Obesity in America:
A Growing Trend

David Todd

Utilizing Geographic Information Systems (GIS) to explore obesity in America, this study aims to determine statistically significant correlations for the growing trend.

Pennsylvania State University

GEOG 586: Geographic Information Analysis

Winter 2011

Dr. Justine Blanford

3/15/2012

I. Introduction

Obesity, according to Merriam-Webster, is a condition characterized by the excessive accumulation and storage of fat in the body. In general terms, obese is the label for a range of weight much greater than what is generally considered healthy for a given height. In adults, obesity is determined by using weight and height to calculate a number called the "body mass index" (BMI). BMI is used because, for most people, it correlates with their amount of body fat. An adult who has a BMI of 30 or higher is considered obese. Obesity has been shown to increase the likelihood of certain diseases and other health problems, e.g. heart disease, diabetes, and hypertension.

While healthy eating and regular exercise is generally considered the best way to combat obesity, there may be other significant correlating factors that can be linked to obesity, e.g. income, education level, or gender. What's interesting about obesity is the rate at which it has risen to epidemic proportions in the United States with no clear cut reason why. Significant correlating factors could help focus on ways to combat obesity across America or which demographics to target with education campaigns.

II. Objectives

This study aims to utilize Geographic Information Systems (GIS) to explore obesity in America by mapping obesity rates in order to visualize any trends and whether or not there is clustering. The objective is to determine statistically significant correlations for the growing trend of obesity in the United States. It will attempt to dissect any patterns that emerge through the 2010 data timeframe and measure spatial autocorrelation. This study looks to answer the question – *Does obesity have statistically significant correlations?*

III. Data

A. Data Sources

The majority of data gathered for this study comes from the Centers for Disease Control and Prevention (CDC). Specifically, the CDC operates the Behavioral Risk Factor Surveillance System (BRFSS), which is the world's largest, on-going telephone health survey system, tracking health conditions and risk behaviors in the United States yearly since 1984. BRFSS data contains over 100 different variables by state, e.g. obesity rates, income levels, education levels, diabetes, alcohol ingestion, race, gender, etc. To aid

with visualization, geographic data sources are also provided by ESRI's base map layers.

B. Data Preparation

The CDC has completed the majority of data preparation. They present the BRFSS data ready to be used for analysis in the form of GIS shapefile maps. The files contain the survey data and documentation, and are available in Zip Archive File (ZIP) format. Once acquired, the data was loaded into ESRI's ArcGIS software for additional preparation. The decision was made to focus on the continental United States (CONUS). The shapefiles were edited to delete the both Alaska and Hawaii, leaving 49 records – 48 States plus the District of Columbia. One additional step for data preparation was required before analysis could begin. Some of the files contained attribute cells that were stored as "Text" data type. Because the data in the fields are numeric in nature, the data type needed to be changed in order to support analysis. To turn the data into a numeric data type, the *.dbf file was opened in Excel so it would be possible to change the format of the Text cells to Number. The file was then saved as an *.xls file. Finally, in ArcMap, the original data layer is joined with the *.xls file, making it possible to perform analysis.

C. Data Analysis

The first step in data analysis for this study is performing a Hot Spot Analysis on obesity, as seen in Figure 1. The output of this tool is a map of z-scores which shows statistically hot spot and cold spot clustering in the data.

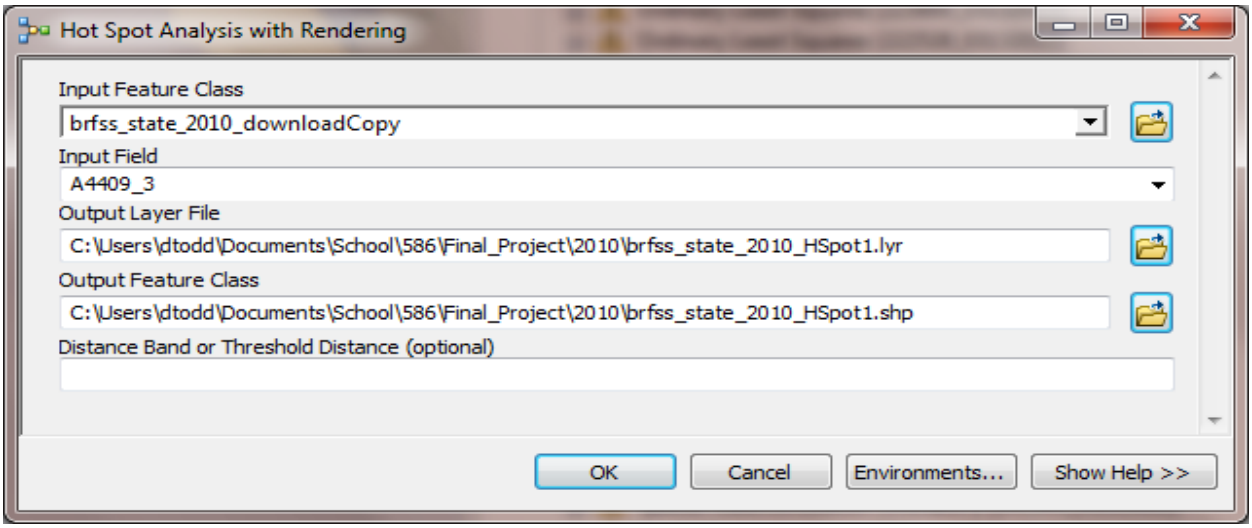


Figure 1: Hot Spot Analysis (Getis-Ord Gi*) tool

In the next steps, data analysis for this study focused primarily on performing regression analysis using the spatial statistics tools in ESRI's ArcGIS software package. Regression analysis is used to attempt to explain phenomena, i.e. obesity, in terms of other variables, e.g. education, income, gender. Utilizing a GIS like ArcGIS, it is possible to perform regression analysis to model and predict these complex phenomena. Regression analysis starts by determining which variables will specify a good regression model.

Finding a good regression model when over 100 variables are at hand is an extremely iterative process. To aid with this time consuming process, the Exploratory Regression tool (Figure 2), provided by ESRI in their supplemental spatial statistics toolbox was used to determine which explanatory variables were the most statistically significant in explaining the dependent variable obesity. This single process took well over 24 hours to determine models of 1 to 5 statistically significant factors. While 24 hours may seem like a long time to process, it would have taken days or weeks to manually run the thousands of possible model variations.

