Quality of Life

A detailed look at 3109 counties in the contiguous United States of America

Arlene Barbas, Marni Cahoon, Toni Charlot, Tina Churchill, Ramsey Dial, Cheryl Little, Ralph Myers, Cynthia Newton, Murshid Saqlain, Kristina Winters, Sean Young

Under the direction of Dr. Raid W. Amin

Abstract:

Although there are many opinions as to what makes an area attractive to live in, there are several quantifiable variables that have a large impact on the livability of communities. Certainly safety, educational attainment, financial well-being, and health are universally accepted as important to quality of life. This paper takes data available at the county level throughout the contiguous United States in each of these categories to determine every county's overall quality of life that it affords its citizens. Univariate, multivariate and factor analysis are performed using SAS and SaTScan software packages. The result is large pockets of good quality of life in many Rocky Mountain, Midwest and Northeast Counties, as well as relatively poor quality of life in many parts of the South and Southeastern areas of the country.

December 3^{rd} 2014

Table of Contents

1. Introduction 3
1.1 Course Goal, General Methodology, and Process 3
1.2 Quality of Life Class Project: An Expedition from Inception to Completion 3

2. Index Development, Application and Discussion 5
2.1 Model for Quality of Life (QoL) 5
2.2 Method 6

3. Results 8
3.1 Cluster Analysis QoL Index 8
3.1.1 Sub-Index Cluster Analysis 10
3.1.1.1 Safety Sub-Index 10
3.1.1.2 Health Sub-Index 11
3.1.1.3 Education Sub-Index 12
3.1.1.4 Economic Sub-Index 13
3.1.2 Individual Variable Cluster Analysis 14
3.1.2.1 Assault Arrests 14
3.1.2.2 Drug Related Arrests 15
3.1.2.3 Murder Arrests 16
3.1.2.4 Rape Arrests 17
3.1.2.5 Robbery Rate 18
3.1.2.6 Median Income 19
3.1.2.7 Percent of Population Living Below the Poverty Level 20
3.1.2.8 Unemployment Rate 21
3.1.2.9 Unemployment Trend from 2009 to 2010 22
3.1.2.10 Percent of Population Over 25 Who Have Less Than a High School Diploma 23
3.1.2.11 Percent of the Population Over 25 Who Have a Bachelor's Degree or Greater 24
3.1.2.12 Percent of Population Under 65 Without Health Insurance 25
3.1.2.13 Percent of Population That Are Obese (Age Adjusted) 26
3.1.2.14 Concentration of Particles Less Than 2.5 Micrometers in Diameter (PM2.5) in the Air 27
3.1.2.15 Percent of Adult Population Who Are Smokers 28
3.2 Multivariate Analysis 29
3.2.1 Multivariate SaTScan 30
3.3 Factor Analysis 30
3.3.1 EFA using the Maximum Likelihood (ML) method 34
3.3.1.1 An Alternate QoL Index using Factor Scores 39
3.3.2 EFA using the Principal Factor Analysis Method 40

4. Application of the QoL 45

5. Conclusion 49
Factor Analysis 51

6. Software Packages 52
6.1 Data Cleaning Software 52
6.2 Data Analysis Software 52

References 55

Appendices:Appendix A

- Cluster analysis SaTScan outputs
Appendix B – Factor Analysis SAS and SaTScan Outputs

1. Introduction

1.1 Course Goal, General Methodology, and Process

Statistics Research is a two semester course. The learning outcomes for Statistics Research established by Dr. Amin was under "the goal to uncover what is not yet known and to learn some modern methodologies." The main goal of the first (Fall 2014) semester was to learn statistical research methods to be used by students for quality statistics research. The main approach to put this into practice was the creation of a Quality of Life Index. Thus, the Quality of Life (QoL) Index is the main focus of this report. This process includes modifying, validating, and further refining preexisting data sets provided at the beginning of the course. Gathering, modifying, validating, and refining additional data sets was also required after research and discussion about the categories for the Quality of Life Index. Tools such as Excel and Google Open Refine were used to sort and modify data sets. Implementation of data into categories and subsequently into the Quality of Life Index will be discussed in more detail later. Other tools like SAS, SaTScan, Google Earth and Tableau took the data and their categories to create the charts, maps, and other data analysis of the Quality of Life Index. Once the Quality of Life Index was established further analysis was done by first focusing on how two counties in Florida, Escambia and Santa Rosa, measured up against the rest of the state. The outcome from this analysis was published by the Pensacola News Journal.

To learn other modern methodologies SAS was used to do Factor Analysis, and SaTScan was used to do Multivariate Analysis to come up with an alternative mapping and analysis of the Quality of Life Index. These other methodologies will be discussed in more detail later. Investigation into using FleXScan and ArcGIS as other possible tools never produced any usable output but has provided a basis for further use in the Statistics Research second semester (Spring 2015). Last, each student will do further analysis on the Quality of Life Index for their assigned region to ensure comprehension on how to do quality statistics research.

Dr. Martin Kulldorff, a professor and biostatistician in the Department of Population Medicine at Harvard Medical and co-creator of SaTScan, has given assistance and praise to Dr. Amin and the Statistics Research graduate students on their work. Other notable local individuals, including Gary Barnes, a retired General Manager of Walmart in Pensacola, have given praise for the Pensacola New Journal article, "Data Detectives," that was published based on the results of the Index.

1.2 Quality of Life Class Project: An Expedition from Inception to Completion

The Quality of Life Index started with previously collected crime data that was researched by another class. The data was analyzed for completeness and accuracy, and the missing values were noted for completing in the future.

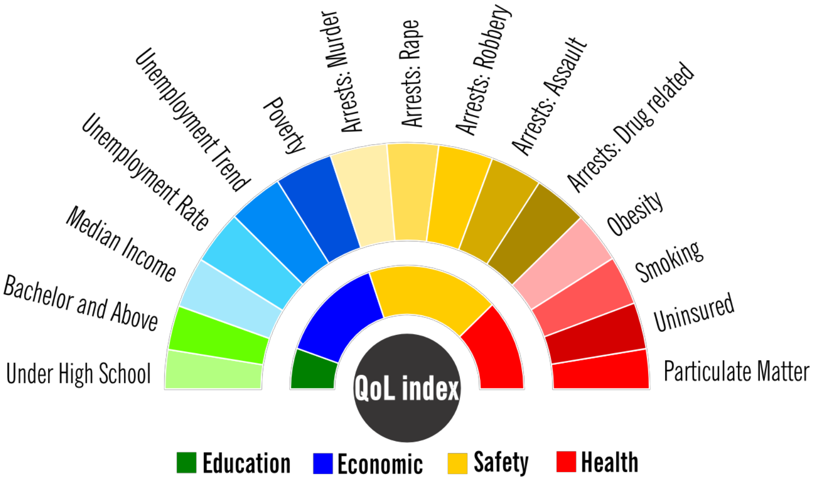
Economic data was also provided by Dr. Amin, including median income and a wealth index. This data helped to form the sub-indexes of safety and economic. The research and review of several articles about quality of life provided insight in having health and education as the other sub-indexes. Gathering data for the quality of life index was determined to be measured and evaluated at the county level instead of the state or city level. Using and documenting valid sources, data was gathered for each category. The data for each subcategory was established as a rate based upon the population for that county. The United States Census Bureau population data was used as the standard to ensure consistency throughout all of the counties. Besides researching the data different methodologies for creating the index was evaluated. This included common variables used in such indexes, typical weights placed on each category, and the motivation behind such Quality of Life representations. As in the name, quality of life indexes are quantitative, categorical approaches to something that can be subjective and qualitative when considered by the individual. Therefore, there is no perfect or “right” quality of life index model. Each person viewing the results and maps must consider the subcategories and weights used in this model. However, since many variables were used, the QoL Index should give a good representation of counties excluding personal preference and other subjectivity. Logically, the more agreed upon variables used, the better the representation.

The class was guided to use SAS to normalize the rates in each category. This process using normalization by BLOM will be discussed later under the model used for the quality of life index. The rates during this process were set to be consistent in direction for either increasing “good” or “bad” values. Next each category index was normalized after any weighting was applied to the normalized rates within the index. The category indices were also set to be consistent in direction for either increasing “good” or “bad” values during this procedure. Each category was weighted for the final index, and the final index was normalized using SAS. The normalized final index was then feed into SaTScan to determine high and low cluster areas using the normal model for the contiguous United States. The high and low cluster results from SaTScan were displayed using Google Earth. These results were also compared to the original normalized Quality of Life Index using Tableau with two colors and their shading variations to show low to high value counties. The sub-indexes were also displayed using SaTScan and Google Earth and were also compared to the normalized sub-indexes in Tableau. Further analysis of the Quality of Life Index data was done using Factor Analysis and Multivariate. Factor Analysis used the data sets equally in SAS without regard to their sub-index. The Factor Analysis model with its factors is given in more detail later on. Multivariate analysis used the sub-indexes in SaTScan to come up with the high and low clusters that would not have typically appeared in a single Quality of Life Index. The details of the Multivariate and its results are given later on.

2. Index Development, Application and Discussion

2.1 Model for Quality of Life (QoL)

In trying to measure quality of life, the class came up with four main sub-indices that we felt were relevant, objective, and non-controversial in terms of their relationship to quality of life. We agreed upon the sub-indices of Safety, Economic, Education and Health. In an effort to prevent one sub-index with more variables from dominating the Quality of Life Index, each sub-index was given equal weight.



Each of these sub-indices with their variables are described below:

Safety: Robbery Rate, Assault Rate, Murder Rate, Rape Rate, Drug Related Crime Rate

Overall crime rates give a good indication of how safe a person will be in any given area, but violent crimes and drug-related crimes are especially relevant to personal safety. We used crime rates (arrests per 1000 people) of robbery, assault, rape, murder and drugs.

Economic: Median Income, Percent of Population Living Below Poverty Level, Unemployment Rate, Unemployment Trend

To determine the overall economic well-being of a county, the class felt that the ability to obtain employment, ability to earn a livable wage, percent of the population that were living below the poverty level, as well as if the economy was improving or worsening were all important factors. Median income gives an indication of overall wages without extremes skewing the numbers. Percent below poverty level and unemployment both give an indication of how many in the county are struggling financially. Employment trend is the unemployment rate in 2009 subtracted from the unemployment rate in 2010. This gives an indication of whether a county's economy is improving or worsening.

Education: Percent of Population Over 25 Who Have Less Than a High School Diploma, Percent of the Population Over 25 Who Have a Bachelor's Degree or Greater

To help determine how well the public schools in the area are performing, we looked at the percentage of people (over 25 years of age) who did not complete high school. This may give an indication of public schools, but we also felt the need to use a variable to determine how much of the population were able to obtain a college education. The percent of the adult population over 25 who had at least a bachelor's degree was used to measure the top end of the counties' educational attainment.

Health: Percent of Population That Are Obese (Age Adjusted), Percent of Adult Population Who Are Smokers, Percent of Population Under 65 Without Health Insurance, Concentration of Particles Less Than 2.5 Micrometers in Diameter ($PM_{2.5}$) in the Air.

Arguably one of the best measurable indicators of overall health for an individual is their weight. We opted to use an age-adjusted obesity rate so that counties with a relatively old or young population were not misrepresented. The factor of percent of smokers was included as an indication of health and attitude towards healthy lifestyles. The percentage of people under 65 who do not have health insurance was used to measure overall access to health care. As air quality has a direct influence on the health of all who live in the county, a factor for air quality was used. The finer particles included in $PM_{2.5}$ measurements (μg/m^3) are considered the most dangerous to health. This is the measurement that was used.

2.2 Method

Most of the data we were working with (such as crime rates) were such that the lower the rate, the better. We needed to be consistent with each variable having that same property. For the few that were such that a high number was more desirable, we simply inverted by multiplying by a negative 1.

Each variable was then normalized in SAS using Normal=BLOM with the default TIES=MEAN. Normal=BLOM computes normal scores from the ranks. The resulting variables appear normally distributed. $\boldsymbol{n}$ is the number of observations of the ranking variable

$$\boldsymbol{y_i}= \Phi^{-1}((\boldsymbol{r_i}-3/8)/(\boldsymbol{n}+1/4))$$

In these formulas, Φ^{-1} is the inverse cumulative normal (PROBIT) function, $\boldsymbol{r_i}$ is the rank of the $\boldsymbol{i}$th observation, and $\boldsymbol{n}$ is the number of observations for the ranking variable.

To get a sub-index score, each variable in a sub-index was then added together and the sum divided by the total number of variables in that sub-index.

$$Sub-Index = \frac{\sum_{i=1}^{n} variable_i}{n}$$

When each county had a sub-index score, the sub-index score was then normalized with the same method in SAS. Each county was then given a Quality of Life Index score by adding the normalized sub-index scores together and dividing by the number of sub-indices, 4.

Quality Of Life Index = (Normalized Safety Sub-Index + Normalized Economic Sub-Index + Normalized Education Sub-Index + Normalized Health Sub-Index)/4

Once the counties were each assigned a quality of life score, the scores were again normalized with the same method is SAS. This result is the final Quality of Life Score given each county.

The scores calculated range from approximately -3.5 to 3.5. Once all counties were assigned a score for each variable, sub-index and index, these scores were input into SatScan. SatScan looked at the scores and found areas throughout the county were either many undesirable scores or desirable scores were found together in close geographical proximity. When this phenomenon was found it was identified as a cluster. These clusters are shown in the next section.

3. Results

3.1 Cluster Analysis QoL Index

The SaTScan map (using Google Earth) shows significant clusters with P-Values of less than 0.05, in order, starting with the most likely cluster. For each map, you'll see a table that will acknowledge each cluster by name, the mean inside the cluster, and a brief meaning of this information.

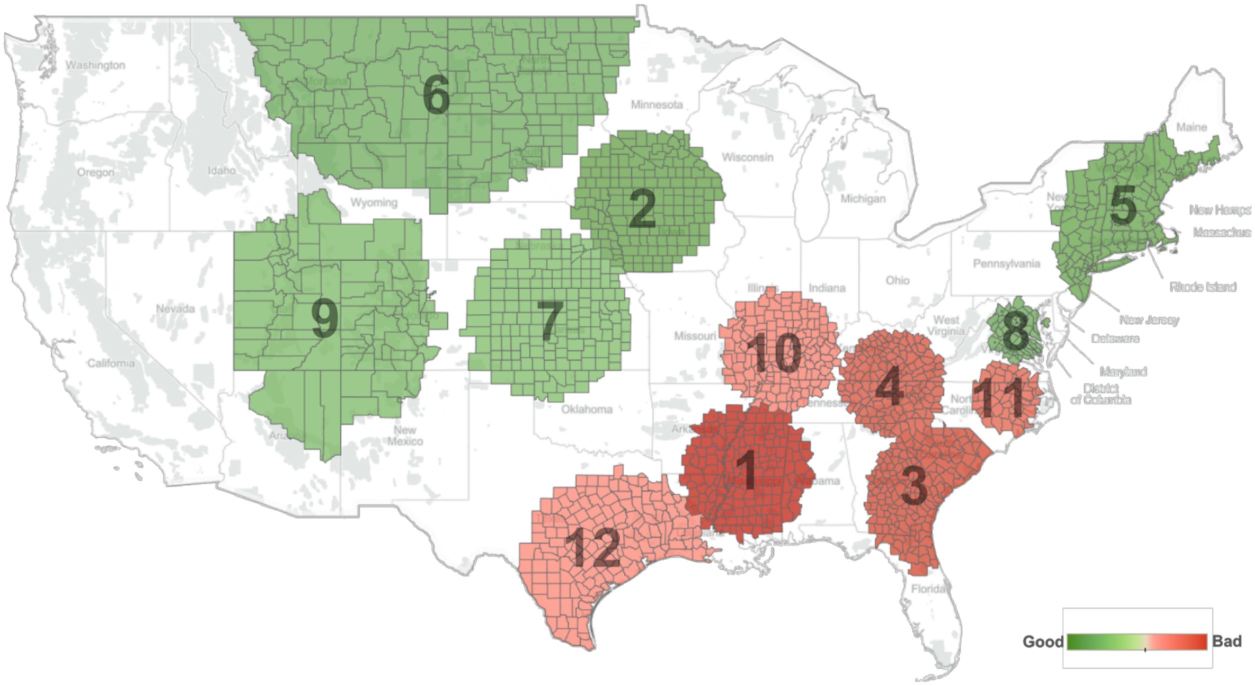


Fig 3.1.1 *Clusters for the QoL Index*

Cluster#	Cluster	Number of Counties	Mean Inside (z-scores)	Good or Bad?
1	Mississippi	155	1.29	Quite Bad
2	Iowa	155	-1.13	Quite Good
3	Georgia/South Carolina	155	1.05	Quite Bad
4	East Kentucky/Tennessee	155	0.95	Quite Bad
5	New England	114	-1.11	Quite Good
6	Montana/Dakota	143	-0.97	Quite Good
7	Kansas	155	-0.77	Moderately Good
8	Virgina	56	-1.03	Quite Good
9	Utah/Colorado	90	-0.80	Quite Good
10	South Illinois	154	0.48	Moderately Bad
11	North Carolina	53	0.80	Quite Bad
12	Texas	155	0.38	Moderately Bad

Figure 3.1.2 was created using Tableau business intelligence software. It reflects how each county 'scored' individually. In dark red are the worst counties, while dark green are the best.

