Spatial Analysis

Andrea Elhajj

Contents

Analysis of Vermont Social Survey and US Census Data

Introduction:

Demographic trends are the statistical foundation for a variety of processes in various sectors, from the marketing industry to the government sphere. Within marketing, demographical data is key for getting a competitive edge in marketing products and services to generate more revenue and higher profit margins. Demography is also vital to helping government and society better prepare for population growth, aging, and migration. In this report, we seek to perform an exploratory analysis on two important Vermont datasets. The first is a social survey of Vermont residents that was conducted in the early 1990's. This survey contains information ranging from the demographic and income of the respondents to their employment status and level of job satisfaction. The SPSS statistical software package has been used to conduct an exploration of this dataset. The second is a spatial dataset of Vermont towns that contains employment and income information acquired from the 1990 and 200 US censuses. It has been analyzed in this report using the ArcGIS software package.

Methods:

The Vermont Social Survey consisted of 641 male respondents and 859 female respondents. The age distribution of the respondents is shown in Figure 1. The data is spread from a range of 18 to 89 years old. The right-skew of the distribution illustrates that most of the respondents are younger than 50 years old. Of the 1500 respondents, 747 respondents work full-time, as depicted by the bar graph in Figure 2. The next most frequent category is retirement, which categorizes merely 231 of the respondents. The most infrequent category is comprised of those that are temporarily not working, which applies to 32 of the respondents. The pie chart depicted in Figure 3 shows that of those that are employed, most of the respondents are either very satisfied or moderately satisfied (43% and 41.3%, respectively). Only 3.5% of those surveyed were very dissatisfied and 8.5% of respondents classified themselves as "a little dissatisfied." The family income trends among the respondents, shown in Figure 4, exhibit an extreme right-skew, with the most frequently occurring income being $20,000. The data ranges from a minimum of $1,000 to a maximum of $395,000. Figure 5 shows a box plot comparing the differences in respondent's income between men and women. The median income of females is below $20,000, while the median income of males is nearly $30,000. The data also shows a wider spread in the range of male income than for females, with a maximum salary of approximately $85,000 for men and approximately $55,000 for women.

Results:

The US Census data shows an average 1999 per capita income for Vermont towns of $19,980.08. The minimum 1999 per capita income by town was $10,472, while the maximum was $37,210, resulting in a range of $26,738. The median household income in Burlington in 1999 was $33,070. The town of Shelburne had the highest median household income, which was $68,091. In 1990, the unemployment in Vermont ranged from 0 to 17.6%. In 2000, this upper limit of the range in unemployment increased to 18.8%. Though the maximum town rate increased, the average state-wide unemployment dropped from 4.8% to 3.61%. The rightmost panel of Figure 6 illustrates the unemployment rates from 2000. The towns that suffered from the highest unemployment rates are depicted in blue (9.3-18.8% unemployment) and purple (4.9-9.2% unemployment). Though these states are also scattered sparsely among central Vermont and there are a few in southern Vermont, most towns that experienced high rates of unemployment are concentrated in northern Vermont. Central and southern Vermont tend to be categorized by low unemployment rates depicted by the majority of white (0.-2.4% unemployment) and orange (2.5-4.8% unemployment) shading of the towns.