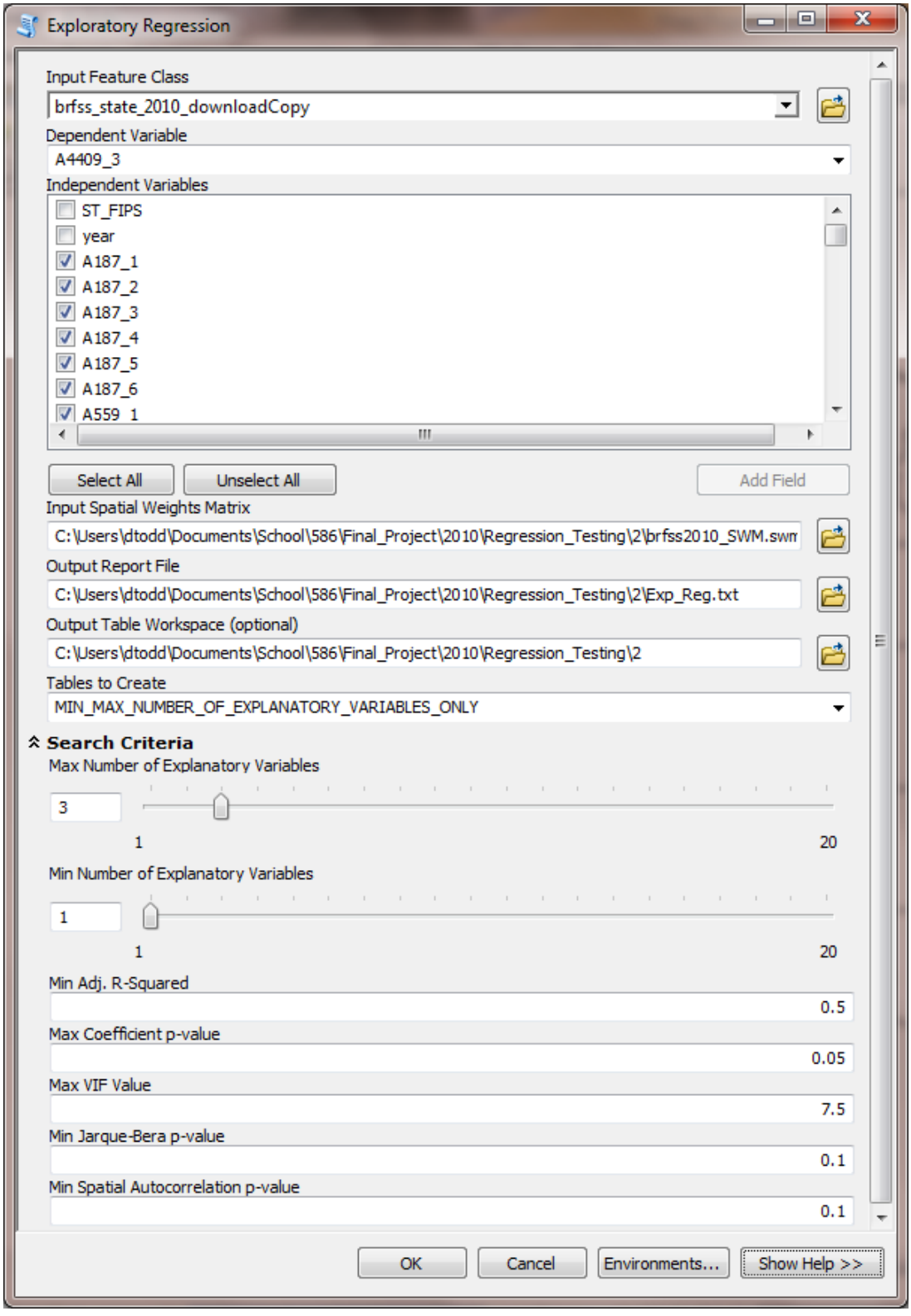


Figure 2: Exploratory Regression Tool in ArcGIS 10.0

Next, as seen in Figure 3, the Scatterplot Matrix tool was used in order to check the possible explanatory variables that were determined by the Exploratory Regression tool. The scatterplots showed whether there was a positive or negative correlation between obesity and the explanatory variables.