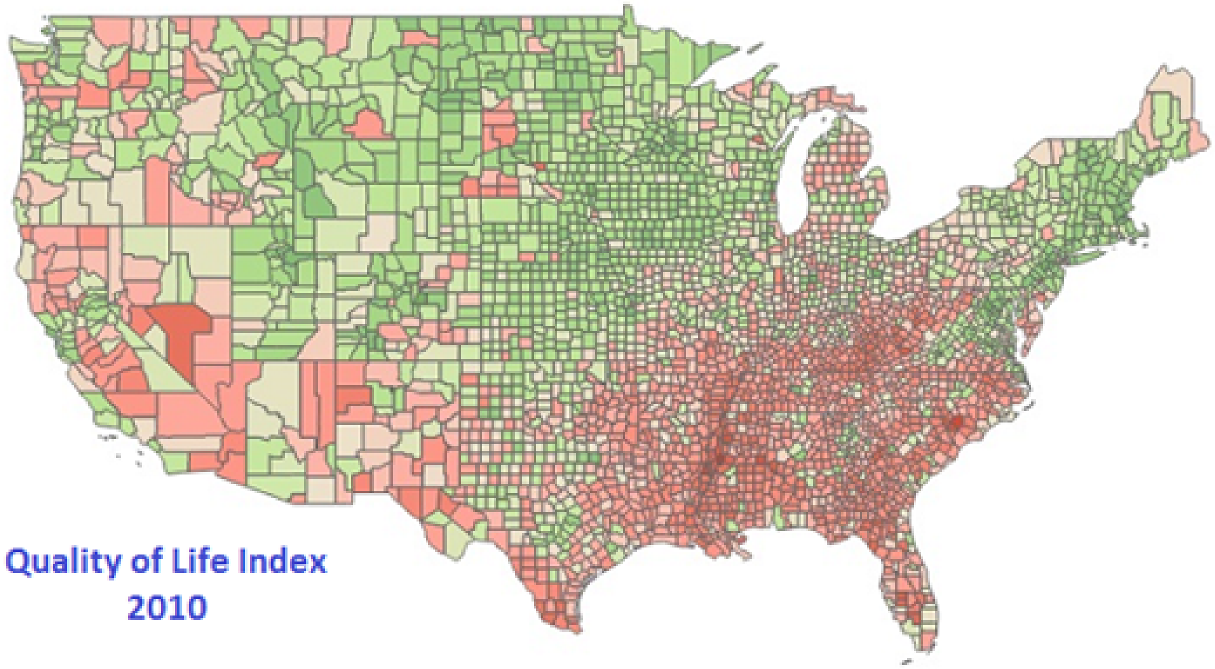


Fig 3.1.2 *Heat map for the QoL Index*

The following sections have both the SaTScan cluster map, Tableau map, and a table for each of the clusters found in SaTScan.

3.1.1 Sub-Index Cluster Analysis

3.1.1.1 Safety Sub-Index

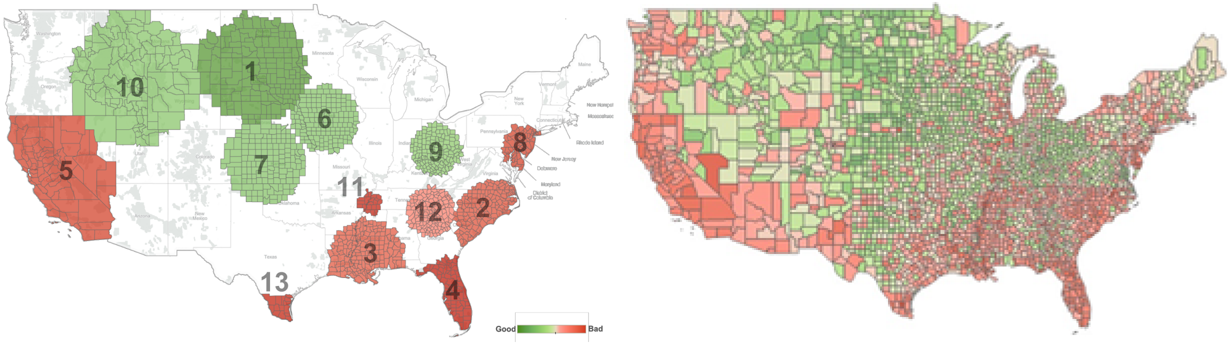


Fig 3.1.1.1 *Clusters and Heat map for the Safety sub-index*

Cluster#	Cluster	Number of Counties	Mean Inside (z-scores)	Good or Bad?
1	North and South Dakota	155	-1.25	Quite Good
2	North and South Carolina	137	0.93	Quite Bad
3	Louisiana/Mississippi	155	0.85	Quite Bad
4	Florida	63	1.32	Quite Bad
5	California/Nevada	73	1.07	Quite Bad
6	Iowa	155	-0.69	Moderately Good
7	Kansas	146	-0.70	Moderately Good
8	NE of Washington D.C.	61	1.05	Quite Bad
9	Ohio	140	-0.60	Moderately Good
10	Rocky Mountain region	113	-0.65	Moderately Good
11	West Tennessee	20	1.28	Quite Bad
12	Northern Georgia	143	0.44	Moderately Bad
13	Southern Texas	14	1.27	Quite Bad

3.1.1.2 Health Sub-Index

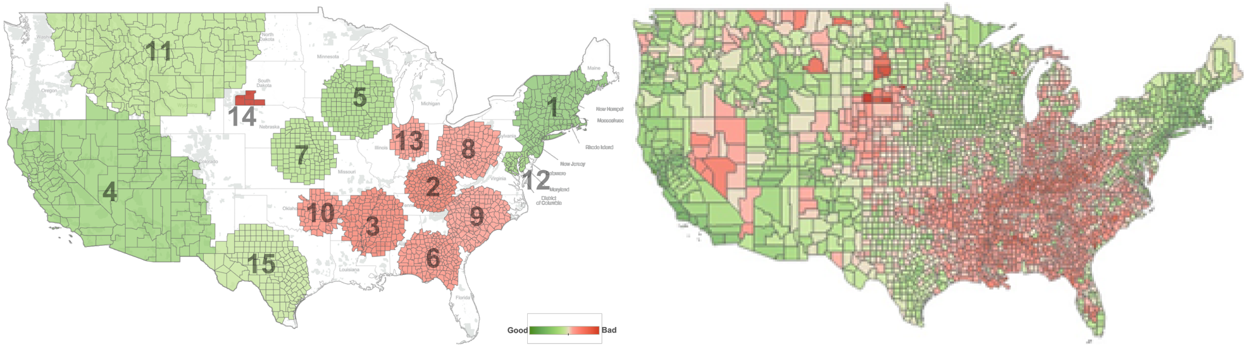


Fig 3.1.1.2 *Clusters and Heat map for the Health sub-index*

Cluster #	Cluster	Number of Counties	Mean Inside (z-scores)	Good or Bad?
1	New England	141	-1.48	Quite Good
2	Kentucky	155	1.24	Quite Bad
3	North Mississippi	155	1.14	Quite Bad
4	South West region	155	-1.11	Quite Good
5	Wisconsin/Iowa	155	-0.91	Quite Good
6	North Florida/South Georgia	154	0.80	Quite Bad
7	Kansas	136	-0.63	Moderately Good
8	Ohio	136	0.62	Moderately Bad
9	North and South Carolina	155	0.56	Moderately Bad
10	East Oklahoma	56	0.86	Quite Bad
11	Montana	151	-0.51	Moderately Good
12	Washington D.C.	33	-1.06	Quite Good
13	Indiana	66	0.68	Moderately Bad
14	South Dakota	4	2.71	Extremely Bad
15	Texas	155	-0.39	Moderately Good

3.1.1.3 Education Sub-Index

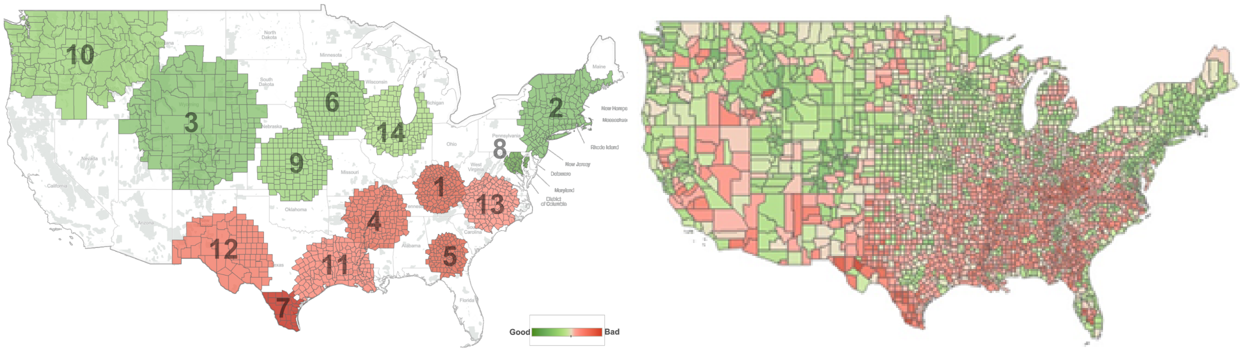


Fig 3.1.1.3 *Clusters and Heat map for the Education sub-index*

Cluster#	Cluster	Number of Counties	Mean Inside (z-scores)	Good or Bad?
1	East Kentucky	142	1.06	Quite Bad
2	New England	141	-0.99	Quite Good
3	Wyoming/Colorado	154	-0.94	Quite Good
4	Northern Mississippi	155	0.89	Quite Bad
5	Georgia	102	1.06	Quite Bad
6	Iowa	141	-0.71	Moderately Good
7	South Texas	25	1.61	Quite Bad
8	Washington D.C.	30	-1.43	Quite Good
9	Kansas	142	-0.64	Moderately Good
10	North West	117	-0.63	Moderately Good
11	Louisiana/Texas	136	0.57	Moderately Bad
12	West Texas	75	0.78	Moderately Bad
13	North Carolina	140	0.45	Moderately Bad
14	Lake Michigan area	137	-0.45	Moderately Good

3.1.1.4 Economic Sub-Index

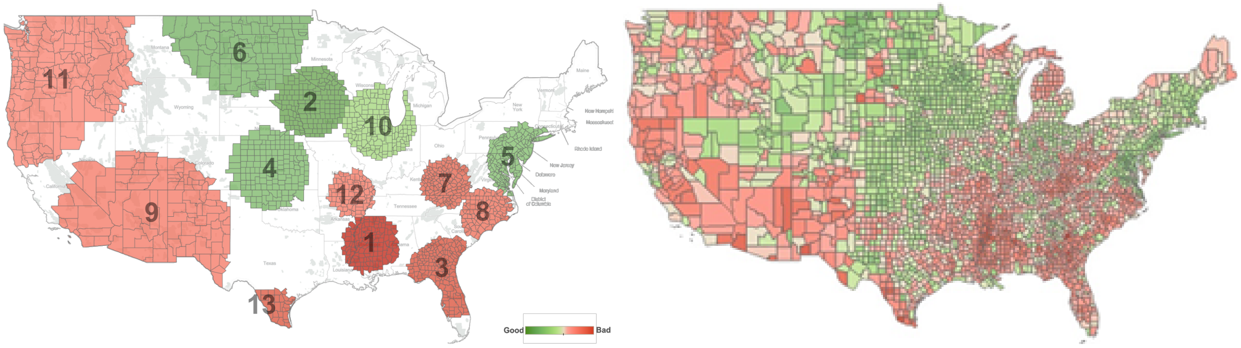


Fig 3.1.1.4 *Clusters and Heat map for the Economic sub-index*

Cluster#	Cluster	Number of Counties	Mean Inside (z-scores)	Good or Bad?
1	Mississippi	123	1.40	Quite Bad
2	Iowa/Minnesota	155	-1.20	Quite Good
3	Florida/Georgia	155	1.05	Quite Bad
4	Kansas	152	-0.90	Quite Good
5	Washington, D.C. area	154	-0.86	Quite Good
6	North Dakota	108	-1.01	Quite Good
7	West Kentucky	137	0.87	Quite Bad
8	North Carolina	94	0.74	Moderately Bad
9	Arizona	102	0.60	Moderately Bad
10	Illinois/Wisconsin	154	-0.48	Moderately Good
11	North West region	137	0.49	Moderately Bad
12	Missouri/Arkansas	73	0.63	Moderately Bad
13	Southern Texas	26	1.01	Quite Bad

3.1.2 Individual Variable Cluster Analysis

3.1.2.1 Assault Arrests

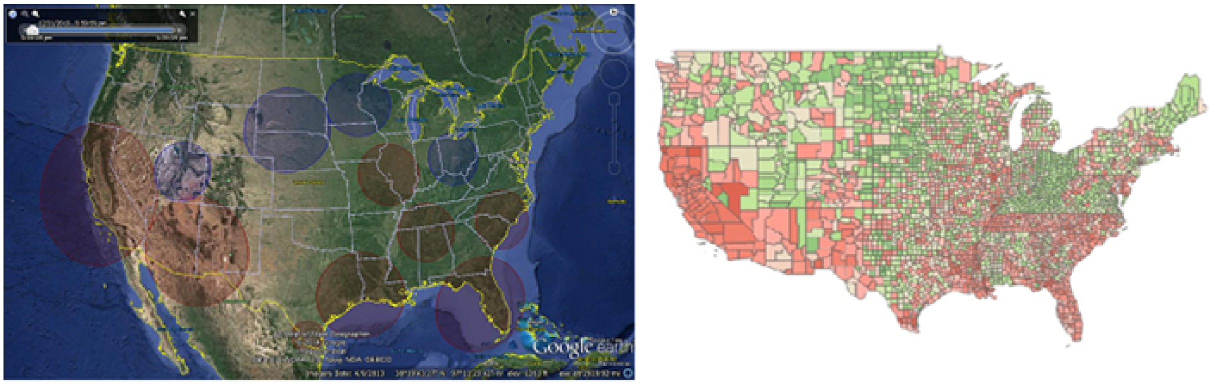


Fig 3.1.2.1 *Clusters and Heat map for Assault arrests only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
S Dakota/Nebraska	155	-1.03	Quite Good
California	74	1.42	Quite Bad
Indiana/Ohio	154	-0.93	Quite Good
N Carolina/S Carolina	110	0.88	Quite Bad
Texas/Louisiana	155	0.73	Moderately Bad
Florida	96	0.79	Moderately Bad
Tennessee/Alabama	147	0.47	Moderately Bad
Illinois/Missouri	153	0.42	Moderately Bad
Arizona/New Mexico	42	0.79	Moderately Bad
Southern Texas	12	1.45	Quite Bad
Utah	28	-0.94	Quite Good
Minnesota/Wisconsin	122	-0.44	Moderately Good

3.1.2.2 Drug Related Arrests

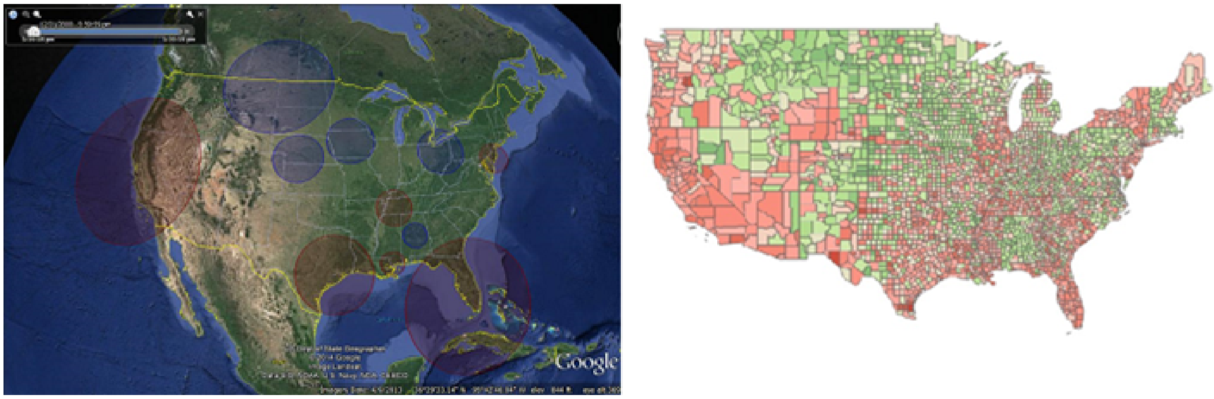


Fig 3.1.2.2 *Clusters and Heat map for Drug related arrests only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Montana/N Dakota	140	-1.12	Quite Good
Iowa	146	-0.85	Quite Good
California/Oregon	114	0.79	Moderately Bad
Texas/Louisiana	151	0.66	Moderately Bad
Florida	100	0.77	Moderately Bad
Nebraska/Kansas	92	-0.8	Quite Good
Ohio	138	-0.54	Moderately Good
New Jersey	47	0.88	Quite Bad
Alabama	29	-1.06	Quite Good
Tennessee	93	0.54	Moderately Bad
Southern Mississippi	23	1.00	Quite Bad

3.1.2.3 Murder Arrests

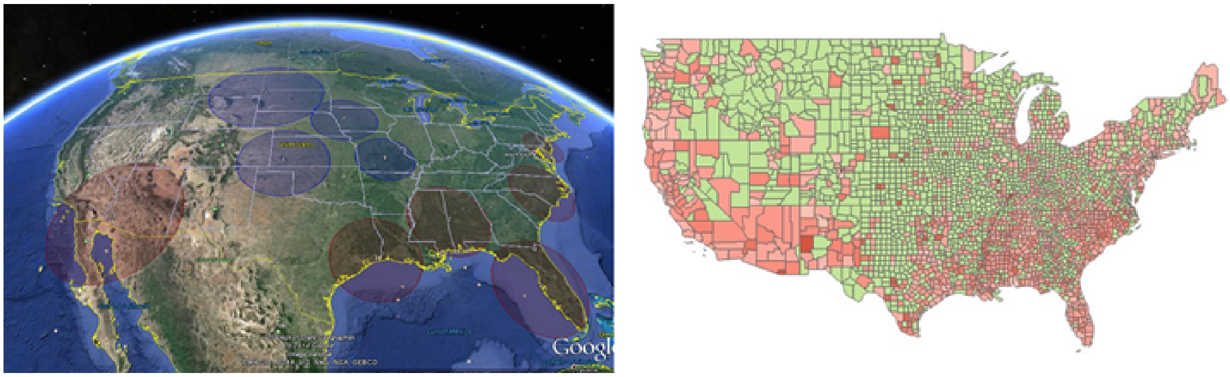


Fig 3.1.2.3 *Clusters and Heat map for Murder arrests only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
North Carolina	137	0.92	Quite Bad
Mississippi/Alabama	154	0.78	Moderately Bad
Montana/South Dakota	155	-0.56	Moderately Good
Iowa	154	-0.50	Moderately Good
Kansas/Oklahoma	136	-0.49	Moderately Good
New Jersey/New York	62	0.67	Moderately Bad
Florida	64	0.65	Moderately Bad
Missouri	99	-0.50	Moderately Good
Texas/Louisiana	87	0.53	Moderately Bad
Arizona/California	28	0.94	Quite Bad