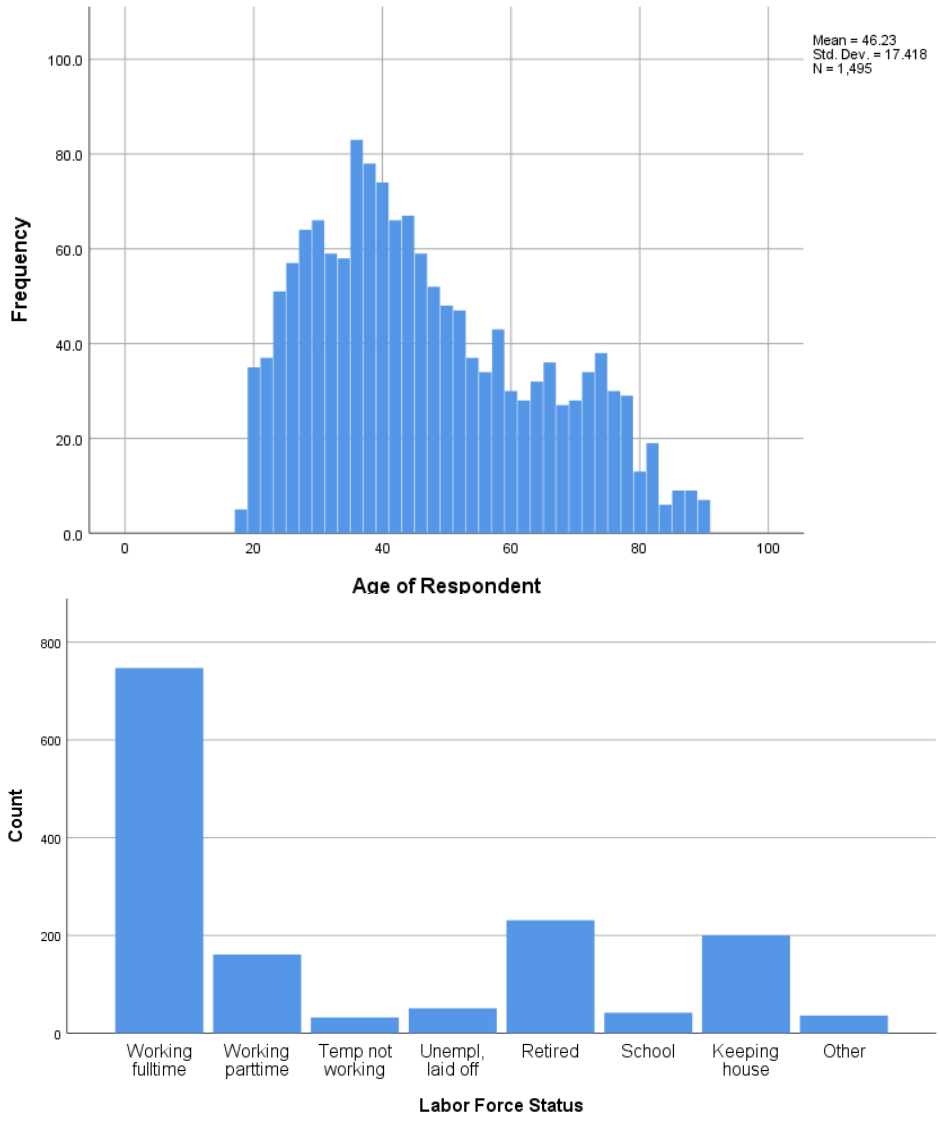

Figure 1: Histogram of Survey Respondent Age

Figure 2: Bar Graph of Survey Respondent Labor Force Status

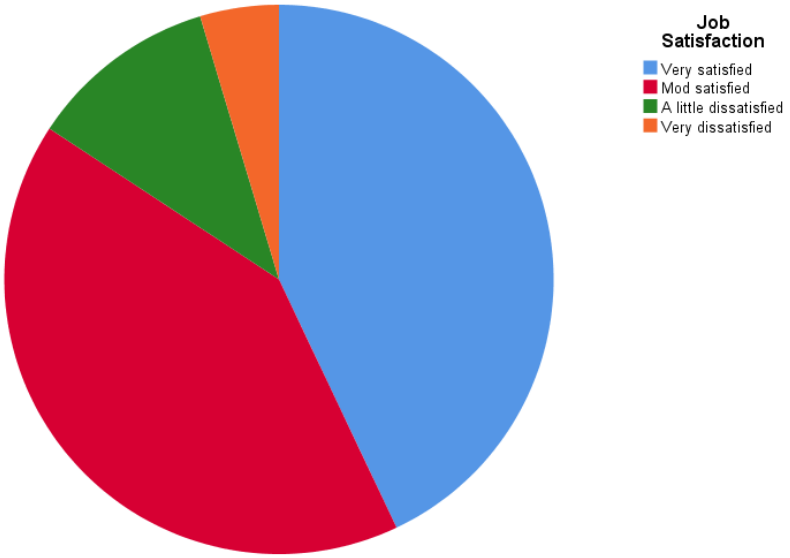

Figure 3: Pie Chart of Survey Respondent Job Satisfaction

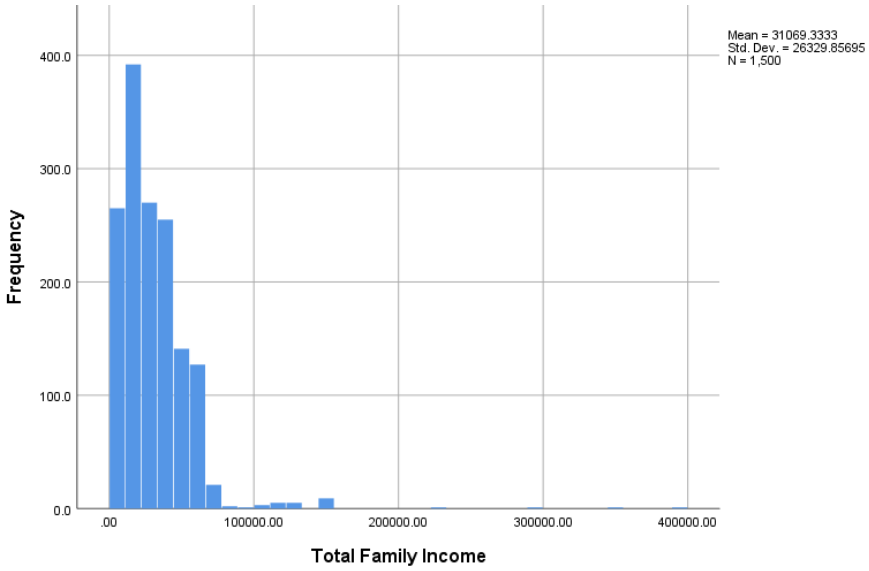

Figure 4: Histogram of Total Family Income of Respondents

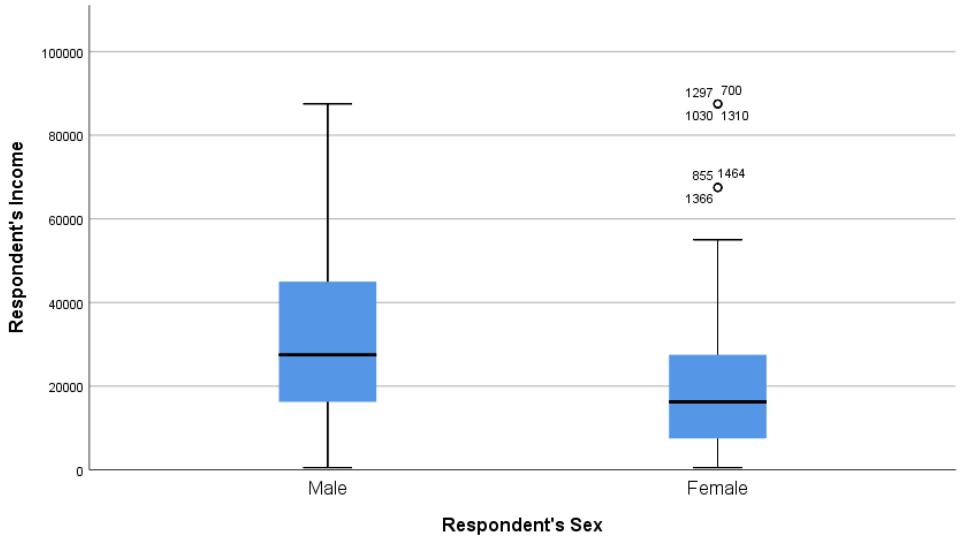

Figure 5: Box Plots of Respondent's Income Between Men and Women

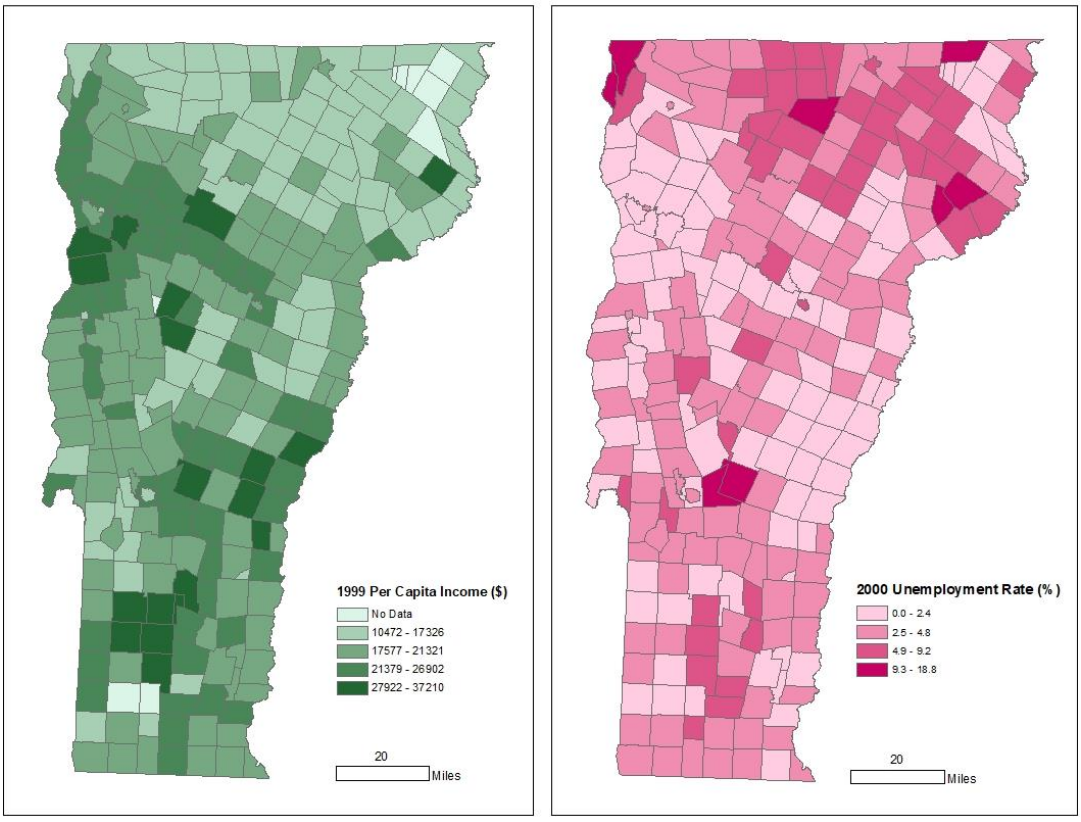

Figure 6: Map Depicting 1999 per Capita Income (Left) and 2000 Unemployment Rate (Right) by Town

Comparison of Cholesterol Levels in Urban and Rural Guatemalans

Introduction:

Socioeconomic status is an important underlying health determinant, affecting factors such as healthcare, environmental exposure, and health behavior. Here, We are interested in exploring the intersection of health and socioeconomic setting. A study has been conducted to examine the cholesterol levels in native Guatemalans. A total of 94 native Guatemalans volunteered for this study, 45 representing residents of an urban setting and 49 representing residents of a rural region. The total serum cholesterol level (in mg/L) of each volunteer was measured. The objective of my analysis is to determine whether there is a difference in the total serum cholesterol levels of urban native Guatemalans versus rural native Guatemalans.

Methods:

My analysis of this data involves a comparison of the mean behavior of the groups using a two independent samples t-test. This parametric test assumes that the data is normally distributed and that there is both independence and equality of variance between the two groups. To ensure that the data is normally distributed, we will plot it with a normal distribution curve transposed on top of the data. We will substantiate the assumption of equality of variance using Levene's Test for Equality of Variance at the 95% level of significance. Independence between the two groups can be assumed as the cholesterol levels of urban residents does not influence the cholesterol levels of rural residents (and vice-versa). With this t-test, we are testing the hypothesis that there is no statistically significant difference between the mean cholesterol level of urban native Guatemalans and the mean cholesterol level of rural native Guatemalans. We will test this hypothesis at the 95% level of significance.

Results:

While the cholesterol levels of the rural residents visually fits a normal distribution better than the cholesterol levels of the urban residents, a normal distribution for both data groups will be assumed for this analysis. Levene's Test for Equality of Variances at the 95% level of significance showed that the variances for cholesterol levels between the two groups were equal ($F = 1.6$, $p = 0.21$). The findings of my two independent samples t-test show that there is indeed a statistically significant difference in mean cholesterol levels between the urban Guatemalan residents and the rural Guatemalan residents ($t = 8.08$, $p < 0.0005$). The boxplots in Figure 1 show that the mean cholesterol level of the urban Guatemalans (216.87 mg/L) is higher than the mean cholesterol level of the rural Guatemalans (157.00 mg/L). The findings of the t-test confirm that this difference in means is significant at the 95% level of significance.

Discussion:

This statistical inference is powerful in that it allows us to conclude that the mean cholesterol level of the urban Guatemalans is higher than that of the rural Guatemalans. This provides enough evidence for further sociological and medical investigation into the variation in lifestyle habits between the two groups. This study is limited in scope. The cholesterol data was acquired as a convenience sample through the basis of volunteers. Because it was not acquired through a random sample, the existence of bias involved in the selection of the sample cannot be ignored or discounted. This data is limited with regards to generalization or inferences about the larger group from which it has been drawn. In the absence of random sampling and random assignment, this analysis only allows us to detect relationships in this sample only, and causality cannot be determined. However, this analysis provides a great basis for an observational study in which a more time-consuming or expensive random sampling method may be explored.