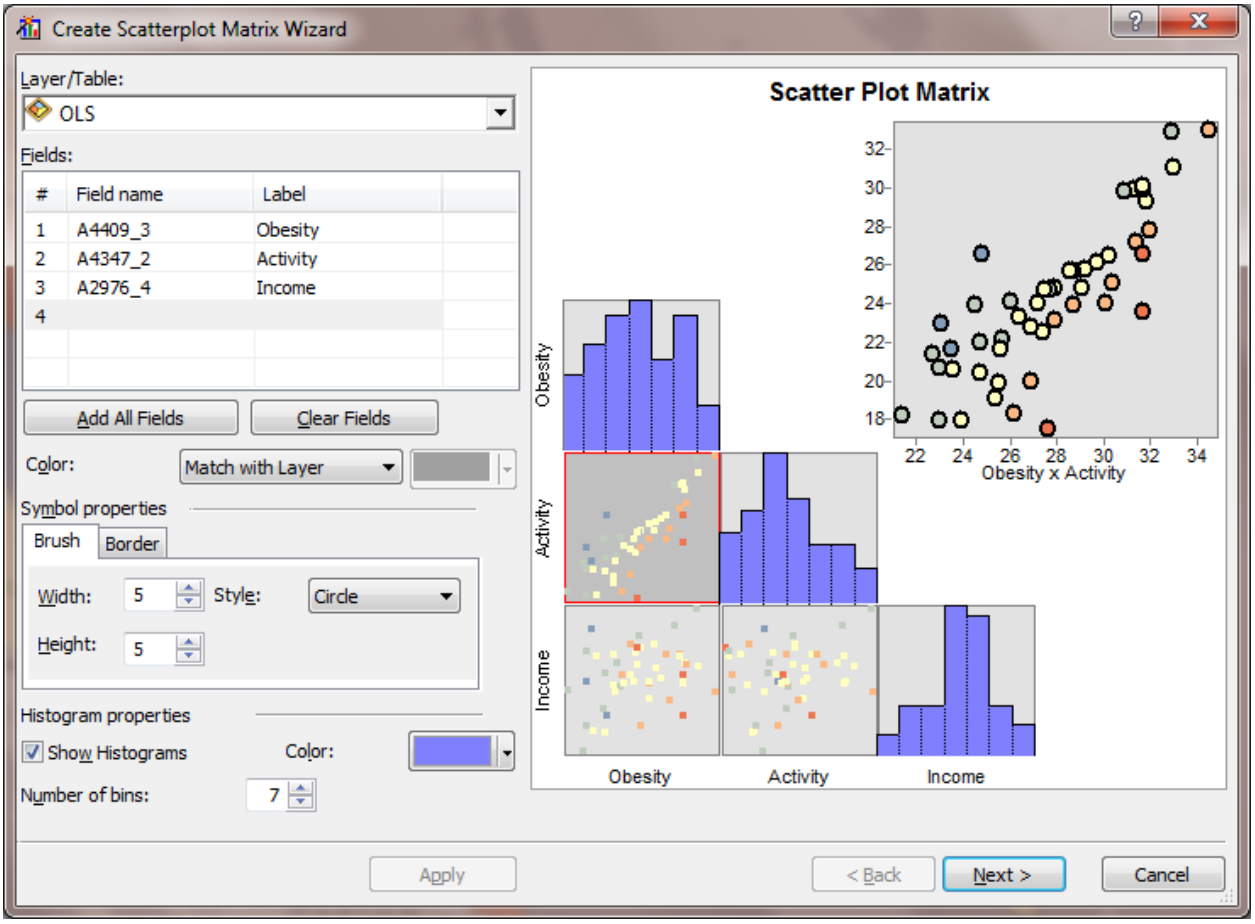


Figure 3: Scatterplot Matrix Tool

While the scatterplot matrix assists in visualizing correlations, it's hard to determine just by eye if the variable is statistically significant, which is another reason the Exploratory Regression tool was used first. The tool allowed the study to immediately focus on the significant variables.

The next step in analysis is to perform Ordinary Least Squares (OLS). The OLS tool (Figure 4) performs linear regression to predictions or to model a dependent variable in terms of its relationships to the explanatory variables. It generates different outputs that include a map of regression residuals and a mostly numeric summary report. The map shows the over and under predictions from your model.

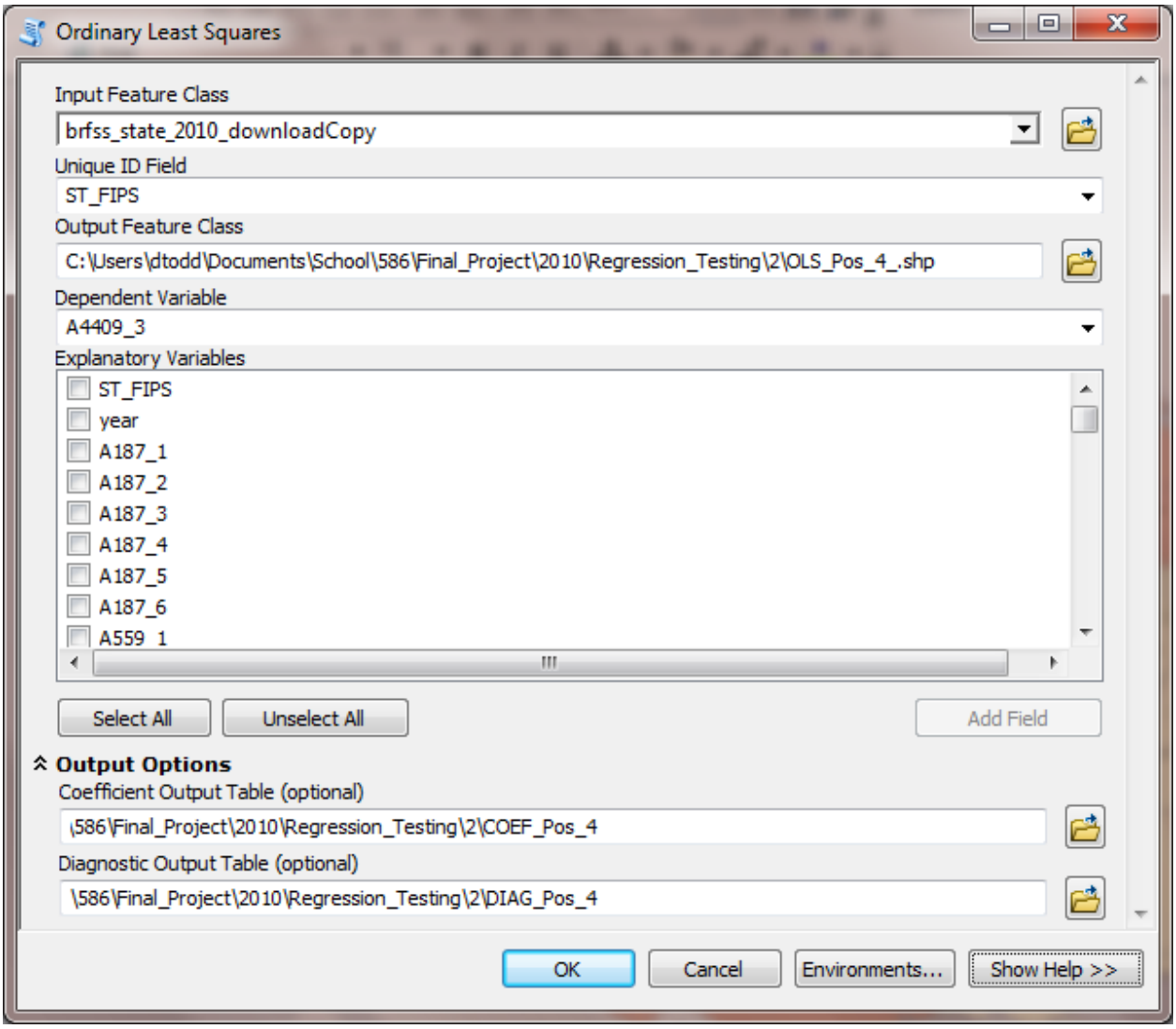


Figure 4: OLS Tool

The other output of OLS is the summary table. It contains a great deal of important information about the explanatory variables in the OLS model. While the summary report has a lot of numeric data, refer to Figure 5, it displays an asterisk next to any variable that is statistically significant. The summary report includes both a probability and robust probability for each variable. The robust probability needs to be used if the spatial relationships vary across the study area. Again, it's easy to determine which probability to use since the summary also includes a Koenker p-value and an asterisk if it's statistically significant. If there is an asterisk, the robust probability should be used. The most notable value in the summary output is the R^2 value. R^2 is a value between 0

and 1 that represents the percentage to which the explanatory variables tell the whole story. The closer to 1 the value is, the better.