3.1.2.4 Rape Arrests

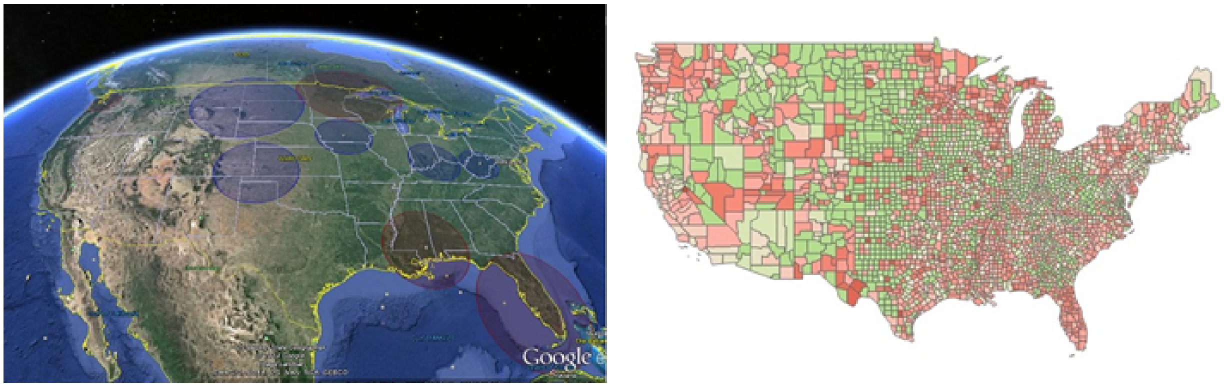


Fig 3.1.2.4 *Clusters and Heat map for Rape arrests only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
South Dakota/Montana	155	-0.68	Moderately Good
Florida	53	1.12	Quite Bad
Minnesota/Wisconsin	148	0.54	Moderately Bad
Iowa	120	-0.52	Moderately Good
Colorado/Kansas	103	-0.55	Moderately Good
Mississippi	134	0.48	Moderately Bad
Indiana/Kentucky	155	-0.42	Moderately Good
Washington	21	1.02	Quite Bad

3.1.2.5 Robbery Rate

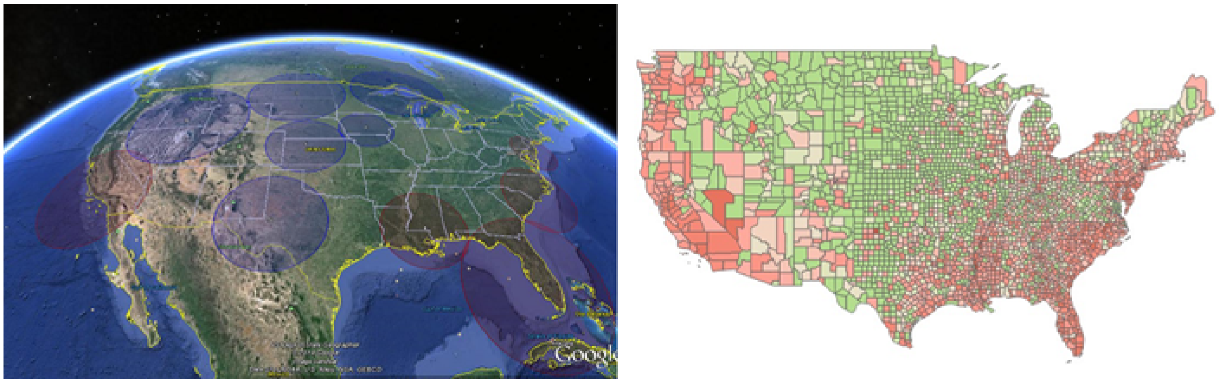


Fig 3.1.2.5 *Clusters and Heat map for Robbery arrests only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
S Carolina/N Carolina	137	1.03	Quite Bad
New Jersey/New York	80	1.28	Quite Bad
S Dakota/N Dakota	155	-0.90	Quite Good
Florida	111	0.91	Quite Bad
Kansas	155	-0.76	Moderately Good
Louisiana/Mississippi	155	0.74	Moderately Bad
Iowa	133	-0.75	Moderately Good
California	50	0.96	Quite Bad
Idaho/Utah	138	-0.55	Moderately Good
Texas	132	-0.54	Moderately Good
Michigan	100	-0.47	Moderately Good

3.1.2.6 Median Income

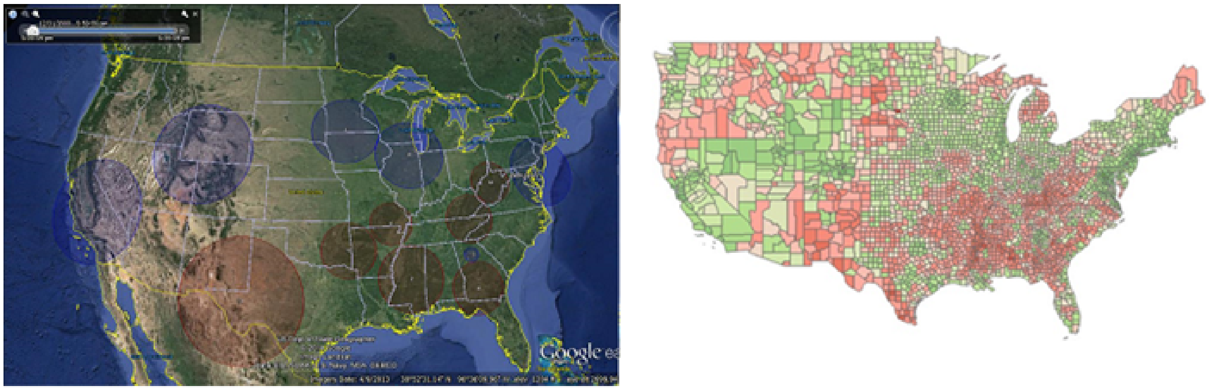


Fig 3.1.2.6 *Clusters and Heat map for Median income only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
New Jersey	147	-1.30	Quite Good
Mississippi	123	1.26	Quite Bad
Eastern Tennessee	123	1.06	Quite Bad
Southern Georgia	115	0.88	Quite Bad
Iowa/Minnesota	153	-0.72	Moderately Good
Southeast Missouri	56	1.14	Quite Bad
Rocky Mountain	87	-0.90	Quite Good
Illinois/Indiana	153	-0.63	Moderately Good
California	55	-1.03	Quite Good
Oklahoma/Arkansas	77	0.73	Moderately Bad
New Mexico/Texas	120	0.55	Moderately Bad
West Virginia	72	0.59	Moderately Bad
Northern Georgia	15	-1.27	Quite Good

3.1.2.7 Percent of Population Living Below the Poverty Level

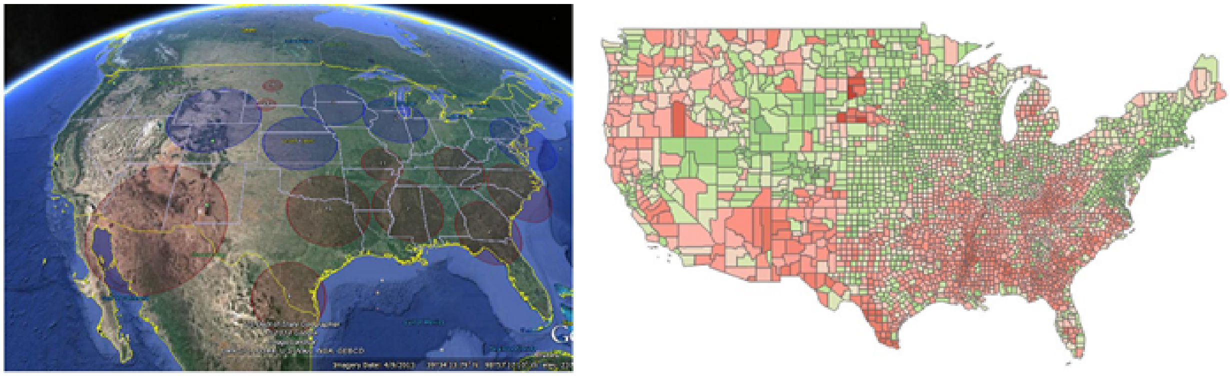


Fig 3.1.2.7 *Clusters and Heat map for Poverty only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Mississippi	139	1.23	Quite Bad
New Jersey/New York	149	-1.06	Quite Good
Iowa/Minnesota	154	-1.02	Quite Good
Georgia/Florida	153	1.01	Quite Bad
Tennessee/Kentucky	155	0.84	Quite Bad
Kansas	141	-0.71	Moderately Good
Illinois/Indiana	151	-0.66	Moderately Good
Wyoming/Colorado	85	-0.84	Quite Good
Southern Texas	26	1.42	Quite Bad
N Carolina/S Carolina	120	0.59	Moderately Bad
Southeast Missouri	51	0.79	Moderately Bad
Arizona/New Mexico	37	0.90	Quite Bad
South Dakota	4	2.65	Extremely Bad
N Dakota/S Dakota	5	2.32	Extremely Bad
Texas/Oklahoma	140	0.41	Moderately Bad

3.1.2.8 Unemployment Rate

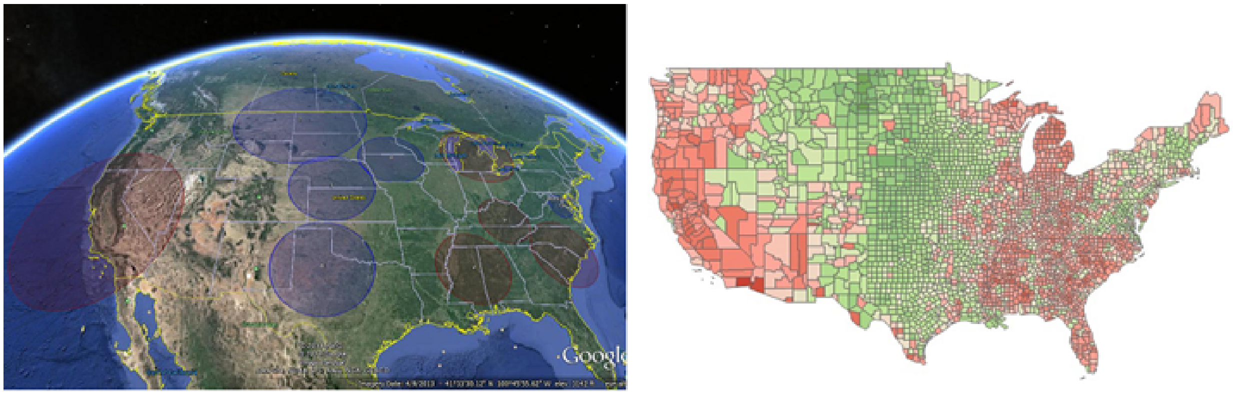


Fig 3.1.2.8 *Clusters and Heat map for Unemployment rate only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Nebraska/Kansas	155	-1.64	Quite Good
Montana/North Dakota	155	-1.47	Quite Good
California/Nevada	87	1.25	Quite Bad
Mississippi	155	0.92	Quite Bad
N Carolina/S Carolina	148	0.94	Quite Bad
Texas	155	-0.90	Quite Good
Michigan	114	1.06	Quite Bad
Iowa/Minnesota	155	-0.85	Quite Good
Kentucky/Tennessee	155	0.70	Moderately Bad
Northern Virginia	53	-0.67	Moderately Good

3.1.2.9 Unemployment Trend from 2009 to 2010

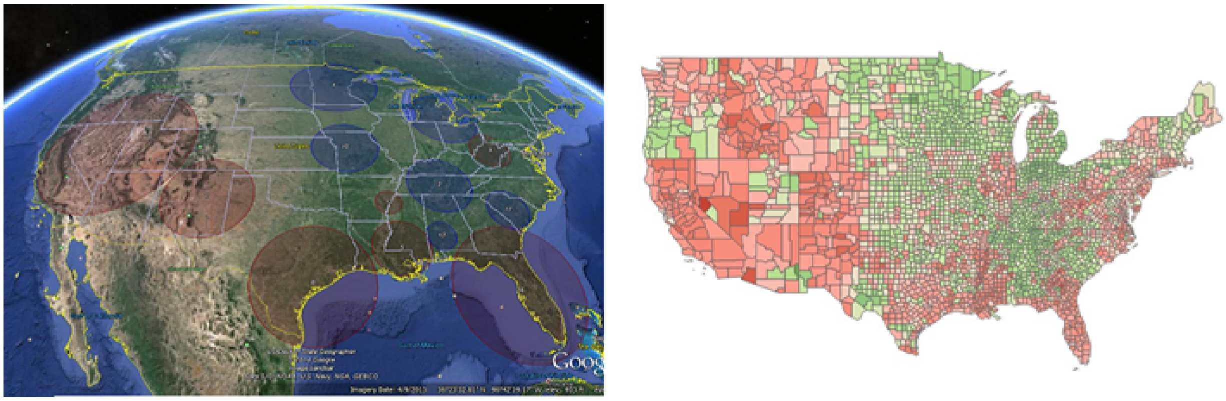
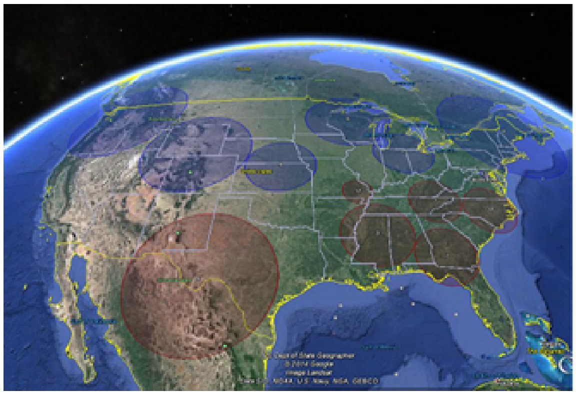


Fig 3.1.2.9 *Clusters and Heat map for Unemployment trend only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Nevada	153	1.17	Quite Bad
Tennessee/Kentucky	155	-1.15	Quite Good
Michigan	138	-1.10	Quite Good
Minnesota	154	-0.96	Quite Good
Mississippi/Louisiana	76	1.18	Quite Bad
Florida	138	0.79	Moderately Bad
New Mexico	63	0.93	Quite Bad
Southern Texas	130	0.54	Moderately Bad
West Virginia	80	0.66	Moderately Bad
Missouri	130	-0.48	Moderately Good
South Carolina	57	-0.73	Moderately Good
Southern Alabama	16	-1.33	Quite Good
Arkansas/Mississippi	17	1.22	Quite Bad

3.1.2.10 Percent of Population Over 25 Who Have Less Than a High School Diploma



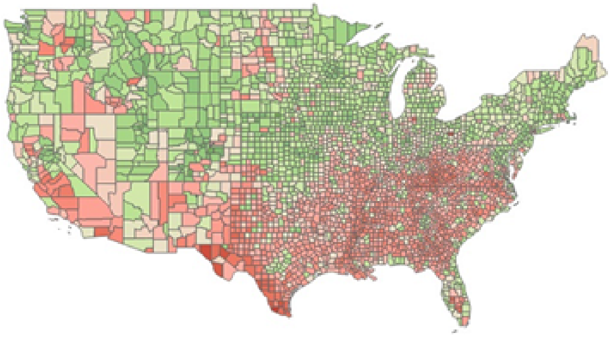


Fig 3.1.2.10 *Clusters and Heat map for those with less than a High School diploma only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
East Tennessee	155	1.06	Quite Bad
Rocky Mountain	143	-1.06	Quite Good
Mississippi	155	0.95	Quite Bad
Nebraska/Kansas	154	-0.93	Quite Good
Texas	145	0.93	Quite Bad
Wisconsin	155	-0.89	Quite Good
Georgia/North Florida	153	0.81	Quite Bad
Northeast	154	-0.69	Moderately Good
Northwest	121	-0.64	Moderately Good
N Carolina	116	0.59	Moderately Bad
Indiana/Ohio/Michigan	152	-0.46	Moderately Good
Southeast Missouri	32	0.94	Quite Bad
Northern Virginia/Maryland	10	-1.55	Quite Good