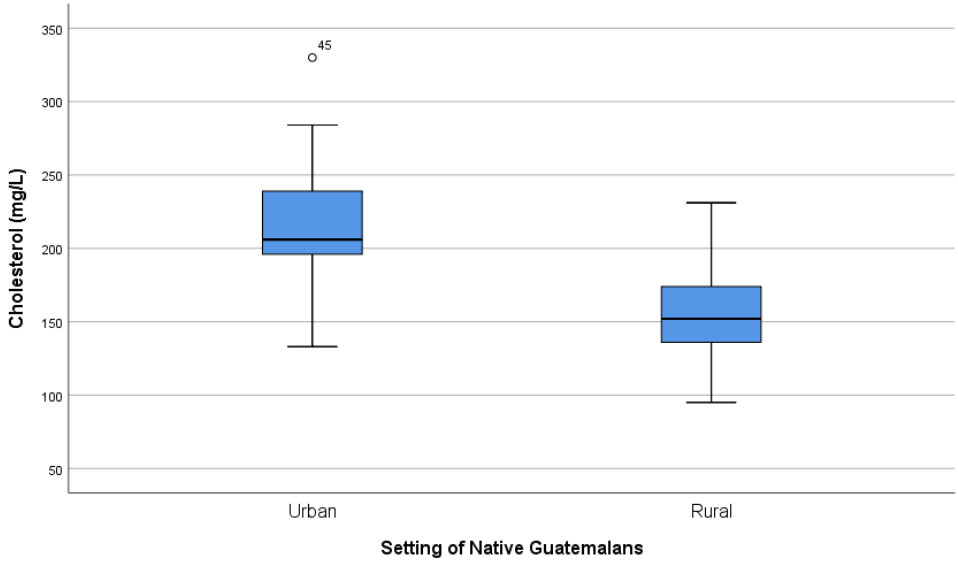

Figure 1: Boxplot of Cholesterol Levels (mg/L) of Urban vs. Rural Guatemalans

Implications of Habitat Conditions on Species Survival

Introduction:

Due to ongoing and potential loss of their sea ice habitat resulting from climate change, polar bears are listed in the United States as a threatened species. This is a well-known and highly publicized example that highlights the concept that habitat conditions are extremely crucial to the survival of a species. Though the northern flying squirrel (*Glaucomys sabrinus*) currently has a stable population conservation status, we are interested in studying the if the presence of certain habitat elements that are used for nesting and refugia lead to extended survival. This analysis seeks to answer the following questions: (a) does the mean lifetime differ for squirrels living in different habitat types?, (b) does lifetime for squirrels differ for those living in a managed woodlot from those living in other forest types?, (c) do older forests (greater than or equal to 125 years) provide improved survival for squirrels?, (d) do mature forests provide for the same survival potential as old-growth forests?

Methods:

To conduct this study, 349 infant squirrels were selected at random and randomly assigned to one of six habitat types, proportional to total acreage available in each habitat type, yielding slightly different group sizes in among the treatments. The squirrels were then monitored for their survival times. This analysis of this data involves a comparison of the mean behavior across the six groups using a one-way ANOVA and a post hoc multiple comparisons test. ANOVA testing assumes that the data is normally distributed and that there is both independence and equality of variance between the groups. Independence between the two groups can be assumed as the survival times of squirrels living in one habitat has no influence on the survival times of squirrels living in another habitat. With these methods, wewill be testing at the 95% level of significance.

Results:

The mean lifetime does indeed differ for squirrels living in different habitat types ($F=57.1$, $p<0.0005$). Lifetime for squirrels living a managed woodlot are less than those living all other forest types we evaluated in this study. More specifically, squirrels living in a managed woodlot live on average 5.3 months less than squirrels living in a 75-year-old mixed-deciduous coniferous forests ($p=0.001$), 12.3 months less than those living in a 175-year-old mature coniferous forests ($p<0.0005$), 14.9 months less than those living in 125-year-old coniferous forests with wind disturbances ($p<0.0005$), 15.5 months less than those living in 125-year-old coniferous forests with fire disturbances ($p<0.0005$), and 17.7 months less than those living in 350-year-old coniferous forests with old growth ($p<0.0005$). Older forests (greater than or equal to 125-years) provide an improved survival for squirrels compared to the younger habitat conditions we studied - namely the managed woodlot and the 75-year-old mixed deciduous-coniferous habitat ($p<0.0005$). For instance, squirrels living in 175-year-old mature coniferous forests lived an average of 12.3 months longer than squirrels living in a managed woodlot ($p<0.0005$) and 7 months longer than squirrels living in the 75-year-old mixed deciduous-coniferous habitat ($p<0.0005$). An important research question we asked was whether mature forests provide for the same survival as old-growth forests. The squirrels living in a 175-year-old mature coniferous forest lived 5.4 months less than squirrels living in the old growth forest, and this difference was statistically significant ($p<0.0005$), but there was no statistically significant difference in the average survival time of the squirrels living in the old growth forest and in a 150-year-old mature coniferous forest with fire disturbances ($p=0.249$) or in the 125-year-old mature coniferous forest with wind disturbances ($p>0.9$).

Discussion:

Since the squirrels were randomly sampled from laboratory breeding pairs and randomly assigned to their habitats, the scope and inference of our study is causal and generalizeable. This broad scope of inference allows us to make inferences about the entire population of northern flying squirrels at large, as well as to make inferences about the cause and effect of factors influencing their lifespan. Therefore, we can infer that older forests (greater than or equal to 125-years) provide an improved survival for northern flying squirrels compared to the younger habitat conditions due to more standing snags and downed trees, which are used for nesting and refugia. We can also infer that squirrels in woodlots will have decreased lifespans than all other unmanaged habitat types due to this lack of standing snags and downed trees. This information has great implications for conservationists in the potential situation that the northern flying squirrel becomes an endangered species in the future.

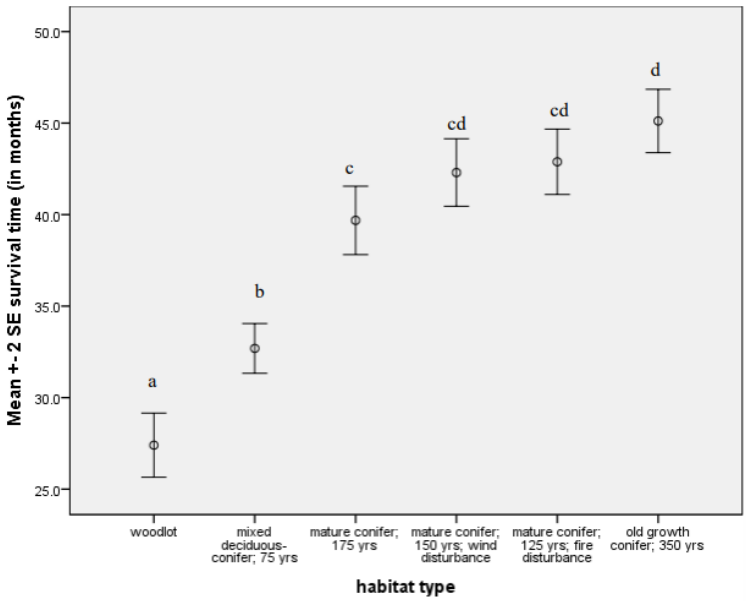

Figure 1: Means (±2 S.E.) of survival time (months) for the six habitat types, with labelled lettering indicating statistically significant groups.

Investigation of Precipitation Lapse Rates in Vermont

Introduction:

You may have heard the popular mantra "location, location, location" when searching for a new home, but this phrase doesn't just come to mind for your next real estate purchase. Determining the suitability of a new site is critical for businesses around the world. A wood products firm that specializes in the production of hybrid poplar for pulpwood is interested in purchasing a tract of land in Vermont's Northeast Kingdom. To determine site suitability, we must estimate the average annual precipitation at the site. The proposed tract is located at an elevation of roughly 1200 feet. We are interested in assessing whether precipitation can be predicted by elevation, and, if so, to create a map of Vermont precipitation that the woods products firm could use to select a tract of land for purchase. This analysis seeks to answer the following questions: (a) is there evidence of a relationship between average annual precipitation and elevation for the stations in Vermont? If so, what is the regression model for the relationship?, (b) How well does the regression model fit the data?, and (c) What does the model suggest as the likely annual precipitation for the site under consideration by the wood products firm?

Methods:

To conduct this analysis, we collected average annual precipitation data at a network of 31 weather stations (with known elevations) in Vermont. The analysis of this data involves a linear regression of precipitation against elevation. Linear regressions assume that the residuals are linear, normally distributed, independent, and homoscedastic. Initial exploration was conducted to assure these criteria. A scatter plot of precipitation versus elevation was created and linearity as well as homoscedasticity were visually determined. Furthermore, a quantile-quantile plot showed that the data was normally distributed. Independence can be assumed as the precipitation measurement at one weather station has no effect on the precipitation measurement at another. With these methods, wewill be testing at the 95% level of significance.

