The Roots of Obesity

*Geographical Weighted Regression as a tool to analyze spatially heterogeneous socio-environmental factors affecting adult obesity tendencies in Texas.*

José Vicente Latorre | UT SoA, MSCR + MSSD Candidate

Fall 2014 | Introduction to Geographic Information Systems | Dec. 11, 2014
Executive Summary

The following study attempts to make use of Geographical Weighted Regression in an analysis of socio-environmental variables affecting adult obesity prevalence in Texas at the county level. The socio-environmental variables taken into consideration are: existing recreational facilities, median household income, educational attainment below high school level, age from 25 to 44 years old, and race. The hypothesis is that the GWR will replicate the socio-environmental tendencies expected of the variables affecting adult obesity at county level in Texas per literature review suggestions. Additionally the study aims at comparing the performance of the GWR model against a Multiple Linear Regression, the GWR model is expected to outperform the MLR.

The results of the GWR clearly outperformed the MLR. Coefficients of the independent variables in the GWR demonstrated considerable variations from the coefficients in the MLR. The variation in obesity prevalence can be better explained by analyzing spatially heterogeneous relationships across variables. However, not all variables behaved as expected, in fact, race yielded a positive correlation between white percentage of population and higher adult obesity prevalence. The data scale of data may be an important factor causing this unexpected correlation; this issue is discussed in the final section of the paper.
Introduction

National Scale

Obesity has evolved into a major public health concern. In the US, obesity is considered a pandemic – 35.7% of adults and 16.9% of children are reportedly obese (Ogden, Carol, Kit, & Flegal, 2010). We have seen an increase in media attention towards the issue, according to the US Center for Disease Control and Prevention (CDC), data from 2012 reveals that more than one-third of the US adult population has obesity (Ogden et al., 2014). If trends persist, within a few years 75% of the population will be overweight, over 40% will be obese (Xu, & Wang, 2014). Refer to the US timeline map layout in the next page.

In the midst of the healthcare reform controversies, obesity has been widely discussed for the medical costs associated with this pandemic: in 2012, obesity accounted for almost 21% of U.S. health care costs (Kelley, 2012). Moreover, the individual yearly medical costs of an obese patient are $2,471 higher than if they were not obese (Kelley, 2012). In fact, research finds obesity to be a significant contributing factor to several leading causes of morbidity and mortality including heart disease, stroke and diabetes (Xu, & Wang, 2014).

Everything is bigger in Texas!

Over the last 20 years, the United States has seen a radical increase in population reportedly obese. Southern states have historically lead obesity prevalence in the US (refer to US maps). In fact, Texas overweight and obesity prevalence have been higher than the national average since 1996 (CDBR, 2010). In fact, Texas prevalence of overweight and obesity has increased by 28.8 percent from 1995 to 2008 – 7.2% than the national percent increase (CDBR, 2010). If we look at the most recent available data at the state level, according to the Behavioral Risk Factor Surveillance System 2013 (BRFSS), 30.9% of adult Texans are reportedly obese. Refer to the TX timeline maps in the next page.

Many factor contribute to the obesity pandemic in Texas. For one, food access and nutrition can be considered one factor. In fact, according to data from the USDA Food Atlas, on average, Texan’s spent $784 in fast food restaurants in 2007. Only Washington DC, Hawaii, and Nevada had higher average rates of fast food expense. The Texas Chronic Disease Burden Report, April 2010, delineates major demographic factors affecting overweight and obesity prevalence. Texan males, African-Americans, Hispanics, and those under 30 years of age were considered significantly less prevalent to be overweight or obese in the cited report.

A complex phenomenon

Literature widely suggests that obesity is much often a consequence of modifiable behaviors including nutrition patterns and physical activities. Such behavioral qualities are challenging to deduce from population demographics, at-large data. Environmental health risk factors such as socio-economic status, access to healthy food, and physical activity infrastructure have been considered essential by some
OBESITY
A NATIONAL PANDEMIC

Mapping the evolution of state obesity prevalence indexes in the U.S. across time. Obesity threshold defined as Body Mass Index (BMI) >30%.

