Percentage of population in Asian
% Nati Hawa and Paci Island	Percentage of population in Native Hawaiian and Other Pacific Islander
% Hispanic or Latino	Percentage of Hispanic or Latino population
(2) Gender, marital status, and age	
% male	Percentage of male population
% married	Percentage of married population age 15 or over
% age 18-29	Percentage of population between age 18 to 29
% age 30-39	Percentage of population between age 30 to 39
% age 40-49	Percentage of population between age 40 to 49
% age 50-59	Percentage of population between age 50 to 59
% age >=60	Percentage of population equal and over age 60
(3) Education	
% <highschool	Percentage of population age 25 or over without high school completion
% >=highschool <university	Percentage of population age 25 or over with high school completion and without bachelor degree
% >=university	Percentage of population age 25 or over with bachelor degree or higher degree
(4) Economic status	
med income	Median household income
% unemployment	Percentage of unemployed labor force population age 16 or over
% below poverty line	Percentage of population below poverty line

% food stamp/SNAP	Percentage of households received food stamp/supplemental nutrition assistance program (SNAP) in the past 12 months
(5) Housing condition	
median value units built	Median value of the house units built (owner-occupied housing units)
median year units built	Median year of the house units built
% renter-occupied housing units	Percentage of renter-occupied housing units

2.3. Multicollinearity diagnosis

After we finish collecting data, we first carry out a series of diagnostic tests to examine whether there exists multicollinearity among the neighborhood-level socioeconomic and demographic variables. To do so, we compute the variance inflation factor (VIF) for the 23 variables, and all variables are standardized by their mean and standard deviation before the analyses. We then gradually remove the variables with the highest VIF values until they are all smaller than the typical cut-off value 5.

2.4. Statistical and machine learning models

We experiment with three different statistical and ML models, namely ordinary least squares (OLS), geographically weighted regression (GWR), and random forest (RF), to examine the effectiveness of the derived variables as a predictor. We use coefficients from statistical models and feature importance from ML models to identify factors associated with the obesity prevalence. Among the four, GWR is spatially explicit models, which can help accommodate spatial heterogeneity typically existing in geographic data (Brunsdon et al., 1998; Georganos et al., 2021). In the following, we briefly describe each model.

Ordinary Least Squares: OLS is a statistical model of analysis that estimates the relationship between multiple input independent variables and the target outcome variable. The OLS model used in this work is in the form of Equation (1):

$$\text{Obesity rate} = \theta_0 + \theta_r r + \theta_a a + \theta_s s + \theta_e e + \theta_h h + \theta_u u + \varepsilon \qquad , \qquad (1)$$

where θ_{r}, θ_{a}, θ_{s}, θ_{e}, θ_{h} , θ_{u} are the coefficients for the five categories of socioeconomic and demographic variables respectively. Note that each of θ_{r}, θ_{a}, θ_{s}, θ_{e}, θ_{h} , θ_{u} contains multiple coefficients for the variables in that category.

Geographically Weighted Regression: GWR fits local OLS models for each geographic unit (i.e, census tract in this study) by taking into account spatial dependence and spatial heterogeneity. Specifically, the GWR model used in this work is in the form of Equation (2):

$$\begin{aligned}\text{Obesity rate} = {} & \theta_0(x_i, y_i) + \theta_r(x_i, y_i)r + \theta_a(x_i, y_i)a + \theta_s(x_i, y_i)s + \theta_e(x_i, y_i)e + \\ & \theta_h(x_i, y_i)h + \theta_u(x_i, y_i)u + \varepsilon_i \qquad , \qquad (2)\end{aligned}$$

where (x_{i}, y_{i}) is the spatial coordinates of the geographic unit *i*. The coefficients have the same meaning as used in OLS, but will vary across different geographic locations capturing the heterogenous local processes.

Random Forest: Random forest is a bagging-based machine learning model that applies an ensemble learning technique by constructing a group of decision trees. Compared with OLS that assumes a linear relation, RF can model nonlinear relations between input features and the target variable. Given this ability, RF has been used in a variety of previous studies in which the input features and the target variable likely have a nonlinear relation.