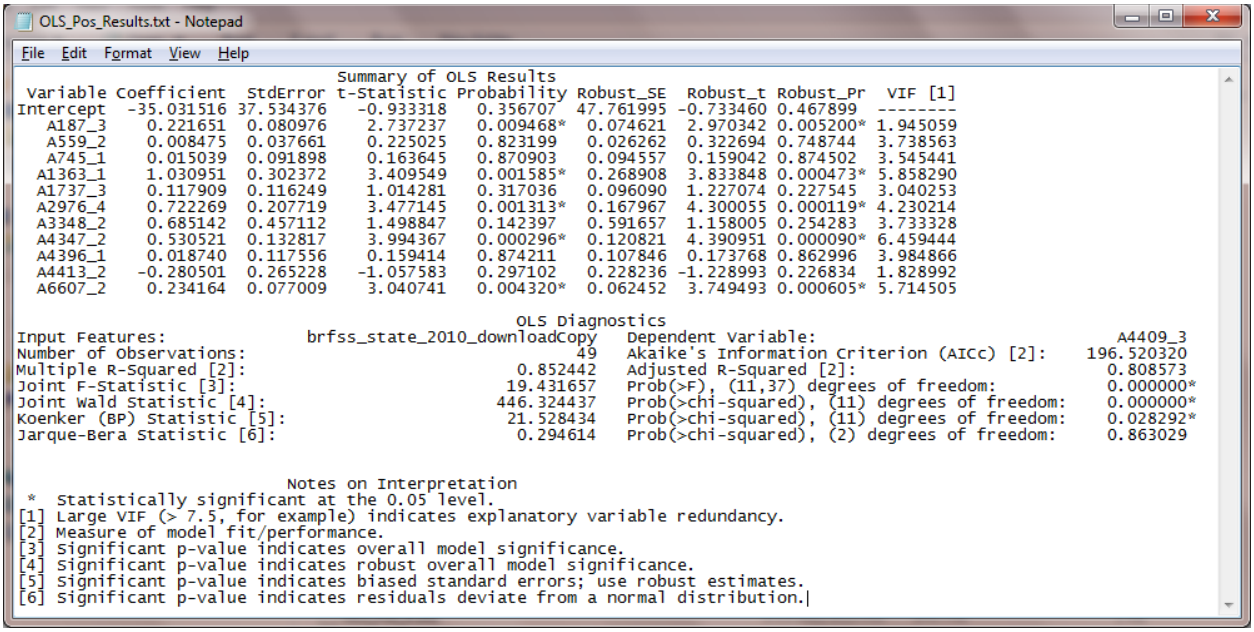


Figure 5: OLS summary output

Once a model has been specified and OLS has been run, the next step is to perform a spatial autocorrelation test on the OLS residuals to determine if the model is biased. This is accomplished using the Spatial Autocorrelation (Morans I) tool, as seen in Figure 6. The tool outputs an *.html file that shows spatial autocorrelation results (z-score) on a bell curve. If the z-score falls to the far left or far right, then it is significant, but if they fall in the middle, they are determined to be no more significant than random chance. If the z-score is significant, the model is missing important explanatory variables.

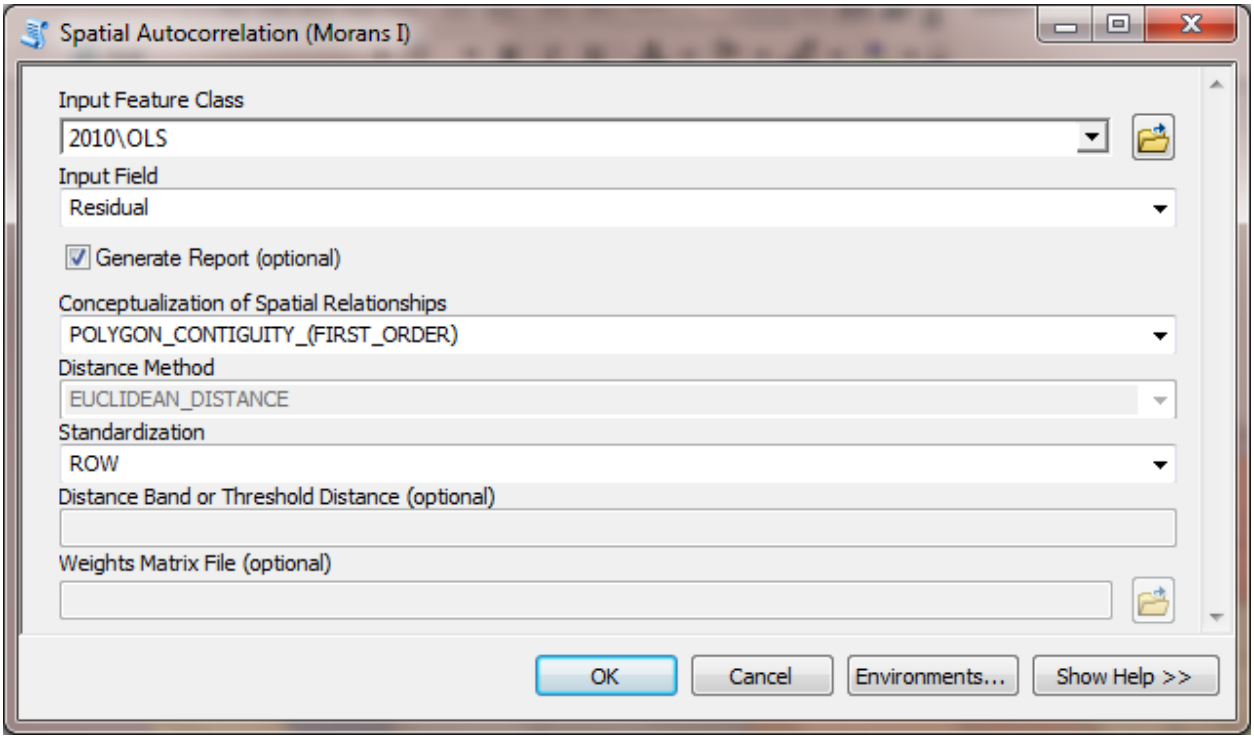


Figure 6: Spatial Autocorrelation (Morans I) Tool

Using the above tools to perform data analysis provided enough information to start to answer the question posed by this study.