Maps and Layout by José Vicente Latorre, 2014. These maps are educational, the information sourced from Center for Disease Control and Prevention (CDC), Obesity Trends 1986-2010. Map Projection and Datum follows NAD 1983 Contiguous USA Albers.
EVERYTHING'S BIGGER IN TEXAS

Mapping the evolution of state obesity prevalence indexes in Texas across time. Obesity threshold defined as Body Mass Index (BMI) >30%.

ASSESSING CHANGE IN OBESITY

Mapping the percent change in obesity prevalence index in Texas Counties from 2004 to 2010.

TX OBESITY PREVALENCE INDEX

% CHANGE

0.01 - 10.00
-5 - 0.00
10.01 - 20.00
20.01 - 30.00
30.01 - 40.00
40.01 - 50.00
50.01 - 60.00

Maps and Layout by José Vicente Latorre, 2014 | These maps are educational, the information sourced from Center for Disease Control and Prevention (CDC), Census Bureau, and USDA Food Atlas | Map Projection and Datum follows NAD 1983 Stateplane TX North Central & Lambert Conformal Conic.
authors (Chi, Grisby, Bradford, Choi, 2013, 2013, Hendrikson, Smith, & Eikenberry 2006). Ecological models are increasingly useful for they examine external factors and individual factors of the obesity pandemic (Chi, Grisby, Bradford, Choi, 2013). In order to study obesity tendencies, one must not only focus on demographic population characteristics, but, much rather, combine socio-economic individual factors and built environment factors that may best explain the individual modifiable behaviors that have been found to be a leading cause of obesity. The analysis of factors leading to obesity trends remains a challenging one. Specially considering that researchers have not found substantial evidence in the last decade of any single psychological, biological, or metabolic phenomena that explains the obesity trends nationwide (Summerbell et al., 2005).

**Goodness of fit**

On the face of things, if obesity was only a matter of at-large demographic population tendencies, and not of individual modifiable behaviors, with little to no consistent reasoning behind them, Multiple Linear Regression would be an appropriate tool to identify the most important variables affecting the obesity pandemic. However, as we have previously established, ecological models are much preferred to study obesity tendencies. Consequently ecological models will be closely tied to spatial characteristics and amenities that can vary significantly in space. Global estimation techniques can hide important spatial variations in model parameters and are not able to deal with spatial autocorrelations existing in variables (Lee, & Schuet, 2014). If global estimation techniques, such as multiple linear regressions, fail to encapsulate spatial relationships of variables, they will probably fail to accurately address models related with ecological models of obesity-related variables.

A variety of studies address the spatial weakness of global estimation techniques by contrasting such models with geographic weighted models (Lee, & Schuett, 2014, Xu, & Wang, 2014, Chi, Grisby, Bradford, Choi, 2013). These studies vary in the issue at hand, from obesity, to public park demand, to transit ridership (Cardozo, Garcia, Guitierrez, 20012), but the varying issues are related to variables that are highly susceptible spatial variation. They discard the use of global prediction models because of these spatial variations; instead, they propose the use of Geographic Weighted Regression (GWR).

ArcGIS defines GWR as “a local form of linear regression used to model spatially varying relationships”. GWR is specially fit for regional-level studies because they allow for accounting of local effect and shows geographical variation in the strength of relationships (Ogneva-Himmelberger, 2009). Research has proven GWR to be an effective tool for analyzing obesity in a geographical context (Xu, & Wang, 2014, Chi, Grisby, Bradford, & Choi, 2013, Chalkias et al., 2013). While the geographic scale of the studies vary, nation, region, or community, they all agree on the relevance of socio-environmental variables in the analysis of obesity. Evermore, they rely on GWR to reflect the spatial character of the diverse socio-environmental factors they consider.
Problem Statement

Given the national relevance of the overweight and obesity pandemic and the above-national obesity prevalence in the southern states and more specifically in Texas, this study aims to analyze the socio-environmental factors related to adult obesity at the county level in the state of Texas.