Results:

The results of our linear regression analysis indicate that there is indeed evidence of a relationship between average annual precipitation and elevation for the 31 weather stations in Vermont ($F = 78.9$, $p < 0.0005$). Approximately 73.1% of the variability in the average annual precipitation is explained by the elevation. This indicates that there is a strong relationship between the two variables. The regression model for the relationship is given by $y = 33.4 + 0.009x$ where y represents the average annual precipitation (in inches) and x represents the elevation (in feet). Figure 1 shows a scatter plot of the average annual precipitation versus the elevation with this fitted regression line. The 95% confidence intervals for individual predictions are also displayed on this figure. The width of the confidence interval increases slightly as x increases away from where the data points are concentrated. This demonstrates that our level of certainty in our predictions decreases as we move to a range where data is sparse. Figure 2, which shows a map of precipitation distribution in Vermont. This map shows that the model predicts a low annual precipitation rate within the 30-40-inch range along the western edge of the state (corresponding to lower relative elevations), precipitation rates of 41-50 inches for most of the state, and a few scattered areas of high precipitation. The proposed tract under consideration by the wood products firm is located at an elevation of roughly 1200 feet. The model suggests that the likely annual precipitation for this site is 44.2 inches.

Discussion:

Regression analyses is an interpolation model by construction and extrapolating beyond the data is not ideal. For this reason, the model can only be used to predict precipitation within the elevation range of 226 ft to 3950

ft. Furthermore, since the data is unique to Vermont, this model can only be used for interpolation purposes within the state. For instance, the model could not be used to predict precipitation for a site in the foothills of the Laurentian mountains in northwestern Quebec. Despite these limitations, the model is highly valuable for other applications beyond its use by the wood products firm. It is a useful tool to predict the precipitation for a proposed ski resort in Vermont at 3500 feet (predicted annual precipitation: 64.9 inches).

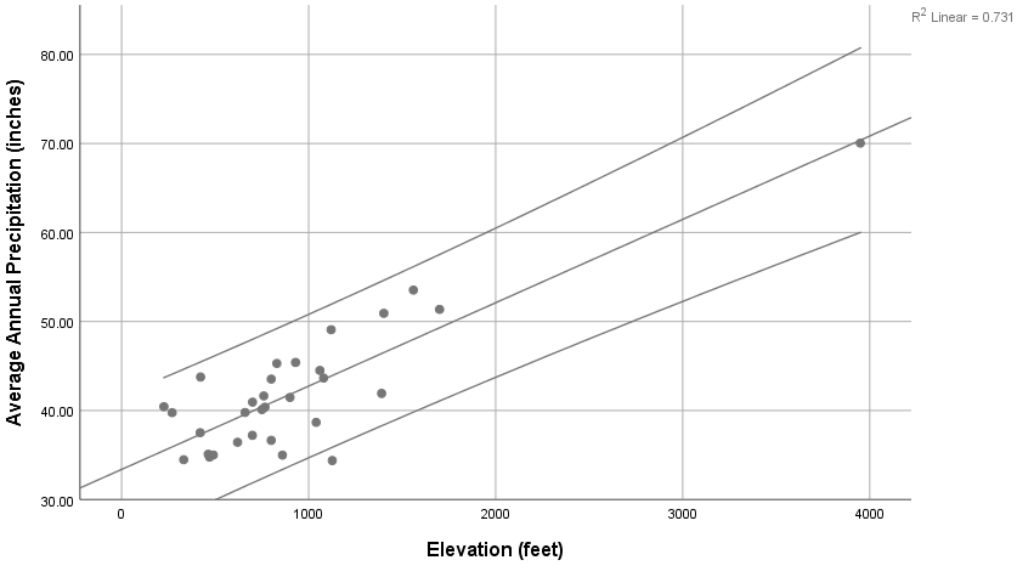

Figure 1: Scatter plot of average annual precipitation versus elevation with fitted regression line and 95% confidence intervals for individual predictions

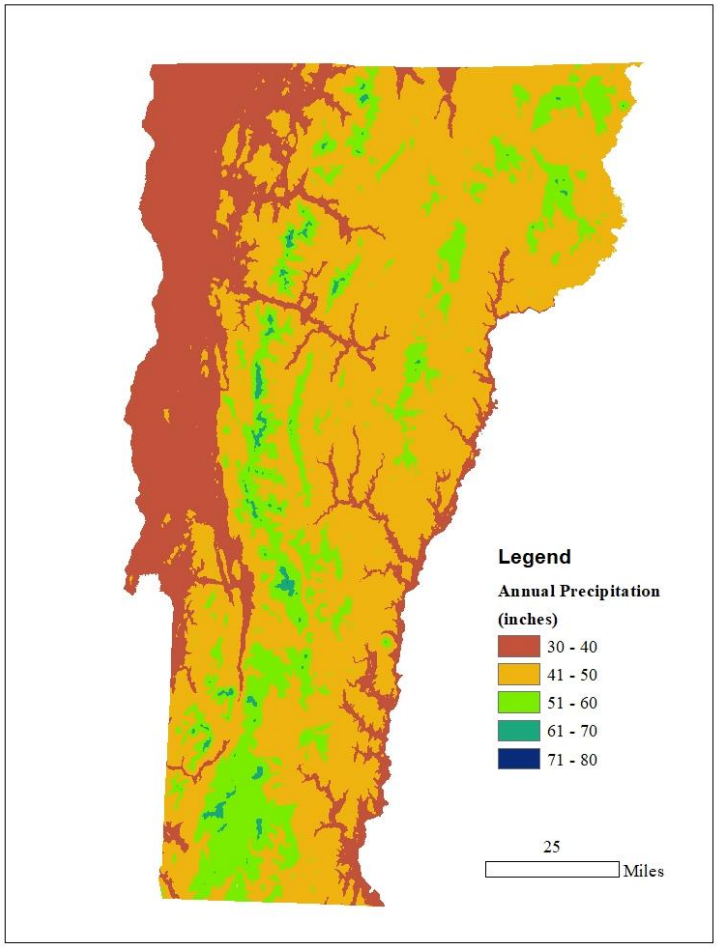

Figure 2: Map of precipitation distribution in Vermont

Predicting Female Life Expectancy

Introduction:

Acciaroli, Italy has long been a location of interest for researchers trying to discover why the town has a disproportionately high number of centenarians living in its population of approximately 2,000. Acciaroli is well-known for its low rates of heart disease and Alzheimer's. Scientists and laymen alike have long been fascinated by the fountain of youth and longevity – whether due to genetics, lifestyle behaviors, or both. Gender also plays an important, yet often overlooked, role in longevity, as well. In the contemporary United States, the mortality rate of males is 60% higher than that of females. This analysis examines female life expectancy and demographic data containing the following variables: percent of the population living in urban areas, number of doctors per 10,000 people, number of hospital beds per 10,000 people, gross domestic product per capita, and number of radios per 100 people. The research seeks to answer the following questions: (a) Which of the five independent variables best demonstrates a linear relationship with female life expectancy? (b) Is a natural log transformation successful in creating a linear relationship for the independent variables that are not linearly related to female life expectancy? (c) Which log transformation model best explains the variability in female life expectancy among countries? (d) What multiple linear regression model best explains the variability in female life expectancy? (e) Among these developed regression models, which does the best possible job of predicting female life expectancy?

Methods:

To conduct this analysis, we collected female life expectancy and demographic data from 122 nations. The analysis of this data involves linear and logarithmic regressions of life expectancy against the explanatory variables, as well as multiple regression analysis. Scatter plots of life expectancy versus each of the independent variables were created and inspection was used to determine which of the five variables most clearly demonstrates a linear relationship with female life expectancy. Next, a natural logarithmic transformation was conducted for each of the independent variables that were not linearly related to female life expectancy. The F-statistic and p-values were used to determine which equations represented statistically significant relationships, and the R^2 value was used to identify the natural logarithmic model that best explains the variability in female life expectancy among countries. Lastly, a multiple linear regression model was created using all five independent variables in the form (untransformed or transformed) that best approximates a linear relationship with female life expectancy using SPSS's stepwise procedure. With these methods, wewill be testing at the 95% level of significance. For modeling purposes, the variables were given abbreviated names as follows: female life expectancy (lifeexpf), percent of the population living in urban areas (urban), number of doctors per 10,000 people (docs), number of hospital beds per 10,000 people (hospbed), gross domestic product per capita (gdp), and number of radios per 100 people (radios).

