Rural vs. Urban Quality of Life

Hannah Robertson

**Abstract**—In this project, I used governmentally curated data on health, census statistics, and demographic information to analyze general quality of health across urban and rural counties by state and overall. I munged this data to create a database of each locality’s traits, and then compiled each of 33 distinct health metrics as measured by the CDC and life expectancy as measured by census data across each state and nationwide. I grouped these comparisons by NCHSUR index, which is a way to describe how urbanized an area is, and by doing so compared these statistics by similarly rural or urban areas.

**Index Terms**— none applicable

Template by the IEEE Computer Society

—————————— ◆ ——————————

# 1 Introduction

My family recently move to a quite rural area in Northern Arizona. While we are very happy there, my mother was involved in a severe automobile accident about a year later. The barriers she faced during her recovery led me to wonder how her recovery would have been if she had lived in a more urban area. This led me to further wonder; Americans think of rural life as idyllic, much more healthful than the rushed, polluted life in cities, but is that really the truth?

Rural areas do have the benefit of cleaner air and better access to nature, but they also have severe health downsides. A much higher percentage of people in rural areas are blue collar workers, access to health care is more difficult due to long distances between health centers, and often average income is lower than in the cities. When presented with the opportunity to create a projet over the course of the semester, I knew that it would be a fantastic opportunity to explore the facts behind this question.

My initial methods were merely to collate every scrap of data I could find about health in the United States by county. The CDC, the Census Commission, and the United States Department of Agriculture. I decided to concentrate on health statistics as measured by county, as measuring only by cities would defeat my idea to attempt to study rural areas, and I couldn’t find all the data I needed when I attempted to measure by census area.

Once I had gathered all my data, I had a lot of questions that I wanted to answer. Is the urban lifestyle or the rural one more healthy? Are there any risk factors that are associated with either lifestyle? Who lives longer, and why? What are the most common causes of death in each?

I decided to try to find out who lived longer, and who was healthier overall. One of the data sets I had found was created by the US Census Bureau covered 33 metrics of health which I decided to use as a baseline.

# 2 Exploratory Data Analysis

My first step was to collate all my data. I had several different documents, each holding information that I needed. All of my documents were .csv or excel files, so they were easy to process. This is advantageous, because most of the time I only needed one or two columns from each file, collecting income from one, life expectancy from another, and so on.

I used a Python dataframe to collate all this information into one file, and used saved that dataframe as a .csv as a base for further calculations. I created one file for each state, and one for the fifty states as a whole.

From each foundational document, I then used Python’s MatPlotLib library to create visualizations for selected data. For each file, I created a histogram for average life expectancy in rural and urban areas. Then, I created two pie charts per health metric, one for rural average, and one for rural average.

The results I got when tabulating, averaging, and graphing my data were surprising.

Rural populations scored one to three percentage points worse on almost every health metric, from preventative ones cancer and blood pressure screenings to outcomes like percentage of the population with cancer and athsma. In the national averages, there was not a single metric that rural populations scored more positively on.

Chart, histogram

Description automatically generatedHowever, despite this discouraging data, rural populations have a higher average life expectancy. Across my averages, people in rural areas were recorded as living longer on average than their urban counterparts.

Figure 1: A histogram comparing average life expecancy in rural and urban areas across the US.

As seen in Figure 1, while a similar percentage of the population of both groups live to the lower seventies, differences start emerging in the data. A lower percentage of rural people only live until eighty, and a higher percentage of them live until eighty five or ninety. Overall, average rural life expectancy is higher; admittedly not by much (in my national average it was only about a month) but higher nonetheless. When plotted out in a histogram, that average becomes even more confusing. Rural life expectancies are more often shortern than their equivalent urban counterparts, but they are also occasionally longer.

# 3 Methodology

To evaluate my data, I decided to make several models. Based on several scatter plots I made, I decided that a linear model would be the best fit.

Due to the limited integer nature of the NCHSUR index, I decided to separate my data sets. I only compared data with other data sets from the same NCHSUR index category.

However, since I was removing one of my model axes, I needed to find another axis. I was curious about two metrics, so I ran models on them both. I chose population density, because it was my next best marker to measure how rural an area was, and I chose income, because a lot of my rural areas tended to have much lower income than urban ones and I was curious to see how much correlation there was.

I decided to use three sets of models, as well. All three were linear regression, but with different methods, in order to attempt to find the best fitting line. My methods were a Least Squares method that I had programmed for an earlier homework, the Lasso and Ridge regression algorithms found in Python Sklearn’s library.

I ran these three models on the training set without modifying it to establish a baseline, and then I separated the set into subsets, each with the same NCHSUR index. I ran my three models on the training set twice, once with the Median Household Income as the other axis, and the second time with the population density.

The most successful model by far was a lasso regression model, the one adjusted for NCHSUR index with Income used as the opposing axis. Generally, this was the best overally model

# 4 Results and Discussion

My lasso regression model using income and health metrics adjusted for NCHSUR index tended to perform the best, though all models did have a fairly high RSS even when it had a low MSE. I believe that I may have used the wrong model in hindsight; perhaps KNN or KMeans would work better for this data.

When I experimented with generating new data, my numbers became even less concrete. I chose to use Income and Population Density to train my models, due to the fact that those two were the factors I had picked out.

When income is varied, Population Density becomes the better predictor of Health Metric, going from about 10% to 30%. However, this is associated with sky high MSE and RSE.

When population density is varied, income is the better fit most of the time, jumping from being best fit about 45% of the time to about 75%. Both MSE and RSE are high, though.

When both trends are varied, Income becomes an even better fit oddly enough, jumping from 45% to 85% of the times as best fit. These data tend to have high MSE but low RSE.

The numbers are all over the board on this one! Some files have high MSE and low RSE, others vice versa, and some have both very high.

In hindsight, I struggled with problems like overfitting a great deal, as well as using the wrong model for this type of dataset.

# 5 Conclusion

While the results of the data do support my initial theory that rural populations live longer (if barely), I was unable to isolate all the other ccontributing factors to health. I know that Median Household Income does play a role in the end result, but how much is part of the scope of a future project. Over winter break, I plan to create a visualization of each state’s health metrics by NCHSUR index.

Another future project may very well be an analysis of this same data using different modelling techniques such as K nearest neighbor or K means.

**Acknowledgment**

The authors would like to thank Prof. Acuna for his help with this project.