IV. Results

Does obesity have statistically significant correlations? Yes, 5 explanatory variables account for 83% of the variance of obesity and they are: people aged 35 – 44 that have diabetes making an income of $35,000 - $49,999 that have had no physical activity in the last 30 days and have not had any permanent teeth extracted.

Mapping obesity rates across the United States clearly allows for visualization of clustering, as can be seen in Figure 7. According to the map, states displayed as red have the highest rates of obesity in the country and appear to be clustered in the South.

Since the simple mapping of the data appears to be clustered, running the Hot Spot Analysis tool will identify significant hot and cold spot clustering. Figure 8 shows that there is a statistically significant hot spot cluster in the South consisting of the states OK, AR, LA, MS, AL, GA, TN, KY, IN, and OH. Likewise, it also shows two cold spot

clusters – one out West consisting of WY, UT, and NV and the second in New England consisting of NY, NJ, CT, RI, MA, NH, VT, and ME.

With the Hot Spot Analysis (Getis-Ord Gi) proving that there are statistically significant hot and cold clusters, the next step is settling on a model that describes and accounts for the clustering of obesity. This is accomplished with a linear regression OLS test of the dependent and explanatory variables. For this OLS, the dependent variable is Obesity (A4409_3) and as a result of the Exploratory Regression tool, the explanatory variables are Age = 35 – 44 (A187_3), Diabetic (A1363_1), Income = $35,000 - $49,999 (A2976_4), No Physical Activity Last 30 Days (A4347_2), and No Permanent Teeth Extracted (A6607_2). Figure 9 displays the resulting map of OLS residuals. The red states indicate that the actual observed values are higher than values predicted by the model. On the other hand, blue states show where actual values are lower than predicted by the model.

The other output of OLS is the summary file. While Figure 9 is a way to visualize the model, the summary table (Figure 10) contains the meat for determining how well suited is the model to the dependent variable. First, all the explanatory variables have a positive coefficient, which means each variable has a positive correlation with obesity. Second, all the explanatory variables appear to be statistically significant at the 0.05 level. Statistical significance is shown with an asterisk, as can be seen in Figure 10. Next is the Koenker Statistic which checks for non-stationarity. Since it's not significant (no asterisk), then probability should be used over the robust probability. Fourth, determine if any of the explanatory variables display redundancy by reviewing the VIF column. The general guideline is if the VIF is above 7.5, then the variable displays multicollinearity, but smaller is better. All values appear to be under the guidelines and therefore multicollinearity is not an issue.

Now that all the explanatory variables have been individually determined to be significant, the model itself needs to be evaluated. This comes in the form of the R^2 value, in this case $R^2 = 0.832578$. This says that the model explains ~83% of the variance of the dependent variable, which is very good. Finally, the summary report also displays the Jarque-Bera Statistic which determines whether or not the model has bias. The statistic is not significant (no asterisk), which means the model residuals are normally distributed.