Results:

As shown in Figure 1, the results of our linear regression analysis indicate that the independent variable that most clearly demonstrates a linear relationship with female life expectancy is the percent of population living in urban areas. The regression model for the relationship is given by: $lifeexpf = 50.58 + 0.32urban$. Only 49.5% of the variability in female life expectancy is explained by the percent of the population living in urban areas. Figure 2 shows that natural log transformation was successful in creating a linear relationship for the following variables: number of doctors per 10,000 people, number of hospital

beds per 10,000 people, gross domestic product per capita, and number of radios per 100 people. The natural logarithmic regressions show that there are indeed individual logarithmic relationships between female life expectancy and each of these four variables. The relationship between female life expectancy and number of doctors per 10,000 people is given by: $lifeexpf = 57.1 + 6.32\ln(docs)$ (F = 402.6, p < 0.0005). About 77.2% of the variability in female life expectancy is explained by the number of doctors per 10,000 people. The relationship between female life expectancy and number of hospital beds per 10,000 people is represented by the model: $lifeexpf = 39.0 + 8.79\ln(docs)$, and 53.2% of the variability in female life expectancy is explained by this variable (F = 129.8, p< 0.0005). The relationship between female life expectancy and gross domestic product per capita is given by: $lifeexpf = 21.5 + 6.18\ln(gdp)$, and 69.3% of the variability in female life expectancy is explained by the gross domestic product per capita (F = 270.5, p < 0.0005). Figure 2 shows that this model does contain heteroscedasticity. The relationship between female life expectancy and number of radios per 100 people is given by: $lifeexpf = 41.5 + 8.20\ln(radio)$, and 48.3% of the variability in female life expectancy is explained by radios per 100 people (F = 112.1, p < 0.0005). A multiple linear regression model using all five independent variables is given by the following equation: $lifeexpf = 41.7 + 4.12\ln(docs) + 1.87\ln(gdp) + 1.68\ln(radio)$, and 81.8% of the variability in female life expectancy is explained by this particular combination of variables (p<0.0005). If wewere interested in using a regression model that did the best possible job of predicting female life expectancy, wewould choose this multiple regression model because it has the highest R^2 value of the models. This means that this combination of these three transformed variables best explains the variability in female life expectancy among countries.

Discussion:

Since regressions are interpolation models, extrapolating beyond the range of the predictor variables is not ideal. The multiple regression model can only be used to predict female life expectancy within the range of 0 to 43 docs per 10,000 people, a GDP of 120 to 22,470, and 3 to 135 hospital beds per 100 people. Because 122 countries out of the 195 in the world were used for this research, the model is not unique to a particular country and could be used to predict female life expectances globally.

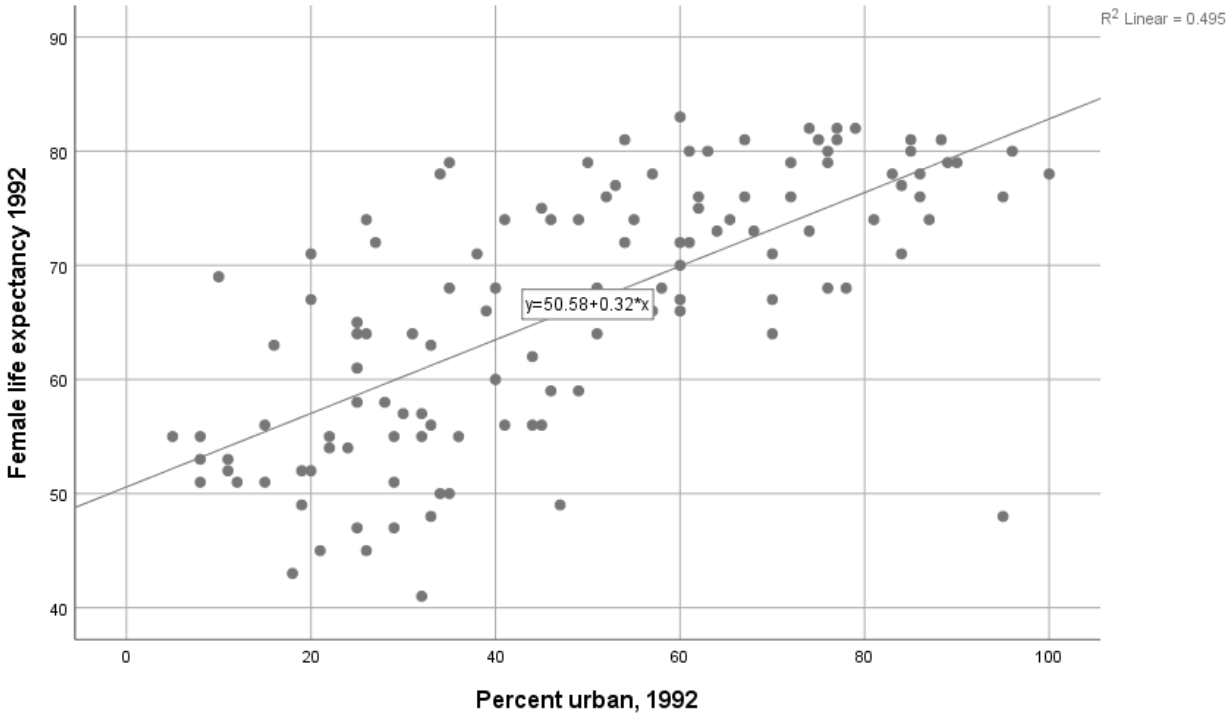

Figure 7: Scatter plot of female life expectancy among 122 countries versus percent of the population living in urban areas.

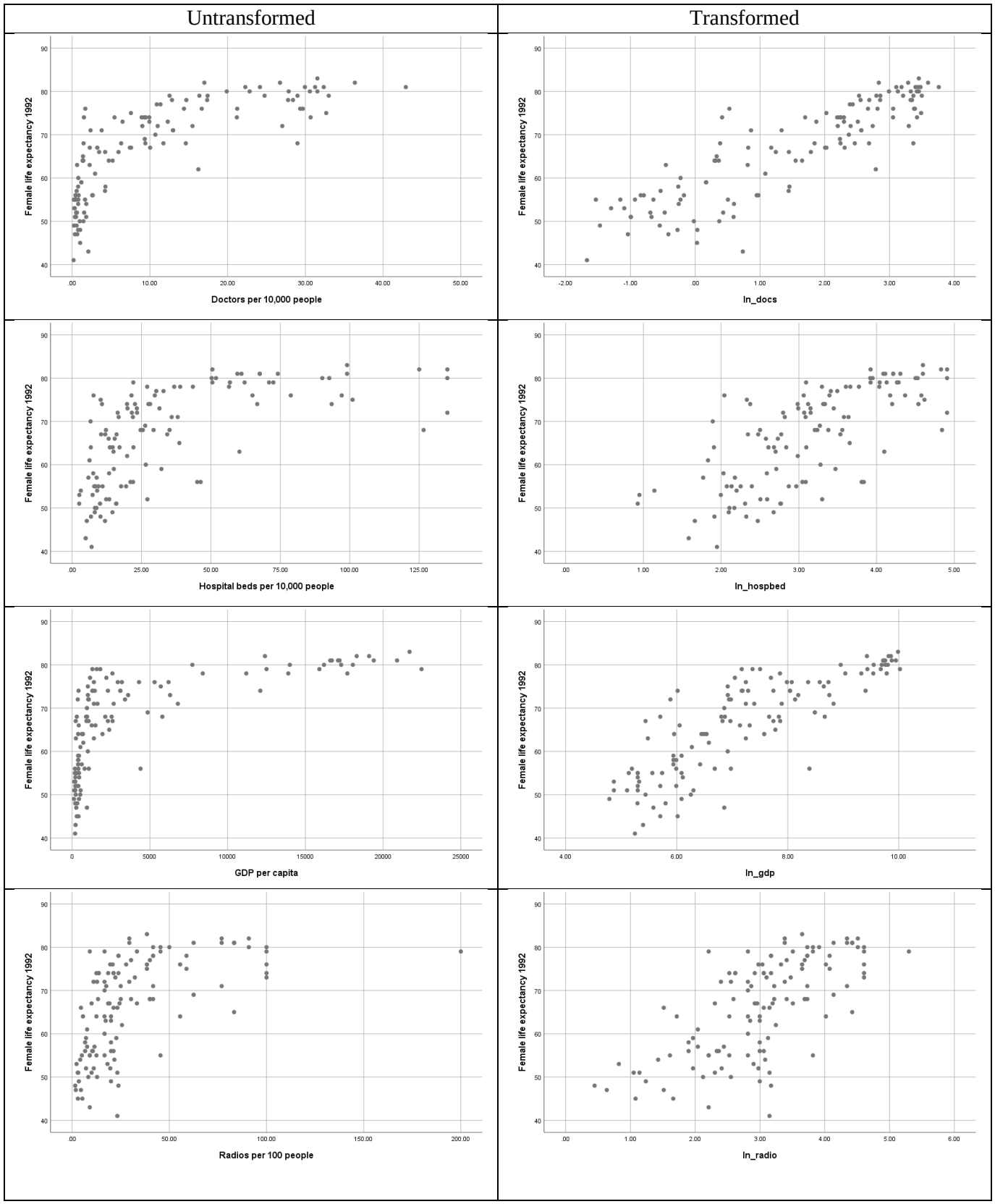

Figure 8: Compilation of untransformed and transformed scatterplots of female life expectancy versus number of doctors per 10,000 people, number of hospital beds per 10,000 people, gross domestic product per capita, and number of radios per 100 people (top to bottom).