Obesity prevalence-related variables are, as precisely explained, not fixed and are tied to individual modifiable behaviors. Socio-environmental variables have been preferred over at-large population variables to analyze obesity. Socio-environmental variables are tied to spatial relationships and variations, Geographic Weighted Regressions (GWR) are best fit to represent the spatial implications of such variables. This study will contrast the results of multiple linear regression (MLR) and GWR on the same set of socio-environmental variables associated with adult obesity prevalence in Texas and in general, as suggested by literature examined: recreational facilities, household income, educational attainment, and race.

Research Questions

Which regression methodology performs better to describe the relationships of the variables of adult obesity prevalence in Texas, MLR or GWR?

Are the effects of socio-environmental variables on adult obesity consistent across the Texas region?

From the selected variables, what is the hierarchy of significance on obesity prevalence in Texas?

Are there any specific areas within the Texas region that must look closely at the socio-environmental variables selected to improve their planning to mitigate the progress of current adult obesity trends in Texas?
Methodology

The process of selecting the variable for the statistical analysis was the most challenging part of this study. The set of variables affecting adult obesity in this project result from the exploration of the literature and case studies cited in the Introduction. However, before the final set was defined, numerous iterations of MLR were performed in order to deliver a model that had an appropriate goodness of fit – always keeping in mind that regressions on a social science context tend to perform poorer than in other fields because human behavior is hard to predict.

We insist on the complexity of the statistical model setup because it must be a factor to consider in any GWR study. A MLR model must be statistically fit in order for it to be worth performing a GWR, and the independent variable of the model must be a result of conscious selection based on research.

The independent variable of the study is the percentage of adult obesity at the county level in Texas. Adult obesity is defined as population over 25 years of age with a BMI (Body Mass Index), relationship of a person’s body fat derived from height and weight, over 30 (CDC, 2013).

The number of recreational facilities per county represents an independent variable strictly attached to physical infrastructure and spatial character of the county – our purely environmental variable. The other variables are more closely tied to social concerns that have spatial implications. Median Household Income not only provides a clue as for the wealth of a county, but it will also vary spatially across counties in the region. Educational attainment in the form of population with less than a high school diploma was selected due to the evidence on relationship between educational attainment and obesity prevalence. As reported by the Texas Department of Health Services, in their Chronic Disease Burden Report, those older than 18 are more prevalent to be overweight and obese, specially the 30 to 44 age cohort. The study defines the age independent variable as the population percentage of those between the ages of 25 and 44, as allowed by ACS data breakdown. The same report cited Texan African-Americans and Hispanics to be significantly more prevalent to be obese than Whites. We included the percent of population white per county. We decided to include the white race variable to study if there was a significant variance in its coefficients from the MLR and the GWR.

Summary of variables:

**Dependent Variable:** adult obesity prevalence (percent) by county in the state of Texas, 2011.

**Independent Variables:**
- Recreational Facilities: number of recreational facilities per county in 2011.
- Median Household Income: median household income per Texas County.
- Educational Attainment: population with no high school diploma per Texas County.
- Age: percent of population in the 25 to 44 years old age cohort per Texas County.
- Race: percent of non-Hispanic White only population per Texas County.

Data Sources:

Data for obesity prevalence (2004 to 2011) by County in Texas was obtained from the Texas Obesity Prevalence data by the Centers for Disease Control and Prevention. Data for the creation of the US obesity maps was also sourced from the CDC.

The data for the independent variables was obtained from two sources the USDA Food Atlas database for the Recreational Facilities, and Census Bureau of Information American Community Survey (ACS) 2008-2012 for the remaining independent variables of the study: age, race, household income, and educational attainment.