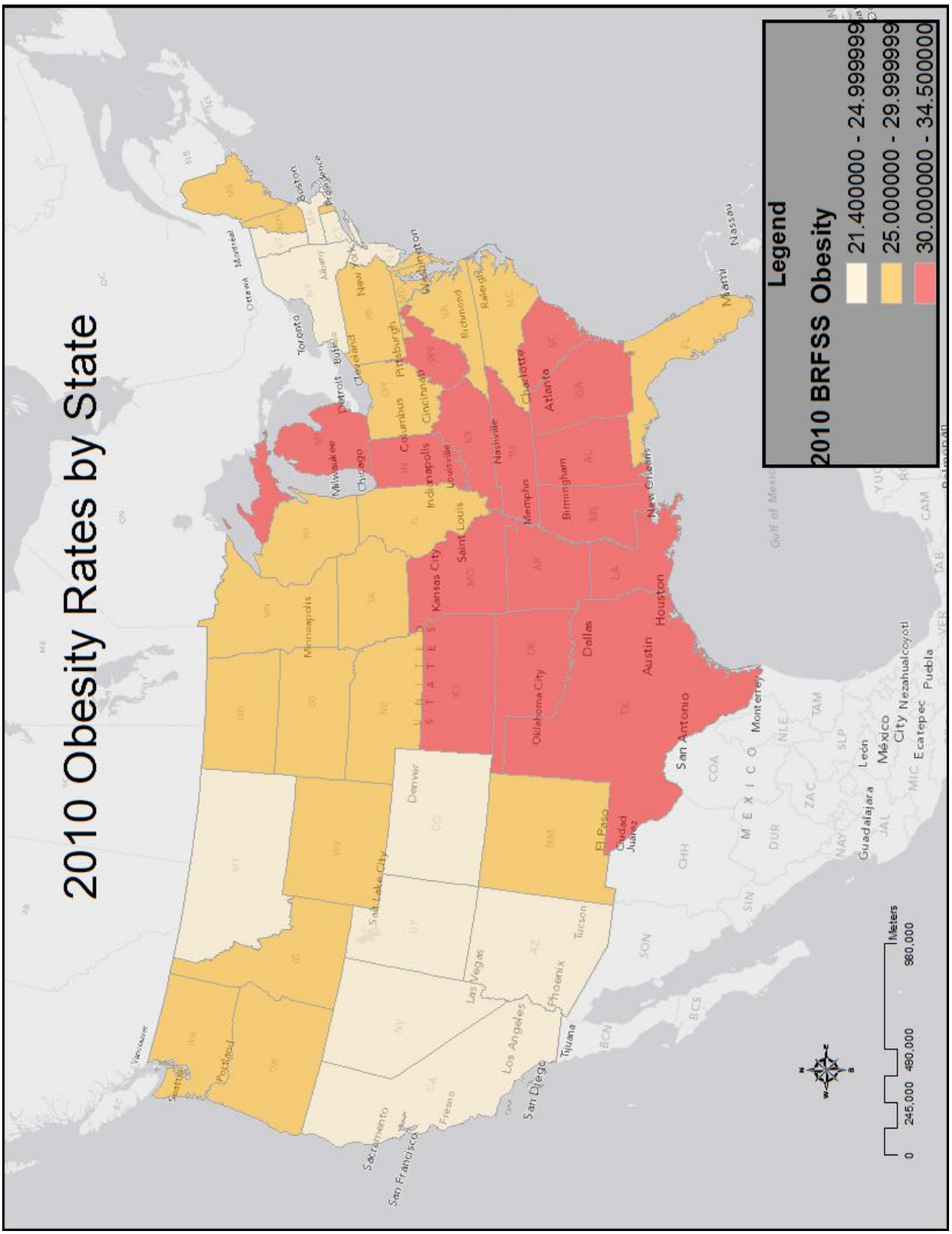


Figure 7: 2010 Obesity Rates by State

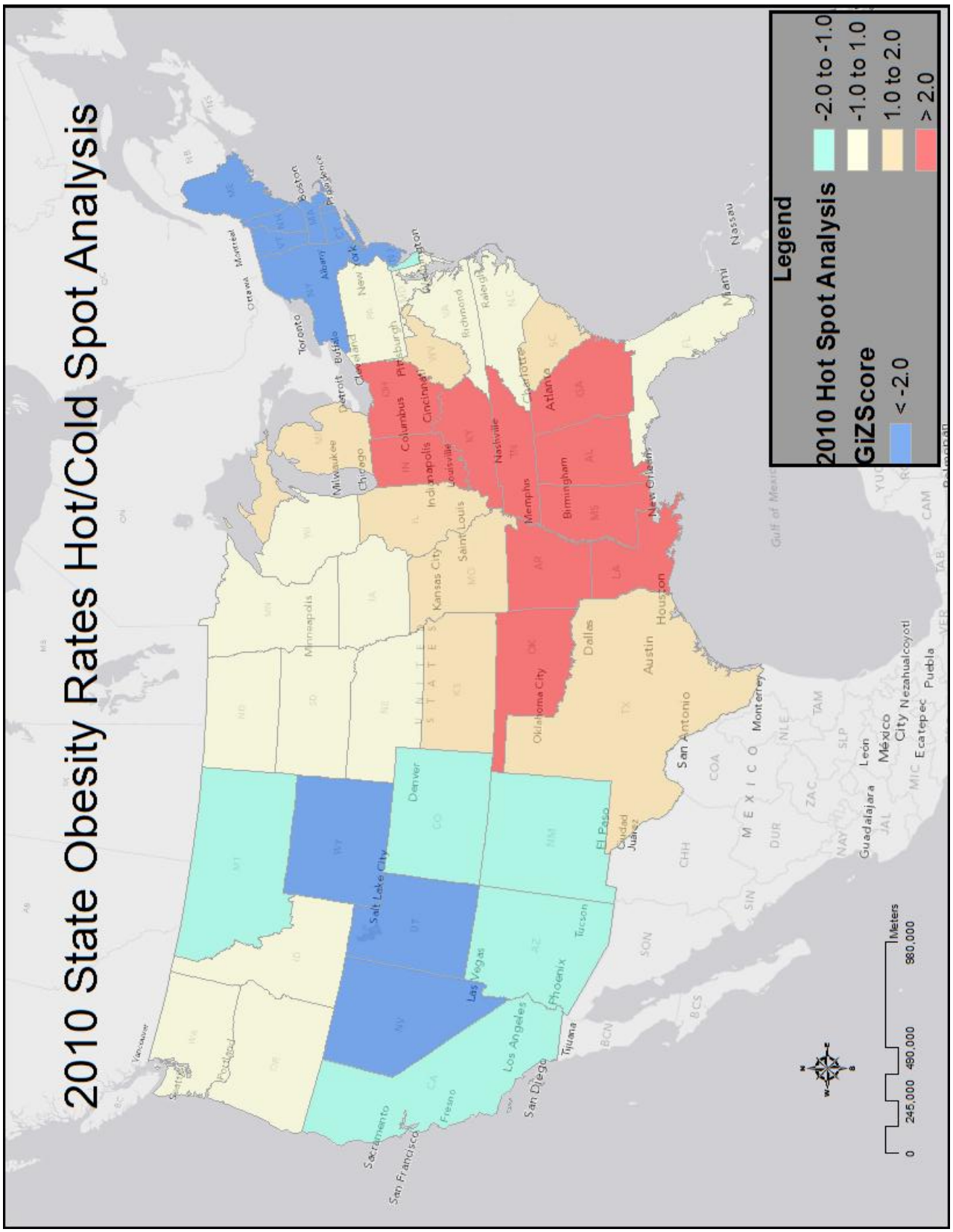


Figure 8: Hot/Cold Spot Analysis of 2010 BRFSS Obesity data by State

With the Hot Spot Analysis (Getis-Ord Gi) proving that there are statistically significant hot and cold clusters, the next step is settling on a model that describes and accounts for the clustering of obesity. This is accomplished with a linear regression OLS test of the dependent and explanatory variables. For this OLS, the dependent variable is Obesity (A4409_3) and as a result of the Exploratory Regression tool, the explanatory variables are Age = 35 – 44 (A187_3), Diabetic (A1363_1), Income = $35,000 - $49,999 (A2976_4), No Physical Activity Last 30 Days (A4347_2), and No Permanent Teeth Extracted (A6607_2). Figure 9 displays the resulting map of OLS residuals. The red states indicate that the actual observed values are higher than values predicted by the model. On the other hand, blue states show where actual values are lower than predicted by the model.

The other output of OLS is the summary file. While Figure 9 is a way to visualize the model, the summary table (Figure 10) contains the meat for determining how well suited is the model to the dependent variable. First, all the explanatory variables have a positive coefficient, which means each variable has a positive correlation with obesity. Second, all the explanatory variables appear to be statistically significant at the 0.05 level. Statistical significance is shown with an asterisk, as can be seen in Figure 10. Next is the Koenker Statistic which checks for non-stationarity. Since it's not significant (no asterisk), then probability should be used over the robust probability. Fourth, determine if any of the explanatory variables display redundancy by reviewing the VIF column. The general guideline is if the VIF is above 7.5, then the variable displays multicollinearity, but smaller is better. All values appear to be under the guidelines and therefore multicollinearity is not an issue.

Now that all the explanatory variables have been individually determined to be significant, the model itself needs to be evaluated. This comes in the form of the R^2 value, in this case $R^2 = 0.832578$. This says that the model explains ~83% of the variance of the dependent variable, which is very good. Finally, the summary report also displays the Jarque-Bera Statistic which determines whether or not the model has bias. The statistic is not significant (no asterisk), which means the model residuals are normally distributed.