2.5. Model evaluation

Two goodness-of-fit measures, R^2 and root mean square error (RMSE), are utilized for assessing the performance of the four models. For the statistical models, namely OLS and GWR, their R^2 and RMSE can be directly obtained from the model fitting results. For the machine learning models, namely RF, its R^2 and RMSE are obtained via a 10-fold cross-validation process, in which the data are divided into 10 non-overlapping folds, and the training and test procedure are iterated 10 times. In each iteration, one of the 10-

fold data is held out as the test data, and the remaining 9-fold data are used for training the model. The mean R^2 of the ten iterations is reported, and the RMSE is calculated by pooling the prediction residuals from the ten iterations. Note that the same random seed is used in the two sets of experiments to separate the data into the exact same 10 folds. Such a 10-fold cross-validation process can help obtain more robust evaluation results and avoid potential biases caused by one particular validation dataset.

3. Results

3.1. Summary of obesity rate

The mean and median value of obesity rate are 34.4% and 34.6%, respectively. Figure 1 shows the box plot and histogram of the obesity rate. As you can see, overall, the obesity rate is similar to normally distributed.

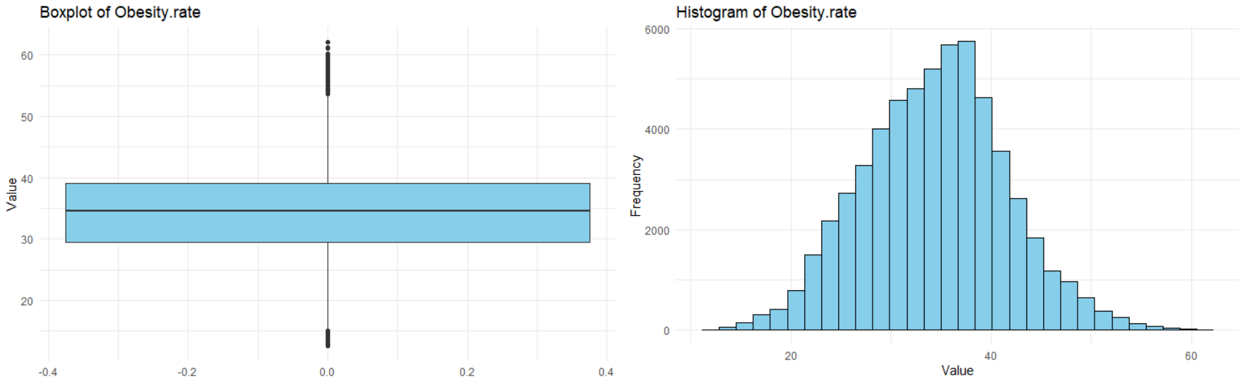

Figure 1. Boxplot and histogram of the obesity rate

3.2. Multicollinearity diagnosis

Before conducting models, we first carry out a series of diagnostic tests to examine whether there exists multicollinearity among the neighborhood-level socioeconomic and demographic variables. To do so, we compute the variance inflation factor (VIF) for the 23 variables, and all variables are standardized by their mean and standard deviation before the analyses. We then gradually remove the variables with the highest VIF values until they are all smaller than the typical cut-of value 5. Based on the result, we removed *% >=university* and *% white* variables.

3.3. Metric results from the three statistical and machine learning models

Table 1 summarizes the R^2 and RMSE of OLS, GWR, and RF for predicting the obesity rate. RF model performs best with R^2 (0.961) and RMSE (1.433), and significantly better than the other two models. For two statistical models OLS and GWR, GWR perform slightly better than OLS, with R^2 increases from 0.768 to 0.770 and RMSE decreases from 3.431 to 3.415.

Table 1. A summary of the metrics for predicting the obesity rate

Fit measures	**OLS**	**GWR**	**RF**
R^2	0.768	0.770	0.961
RMSE	3.431	3.415	1.433

3.4. Coefficients from OLS and GWR

We present the regression coefficients output by the OLS and GWR model in order to understand the roles played by different independent variables in predicting the neighborhood-level obesity rate. As you can see Figure 2, overall, the coefficients of OLS and GWR are similar. *% >=highschool <university*, *%*

<*highschool*, *% Black*, and *% below poverty line* show strongly positive correlation with obesity rate. *median value units built*, *% Asian*, *% age 18-29*, *% age >=60*, and *% Hispanic or Latino* show strongly negative correlation with obesity rate.