The shapefiles for the creation of the US maps and the TX maps were obtained from the US Census Bureau TIGER/Line Shapefiles database and projected appropriately to the geographic region analyzed.

Analysis Methodology:

- Downloaded variables data from the previously cited sources.
- Sorted through the Excel files containing the downloaded data to extract the relevant variables into a single database to be used throughout the project. Relevant data from the USDA Food Atlas and American Community Survey were consolidated into one file.
- For the US historical obesity prevalence data, the available data had to be transcribed into a new database manually.
- Using the database previously created, conducted a Multiple Linear Regression using SPSS to serve as a global estimation method in the study.
- Time to step into ArcGIS and produce a series of descriptive maps about obesity in the US and Texas, and analytical maps from the Geographic Weighted Regression (GWR).
- Check shapefiles (US and TX) for appropriate datum and projections.
- Project both US and TX maps to projections that are accurate representations of their geography.
- Create an ArcMap file for the US historic obesity prevalence maps.
- Place the projected shapefile in the file, join the data from the transcribed US historic obesity data by joining the StateName field in both files.
- Create enough dataframes to display the desired number of years in the layout.
- Unify the symbology of the maps into one common scale of measure as for the values of the data previously joined.
- Insert legend and scale in the layout.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
- Create an ArcMap file for the TX historic obesity prevalence maps.
- Place the TX projected shapefile on the file.
- Join the TX historic obesity data previously cleaned up using the Geo_FIPS field in both files.
- Create enough dataframes to display the 2004-2010 obesity prevalence TX maps on the layout.
- Create a new dataframe for the obesity prevalence percent change map.
- On the attribute table, create a new field for percent change in obesity prevalence per county.
- Use the Field Calculator to populate the new field with the percent change of obesity prevalence in each county in TX from 2004 to 2010.
- Symbolize the new values in the map with graduated colors and 10% increases of percent change.
- Insert legend and scale in the layout.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
- Create an ArcMap file for the TX Obesity Prevalence GWR.
- Place the TX projected shapefile on the file.
- Join the TX dependent and independent variable data previously cleaned up using the Geo_FIPS field in both files.
- Use the search tool to find the GWR tool in ArcGIS.
- Run the GWR and save the result statistics and the new shapefile.
- Create data frames for Standard Residual map and Local R2 map.
- Symbolize Local R2 in 5 classes using Natural Breaks, preserve the symbology of the GWR output.
- Insert legend and scale in the layout.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
- Create a new ArcMap file for variables layout.
- Create a dataframe including the shapefile of the original data with variables and the shapefile with the GWR results.
- Create a layout of representative maps: one map for each variable, simply displaying each variable’s values for the counties in TX.
- Create a second layout of analytical maps: one map for each variable, each map displaying the variable’s coefficient values resulting from the GWR.
- Insert legend and scale in both layouts.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
Findings

Multiple Linear Regression

<table>
<thead>
<tr>
<th>v</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>.425*</td>
<td>.181</td>
<td>.164</td>
<td>1.7994378</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td>(Constant)</td>
<td>25.744</td>
<td>1.032</td>
</tr>
<tr>
<td>RECFAC</td>
<td>-0.071</td>
<td>.014</td>
</tr>
<tr>
<td>MEDHHINC</td>
<td>-0.000038</td>
<td>.000</td>
</tr>
<tr>
<td>EDU_NOHS</td>
<td>0.000033</td>
<td>.000</td>
</tr>
<tr>
<td>AGE_2544PERC</td>
<td>.112</td>
<td>.022</td>
</tr>
<tr>
<td>RACE_WHITE</td>
<td>.030</td>
<td>.007</td>
</tr>
</tbody>
</table>

\[ \text{ObesityPrevalence} = 25.744 - 0.071(\text{RECFAC}) - 0.000038(\text{MEDHHINC}) + 0.000033 \ (\text{EDU_NOHS}) + 0.112 \ (\text{AGE_2544PERC}) + 0.30 \ (\text{RACE_WHITE}) \]

The regression was performed to examine significance across variables in the diverse datasets and possible multicollinearity. After thorough experimentation with variable combinations the model above is presented as the model for the research study.