3.1.2.11 Percent of the Population Over 25 Who Have a Bachelor’s Degree or Greater

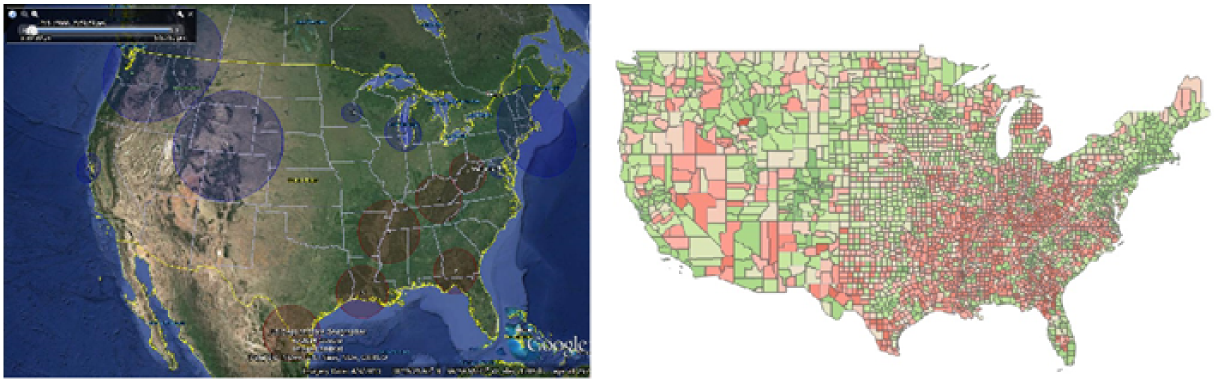


Fig 3.1.2.11 *Clusters and Heat map for those with at least a Bachelor’s Degree only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Northeast	120	-1.22	Quite Good
Kentucky	155	0.84	Quite Bad
Florida/Georgia	98	1.03	Quite Bad
Maryland/Virginia	34	-1.64	Quite Good
Arkansas/Mississippi	155	0.75	Moderately Bad
Rocky Mountain	154	-0.73	Moderately Good
West Virginia	78	0.73	Moderately Bad
California	11	-1.78	Quite Good
Northwest	117	-0.49	Moderately Good
Louisiana	58	0.69	Moderately Bad
Lake Michigan	40	-0.8	Quite Good
Southern Texas	25	0.99	Quite Bad
Minnesota	18	-1.14	Quite Good

3.1.2.12 Percent of Population Under 65 Without Health Insurance

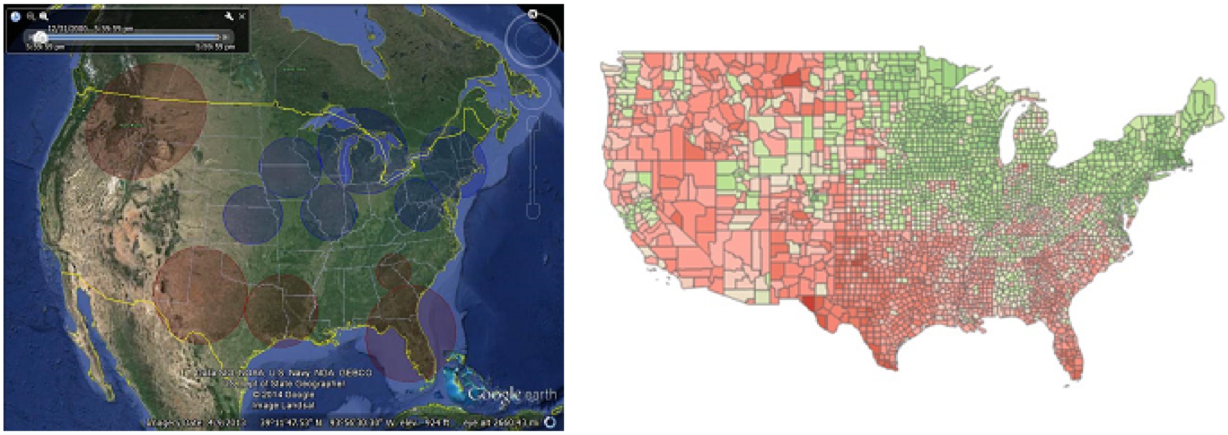


Fig 3.1.2.12 *Clusters and Heat map for those under 65 without health insurance only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
West Texas	154	1.58	Quite Bad
Iowa/Minnesota	155	-1.45	Quite Good
Northeast	155	-1.30	Quite Good
East Texas/Louisiana	149	0.96	Quite Bad
Florida	155	0.86	Quite Bad
Michigan	153	-0.74	Moderately Good
Illinois	148	-0.74	Moderately Good
Northwest	138	0.76	Moderately Bad
Kansas/Nebraska	129	-0.56	Moderately Good
Pennsylvania/W Virginia	154	-0.50	Moderately Good
Northern Georgia	79	0.70	Moderately Bad

3.1.2.13 Percent of Population That Are Obese (Age Adjusted)

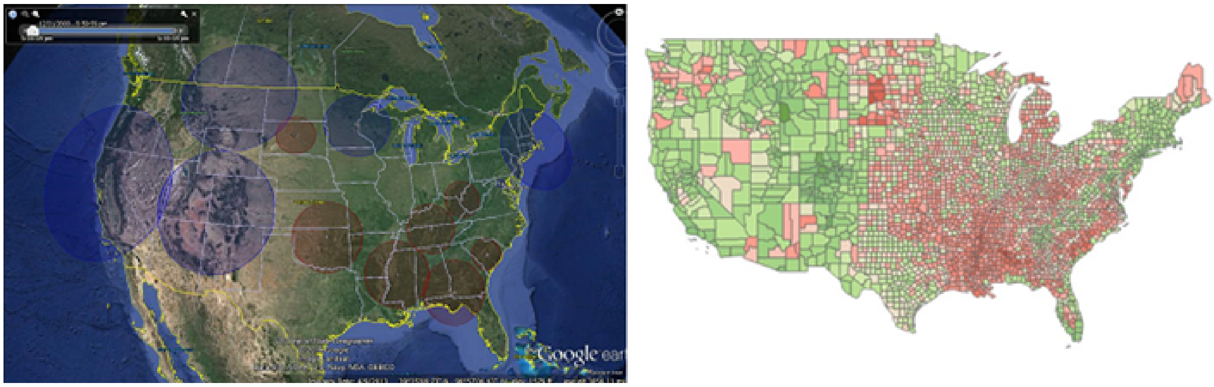


Fig 3.1.2.13 *Clusters and Heat map for Obesity only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Mississippi	155	1.46	Quite Bad
Colorado/Utah	134	-1.50	Quite Good
Northeast	114	-1.22	Quite Good
California/Oregon	112	-1.10	Quite Good
West Virginia/Ohio	85	0.89	Quite Bad
Oklahoma/Kansas	143	0.66	Moderately Bad
Montana	96	-0.78	Moderately Good
South Carolina	39	1.21	Quite Bad
Florida/Georgia/Alabama	153	0.60	Moderately Bad
Minnesota/Wisconsin	124	-0.58	Moderately Good
Tennessee/Kentucky	143	0.53	Moderately Bad
Virginia/Maryland	10	-1.70	Quite Good
South Dakota	32	0.92	Quite Bad

3.1.2.14 Concentration of Particles Less Than 2.5 Micrometers in Diameter (PM$_{2.5}$) in the Air

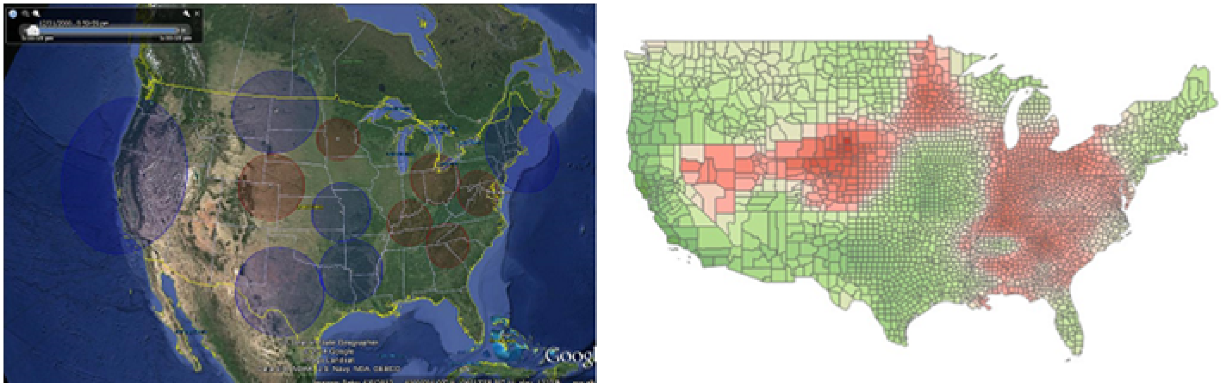


Fig 3.1.2.14 *Clusters and Heat map for PM$_{2.5}$ only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
West Texas	155	-1.56	Quite Good
Tennessee/Kentucky	155	1.42	Quite Bad
Nebraska/Colorado	97	1.75	Quite Bad
California/Oregon	129	-1.49	Quite Good
Kansas/Missouri	155	-1.27	Quite Good
Georgia/South Carolina	155	1.1	Quite Bad
Ohio	155	1.03	Quite Bad
Minnesota	78	1.25	Quite Bad
Arkansas/Louisiana	155	-0.59	Moderately Good
Virginia/Maryland	154	0.53	Moderately Bad
Northeast	127	-0.58	Moderately Good
Montana/North Dakota	90	-0.64	Moderately Good

3.1.2.15 Percent of Adult Population Who Are Smokers

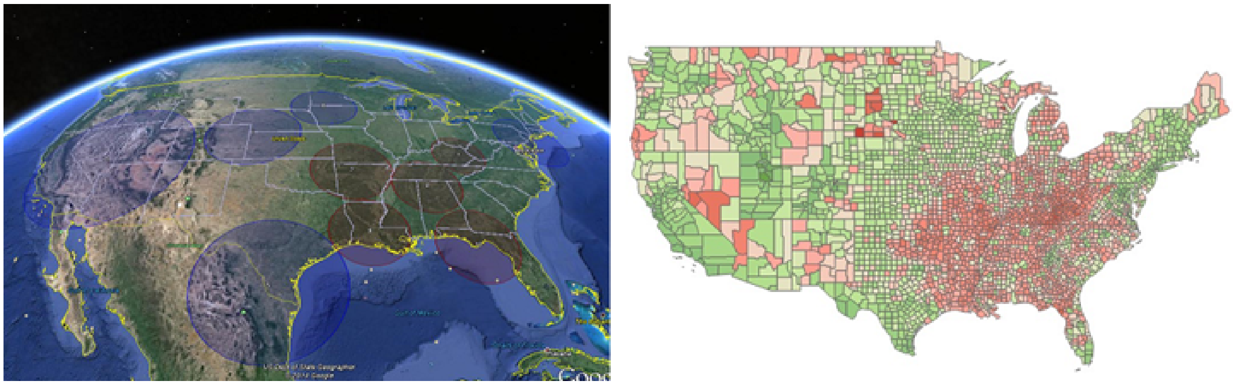


Fig 3.1.2.15 *Clusters and Heat map for Smokers only*

Cluster	Number of Counties	Mean Inside	Good or Bad?
Eastern Kentucky	153	1.40	Quite Bad
California/Utah	155	-1.08	Quite Good
Arkansas/Missouri	152	1.07	Quite Bad
Northeast	155	-0.86	Quite Good
Southern Texas	86	-0.94	Quite Good
Iowa/Minnesota	151	-0.69	Moderately Good
Colorado/Nebraska	153	-0.62	Moderately Good
Tennessee/Alabama	152	0.58	Moderately Bad
Louisiana	99	0.68	Moderately Bad
Florida	95	0.68	Moderately Bad

3.2 Multivariate Analysis

Using SaTScan to analyze one combined data set, as was performed in our QoL index, may sometimes lead us to miss a cluster that may be present in one of the subgroups. However, as Dr. Kulldorff states in the SaTScan User Manual, if the sub-indexes are analyzed separately in SaTScan, there is a loss of power if there is a true cluster that is equally strong in multiple indexes. So, a multivariate analysis was performed on the four sub-indexes in the QoL index: education, economic, safety, and health.

A multivariate analysis is able to detect clusters due to only one of the sub-indexes, or two or more combined [1]. These clusters provide an additional insight on where the Quality of Life has a higher rank based on some or all our sub-indexes.

To simultaneously search for clusters in multiple data sets, Dr. Kulldorff [1] suggests the following approach:

1. The log likelihood ratios are calculated for each data set and it is noted whether the observed number of cases is larger or smaller than expected.
2. The log likelihood ratios are summed up for the data sets with more than expected number of cases. Similarly, the log likelihood ratios are summed up for the data sets with the less than expected number of cases.
3. The maximum value of all the summed log likelihood ratios reveals the most likely cluster.

According to Dr. Kulldorff, the scan statistic can be written as:

$$T = \max_{z}\ \max\left(\sum_{i} LLR_i(high, z), \sum_{i} LLR_i(low, z)\right)$$

for cylinder *z*, in data set *i*.

If multiple data sets are fed into SaTScan, the software runs the above calculations and provides us with the clusters that are due to one more data sets.

To perform our Multivariate Analysis, we took the normalized scores for each of our sub-indexes and used those data sets in SaTScan 9.3. The following settings were used:

- Type of Analysis: **Purely Spatial**
- Probability Model: **Normal**
- Scan For Areas With: **High or Low Values**
- Maximum Spatial Cluster Size: **5.0**
- Maximum number of Monte Carlo Replications: **999**

3.2.1 Multivariate SaTScan

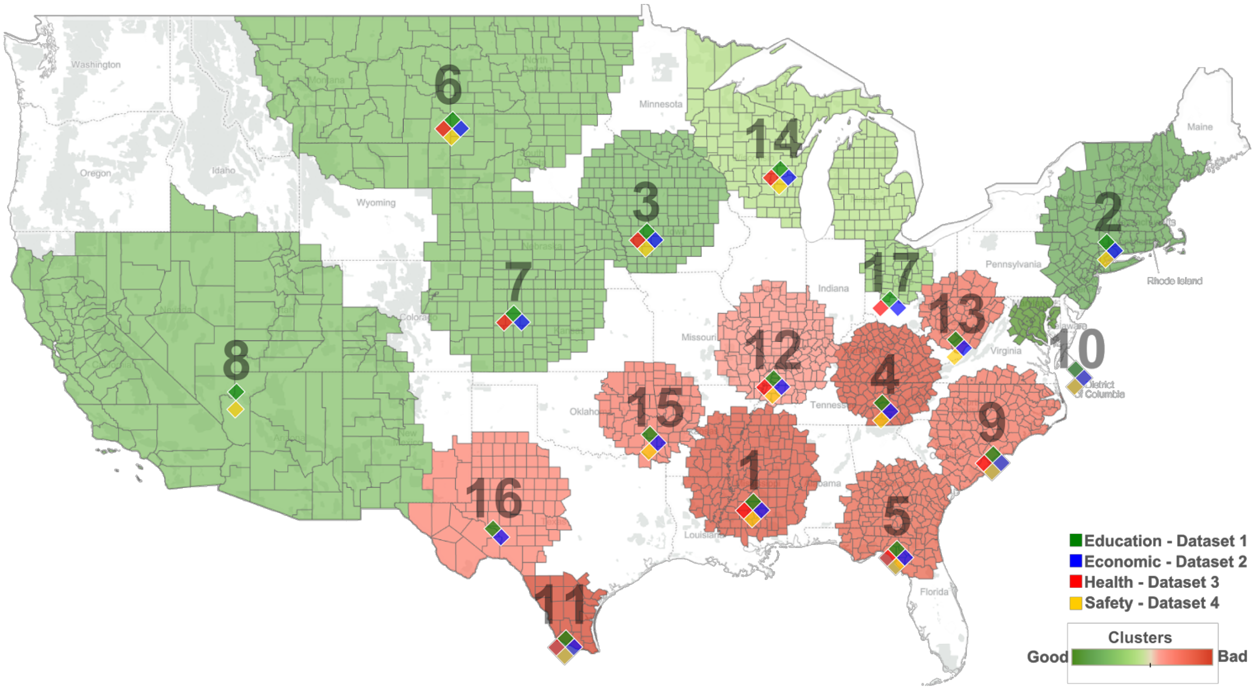


Fig 3.2.1 *Clusters resulting from the multivariate analysis*

Multivariate SaTScan interpretations on the 2010 QoL Index
Data Set 1 : Education (Holds Bachelor's Degree or Higher, Holds Less than High School diploma)
Data Set 2: Financial (Median Income, Unemployment, Poverty, Unemployment change from 2009)
Data Set 3: Health (PM2.5, Obesity, Smoking, Health Insurance)
Data Set 4: Safety (5 crime arrest rates)

Cluster 1: (The Mississippi cluster) High cluster. All 4 data sets make this a high cluster. The means inside range from 0.6 to 1.2. This would mean that this area is typically lowly educated, low financial means, unhealthy environment and lifestyle, and high in crime arrests.

Cluster 2: (The New England cluster) Low cluster. Sets 1, 2, and 4 make this low cluster appear. The means inside range from -1.5 to -0.7. This would mean that this area is typically highly educated, high financial means, and low in crime arrests. Health does not influence this cluster so either is average or slightly unhealthy compared to the rest of the US.

Cluster 3: (The Iowa cluster) Low cluster. All four data sets make this a low cluster. The means inside range from -1.17 to -0.63. This would mean that this area is typically highly educated, high financial means, low in crime arrests, and a healthy environment and lifestyle.

Cluster 4: (The East Kentucky/East Tennessee cluster) High cluster. Data sets 1, 2, and 4 make this a high cluster. The means inside range from 0.67 to 1.18. This would mean that this area is typically lowly educated, low financial means, and high in crime arrests. Health does not influence this cluster so either is average or slightly healthy compared to the rest of the US.

Cluster 5: (The Florida/Georgia cluster) High cluster. All 4 data sets make this a high cluster. The means inside range from 0.5 to 1.03. This would mean that this area is typically lowly educated, low financial means, unhealthy environment and lifestyle, and high in crime arrests.

Cluster 6: (The Plains cluster – Dakotas, Montana, Wyoming) Low cluster. All four data sets make this a low cluster. The means inside range from -1.23 to -0.38. This would mean that this area is typically highly educated, high financial means, low in crime arrests, and a healthy environment and lifestyle.

Cluster 7: (The Nebraska/Kansas/Colorado cluster) Low cluster. Data sets 1, 2, and 3 make this a low cluster. The means inside range from -0.86 to -0.55. This would mean that this area is typically highly educated, high financial means, and a healthy environment and lifestyle. Safety does not influence this cluster so either average or slightly high compared to the rest of the US.