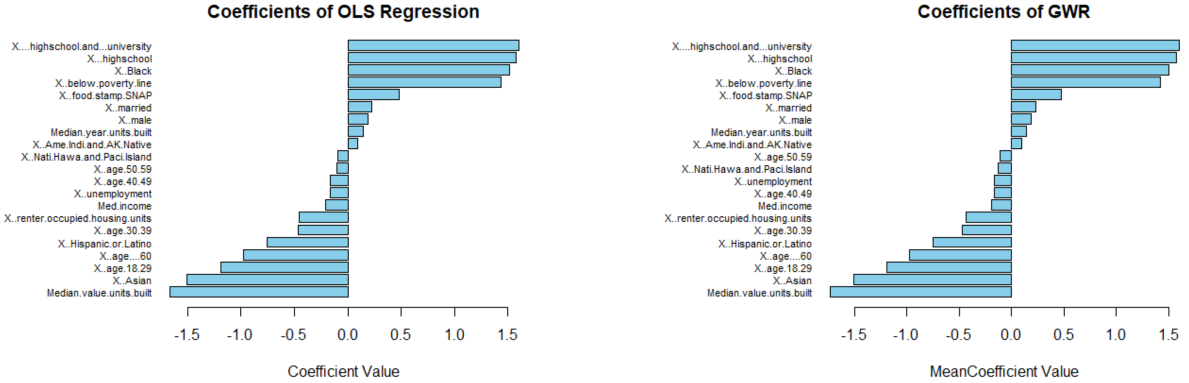

Figure 2. Coefficients from OLS and GWR for predicting the obesity rate

3.5. Feature importance from RF

The RF model provides feature importance values indicating the relative importance of different input variables for helping the RF model predict neighborhood-level obesity rate. Figure 3 shows the feature importance values. Since we have used 10-fold cross-validation, 10 RF models are trained which result in 10 sets of feature importance values. Figure 3 shows the mean importance value for each variable. As you can see, *median value units built*, *% >=highschool <university*, *med income*, *% Black*, *% Asian*, and *% food stamp/SNAP* play important roles in helping the RF model predict the obesity rate.