Geographic Weighted Regression

<table>
<thead>
<tr>
<th>Residual Squares</th>
<th>721.450</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective Number</td>
<td>13.794</td>
</tr>
<tr>
<td>Sigma</td>
<td>1.733</td>
</tr>
<tr>
<td>AICc</td>
<td>1011.040</td>
</tr>
<tr>
<td>R2</td>
<td>0.264</td>
</tr>
<tr>
<td>Adjusted R2</td>
<td>0.225</td>
</tr>
</tbody>
</table>

For analysis on the results of the Geographic Weighted Regression displayed in the table above refer to the Analysis & Conclusions section. Please refer to the next 3 attached map layouts for reference to the GWR map representations.
GEOGRAPHIC WEIGHTED REGRESSION

Maps and Layout by José Vicente Latorre, 2014 | These maps are educational, the information sourced from Center for Disease Control and Prevention (CDC), Obesity Trends 1986-2010 | Map Projection and Datum follows NAD 1983 Stateplane TX North Central & Lambert Conformal Conic.

Adult Obesity Prevalence as independent variable. Recreational Facilities, Median Income, Education, Age and Race as explanatory variables.

GWR Std. Resid.

-2.5 - -1.5 Std. Dev.
-1.5 - -0.5 Std. Dev.
-0.5 - 0.5 Std. Dev.
0.5 - 1.5 Std. Dev.
1.5 - 2.5 Std. Dev.
> 2.5 Std. Dev.

GWR Local R2

0.227 - 0.242
0.212 - 0.226
0.201 - 0.211
0.182 - 0.200
0.173 - 0.181 (performed unde OLS)
VARIABLE VALUE DISTRIBUTION
DESCRIPTIVE MAPS

Six maps representing the distribution of variable values across Texas counties.

RECREATION FACILITIES
Number of recreational facilities per county.

INCOME
Median Household Income (US$).

EDUCATION
Population with educational attainment under high school degree.

AGE
Percent of total population between ages 25 and 44.

RACE
Percent of white only non-hispanic population per county.
VARIABLE LOCAL COEFFICIENTS
GWR RESULTS

Six maps representing the local coefficient values resulting from the GWR.

RECREATION FACILITIES
Number of recreational facilities per county.

INCOME
Median Household Income (US$).

EDUCATION
Population with educational attainment under high school degree.

AGE
Percent of total population between ages 25 and 44.

RACE
Percent of white only non-hispanic population per county.

INTERCEPT
Intercept values as per GWR results.

Coefficient Values

Coefficient Values

Coefficient Values

Coefficient Values

Coefficient Values

Coefficient Values

Maps and Layout by José Vicente Latorre, 2014 | These maps are educational, the information sourced from Center for Disease Control and Prevention (CDC), Census Bureau, and USDA Food Atlas | Map Projection and Datum follows NAD 1983 Stateplane TX North Central & Lambert Conformal Conic.
Analysis & Conclusions

MLR
According to the R2 value, 18.1% of the variation on Obesity Prevalence in Texas Counties can be explained by the independent variables in the model. All variables in the model are significant at the 95% level. Only RECFAC and EDU_NOHS presented VIF levels higher than 5.

If we consider the significant variables in the 95% level (p-value < .05) in the MLR model they can be interpreted as follows:

- For every new Recreational Facility in a county, obesity prevalence in the county will be reduced by 0.07%.
- For every $1,000 increment in the mean household income of a county, obesity prevalence of the jurisdiction will be reduced by 0.031%.
- For every 1,000 people with no high school education attainment, obesity prevalence in the county will increase by 0.033%.
- For every 10% increase in population in the 25-44 years old cohort in a county, obesity prevalence will increase by 1.12%.
- For every 10% increase in White population in a county, obesity prevalence will increase by 0.3%.