Spatial Analysis of Housing Values in Baltimore County, Maryland

Introduction:

The housing market is intricately linked to the economy, acting as a money multiplier and affecting many aspects of our daily lives. For instance, the "wealth effect" is a commonly observed phenomenon in the US in which individuals and families with residential real estate feel an increase in financial security and tend to spend more when their home values increase. Alternatively, when home values are low, consumer spending is less, causing a negative effect on the GDP. In this way, housing value is quite powerful. Adding to the complexity of housing value, not only does it vary over time, but it also varies in space. One factor may increase a home's value in one location yet decrease it when it is present in another location in Baltimore, Maryland. Through this research, we seek to explore relationships correlating housing values to various independent variables. Then, we will investigate whether a multiple regression model and/or a geographically weighted regression (GWR) model improves upon these bi-variate relationships.

Methods:

To conduct this analysis, we used block group property data from the US Census for Baltimore county. This geodata set contains median house values, percentage of single family detached homes, percent owner occupied housing, percentage vacant housing, median number of rooms per house, median years old for housing, distance to downtown from census block centroid, distance to protected area from census block centroid, mean tree canopy coverage for Baltimore county census blocks. Using SPSS statistical software package, median house values were linearly regressed against three variables of interest to build a model that could be used to predict housing prices. These three variables are: canopy cover, distances to downtown (from census block centroid), and percentage of vacant housing. Next, GWR4, an open source package, was used to develop a GWR of median house value. The median house value was regressed against percent single detached homes, percent owner occupied housing, percent vacant housing, median number of rooms per house, median years old for housing, distance to downtown from census block centroid, distance to protected area from census block centroid, and mean tree canopy coverage for the block group to build a model that could be used to predict housing prices. ArcGIS, a geographic information system for working with maps and geographic information, was then linked to the GWR data to create paneled maps that illustrate the results.

Results:

The median house values across block groups in Baltimore county range from $0 to $599,400 with a mean of $69,204 and a standard deviation of $46,782.69. The percent canopy cover across block groups in Baltimore county ranges from 0% to 76.3% with a mean of 7.3% and a standard deviation of 11.7%. As shown in the first panel of Figure 1, the results of our linear regression analyses indicate that there is indeed evidence of a relationship between median housing value and mean tree canopy cover in Baltimore ($F = 201.97$, $p < 0.0005$). Only 22.2% of the variation in median house value is linearly explained by mean tree canopy cover ($R^2=0.22$). The second and third panels of Figure 2 show that there is a very weak linear correlation between median housing value and distances to downtown as well as between median housing value and percent vacant housing. A mere 6.1% of the variation in median house value is explained by distance to downtown and a meager 5.5% of the variation in median house value is explained by percentage of vacant housing ($R^2=0.061$, 0.055). Of the three relationships, the mean tree canopy cover has the highest predictive power. However, these bi-variate relationships do not do very well in predicting median house value by census block group in Baltimore in general.

The summary statistical output generated by GWR4 shows an adjusted R^2 value for the single (global) multiple regression model is 0.38. For the GWR model, the adjusted R^2 value is 0.58. Thus, we can conclude that a multiple regression and a GWR improves upon the simple bi-variate relationship explored using SPSS as both R^2 values are higher than the R^2 value of the bi-variate relationship between median housing value and mean tree canopy. Between the two models, the GWR model outperforms the global model. The GWR of median house value was visualized through the creation of Figure 2. The leftmost panel of this map displays the t-statistic for canopy cover, while the rightmost panel shows the local R^2 values for the GWR model for each block group. From the left pane of Figure 2, we can see that the palest shade of pink depicts the regions of Baltimore where mean tree canopy cover is not a significant predictor in the outcome of median house value. These regions are concentrated in two distinct spots in inner Baltimore. Alternatively, the center of Baltimore's northern edge contains a region where mean tree canopy cover is a very significant predictor of median house value. The right pane of Figure 2 shows that the global model is the strongest in the northwest region of Baltimore and weakest in the inner city and southern regions.

Discussion:

GWR is the optimal method to look at controls on house values in this area and it is intuitive that the relationship between house value and any control would not be constant from one location to the next. This adds a layer of complexity that is important to consider while seeking to purchase a home. A prospective home buyer from a rural area may not be attuned to this nuance, instead believing that a surplus of canopy cover would partially explain the inflated asking price of a home in the inner city. While the canopy cover may add to the value in a rural region, it could provide justification to negotiate a lower price as it may prove to attract illicit activity or crime. In this way, a GWR model is an extremely powerful tool.

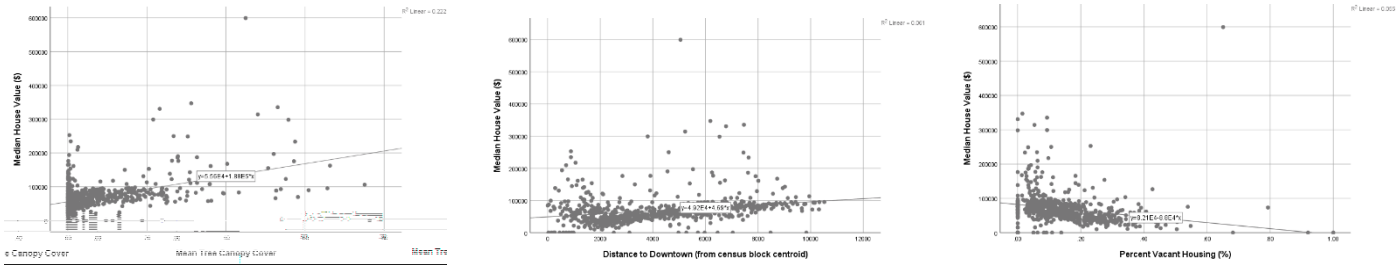

Figure 1: Linear Regression of Median House Value ($) versus Mean Tree Canopy Cover, Median House Value ($) versus Distance to Downtown, and Median House Value ($) versus Percent Vacant Housing

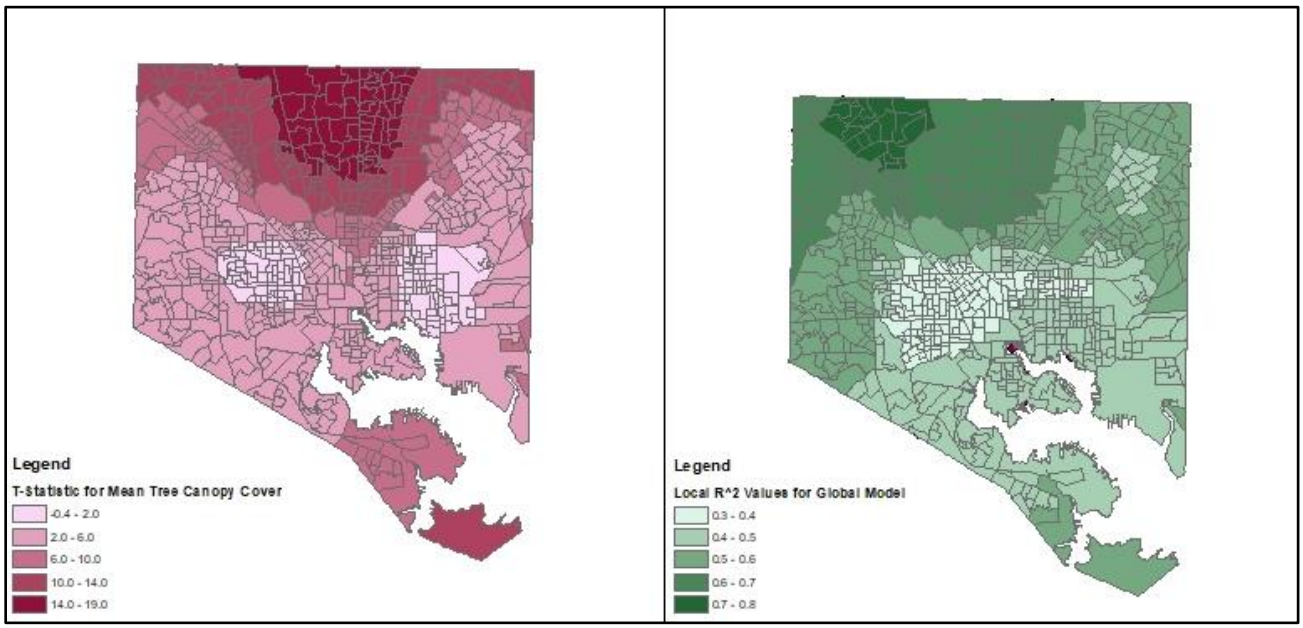

Figure 2: Map Depicting T-Statistics for Mean Tree Canopy Cover (Left) and Local R^2 Values for the Global Model (Right)

Autocorrelation and Interpolation of Contaminated Sites

Introduction:

Tetraethyl lead (TEL) is a compound that was combined with gasoline to boost octane ratings and reduce engine knocking by raising the temperature and pressure that auto-ignition occurs across the world since the 1920s. However, discoveries of its toxicity on human health and the environment have caused it to be phased out of gasoline in many areas of the world, with the 2002 Earth Summit pushing for its total global ban. To date, Algeria, Iraq, Yemen, Myanmar, North Korea, and Afghanistan still use leaded fuel due to economic reasons. Through this research, we seek to examine leaded gasoline contaminated sites in Myanmar to discover whether the data exhibits spatial autocorrelation and/or anisotropy as well as to compare two different map interpolation methods.