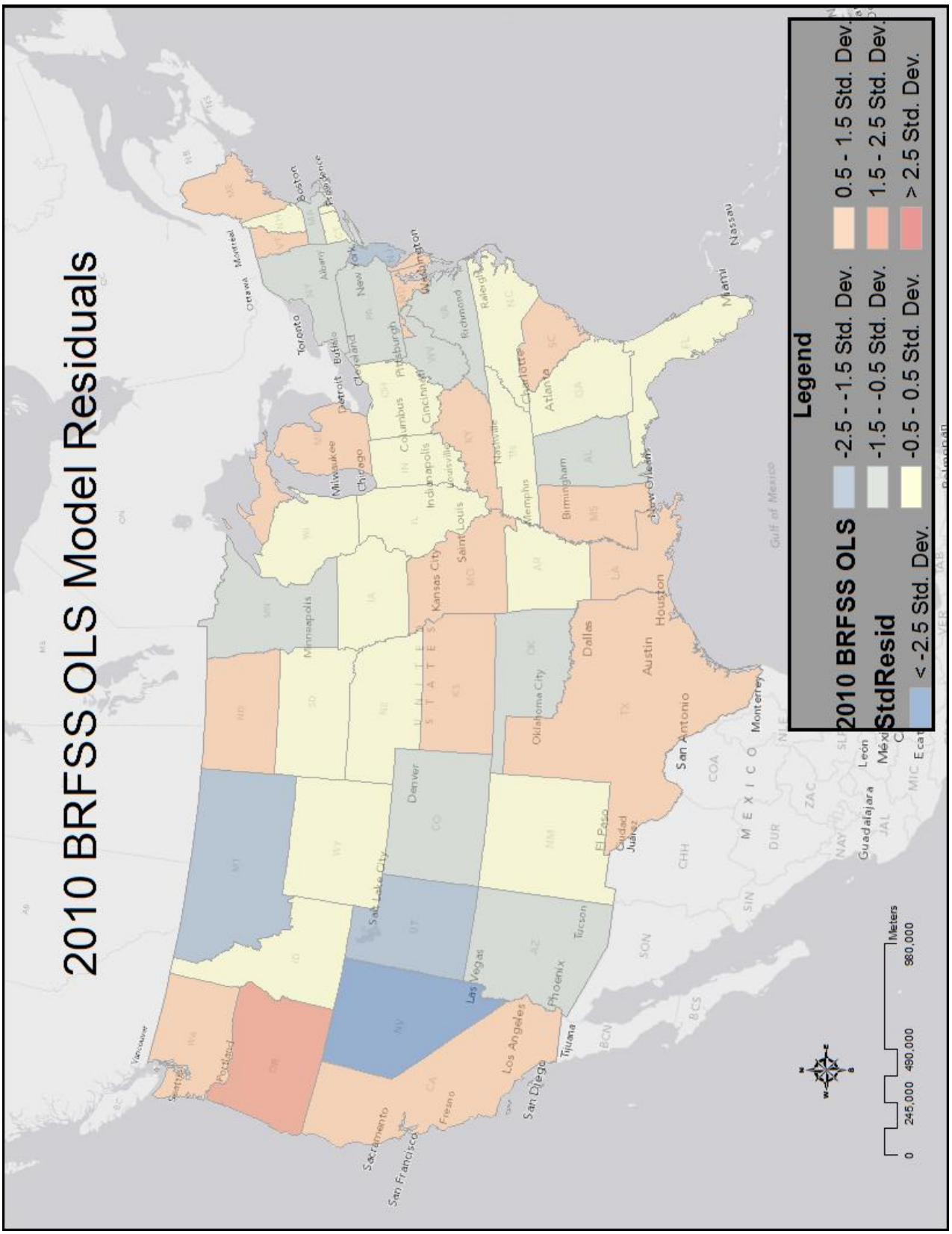


Figure 9: OLS Model of 2010 BRFSS data

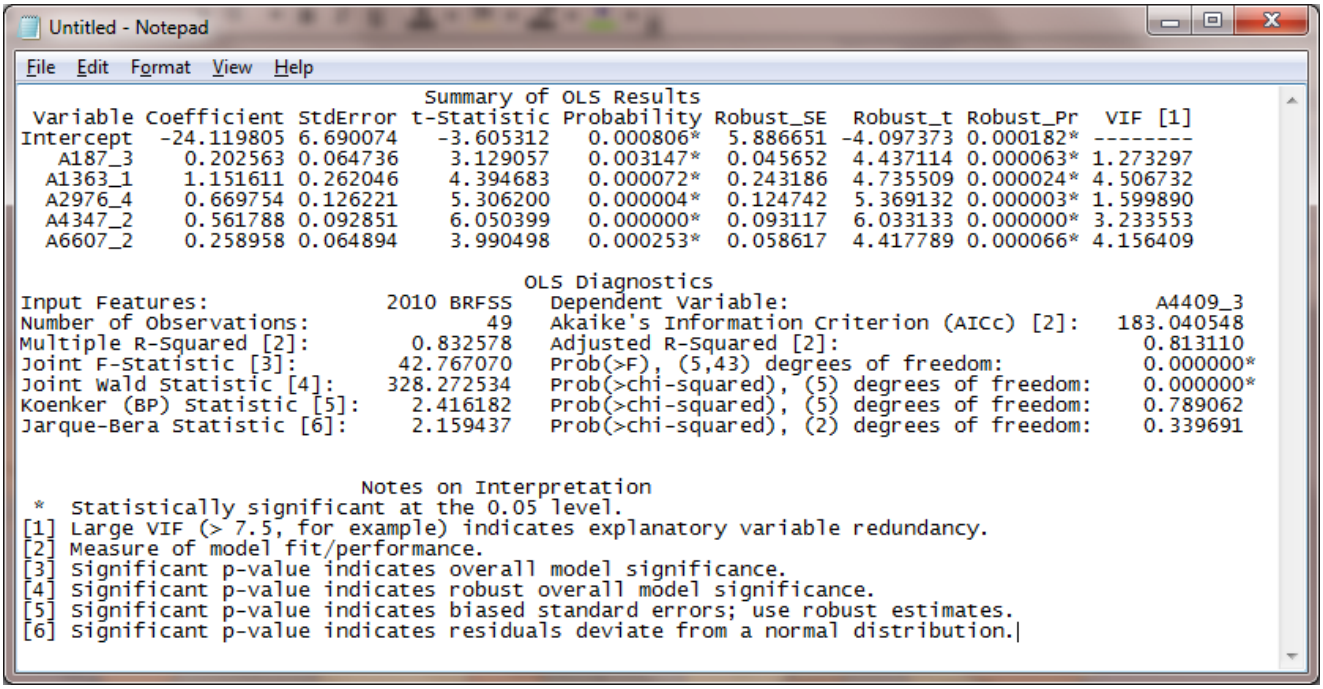
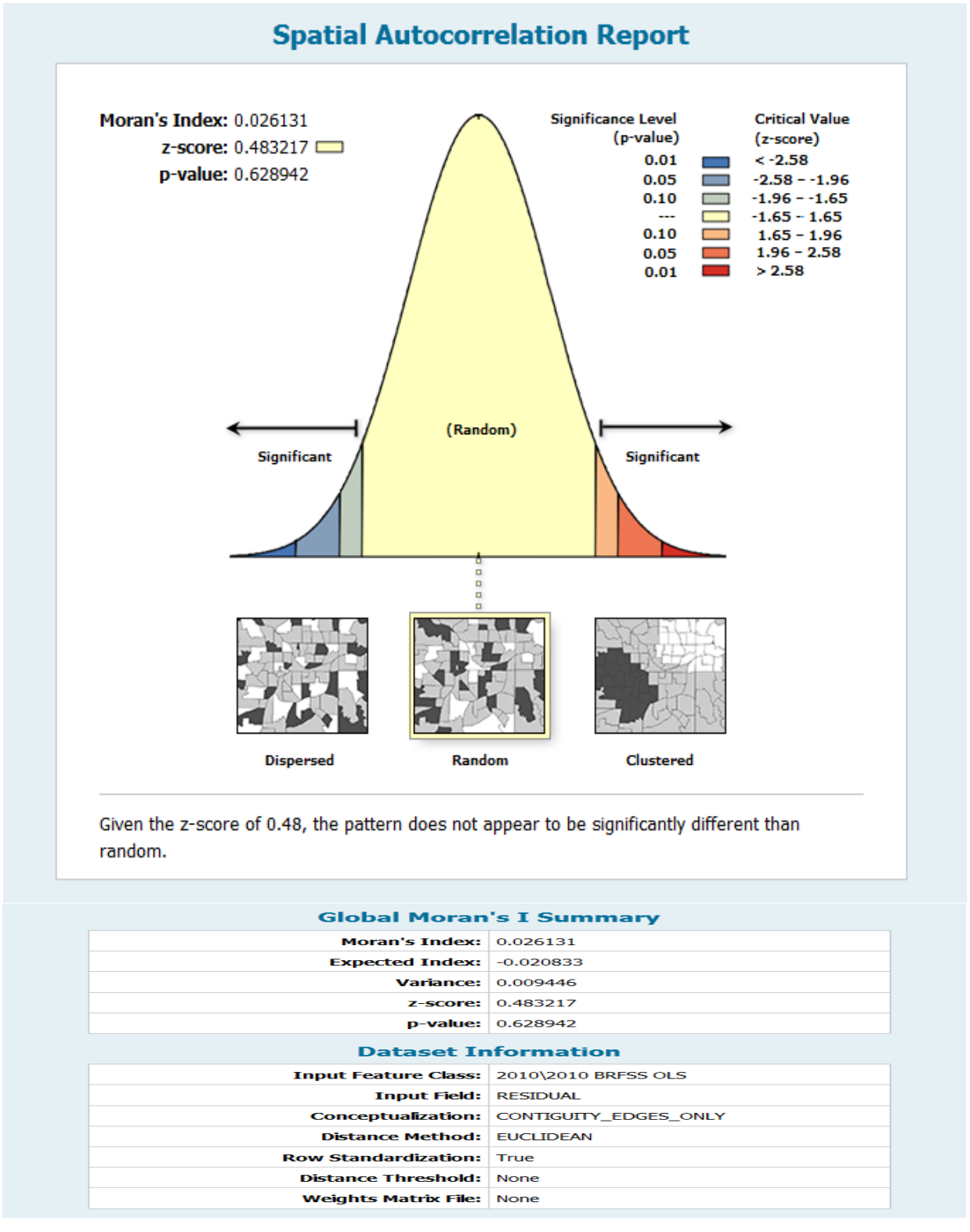


Figure 10: OLS Summary Results

Since the OLS model has been determined to be a good fit, there just needs to be a final process that checks the regression model's over and under predictions are not clustered. As seen in Figure 11, the output of the Spatial Autocorrelation (Morans I) tool shows that the residuals display a random spatial pattern and are not clustered. Taking all these factors into account, it appears that the OLS model is well specified.

So *Does obesity have statistically significant correlations?* Yes and the variables hint at who is most likely to be obese and that conventional wisdom still stands – physical activity is important to control obesity. The two most surprising results are the income level and permanent teeth extraction results. One hypothesis about income level is that people making less than $35,000 may receive some Federal assistance, e.g. food stamps, that cannot be used at fast food places. Likewise, people that make $50,000 or more might be more inclined to spend more on fresh fruits and vegetables. The no permanent teeth extraction variable is still a mystery.



Global Moran's I Summary

Moran's Index:	0.026131
Expected Index:	-0.020833
Variance:	0.009446
z-score:	0.483217
p-value:	0.628942

Dataset Information

Input Feature Class:	2010\2010 BRFSS OLS
Input Field:	RESIDUAL