Cluster 8: (The SouthWest cluster – California, Nevada, Utah, Arizona, New Mexico) Low cluster. Data sets 1 and 4 make this a low cluster. The means inside range from -1.11 to -037. This would mean that this area is typically highly educated, and low in crime arrests. Financial and a healthy environment and lifestyle does not influence this cluster so either average or slightly low or unhealthy compared to the rest of the US.

Cluster 9: (The Carolina cluster) High cluster. All 4 data sets make this a high cluster. The means inside range from 0.41 to 0.95. This would mean that this area is typically lowly educated, low financial means, unhealthy environment and lifestyle, and high in crime arrests.

Cluster 10: (The Washington DC cluster) Low cluster. Data sets 1, 2, and 4 make this a low cluster. The means inside range from -1.69 to -1.03. This would mean that this area is typically highly educated, high financial means, and low in crime arrests. Health does not influence this cluster so either is average or slightly unhealthy compared to the rest of the US.

Cluster 11: (The southern Texas cluster) High cluster. Data sets 1, 2, and 3 make this a high cluster. The means inside range from 0.65 1.61. This would mean that this area is typically lowly educated, low financial means, and a unhealthy environment and lifestyle. Safety does not influence this cluster so either average or slightly low compared to the rest of the US.

Cluster 12: (The south Illinois cluster) High cluster. All 4 data sets make this a high cluster. The means inside range from 0.015 to 0.64. This would mean that this area is typically lowly educated, low financial means, unhealthy environment and lifestyle, and high in crime arrests.

Cluster 13: (The West Virginia cluster) High cluster. Data sets 1, 2, and 4 make this a high cluster. The means inside range from 0.45 to 0.83. This would mean that this area is typically lowly educated, low financial means, and high in crime arrests. Health does not influence this cluster so either is average or slightly healthy compared to the rest of the US.

Cluster 14: (The Great Lakes cluster) Low cluster. All four data sets make this a low cluster. The means inside range from -0.45 to -0.052. This would mean that this area is typically highly educated, high financial means, low in crime arrests, and a healthy environment and lifestyle.

Cluster 15: (The west Arkansas/east Oklahoma cluster) High cluster. Data sets 1, 2, and 4 make this a high cluster. The means inside range from 0.35 to 0.76. This would mean that this area is typically lowly educated, low financial means, and high in crime arrests. Health does not influence this cluster so either is average or slightly healthy compared to the rest of the US.

Cluster 16: (The west Texas cluster) High cluster. Data sets 1 and 2 make this a high cluster. The means inside range from 0.15 to 0.7. This would mean that this area is typically lowly educated and low financial means. Safety and health does not influence this cluster so either is average or crime slightly low or healthy environment compared to the rest of the US.

Cluster 17: (The Ohio cluster) Low cluster. Data sets 1, 2, and 3 make this a low cluster. The means inside range from -0.69 to -0.24. This would mean that this area is typically highly educated, high financial means, and a healthy environment and lifestyle. Safety does not influence this cluster so either average or slightly high compared to the rest of the US.

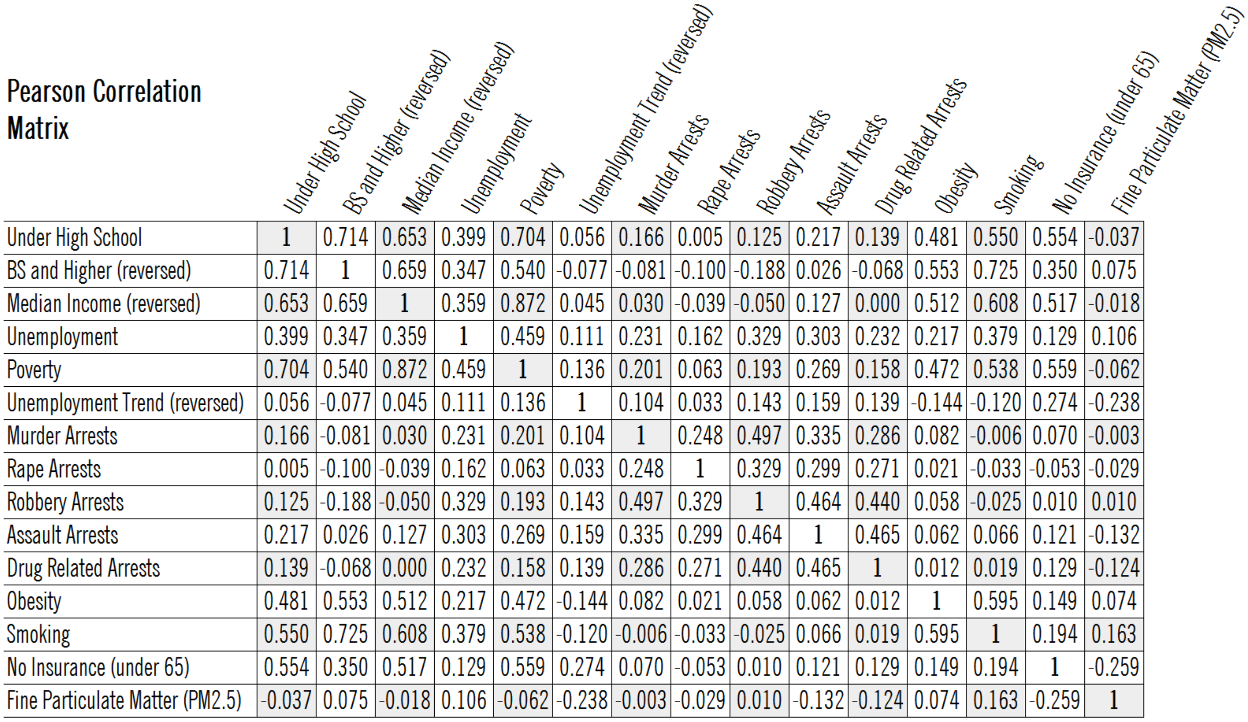
3.3 Factor Analysis

.

A factor analysis was performed on the variables in the QoL index. For this process each variable is examined independently rather than as part of a sub-index. We chose to do an exploratory factor analysis rather than a principal component analysis because an exploratory factor analysis looks for the latent variables underneath the actual variables being measured, while a principal component analysis works best for variables that are measured directly. Our QoL index was designed to measure the latent variables under the indexes, therefore, it lends itself nicely to an exploratory factor analysis approach. An example of one latent variable in the index is *health* which is comprised of the measurable variables smoking rate, obesity rate, PM2.5 air particulates, and percent of people under age 65 without health insurance.

Two independent exploratory factor analyses were performed on the QoL variables, one using the Most Likelihood method and one using a Principal Factor Analysis. Both were performed using SAS version 9.3 and the factor results were put through a cluster analysis using SaTScan version 9.3. This paper will compare and contrast the two exploratory factor analyses and their corresponding results.

Prior to beginning either factor analysis, a ***PROC CORR*** was run in SAS to determine how correlated the variables in the QoL index are (see Table 3.3.1). The Pearson's Correlation Coefficients were then examined. While most of the variables in the index are highly correlated at least a few other variables, it must be noted that the *change in unemployment rate* and *PM2.5 air quality* variables have very little correlation to any other variable in the index.



Pearson Correlation Matrix	Under High School	BS and Higher (reversed)	Median Income (reversed)	Unemployment	Poverty	Unemployment Trend (reversed)	Murder Arrests	Rape Arrests	Robbery Arrests	Assault Arrests	Drug Related Arrests	Obesity	Smoking	No Insurance (under 65)	Fine Particulate Matter (PM2.5)
Under High School	1	0.714	0.653	0.399	0.704	0.056	0.166	0.005	0.125	0.217	0.139	0.481	0.550	0.554	-0.037
BS and Higher (reversed)	0.714	1	0.659	0.347	0.540	-0.077	-0.081	-0.100	-0.188	0.026	-0.068	0.553	0.725	0.350	0.075
Median Income (reversed)	0.653	0.659	1	0.359	0.872	0.045	0.030	-0.039	-0.050	0.127	0.000	0.512	0.608	0.517	-0.018
Unemployment	0.399	0.347	0.359	1	0.459	0.111	0.231	0.162	0.329	0.303	0.232	0.217	0.379	0.129	0.106
Poverty	0.704	0.540	0.872	0.459	1	0.136	0.201	0.063	0.193	0.269	0.158	0.472	0.538	0.559	-0.062
Unemployment Trend (reversed)	0.056	-0.077	0.045	0.111	0.136	1	0.104	0.033	0.143	0.159	0.139	-0.144	-0.120	0.274	-0.238
Murder Arrests	0.166	-0.081	0.030	0.231	0.201	0.104	1	0.248	0.497	0.335	0.286	0.082	-0.006	0.070	-0.003
Rape Arrests	0.005	-0.100	-0.039	0.162	0.063	0.033	0.248	1	0.329	0.299	0.271	0.021	-0.033	-0.053	-0.029
Robbery Arrests	0.125	-0.188	-0.050	0.329	0.193	0.143	0.497	0.329	1	0.464	0.440	0.058	-0.025	0.010	0.010
Assault Arrests	0.217	0.026	0.127	0.303	0.269	0.159	0.335	0.299	0.464	1	0.465	0.062	0.066	0.121	-0.132
Drug Related Arrests	0.139	-0.068	0.000	0.232	0.158	0.139	0.286	0.271	0.440	0.465	1	0.012	0.019	0.129	-0.124
Obesity	0.481	0.553	0.512	0.217	0.472	-0.144	0.082	0.021	0.058	0.062	0.012	1	0.595	0.149	0.074
Smoking	0.550	0.725	0.608	0.379	0.538	-0.120	-0.006	-0.033	-0.025	0.066	0.019	0.595	1	0.194	0.163
No Insurance (under 65)	0.554	0.350	0.517	0.129	0.559	0.274	0.070	-0.053	0.010	0.121	0.129	0.149	0.194	1	-0.259
Fine Particulate Matter (PM2.5)	-0.037	0.075	-0.018	0.106	-0.062	-0.238	-0.003	-0.029	0.010	-0.132	-0.124	0.074	0.163	-0.259	1

Table 3.3.1 *Pearson's Correlation Matrix*

3.3.1 EFA using the Maximum Likelihood (ML) method

A Maximum Likelihood Exploratory factor analysis was performed on the 15 selected social indicators using the following code in SAS:

```
PROC FACTOR method=ML prior=SMC rotate=Varimax
```

The varimax parameter causes an orthogonal rotation of the factors which is easier to interpret than the original factor results. It is easier to interpret because the rotation aims to weigh each variable heavily in one factor. Furthermore, an orthogonal rotation results in independent factors.

The function converged after 15 iterations. After the factors are extracted, the ML method performs a Chi-square test with 2 separate hypothesis on the significance of the residuals, the results is as follows:

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	105	22939.8739	<.0001
HA: At least one common factor			
H0: 3 Factors are sufficient	63	2387.4666	<.0001
HA: More factors are needed			

Table 3.3.2 *Significance Tests Based on 3109 Observations*

The null hypothesis cannot be rejected for both test ($p<0.05$). For the first test, a nonsignificant value would imply that the indicators are not intercorrelated enough to perform factor analysis since the results would be difficult to reproduce [2].

	Eigenvalue	Difference	Proportion	Cumulative
1	102.432324	90.035307	0.8667	0.8667
2	12.397016	9.045680	0.1049	0.9716
3	3.351336	2.284937	0.0284	**1.0000**
4	1.066399	0.721735	0.0090	1.0090
5	0.344665	0.109111	0.0029	1.0119
6	0.235554	0.200647	0.0020	1.0139
7	0.034907	0.041602	0.0003	1.0142
8	-0.006695	0.071624	-0.0001	1.0142
9	-0.078318	0.039118	-0.0007	1.0135
10	-0.117437	0.019562	-0.0010	1.0125
11	-0.136999	0.087094	-0.0012	1.0114
12	-0.224093	0.066362	-0.0019	1.0095
13	-0.290454	0.058151	-0.0025	1.0070
14	-0.348605	0.130323	-0.0029	1.0041
15	-0.478929		-0.0041	1.0000

Table 3.3.3 *Eigenvalues of the Weighted Reduced Correlation Matrix*

We observed that the ML method was able to explain 100% of the common variance with 3 factors. We also notice that the eigenvalues for the first three factors were well above one.

	Factor1	**Factor2**	**Factor3**
Under High School	0.74911	0.28469	0.21331
BS and Higher (reversed)	0.63176	0.72909	0.10026
Median Income (reversed)	0.89648	0.15112	-0.16884
Unemployment	0.47187	0.01007	0.37955
Poverty	0.9877	-0.11263	-0.02059
Unemployment Trend (reversed)	0.11351	-0.23001	0.04858
Murder Arrests	0.1764	-0.3242	0.45396
Rape Arrests	0.04554	-0.22682	0.35651
Robbery Arrests	0.15432	-0.47417	0.64871
Assault Arrests	0.25408	-0.26125	0.49903
Drug Related Arrests	0.13951	-0.28126	0.48285
Obesity	0.51591	0.31012	0.125
Smoking	0.59978	0.46399	0.12713
No Insurance (under 65)	0.56466	-0.01235	-0.08477
Fine Particulate Matter (PM2.5)	-0.04611	0.15134	0.05776

Table 3.3.4 *Factor Pattern (before rotation)*

	Factor1	**Factor2**	**Factor3**
Under High School	0.80654	0.1812	-0.0661
BS and Higher (reversed)	0.86717	-0.13519	-0.41289
Median Income (reversed)	0.88005	-0.07223	0.27443
Unemployment	0.44939	0.40551	-0.02089
Poverty	0.86621	0.19048	0.4495
Unemployment Trend (reversed)	0.01699	0.16348	0.20283
Murder Arrests	0.04984	0.57258	0.10946
Rape Arrests	-0.03586	0.42236	0.03068
Robbery Arrests	-0.02316	0.80738	0.13066
Assault Arrests	0.147	0.59668	0.06476
Drug Related Arrests	0.03321	0.57298	0.04809

Obesity	0.59898	0.05468	-0.12725
Smoking	0.73587	0.00146	-0.22286
No Insurance (under 65)	0.51331	0.02142	0.24946
Fine Particulate Matter (PM2.5)	0.01766	-0.02395	-0.16578

Table 3.3.5 *Factor Pattern (after VARIMAX orthogonal rotation)*

Tables 3.3.4 and 3.3.5 shows the significance of rotating the pattern. The results in table 3.3.4 were a bit difficult to label whereas table 3.3.5 makes it a bit clearer. We used a 0.4 rule of thumb. If the absolute value of a given factor loading is 0.4 and above, we regarded it as significant. The significant loadings are highlighted in table 2.4. We use this rule of thumb to name the factors. This labeling of the factors is important as it will be key in understanding the meaning of the factor scores and ultimately in interpreting the meaning of high and low clusters when cluster analysis is performed on those scores.

The factors are described and labeled below:

- ***Factor 1***
  This factor is comprised of *low HS graduation rate, low Bachelor degree completion rate, low median income, high poverty, high obesity rates, high smoking rates,* and *high levels of uninsured people*. The *unemployment rate* also contributed to this factor, though not as highly as the other variables. This factor has the highest eigenvalue, accounting for 76.72% of the common variance. Let us call this the **uneducated, unhealthy poor factor**.
- ***Factor 2***
  This factor is comprised primarily of the crime variables: *murder arrests, rape arrests, robbery arrests, aggravated assault arrests* and *drug related arrests*. The unemployment rate also contributes to this factor. This factor accounts for 16.47% of the common variance. We can probably call this one the **unsafe factor**.
- ***Factor 3***
  This is the factor with the lowest eigenvalue of the three, accounting for only 8.46% of the common variance among the indicators. It is comprised of *high bachelor degree completion and high poverty rates*. We can call this one the **educated but poor factor**.

All the social indicators, with the exception of *Fine Particulate Matter (PM2.5) air quality* and *unemployment trend*, contributed to at least one factor with a weight greater than 0.40, as indicated by the highlighted factor loadings in Table 3.3.5. The *unemployment rate* contributed to Factor 1 and Factor 2 about equally. This seems reasonable since high unemployment would be associated with areas that are either high in poverty or crime.

	Factor1	**Factor2**	**Factor3**
Under High School	0.05204466	0.11159639	-0.1208795
BS and Higher (reversed)	0.46106705	-0.0501745	-0.8886719
Median Income (reversed)	0.0777503	-0.2607964	0.08481199
Unemployment	0.01093686	0.12196212	-0.0635351
Poverty	0.47729523	0.09781251	1.09224743
Unemployment Trend (reversed)	-0.0057099	0.01895232	0.00962826
Murder Arrests	-0.0074274	0.15652556	-0.0443644
Rape Arrests	-0.0047143	0.0973608	-0.0303957
Robbery Arrests	-0.0247092	0.44572296	-0.1251768
Assault Arrests	-0.0033753	0.17807599	-0.0608211
Drug Related Arrests	-0.0056691	0.1605566	-0.052414
Obesity	0.02308162	0.02538739	-0.0490564
Smoking	0.04785624	0.02748603	-0.0972004
No Insurance (under 65)	0.00613354	-0.023556	0.01763322
Fine Particulate Matter (PM2.5)	0.00445646	0.00678054	-0.016009

Table 3.3.6 - *Standardized scoring Coefficients*

To calculate the new latent variables or the factor scores, SAS ran PROC SCORE using the standardized scoring coefficient matrix output from PROC FACTOR, Table 3.3.6 above. The following equations was used: [2]

$$f_{i,k} = v_{i,1}l_{k,1} + v_{i,2}l_{k,1} + \ldots + v_{i,15}l_{k,15} \quad \textit{where} \;\; k=1,2,3;\; i=1, \ldots, 3109 \qquad \textbf{(3.3.1)}$$

k is the factor number, 3 factors. i is the county, 3109 counties. $f_{i,k}$ is the kth factor score for the ith county. v is the z-score for the social indicator, we have 15 of them. e_i is the residual of f_i.