Method:

Two datasets were used for this research as two separate studies Myanmar were evaluated. The first data set was collected by testing 100 sampling sites for soil lead concentrations on a regularly sampled grid of a gas station in Naypyidaw at 1 meter spacing. The second data set is a geodatabase of randomly selected soil samples in a 35 x 35 km area at The Taunggyi Bird Sanctuary, a protected area of wetlands in the Shan Hills. Manual calculations using Excel were conducted on the first dataset to calculate and plot an experimental variogram for two transects, one oriented North-South and the other oriented East-West. The variogram was calculated out to lag distances of 7 meters for each transect. The second dataset was evaluated using ArcMap to interpolate lead concentrations using both the Inverse Distance Weighting (IDW) method as well as the Kriging method. The results were then mapped to compare the results of the two interpolation methods.

Results:

The resulting graphs of the manual calculations using Excel are shown in Figures 1 and 2. The theoretical variogram that most suitably fits the North-South transect is a spherical trend, while an exponential fits the theoretical variogram of the East-West oriented transect the best. The approximate value of the sill for Figure 1 is 6.9, the range is 5, and the nugget is estimated to be 0.5. Figure 2 shows that the sill is 6, the range is 3, and the nugget is 1.5. This indicates that the data for the North-South transect is correlated from 0 to 5m, while the East-West transect is correlated from 0 to 3m. The interpolated maps for the lead geodatabase, based on the IDW and Kriging methods are shown in Figure 3. A comparison of the two methods shows that the Kriging method creates a more detailed gradient of the interpolated lead contaminant levels than the IDW method. This makes sense as the Kriging method relies on spatial autocorrelation, while the IDW does not use statistical models.

Discussion:

The results of the semi-variance analysis would be useful in designing a sampling protocol to select independent soil samples on this site. The advantage of using the Kriging method over the IDW method for interpolation is that it can provide higher level gradients for more precise and detailed interpolations.

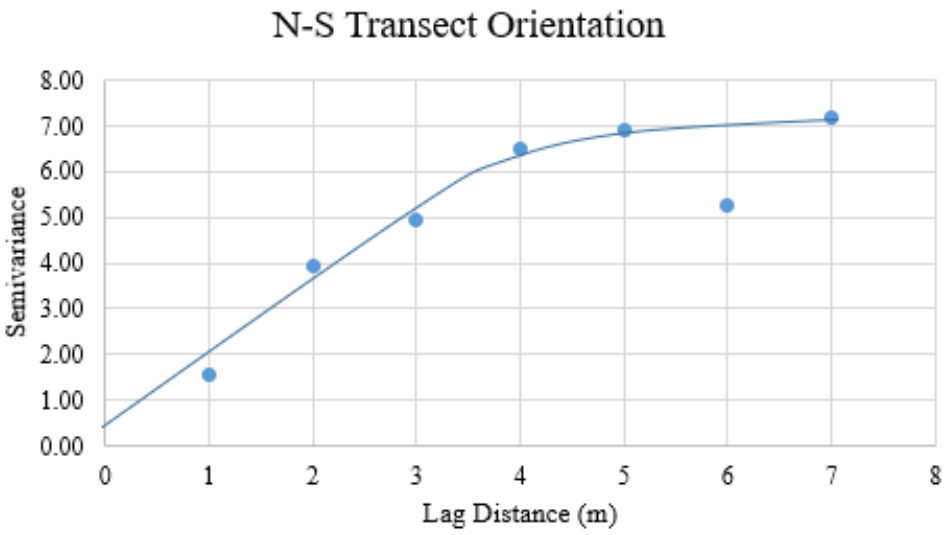

Figure 9: Variogram of North-South Transect with Spherical Fit

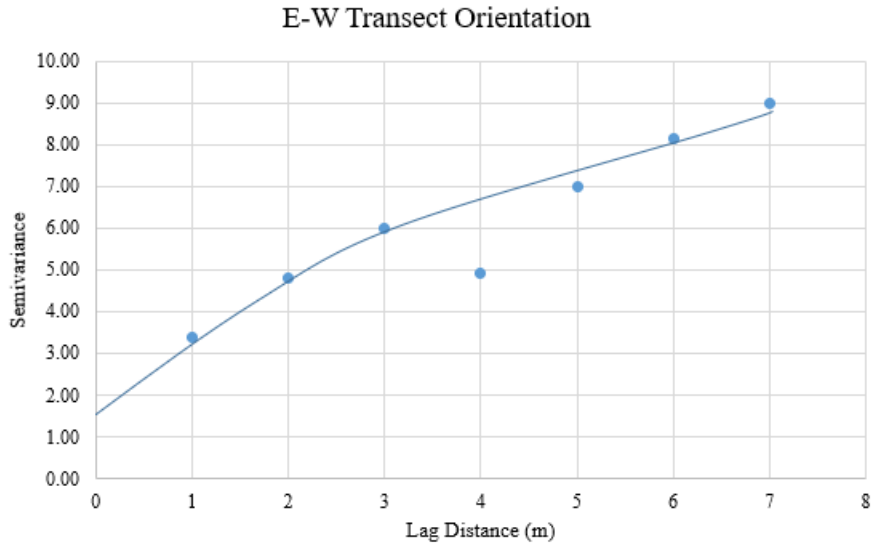

Figure 10: Variogram of East-West Transect with Exponential Fit

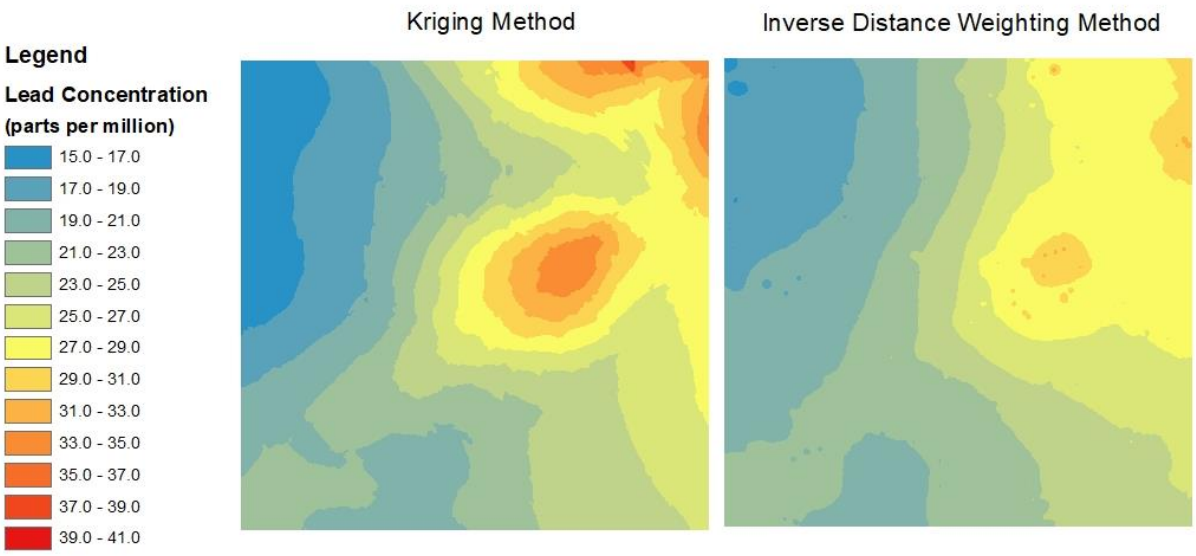

Figure 11: Lead Concentration Interpolation Schemes Using Kriging (left) and Inverse Distance Weighting (right) Methods

Predicting Erosion from Forestry Practices

Introduction:

A well-known and leading cause of soil erosion around the world are the logging and clear-cutting of rain forests. In the Amazonian rainforests of Brazil, a football field-sized area is cut each second, leaving large spans of land susceptible to wind, rains, and floods that cause it to erode. Erosion is a large issue as once plant cover is gone, there are no longer roots to hold the soil together and retain water and nutrients vital to future vegetation. Minimizing the effects of forestry on erosion is a key concern not just in the rainforests, but in any steep, forested landscape. Logging roads on steep hills exacerbate erosion. We are interested in studying factors leading to erosion associated with logging roads and the tendency of hill slopes to gully immediately below culverts. Through this research, we seek to develop a model that could be used to predict the odds of gullying at a site based on some combination of road length, road grade, and slope steepness.