The effect of the variables in increasing or decreasing obesity prevalence is according to the expected influences as seen in the research framework. However, the race variable did not conform to the expected result. We expected RACE_WHITE to negatively affect obesity prevalence; the more whites the less obesity prevalence in a county. At least that is what the Texas Department of Health suggested when they reported African Americans and Hispanics to be more prevalent to be overweight and obese.

In exploring the data closer we find certain counties of high obesity prevalence to be also predominantly white. This condition may be one reason contributing to the MLR result on race. Another may be a limitation in the building of the dataset for this study. Hispanic/Non-Hispanic Race data on ACS is not categorized by age. Therefore, in our study we extrapolated the percent white to the age cohort 25+. The extrapolation of race onto age may be affecting the regression. The adult population does not necessarily follow the same racial breakdown as the total population of the counties – we perform this study accepting this assumption.

Despite the unusual result of RACE_WHITE variable, we did not remove the variable from the model in hopes that GWR would shed light on whether or not this result could be further explained by the geographic distribution of the white demographic across the state.

GWR
In general, the GWR outperformed the MLR. Typically, the value considered to compare OLS (Ordinary Least Squares model – another term for the MLR) and
GWR models is R2. In this case, the GWR R2 value is 0.264 and the OLS value 0.181. The result is persistent with the previous case studies that compare GWR and OLS models for regional analysis of socio-environmental related issues.

The differences in obesity prevalence in the Texas, at the county level, can be better explained through a regional approach (GWR) than a statewide, general, approach (MLR) – meaning that the variation in obesity prevalence can be better explained by analyzing spatially heterogeneous relationships across variables. A word of warning is in place. GWR did not always outperform MLR. The local R2 values (GWR yields a R2 value for each unit of measure, in this case, each county) mostly out performed the MLR R2, but in the South of Texas it did not. The fact that southern Texas is not performing better than the OLS model may reflect the fact that the explanatory variables selected for the study are not relevant to the border area. This outcome suggests that further research in this area must rely on a different set of variables to address the adult obesity concern.

GWR: local coefficients of explanatory variables

Recreational Facilities

The potential effect of recreational facilities in affecting adult obesity prevalence levels is limited, according to the OLS model, 10 new recreational facilities could reduce obesity prevalence in 0.71%. The coefficients of the GWR reveal that recreational facilities can have both a greater impact in reducing obesity. In fact, the results suggest that in most counties increasing the number of recreational facilities may result in obesity reductions greater than -0.71% per 10 facilities – in some cases up to -1.05% per 10 facilities. The effect of recreational facilities on reducing obesity prevalence is higher in the Northeast of Texas.
Median Household Income

The degree of effect income exerts on adult obesity prevalence in this model specifically is very limited, as illustrated by the interpretation of the OLS model results. The results in the GWR highlight the Eastern and Central areas of Texas to be the only areas where the effect of increased household incomes on reducing obesity prevalence can be greater than those expressed by the OLS model. The areas that outperform the OLS model include major urban centers in Texas, as well as the coast. One reasoning behind the increased effect of income on obesity reduction can be related to the availability of healthy food choices and amenities in urban areas. The wealthier people become in these areas the more access they may have to a healthier standard of living. In contrast, Western areas of the state do not offer the amenities for healthier living that more developed areas offer and, therefore, do not provide an environment on which more income necessarily translate in healthier living.
Educational Attainment

The analysis of this explanatory variable is subject to the overall population of each county. In hindsight, the values should have been normalized to the total population in the county. Generally, coefficients in the GWR model are greater than those in the OLS model. High School completion has a greater effect in reducing adult obesity prevalence in most of the counties in Texas than what a general regression model would reveal. On this model, the magnitude of the effect educational attainment can have on obesity prevalence will be reflected to be strictly related to the county’s total population.
Age