The factor scores were normalized in SAS using ***`PROC RANK`*** with parameter **`normal=blom`**. A purely spatial cluster analysis was performed on the normalized factor scores using the normal probability model with a 5% window size in SaTScan. The "low" clusters in SaTScan show areas where this index indicates favorable results, while the "high" clusters show areas with undesirable conditions. Tableau was also used to produce a heat map of each index. In these maps, green areas correspond to positive conditions (low values) while red areas correspond to undesirable conditions (high values). The clusters shown are all statistically significant with p-values less than 0.05. The log likelihood ratio is indicated by the cluster number where 1 is most likely (highest log likelihood ratio). The intensity of the colors indicates how far the mean inside a particular cluster is away from the mean of all the counties.

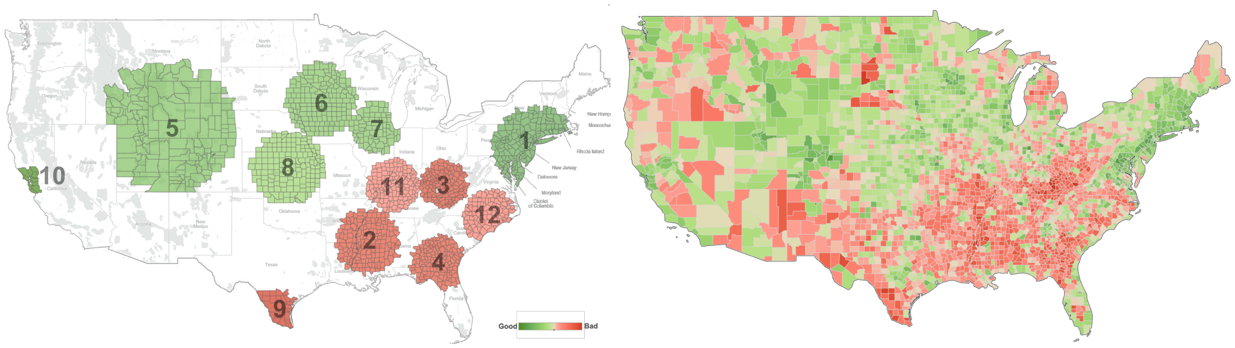


Fig 3.3.1 *Clusters and Heat map for the poor, uneducated and unhealthy (Factor 1)*

In looking at these maps, figure 3.3.1, it is apparent that Factor 1 explains a large percentage of the common variance in the index. The South East shows as areas with lower education and higher poverty and more unhealthy behaviors, and the North East, areas around the great lakes and the Midwest have areas with higher education, lower poverty, lower obesity and lower smoking rates. The clusters in the southern area of the country in the Factor 1 map are nearly identical to those in the QoL cluster maps shown in section 3.1, figure 3.1.1. The same can be said of the green clusters in the North East while the ones near the Great Lakes.

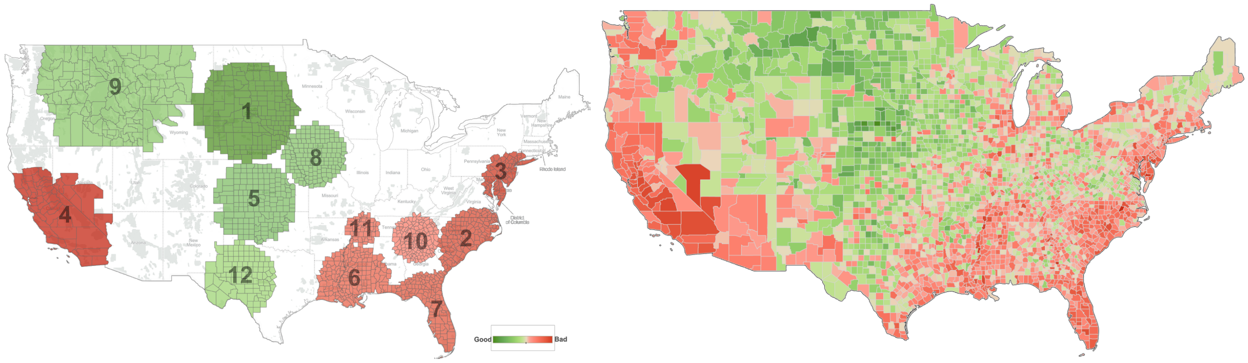


Fig 3.3.2 *Clusters and Heat map for unsafe factor (Factor 2)*

The maps for Factor 2 also compared favorably to the QoL index as a whole (See figure 3.1.1). The green clusters in the South correspond quite closely to those in the QoL index, showing that particular area of the country is being impacted negatively by all the variables in both factors. The areas in Kansas, Nebraska, Idaho and both North and South Dakota show as red clusters again showing that these areas of the country are affected by both factors. Interestingly, the green cluster in the North East is not on the Factor 2 map, showing that this area of the country has some crime and the Factor 1 scores are driving it to perform well on the QoL index. The clusters are also similar to those identified using the Safety Subindex, figure 3.1.1.1. The California cluster and cluster #4 as well as the red clusters in the Southeast seem very close.

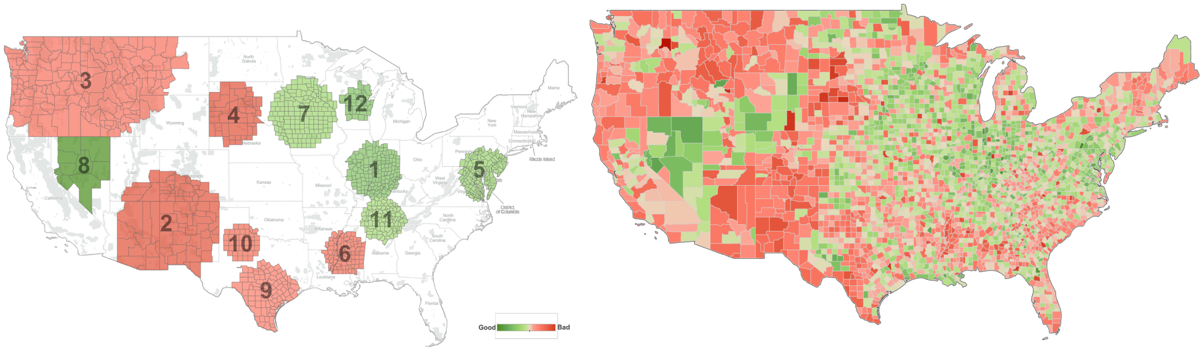


Fig 3.3.3 *Clusters and Heat map for the educated but poor factor (Factor 3)*

Although Factor 3 explained the lowest percentage of the overall common variance, it came up with a rather interesting latent concept. As expected, the clusters here are very different for any of the maps shown in section 3.1, they represent an irony. The red clusters are areas with high education and high poverty whereas the green clusters represent areas the opposite phenomenon. It seems as though there is an area extending through the middle of the country where things are not necessarily as they seem. These areas scored relatively well in the QoL Index as well as in the subindices but the latent subindex uncovered through exploratory factor analysis shows an alternate view. This is really a testament to the kind of information that can come out of Factor Analysis.

3.3.1.1 An Alternate QoL Index using Factor Scores

The factor scores are linear combinations of the social indicators with the factor loadings. See equation 3.3.1. These factor scores can be viewed as latent sub-indexes analogous to the sub-indexes we described in section 2.1. The added benefit here is that these latent sub-indexes or factor scores are inherently uncorrelated. They're telling the same story from a different angle, as we've seen with the cluster analysis on the individual factor scores. Furthermore, the weights associated with each latent sub-index are given by the percentage of the variance explained by each factor. From Table 3.3.3 we can see that Factor 1 accounts for 86.67% of the total common variance, Factor 2 for 10.49% and Factor 3 for 2.84%. We use these values as weights in calculating the alternate QoL Index as follows:

$$\textit{QoL Index} = 0.866f_1 + 0.1049f_2 + 0.0284f_3 \tag{3.3.2}$$

where f_1, f_2, and f_3 are the factor scores associated with each county.

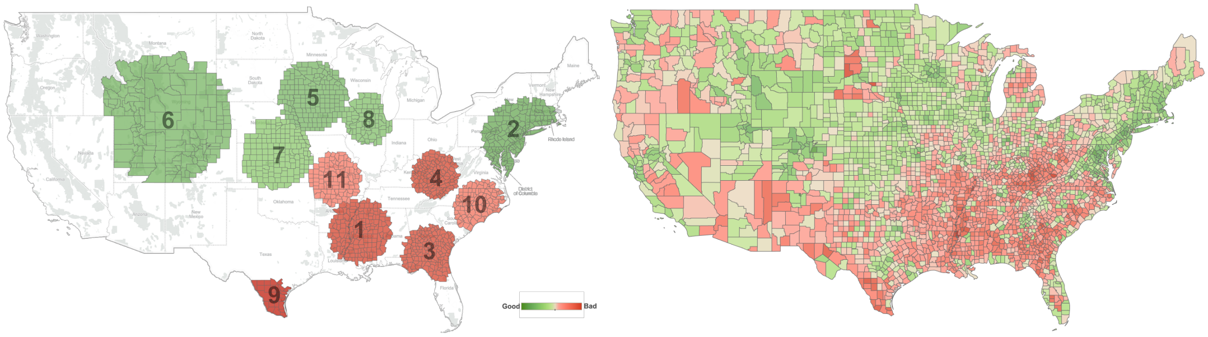


Fig 3.3.4 *Clusters and Heat map for the alternate QoL Index*

A purely spatial cluster analysis was performed on this alternate QoL Index. Interestingly enough, these results are not that different from the earlier results found in section 3.1, figures 3.1.1 and 3.1.2. The high(red) and low (green) clusters identified are nearly identical with a slight shift in the maximum likelihood ratio.

It turns out that using Factor Analysis to weigh components in a Quality of Life application is not such a foreign idea. Money magazine's "Best Places to Live" does just that in combination with surveys from their readers [15]. Estes's (1988) Index of Social Progress also uses factor analysis with a varimax rotation on 40 social indicators [15].

3.3.2 EFA using the Principal Factor Analysis Method

A Principal Factor Analysis was performed independently of the Most Likelihood Factor Analysis. Prior to performing the Principal Factor Analysis, we decided to remove the *PM2.5 air quality* and *unemployment rate change* variables from the set of variables used for the factor analysis because they are not correlated to any other variables in the set. The analysis was run in SAS using the following code:

```
Proc FACTOR  method=p n=5 prior=SMC rotate=varimax
```

Five factors were retained due to the n=5 parameter, but it appears that three factors account for 100% of the common variance among the variables (see Table 3.3.7). Table 3.3.8 shows the rotated factors.

	Eigenvalue	Difference	Proportion	Cumulative
1	4.36619302	2.34658273	0.6506	0.6506
2	2.01961029	1.39746754	0.3009	0.9515
3	0.62214274	0.35806061	0.0927	1.0443

4	0.26408213	0.12242867	0.0394	1.0836
5	0.14165346	0.05298876	0.0211	1.1047
6	0.08866470	0.11184169	0.0132	1.1179
7	-0.02317699	0.04909914	-0.0035	1.1145
8	-0.07227613	0.02619447	-0.0108	1.1037
9	-0.09847061	0.01325799	-0.0147	1.0890
10	-0.11172859	0.02911804	-0.0166	1.0724
11	-0.14084663	0.01933529	-0.0210	1.0514
12	-0.16018193	0.02452242	-0.0239	1.0275
13	-0.18470435		-0.0275	1.000

Table 3.3.7 *Eigenvalues of the Reduced Correlation Matrix*

	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
Under High School	0.61522	0.16757	0.58492	-0.15184	0.00632
BS and Higher (reversed)	0.81305	-0.15878	0.31507	-0.11723	-0.14523
Median Income (reversed)	0.62269	-0.03033	0.59533	0.36853	-0.02316
Unemployment	0.41254	0.41093	0.12481	0.10553	-0.15425
Poverty	0.53962	0.24168	0.65256	0.34879	0.04162
Murder Arrests	0.02169	0.55884	0.06883	0.00390	0.18788
Rape Arrests	-0.01141	0.44421	-0.07616	0.04322	-0.01250
Robbery Arrests	-0.01262	0.75885	-0.00906	0.02461	0.12539
Assault Arrests	0.06597	0.62558	0.13245	0.00687	-0.11464
Drug Related Arrests	-0.02175	0.58757	0.10303	-0.06335	-0.11298
Obesity	0.67228	0.03966	0.10238	0.05409	0.16910
Smoking	0.81445	-0.00879	0.11890	0.06332	-0.06056
No Insurance (under 65)	0.15105	0.03588	0.70517	-0.03214	-0.01570

Table 3.3.8 *Rotated Factor Pattern from the Principal Factor Analysis*

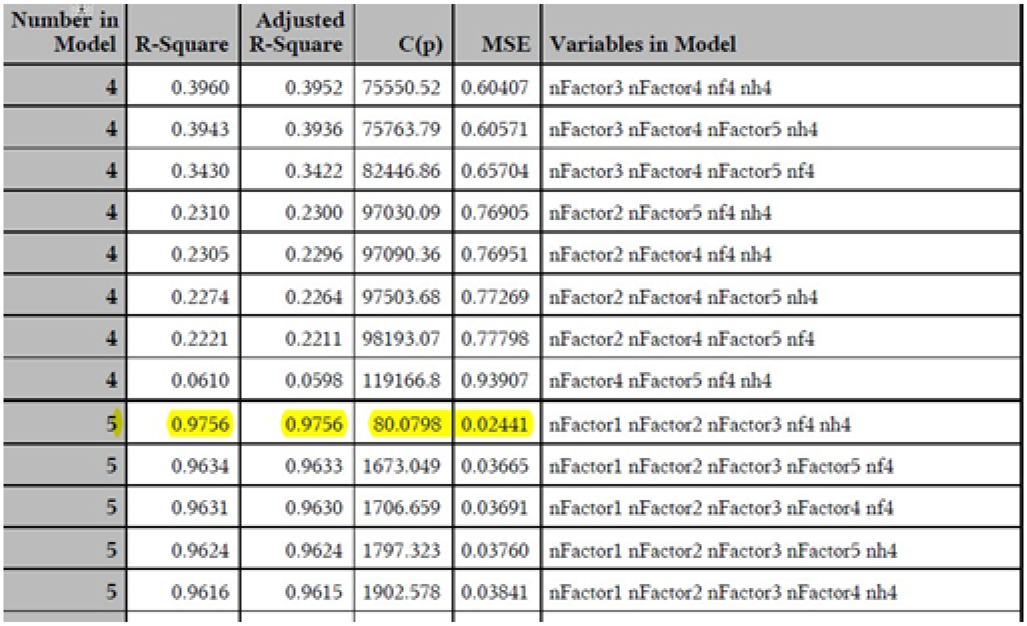
An analysis of the rotated factors found the following:

- ***Factor 1***
  The results of this factor are strikingly similar to those produced by the ML analysis described earlier, and it will also be referred to as the **uneducated, unhealthy poor factor.** The only difference in between the results produced using the ML method versus this Principal Factor Analysis, is that the ML included *percent of uninsured* in Factor 1, while it is not included in this analysis. Factor 1 is the strongest factor in this analysis, having an eigenvalue of 4.366 and accounting for 65% of the common variance among the variables.
- ***Factor 2***
  This factor included all the same variables as those found using the ML analysis and will be called the **unsafe factor**. In the Principal Component Analysis, Factor 2 has an eigenvalue of 2.019 and accounts for 30% of the common variance.
- ***Factor 3***
  This factor includes the variables: *low high school graduation, low median income, high poverty,* and *high uninsured rates*. Let us call this factor the **uneducated, uninsured, poor factor.** Factor 3 has an eigenvalue of .622 and accounts for 9% of the common variance. Factor 3 will be retained even though its eigenvalue is less than 1 because, along with Factor 1 and Factor 2, it will explain 100% of the common variance in the variables.
- ***Factors 4 & 5***
  In examining the rotated factor pattern, it does not appear that anything new is added to the analysis from these two factors. There is not a single variable in either of these factors that weighs in at .40 or higher. Therefore, these factors do not appear to be needed.

In order to validate the decision to keep Factor 1, Factor 2, Factor 3, the *PM2.5 air quality (nh4)*, and the *unemployment percent change (nf4)* variables as the five factors in this analysis, a regression was performed to see if these variables produce an acceptable model of the QoL Index. The regression was run in SAS using the following code:

```
Proc RSQUARE adjrsq mse cp
Model nQoL=nFactor1-nFactor5 nf4  nh4
```

The results of the **RSQUARE** procedure show that these variables do fit the QoL index with a fit of 97.56% and an MSE of 2.44% (see Table 3.3.9). It must be noted that the C(p)=80.08 which is higher than would be desired in building a predictive model. Since we are just using this information to determine which factors to keep as part of our analysis, this value should not negatively impact our results.



Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
4	0.3960	0.3952	75550.52	0.60407	nFactor3 nFactor4 nf4 nh4
4	0.3943	0.3936	75763.79	0.60571	nFactor3 nFactor4 nFactor5 nh4
4	0.3430	0.3422	82446.86	0.65704	nFactor3 nFactor4 nFactor5 nf4
4	0.2310	0.2300	97030.09	0.76905	nFactor2 nFactor5 nf4 nh4
4	0.2305	0.2296	97090.36	0.76951	nFactor2 nFactor4 nf4 nh4
4	0.2274	0.2264	97503.68	0.77269	nFactor2 nFactor4 nFactor5 nh4
4	0.2221	0.2211	98193.07	0.77798	nFactor2 nFactor4 nFactor5 nf4
4	0.0610	0.0598	119166.8	0.93907	nFactor4 nFactor5 nf4 nh4
5	0.9756	0.9756	80.0798	0.02441	nFactor1 nFactor2 nFactor3 nf4 nh4
5	0.9634	0.9633	1673.049	0.03665	nFactor1 nFactor2 nFactor3 nFactor5 nf4
5	0.9631	0.9630	1706.659	0.03691	nFactor1 nFactor2 nFactor3 nFactor4 nf4
5	0.9624	0.9624	1797.323	0.03760	nFactor1 nFactor2 nFactor3 nFactor5 nh4
5	0.9616	0.9615	1902.578	0.03841	nFactor1 nFactor2 nFactor3 nFactor4 nh4

Table 3.3.9 *Results of the RSQUARE Procedure*

The factors found using the Principal Factor Analysis were normalized in SAS using:

```
Proc Rank normal = blom
```

These normalized factors were then used in SaTScan to do a cluster analysis. SaTScan was run using a normal model, circular clusters, and a 5% population maximum size for each cluster. Maps were produced using Tableau, and all show green areas as having a more desirable result than the red areas.