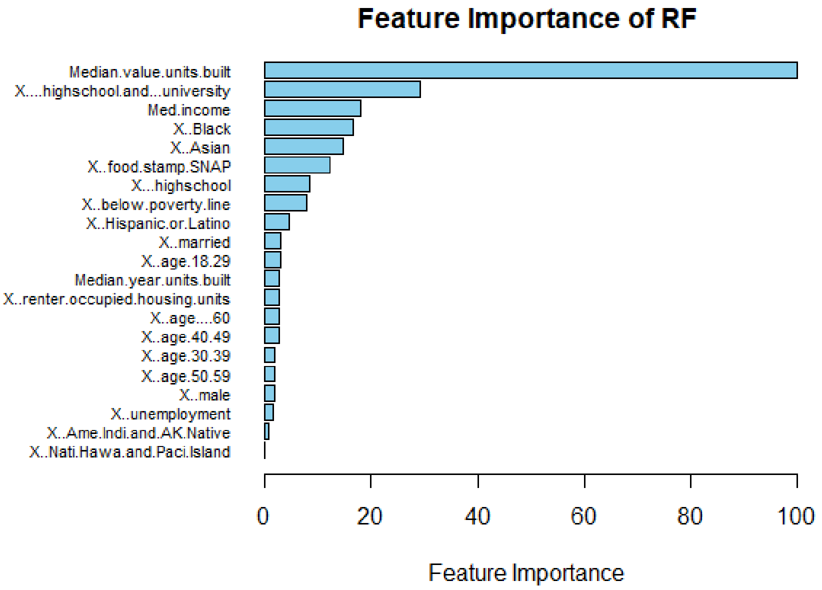

Figure 3. Feature importance from RF for predicting the obesity rate

4. Summary of contributions

This work has two contributions. First, we collect socioeconomic and demographic variables related to obesity at the neighborhood level in large geographic scale studies. Second, we conduct experiments to understand to what extent the derived variables can estimate obesity prevalence at the neighborhood level. We utilize three different statistical and ML models, including one spatially explicit model, to understand the role of the socioeconomic and demographic factors for obesity estimation across the USA. Note that the risks and payoffs are based on the data quality and the right matter about the model conduction (e.g., control randomness of models).

5. Project Timeline

Task	*4/1/2024 - 5/12/2024 (6 weeks)*					
	1	***2***	***3***	***4***	***5***	***6***
Project plan proposal	zhenqizh	zhenqizh				
Collect data			Pvarhadp gokhale7 zhenqizh			
Progress report				zhenqizh		
Conduct models				Pvarhadp gokhale7 zhenqizh	Pvarhadp gokhale7 zhenqizh	Pvarhadp gokhale7 zhenqizh
Final report and poster						zhenqizh

All team members have contributed a similar amount of effort.

References

Akil, L., Ahmad, H.A., 2011. Effects of socioeconomic factors on obesity rates in four southern states and Colorado. Ethn. Dis. 21, 58.

Ball, K., Mishra, G., Crawford, D., 2002. Which aspects of socioeconomic status are related to obesity among men and women? Int. J. Obes. 26, 559–565.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. Ser. Stat. 47, 431–443.

Cawley, J., Meyerhoefer, C., 2012. The medical care costs of obesity: an instrumental variables approach. J. Health Econ. 31, 219–230.

Cui, T., Chen, Y., Wang, J., Deng, H., Huang, Y., 2021. Estimation of Obesity Levels Based on Decision Trees, in: 2021 International Symposium on Artificial Intelligence and Its Application on Media (ISAIAM). Presented at the 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM), IEEE, Xi'an, China, pp. 160–165. https://doi.org/10.1109/ISAIAM53259.2021.00041

Finkelstein, E.A., Trogdon, J.G., Cohen, J.W., Dietz, W., 2009. Annual Medical Spending Attributable To Obesity: Payer-And Service-Specific Estimates: Amid calls for health reform, real cost savings are more likely to be achieved through reducing obesity and related risk factors. Health Aff. (Millwood) 28, w822–w831.

Fitzpatrick, K.M., Shi, X., Willis, D., Niemeier, J., 2018. Obesity and place: chronic disease in the 500 largest US cities. Obes. Res. Clin. Pract. 12, 421–425.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2021. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 36, 121–136.

Hales, C.M., 2020. Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017–2018 8.

PLACES: Census Tract Data (GIS Friendly Format), 2021 release | Chronic Disease and Health Promotion Data & Indicators [WWW Document], n.d. URL https://chronicdata.cdc.gov/500-

Cities-Places/PLACES-Census-Tract-Data-GIS-Friendly-Format-2021-/yjkw-uj5s (accessed 5.10.22).

Puciato, D., Rozpara, M., 2020. Demographic and Socioeconomic Determinants of Body Mass Index in People of Working Age. Int. J. Environ. Res. Public. Health 17, 8168.

Robert, S.A., Reither, E.N., 2004. A multilevel analysis of race, community disadvantage, and body mass index among adults in the US. Soc. Sci. Med. 59, 2421–2434.

Shaikh, R.A., Siahpush, M., Singh, G.K., Tibbits, M., 2015. Socioeconomic status, smoking, alcohol use, physical activity, and dietary behavior as determinants of obesity and body mass index in the United States: findings from the National Health Interview Survey. Int. J. MCH AIDS 4, 22.

Wang, Y., Beydoun, M.A., 2007. The Obesity Epidemic in the United States Gender, Age, Socioeconomic, Racial/Ethnic, and Geographic Characteristics: A Systematic Review and Meta-Regression Analysis. Epidemiol. Rev. 29, 6–28. https://doi.org/10.1093/epirev/mxm007