The age explanatory variable presents a case in which GWR reveal lower influence of the variable on obesity prevalence in a majority of the counties. Only regions in the Northeast and Northwest of Texas retain a relationship approximately similar to that of the OLS coefficient – for every 10% increase in population of the 25-44 year old cohort, obesity prevalence will tend to increase 1.12%. The share of population in the specified age cohort less proportionally affects adult obesity prevalence in counties of central Texas and in the southern tip of the state. In the areas with greater influence of young adults over increased adult obesity prevalence efforts to reduce obesity must be tailored more specifically to this age cohort. Analyzing age groups in relationship to obesity prevalence can help identify target populations for efforts to reduce obesity in specific regions. In this case, we now know that programs geared to the 25 to 44 age group will be more effective in reducing adult obesity prevalence in areas in the north, east, and west tips of Texas.
Race

We previously stated that we hoped GWR would clarify the unsettling results about race effect on adult obesity prevalence from the OLS. GWR, with the same limitations on racial data explained before, did not confirm the Texas Department of Health racial prevalence trends. In order to do so, we would have expected to see negative in some of the local coefficient values for the variable in the GWR. We didn't; the GWR did, however, reveal that, in fact, the proportion of white population in a county can only have a lesser impact in increasing adult obesity prevalence than the one explained by the OLS model: for every 10% increase in White population in a county, obesity prevalence will increase by 0.3%. This tendency partially suggests the fact that racial minorities may have increased obesity prevalence in some counties – mostly in the northern central portion of the state. Interestingly enough the counties in this area are among the most white-predominant counties in the state. If valid, this GWR model’s results suggest that in areas with high percentage of white population, minorities are more likely to be considered obese.

Reflections on the Study

For the most part, the idea of GWR being able to represent the spatial relationship of explanatory variables is fascinating. Despite the limitations of the data, which had to be built in order to be descriptive of the “adult” component of the obesity prevalence being analyzed, the results yielded interesting results that can be interpreted in diverse ways.

On performing a study using GWR, the fitness of the data to achieve credible results is as important as being able to effectively use ArcGIS or create map
layouts. The quality of the dependent variables to effectively describe the dependent variable is of essence – that was the main limitation of this study.

Another issue is scale. The more numerous and heterogeneous the samples the more interesting the GWR results will be. In this case, the study focuses on TX, most of the dependent variables vary slightly from county to county. We intended to perform this study in a census tract level, where contrasts on independent variable geographical variations would be more evident – contrasts in household income or race in a city are far more spatially placed than in the county level. The main obstacle to perform a census-tract level study was the lack of obesity prevalence data at a census tract level. The Texas Department of Health was contacted in regards to the Behavioral Risk Factor Surveillance System (BRFSS), from which the CDC bases of its obesity prevalence data, but they do not have publicly available census tract level data for the BRFSS survey outputs.
Appendix
References