The maps for Factor 1 **(uneducated, unhealthy poor)** are shown in Figure 3.3.5. They are similar to the results produced using Factor 1 in the ML analysis, with a couple of notable differences. The Principal Factor Analysis included a large positive area in Texas and excluded a positive area near the Great Lakes. The other clusters are very similar with only subtle differences. The heat maps produced by using the Principal Factor results are similar to those using the ML results, except for areas around southern Texas, California and the very northern part of the country, which rate higher using the Principal Factor method on only 13 variables. The clusters found using the results from Factor 1 are similar to those shown in the Quality of Life Index maps. Again, the most notable difference is the positive area in Texas, while areas such as the Midwest, the South, and the Northeast are very similar.

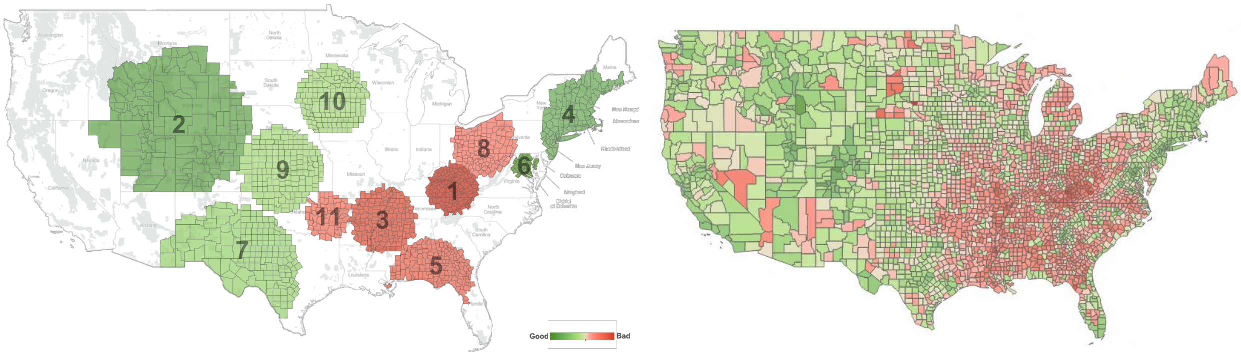


Figure 3.3.5: *Cluster and Heat Maps for Factor 1 (**uneducated, unhealthy poor**)*

The maps for Factor 2 **(unsafe)** match the maps for the ML Factor 2 results with the exception of a cluster in Texas. The ML method showed an area of Texas having positive results in this factor, while the Principal Factor Analysis results don't show that cluster (See Figure 3.3.6). Interestingly, this is the same area where the Principal Factor Analysis found a positive cluster using the Factor 1 results, while the ML method did not. Since the ML results have already been compared to the QoL Index, that same analysis will not be repeated here.

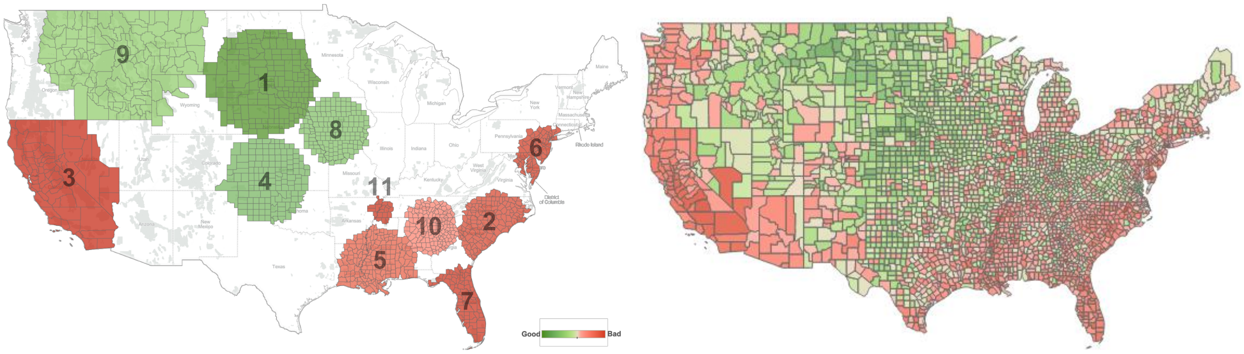


Figure 3.3.6: *Cluster and Heat Maps for Factor 2 (**unsafe)***

The variables making up Factor 3 in the Principal Factor Analysis differ significantly from the variables making up Factor 3 in the ML analysis, so there is no benefit in comparing the cluster maps developed from these factors to each other. The strongest variable in Factor 3 **(uneducated, uninsured poor)** is the *percent of uninsured*. It's not surprising that the South performs negatively in this factor while the Great Lakes area and north-east perform well. Even though this factor only accounts for about 9% of the variance in the variables, it still does a really good job of reflecting the QoL Index as a whole (see Figure 3.3.7).

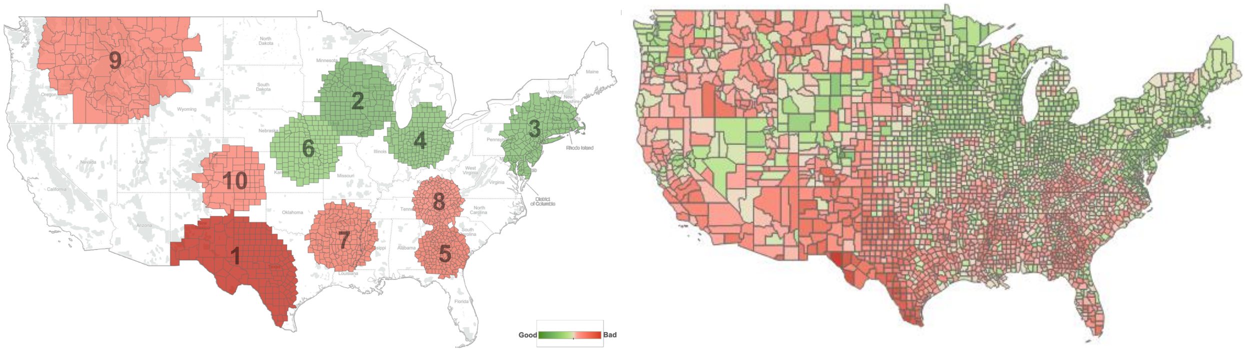


Figure 3.3.7 *Cluster and Heat Maps for Factor 3* ***(uneducated, uninsured poor)***

The variables *PM2.5 air quality* and *unemployment percent change* also explain some of the QoL Index. The impact of these variables has been analyzed earlier in this paper; therefore, that analysis will not be repeated here.

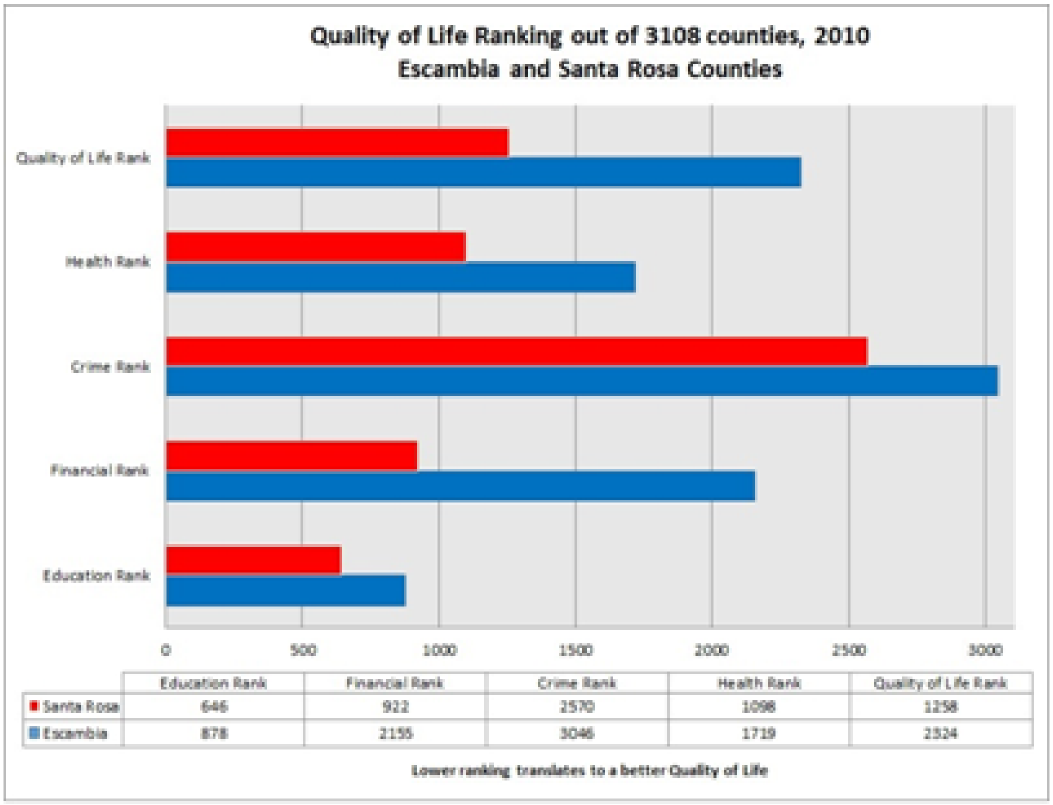
4. Application of the QoL

During the development of QoL index, The Pensacola News Journal showed interest in the rankings of Santa Rosa and Escambia counties in Florida and how they fared in our Quality of Life Index. We utilized the QoL Index in progress to provide a comparison. At this time the QoL Index consisted of 11 variables (Table 4.1) where the final QoL Index includes 15. Using the same normalized scores from each of our sub-indexes and our QoL index, we ranked all the counties in an order such that lower rank implies better (Figure 4.1). The percentile rank of Santa Rosa and Escambia were calculated as well (Figure 4.2). A Tableau map was generated for the QoL index for 2010 (Figure 4.4)

The sub-indexes in our QoL index included the following:

Education	Age 25 and older with less than a high school diploma
	Age 25 and older with a Bachelor's degree or higher
Financial	Median Income
	Unemployment
	Poverty
Safety	Murder Rate
	Rape Rate
	Robbery Rate
	Assault Rate
Health	Obesity Rate
	Smoking Rate

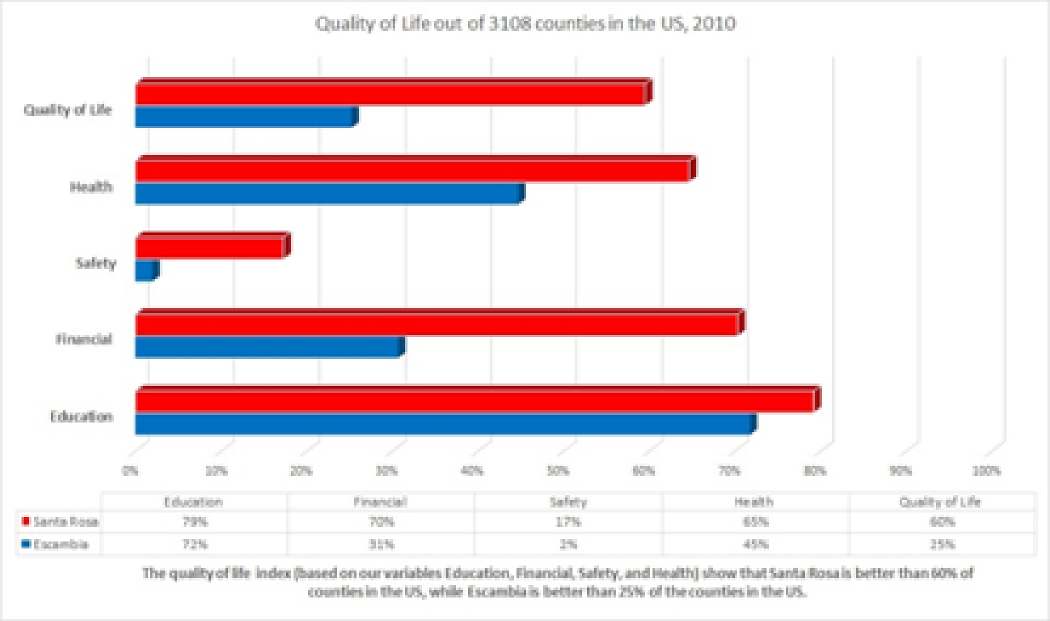
Table 4.1 *Table of variables used for the comparison of Santa Rosa and Escambia*



	Education Rank	Financial Rank	Crime Rank	Health Rank	Quality of Life Rank
🟥 Santa Rosa	646	922	2570	1098	1258
🟦 Escambia	878	2155	3046	1719	2324

Lower ranking translates to a better Quality of Life

Fig 4.1 *Ranking of Escambia and Santa Rosa Counties in QoL Index*



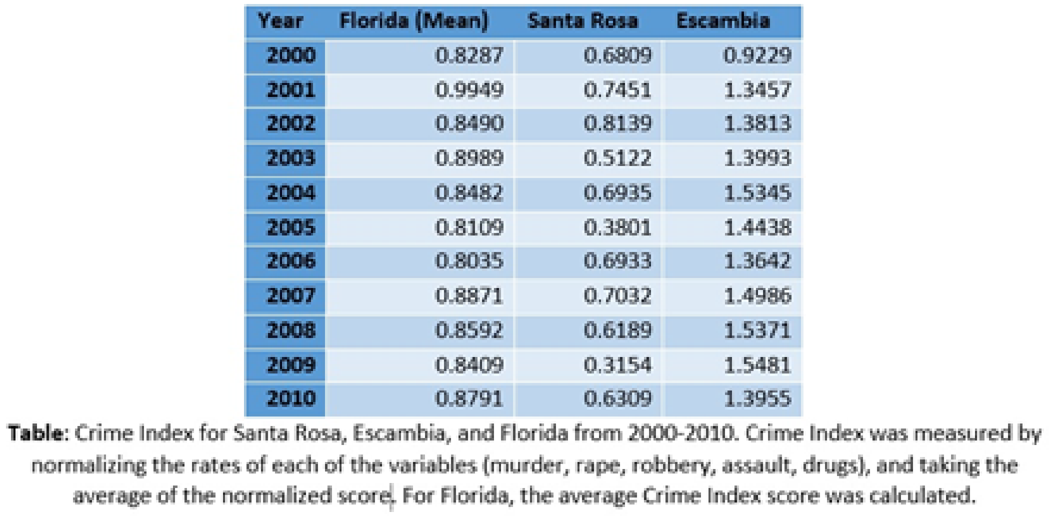
	Education	Financial	Safety	Health	Quality of Life
🟥 Santa Rosa	79%	70%	17%	65%	60%
🟦 Escambia	72%	31%	2%	45%	25%

The quality of life index (based on our variables Education, Financial, Safety, and Health) show that Santa Rosa is better than 60% of counties in the US, while Escambia is better than 25% of the counties in the US.

Fig 4.2 *Percentile ranking of Escambia and Santa Rosa in Qol Index*

It was seen that Santa Rosa ranks 646 in the education sub-index and is higher in the ranks than 79% of the counties in the US; whereas, Escambia is ranked at 878 and in the 72th percentile. In terms of the Economic sub-index, Santa Rosa ranked 922 and in the 70th percentile, while Escambia ranked 2155 and is in the 31st percentile, implying that Santa Rosa is much better off than Escambia is. Both the counties ranked poorly in the safety index: Santa Rosa ranked 2570 and in the 17th percentile, while Escambia ranked 3046 and in the 2nd percentile; this implies that 98% of the counties in the United States are better than Escambia county in the safety index. In the health index, Santa Rosa ranked at 1098 and in the 65th percentile, while Escambia ranked 1719 and in the 45th percentile. When it came to the QoL index, Santa Rosa's rank is 1258 and Escambia's rank is 2324; Santa Rosa is better than 65% of the counties in the QoL index for 2010, whereas Escambia is better than 25% of the counties. According to our QoL index, Santa Rosa is better than Escambia in each of the sub-indexes as well as in the overall QoL index.

In addition to that, we used the same methodology that was used to calculate our QoL index to find the safety sub-index for all the counties for the years 2000-2010. That is, the rates of each of the crime variables (murder, rape, robbery, assault, and drugs) were normalized and the average of these normalized scores was taken for each county. From this index, we extracted the normalized score for Santa Rosa and Escambia (Table 4.2) for each year. The average normalized score was also calculated for Florida so that we may compare how these two counties rate in the crime index (Figure 4.3).



Year	Florida (Mean)	Santa Rosa	Escambia
2000	0.8287	0.6809	0.9229
2001	0.9949	0.7451	1.3457
2002	0.8490	0.8139	1.3813
2003	0.8989	0.5122	1.3993
2004	0.8482	0.6935	1.5345
2005	0.8109	0.3801	1.4438
2006	0.8035	0.6933	1.3642
2007	0.8871	0.7032	1.4986
2008	0.8592	0.6189	1.5371
2009	0.8409	0.3154	1.5481
2010	0.8791	0.6309	1.3955

Table: Crime Index for Santa Rosa, Escambia, and Florida from 2000-2010. Crime Index was measured by normalizing the rates of each of the variables (murder, rape, robbery, assault, drugs), and taking the average of the normalized score. For Florida, the average Crime Index score was calculated.