Conceptualization:	CONTIGUITY_EDGES_ONLY
Distance Method:	EUCLIDEAN
Row Standardization:	True
Distance Threshold:	None
Weights Matrix File:	None

Figure 11: Spatial Autocorrelation (Morans I) of the OLS Residuals

V. Summary

Upon reflection, I feel this study was a practical use of the knowledge and tools learned throughout the course. Finding the Exploratory Regression tool was hugely helpful in checking variables contained within the data set that I would not have had time to do otherwise. I had hoped to continue back in time to provide trending analysis, but between inconsistencies in the data and the amount of time needed to determine variables, it was not to be in this study. Probably the biggest disappointment with this study was not being able to obtain finer level data to do more detailed distance analysis, i.e. distance to grocery store, proximity to fast food restaurants, etc. Time permitting, this study could certainly be extended and supplemented with more detailed data over a longer period of historical time. It could be effective in creating obesity education programs, as well as knowing exactly where to concentrate those education efforts.

VI. References

O'Sullivan, David and Unwin, David. Geographic Information Analysis – 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

GEOG 586: Geographic Information Analysis. https://www.e-education.psu.edu/geog586/

CDC BRFSS GIS Data Files. http://www.cdc.gov/brfss/maps/gis_data.htm

Merriam-Webster. http://www.merriam-webster.com/dictionary/obesity