Method:

In our study, we examined 275 culvert outlets throughout Vermont and recorded at each whether gullying of the hill slope had occurred. In addition, we measured the length of the road draining to each culvert, the grade of the road, the hill slope steepness on which the road was constructed (classified as 0 if less than a 40% slope angle, 1 if greater than or equal to 40% slope angle), and whether the culvert spacing exceeded currently adopted design standards. Several tools using SPSS software were used to explore relationships between the variables. For a comparison of two categorical variables, contingency analysis is used. Logistic regression models are used to explore relationships between a binary response variable and one or more explanatory variables, when at least one explanatory variable is continuous. When testing logistic regression models, we evaluated the chi-square statistic associated with the model to test whether the explanatory variables predict the odds of gullying. With these methods, we will be testing at the 95% confidence level.

Results:

Best management practices suggest that erosion of hillslopes would be minimized when culvert spacing followed recommended design guidelines. However, there is not enough evidence in our data to suggest that there is a relationship between gullying at a site and exceedance of design standards for culvert spacing ($\chi 2 = 2.64$, $p = 0.107$). The evidence suggests that there is, in fact, a relationship between gullying and road grade ($\chi 2 = 20.17$, $p = 0.165$). There is not enough evidence to suggest that there is a relationship between gullying and road length ($\chi^2 = 35.27$, $p = 0.064$). Furthermore, our data shows that there is also not enough evidence to suggest that there is a relationship between gullying and road length ($\chi^2 = 35.27$, $p = 0.064$). Table 1 shows constructed models that could be used to predict the odds of gullying at a site based on various combinations of road length, road grade, and slope steepness. Of these developed logistic regression models for odds of gullying, the evidence suggests that the best model has three explanatory variables as shown below between length and slope ($\chi^2 = 29.288$, $p < 0.0005$):

$$\ln(y) = -0.008(x_1) - 0.081(x_2) + 0.16(x_1 * x_2) - 0.371$$

where:

y = gullying occurring or not

x_1 = length

x_2 = slope

We concluded that this is the best model because it contains the highest chi-squared value as well as the highest qualitative measure of percent correctly classified. This model correctly classifies 68% of cases. A drop-in-deviance test was conducted to compare a model with two explanatory variables (Length + Slope) with a model with three explanatory variables (Length + Slope + (Length*Slope)). The results of this test show that the fuller model with three explanatory variables is better than the model with two explanatory variables at the 95% level of significance and with 1 degree of freedom difference ($\chi 2 = 29.288$, $p = 0.005$).

Discussion:

Having a model that can be used to predict the odds of gullying at a site based on this combination of road length and slope steepness is an incredibly useful tool that can be used to minimize the effects of forestry operations in steep, forested landscapes. Because these culvert locations are all from the Vermont area, this prediction model is specific to the state and it would be unwise to apply this model for predictions in other locations.

Table 1: Developed models that could be used to predict the odds of gullying at a site based on various combinations of road length, road grade, and slope steepness.

Variable(s) in model	Chi-squared	df	p	-2 log likelihood	% correctly classified
models with one explanatory variable:					
Length	0.48	1	0.827	363.684	62.5
Grade	2.397	1	0.123	361.334	62.5
Slope	21.430	1	<0.0005	-	-
Models with two explanatory variables:					
Length + Grade	2.491	2	361.241	361.241	62.5
Length + Slope	21.397	2	342.335	342.335	65.8
Grade + Slope	22.524	2	342.335	342.335	65.8
Models with three explanatory variables					
Length + Grade + (Length * Grade)	3.482	3	360.250	360.250	60.7
Length + Slope + (Length * Slope)	29.288	3	334.444	334.444	68
Grade + Slope + (Grade * Slope)	23.116	3	340.616	340.616	65.8

Spatially Distributed Regression Modeling in GIS

Introduction:

Biologists around the world are interested in studying the distribution patterns of species. The three basic types of population distribution on a regional range are clumped, random, and uniform distributions. Clumped distributions are characterized by minimal distances between neighboring individuals, random distributions are patterns in which the position of an individual is independent of another individual, and uniform distributions involve individuals that are evenly spaced. Species distribution is especially important in identifying changing patterns that may indicate endangerment, as well as in predicting habitat areas. We are interested in studying the wolf distribution within Yellowstone National Park to predict habitat most favorable for a rescued wolf that is planned to be introduced to the park so that it may more easily adjust to its unfamiliar environment.

Method:

Through our research, we have mapped existing wolf dens and associated these locations with several possible explanatory factors, including elevation, vegetation communities, known elk habitat (the primary food source for wolves), existing wolf pack territories, and distance to sites occupied by humans developed sites (roads, campgrounds). Utilizing ArcMap, these mapped layers were joined to apply the following previously developed logistic regression model and predict wolf habitat:

$$\ln(odds_{wolf}) = -0.847 + (1.386 * elk) + (0.201 * humans) - (2.197 * wolf)$$

Where the predictor variables, elk, $humans$, and $wolf$ measures the proximity to humans, elk, and the distance from other wolf packs, respectively. The dependent variable, $odds_{wolf}$, represents the odds of finding a favorable wolf habitat at a given location.

Results:

The range in suitable wolf habitat probabilities was found to be within the range of 5.5% to 82.4%. Table 1 describes the distribution of the areas in different probability classes. As researchers, it is ideal that the areas of Yellowstone National Park with the greatest probabilities of containing a favorable wolf habitat are chosen. The 81% to 100% range class contains a total area of 1,706.91km^2. This area is illustrated in Figure 1 in the darkest shade of the legend, and these regions are scattered throughout the area. The map indicates that top left and right corners as well as the middle bottom area of Yellowstone should be avoided for releasing the wolf.

Discussion:

The translation of a regression model into a spatial representation is a powerful tool for predictions. In this context, it allows us to recommend areas that are favorable for releasing the wolf in order to ensure its survival. These areas are at an optimal distance from sites occupied by humans (roads, compgrounds), close to known elk habitats, close to existing wolf pack territories for integration, and situation at an ideal elevation with sufficient vegetation.

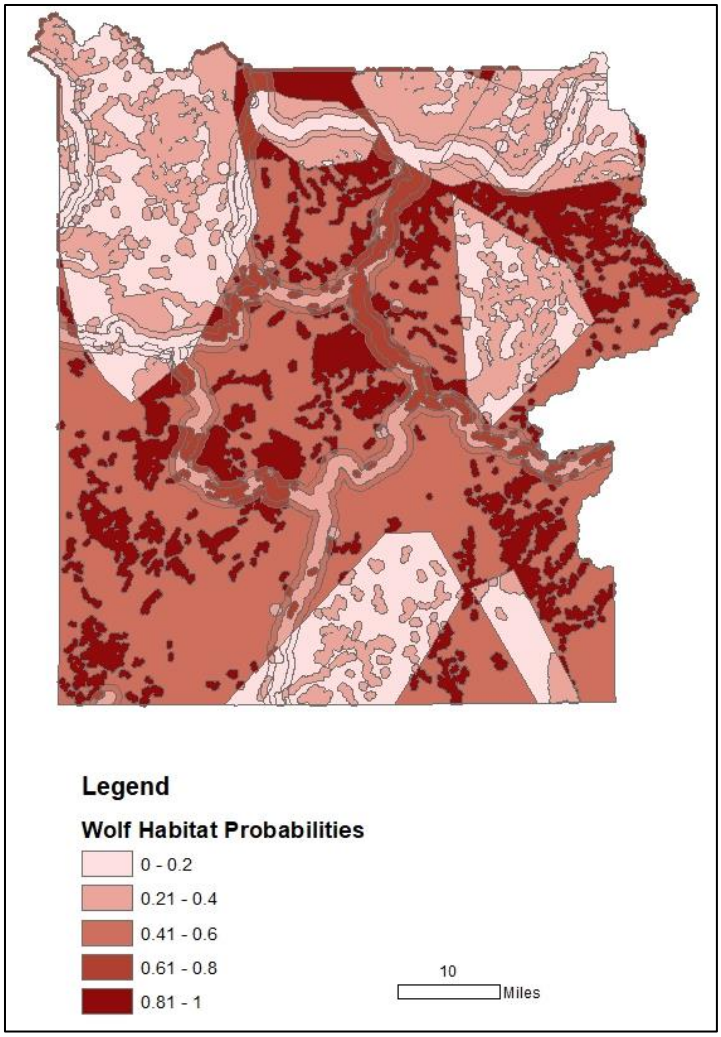

Figure 12: Mapped logistic regression model of wolf habitat in Yellowstone National Park

Table 2: Summaries of areas in different probability ranges of wolf habitats

Probability Range	Area (km^2)
0 – 0.2	1,779.80
0.21 – 0.4	2,002.63
0.41 – 0.6	3,065.24
0.61 – 0.8	497.58
0.81 – 1	1,706.91