Table 4.2 *Comparison of Santa Rosa and Escambia from 2000-2010 in the crime Index*

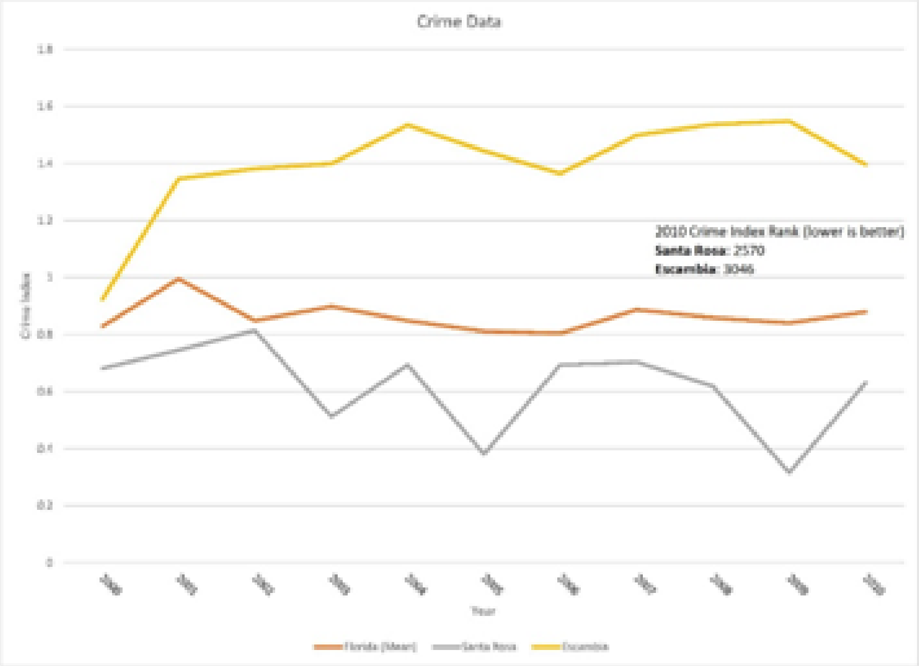


Figure 4.3. *Safety Index normalized score for Santa Rosa, Escambia and Florida (mean)*

As seen in Figure 4.3, Florida is between 0.8 and 1.0 standard deviations above the mean throughout 2000-2010, that is Florida has been worse off than about 80% of the states throughout these years. The figure also shows the line for Santa Rosa below the Florida mean line, implying that Santa Rosa is better than the Florida average safety index scores. On the other hand, Escambia county is always higher than the Florida mean, implying that it's safety index is rather poor compared to the rest of Florida. The average Florida safety index has remained fairly the same, between 0.8 and 1.0, showing little change. Santa Rosa shows some fluctuations between the years. Escambia county saw an increase in its safety index throughout the years, although, in 2010, it did see a drop.

This application shows that using the methodology derived in our QoL index, we may study the ranks of any county of interest to compare how well it is doing in either the overall QoL index or any of the sub-indexes.

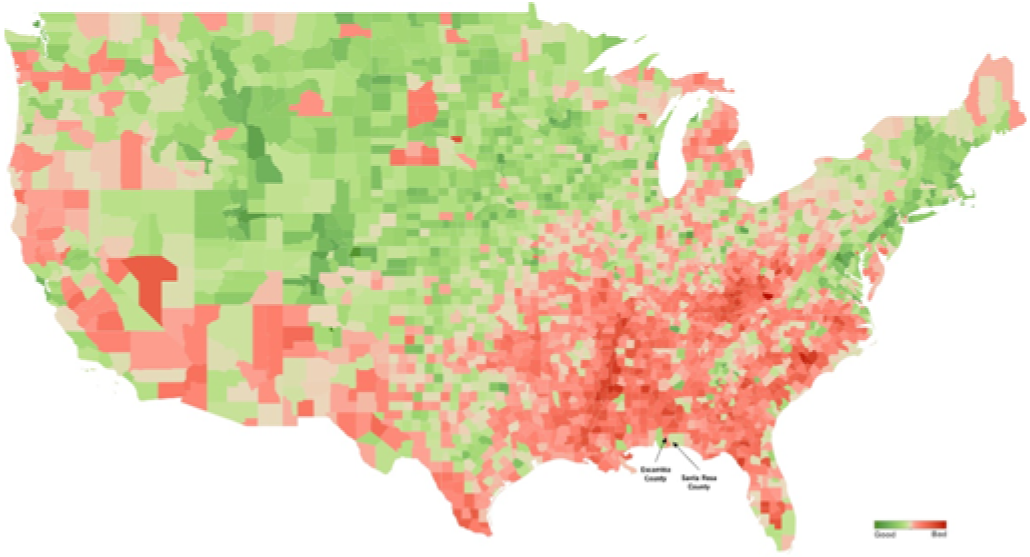


Fig 4.4 *QoL map for 2010 using the variables shown in Table **Table 4.1**. Escambia and Santa Rosa are highlighted on the map*.

5. Conclusion

The maps shown in SaTScan and Tableau was useful in different ways in determining which areas have the better or worse Quality of Life. SaTScan showed the performance of a general area and Tableau showed the performance of each county individually.

Both the originally QoL Index maps and the Multivariate analysis map provided similar results. The QoL Index for SaTScan shows the Mountain West (Colorado, Wyoming, Montana, North Dakota, South Dakota, etc.) performed the best while the South (Mississippi, Louisiana, Texas, Georgia, Kentucky, and North Carolina) performed the worst. The Multivariate Analysis (SaTScan analysis of all variables together) results showed California, Mountain West, and Midwest, and the Northeast performed the best, while the South performed the worst.

In the Safety Sub-Index (Murder Rate, Rape Rate, Robbery Rate, Aggravated Assault Rate, and Drug Related Crimes Rate), the Mountain West and Ohio performed the best, while California and the South performed the worst. Most of the crime variables do show the South and California as the worst and the Mountain West as the best. These are the differences in the variables from the Sub-Indexes:

- Alabama, a southern state, showed as one of the best for drug related crimes.
- It also showed the New York and New Jersey areas as one of the worst in drug related crimes.

- It also had the Northwest (Oregon and Washington) among the worst in rape rate with Minnesota and Wisconsin.
- It had the New York and New Jersey areas among the worst in murder and robbery
- West Texas was shown to be among the best in robbery rate.

In the Economic Sub-Index (Unemployment Rate, Poverty Rate, Unemployment Trend, and Median Household Income), the Upper Midwest (South Dakota, North Dakota, Iowa, Minnesota, Kansas, and Illinois) and the East Coast between Washington DC and New York showed the best results. The South, Southwest (New Mexico, Arizona, Southern California) and the Pacific Northwest (Oregon and Washington) showed the worst results. There were some differences in each variable than there the results from the Economic Sub-Index map:

- The Pacific Northwest does not show as among the worst in any of the variables, but show as among the worst in the Economic Sub-Index.
- The Atlanta area in Georgia, in the South, showed as one of the best for median income
- California showed a one of the best for median income despite showing among the worst in the Sub-Index map
- Some parts of South Dakota and North Dakota showed as the worst in poverty rate. This could be the Indian reserves in those states.
- Michigan, a Midwest state, showed as among the worst in unemployment rate and showed the best for unemployment trend.
- West Texas showed among the best in unemployment rate.
- Three southern states (Alabama, Tennessee, and South Carolina) showed among the best for unemployment trend.

In the Education Sub-Index (High School Dropout, Bachelor’s Degree or higher), the northern half of the country performed the best and the southern half performed the worst. The SaTScan and Tableau of each variable did not show any big differences.

In the Health Sub-Index (Adult Obesity Percent, Adult Smoking Percent, PM2.5, and Percent Uninsured), the Northeast, West Coast, Mountain West, and Minnesota showed the best, while the South and Midwest showed the worst. There were many differences shown in this sub-index as the PM2.5 showed different results than the other health variables. The obesity, smoking, and percent without insurance mainly showed the South as the worst and the Northeast and the West Coast as the best. The PM2.5 shows different results than the remaining of the variables. The South, Midwest and Colorado shows the worst. The East Coast and West Coast shows the best values.

Factor Analysis

Factor analysis was important tool and analyzing correlated data. We performed an exploratory factor analysis for with all 15 variable. It was performed under the maximum likelihood method with varimax rotation with 3 factors. PM2.5 was removed because the eigenvalues for all factors were low and it was not correlated to any of the other variables. This is the breakdown of all of the factors:

- Factor 1 had more emphasis on the educational, economic, and health variables. The only variables that were missing on this factor from that sub-index were the PM2.5 values and the unemployment growth percent. All variables from those sub-index were positive.
- The crime variables with income and unemployment percent were more emphasized in Factor 2.
- Factor 3 shows more emphasis on the poverty rate and the bachelor's degree or higher percent. It has a positive eigenvalue for the poverty rate and a negative eigenvalue for the bachelor's degree or higher percent. According to UCLA Institution for Digital Research and Education, the negative eigenvalue occurs because the common variance is less than the total variance [3]. Factor analysis measures the common variance [3].

Factors 1 and 2 showed a similar SaTScan and Tableau map as the original QoL Index map and their respective variables. Factor 3 showed a different map. It had Texas, Mississippi, and the Northwest as the worst and Tennessee, Nevada, Indiana, Illinois, Wisconsin and Minnesota as the best.

The QoL Index with latent factors showed a similar map as the one shown on the original QoL Index map, but there were some differences between the two maps. For example, the original QoL Index map showed Escambia County, FL for dark red (significantly worse than normal) and the QoL Index from Factor Analysis shows light red (slightly worse than average). The difference between the original QoL Index and the QoL Index with latent factors is that the QoL Index done by latent factors was done by correlation analysis. This placed each variable in a particular factor based on the correlation and more accurately weighs each factor.

6. Software Packages

There were many software packages that were used to analyze the Quality of Life (QoL) index. There were three types of software used for data cleaning (OpenRefine, Microsoft Excel, and Microsoft Access) and three types of software used for data analysis (SAS, SaTScan, and Tableau).

6.1 Data Cleaning Software

Open Refine

According to openrefine.org, OpenRefine was founded in May 2010 to support Freebase Gridworks software in data cleaning, reconciliation, and upload [4]. OpenRefine was originally Google Refine until 2012 [4]. When Google stopped supporting Google Refine, it was rebranded to OpenRefine [4]. OpenRefine was first used to create a master file. It also was used to clean and sort data that was in a different order from the master file. Finally, also assisted with the geographic data file for SaTScan.

Microsoft Excel

Microsoft Excel was founded in 1985 for Macintosh and was introduced to windows in 1987 [5]. According to Haresoftware.com, “It features an intuitive interface and capable calculation and graphing tools which, along with aggressive marketing, have made Excel one of the most popular microcomputer applications to date [6].” Microsoft Excel can analyze large numerical data [7]. Excel was mainly used to store the results received from the SAS program. It also can be changed to a csv file to make it more suitable for analysis for SaTScan.

Microsoft Access

According to brighthub.com, Microsoft Access was founded in 1992 [8]. Microsoft Access is similar to Microsoft Excel, but there are a few differences. First, Microsoft Access is used to store data compared to data analysis of Microsoft Excel [7]. Microsoft Access also works better with text files compared to numerical files for Microsoft Excel [7]. Finally, Microsoft Access makes data easier to find [8].

6.2 Data Analysis Software

SAS

Founded in 1976 at North Carolina State University [9], Statistical Analysis System (SAS) is a statistical program that allows you to do statistics based functions with datasets. According to sas.com, SAS can do many statistical functions such as analysis of variance, Bayesian analysis, categorical data analysis, cluster analysis, exact methods, mixed models, multivariate analysis, nonparametric analysis, psychometric analysis, regression, survey sampling and analysis, survival analysis, statistical graphics and more

[10]. SAS has assisted us with the QoL index by normalizing the data, ranking the data, and factor analysis.

SAS has been used in many ways in the research of the QoL index. First, it allowed us to form the QoL index with the ability to normalize variables and subcomponents (crime, health, economic, educational) using the blom method (proc rank; normal=blom). Second, it allowed us to rank the QoL index and the subcomponents (proc rank; ties=low or ties=mean). Ties=low and ties=mean yielded the same rankings. That was important in determining where Escambia and Santa Rosa counties stand compared to the rest of the country and the state of Florida. Finally, it allow us to perform a rotated factor analysis and correlation analysis of the variables. The factor analysis was used with the maximum likelihood method and rotating with varimax (proc factor; method=ml; rotate=varimax).

SaTScan

SaTScan was founded by Martin Kulldorff and Information Management Systems inc [11]. According to the satscan.org, “SaTScan is a free software that analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics” [11]. This program uses the data provided from a csv or a text file. It can then analyze one or more variables (for example, murder rate) and provides a cluster with the highest and lowest p-value of each variable. It also provides the radius and log-likelihood ratio values of each cluster. The SaTScan program has been helpful in analyzing spatial data in the QoL Index.

SaTScan clusters can also be shown in the form of a map using Google earth. In the SaTScan results section, there is an option that allows you to provide a Google earth map a KML file. From google earth, that map can be copied and saved as a jpeg file.

SaTScan is very useful in the cluster analysis of in three different areas. First, it is useful in analyzing clusters for education, economic, crime, and health subcomponents and variables included in each subcomponent. Second, it helped determine which areas are the best and worst QoL. Finally, it was used for multivariate analysis. This is when all the variables were combined to determine what areas performed the best and the worst in all the variables in the study.

There is one main advantage and disadvantage to using the SaTScan program. According to MGIS Capstone proposal at Penn State, the advantage of SaTScan is that it can provide an accurate circular cluster [12]. The disadvantage of SaTScan is that it not every cluster is circular [12]. Some areas (county, zip code, etc.) inside the circular cluster can be performing far better or worse than other areas inside the cluster. Therefore, a more specific map will be needed to analyze each county one by one.

Tableau

Tableau was founded in 2003 by Chris Stolte, Christian Chabot, and Pat Hanrahan at Stanford University [13]. According to tableausoftware.com, it is a “business intelligence” software that allows you to turn data into many different interactive tables, graphs, and maps [14]. You can use data from a Microsoft Excel file, text file or from a server such as MySQL just to name a few. Tableau was used to provide a map of how each county performed and a line graph comparing counties in and around Pensacola to the USA and the state.

A Tableau map is an improvement from the SaTScan map. It provided a more specific map of which counties were performing well or poorly in the QoL index and the different subcategories. It will provide a map of each county individually and color code of each county based on their performance in the QoL index and all variables and subcomponents. We put both the SaTScan and Tableau map side-by-side to show how a general area is performing and also which specific counties were performing better or worse.

We also did a comparison line graph and bar graph using Tableau. This was used to compare the QoL index and each subcomponent indices in Escambia and Santa Rosa County Florida (Pensacola Metropolitan Area) to the national rate and the state of Florida. It also showed year-by year data of each county in comparison to the national average and the state of Florida.

References

[1] Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. *Multivariate spatial scan statistics for disease surveillance*. Statistics in Medicine, 2007, 26:1824-1833. <http://satscan.us/papers/k-M2005.pdf>

[2] Factor Analysis Using SAS PROC FACTOR. UCLA: Statistical Consulting Group.
(1995, June 26). Retrieved November 27, 2014, from http://www.ats.ucla.edu/stat/sas/library/factor_ut.htm

[3] *SAS Annotated Output Factor Analysis*. UCLA Institute for Digital Research and Education, 2014. Web. 27 Nov 2014. <http://www.ats.ucla.edu/stat/sas/output/factor.htm>

[4] Magdinier, Martin. *OpenRefine History*. OpenRefine, 12 Oct 2013. Web. 25 Nov 2014. <http://openrefine.org/2013/10/12/openrefine-history.html>

[5] Kaul, Ankit. *History of Microsoft Excel 1978 – 2013 [Infographic]*. Excel Trick, 2012-2013. Web. 25 Nov 2014. <http://www.exceltrick.com/others/history-of-excel/>

[6] History of Microsoft Excel. HareSoftware, 1999-2008. Web. 25 Nov 2014. <http://www.haresoftware.com/ExcelHistory.htm>

[7] *More Questions, More Clarity*. Microsoft, 2014. Web. 26 Nov 2014. <http://office.microsoft.com/en-us/training/more-questions-more-clarity-RZ010253998.aspx?section=7>

[8] Hewitt, J, Richter, L. *Microsoft Access: A Walk Through History*. Bright Hub, 29 Oct 2008. Web. 26. Nov 2014. <http://www.brighthub.com/office/collaboration/articles/13190.aspx>

[9] *History*. SAS, n.d. Web. 19 Nov 2014. <http://www.sas.com/en_us/company-information.html#history>

[10] *Features*. SAS, n.d. Web. 19 Nov 2014 <http://www.sas.com/en_us/software/analytics/stat.html#m=features>

[11] *SaTScan, Software for the Spatial, Temporal, and Space-Time Statistics*. SaTScan, 2005. Web. 15 Nov 2014. <http://www.satscan.org>

[12] *Cluster Detection Comparison in Syndromic Surveillance*. Pennsylvania State University, 8 Jul. 2008. Web. 17 Nov. 2014. < https://gis.e-education.psu.edu/sites/default/files/capstone/>

[13] *How to Get a 20 Million Dollar Pre-Money Valuation for Series A: Tableau Software CEO Christian Chabot (Part 3)*. One Million by One Million Blog, 5 Mar. 2010. Web. 20 Nov. 2014. <http://www.sramanamitra.com/2010/03/05/how-to-get-a-20-million-pre-money-valuation-for-series-a-tableau-software-ceo-christian-chabot-part-3/>

[14] *Tableau Business Intelligence*. Tableau, 2003-2014. Web. 19 Nov. 2014. <http://www.tableausoftware.com/business-intelligence>

[15] Hagerty, M., & Land, K., 2007. Constructing Summary Indices of Quality of Life: A Model for the Effect of Heterogeneous Importance Weights. *Sociological Methods & Research*, 455-496.

Appendices

Appendix A - Cluster analysis SaTScan outputs

Appendix B – Factor Analysis SAS and SaTScan Outputs