Methodology

- Downloaded variables data from the previously cited sources.
- Sorted through the Excel files containing the downloaded data to extract the relevant variables into a single database to be used throughout the project. Relevant data from the USDA Food Atlas (http://www.ers.usda.gov/data-products/food-environment-atlas.aspx) and American Community Survey were consolidated into one file. ACS Data was obtained using Social Explorer (http://www.socialexplorer.com/explore/tables).
- For the US historical obesity prevalence data, the available data had to be transcribed into a new database manually. Historical data for the US maps and 2010 data for obesity prevalence (dependent variable) were obtained from the Center for Disease Control (www.cdc.gov/diabetes/atlas/countydata/County_ListofIndicators.html).
- Race data for Hispanic/Non-Hispanic breakdown is not available by age. Therefore, the percentage of total population white only non-hispanics was extrapolated to each counties 25+ years old population.
- Age for the 25-44 cohort was used in percentage of population.
- Median Household Income was used in place of average household income in aims to avoid distortion of income data common in averages.
- Educational data used was the magnitude of population over 25 years old with educational attainment lower than high school diploma.
- Using the database previously created, conducted a Multiple Linear Regression using SPSS to serve as a global estimation method in the study.
- Export the results and briefly analyze them.
- Time to step into ArcGIS and produce a series of descriptive maps about adult obesity prevalence in the US and Texas, and analytical maps from the Geographic Weighted Regression (GWR).
- Check shapefiles (US and TX) for appropriate datum and projections.
- Project both US and TX maps to projections that are accurate representations of their geography. Maps of the US were projected onto WGS NAD 1984 World Mercator, and maps of Texas were projected onto NAD 1983 Stateplane TX North Central & Lambert Conformal Conic.
- On the original US map shapefile obtained from the Tiger Shapeline database, eliminate outlying states and territories by using the select tool and clicking on the mainland states.
- Once the states are selected export the selection onto a new shapefile.
- Create an ArcMap file for the US historic obesity prevalence maps.
- Place the projected shapefile in the file; join the data from the transcribed US historic obesity data by joining the StateName field in both files.
- Create enough dataframes to display the desired number of years in the layout, in this case one map for each year, from 1990 to 2010.
- Unify the symbology of the maps into one common scale of measure as for the values of the data previously joined.
- Insert legend and scale in the layout.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
- Create an ArcMap file for the TX historic obesity prevalence maps.
- Place the TX projected shapefile on the file.
- Join the TX historic obesity data previously cleaned up using the Geo_FIPS field in both files. This shapefile contains the necessary data for both the representative maps and the GWR.
- Create enough dataframes to display the 2004-2010 obesity prevalence TX maps on the layout.
- Create a new dataframe for the obesity prevalence percent change map.
- On the attribute table, create a new field for percent change in obesity prevalence per county.
- Use the Field Calculator to populate the new field with the percent change of obesity prevalence in each county in TX from 2004 to 2010.
- Symbolize the new values in the map with graduated colors and 10% increases of percent change.
- Insert legend and scale in the layout for both, the timeline maps and the main obesity prevalence percentage change map.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
- Create an ArcMap file for the TX Obesity Prevalence GWR.
- Place the TX projected shapefile on the file. Remember we had previously created this file with all the data necessary for running the regression.
- Now, since the regression exercise is an iterative process, in several occasion we had to add new variables into the shapefile’s attribute table. At any moment we needed to add new variables we joined the TX dependent and independent variable data previously cleaned up using the Geo_FIPS field in both the shapefile and the excel CSV file.
- This step was the longest and most extenuating process of the study. Countless iterations of variable combinations were tested for statistical fit until the proposed set of variables was found to be the best fit combination of variables available for the study.
- Use the search tool to find the GWR tool in ArcGIS.
- Run the GWR and save the result statistics and the new shapefile. Following similar previous research on similar regression dealing with polygons of varying sizes at the county level, Kernel Type was set up to Adaptive and Bandwidth method was left in AICc per recommended by the cited case studies.
- Create data frames for Standard Residual map and Local R2 map.
- Symbolize Local R2 in 5 classes using Natural Breaks, preserve the symbology of the GWR output.
- Insert legend and scale in the layout.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.
- Create a new ArcMap file for variables layout.
- Create a dataframe including the shapefile of the original data with variables and the shapefile with the GWR results.
- Create a layout of representative maps: one map for each variable, simply displaying each variable’s values for the counties in TX. For instance, recreational facility was setup to display its numerical value under its properties > symbology. All representative maps are symbolized using 5 classes and natural breaks.
- Create a second layout of analytical maps: one map for each variable, each map displaying the variable’s coefficient values resulting from the GWR. By simply adding the GWR shapefile to the dataframes on the variables and symbolizing them according to each variable’s dataframes, we are capable of turning on an off one set or the other (representative map or GWR coefficient maps for each variable). By doing so we can have both layouts within one same Arcmap file.
- Insert legend and scale in both layouts.
- Export the layout in AI file mode.
- Edit the AI file in Illustrator.