Multivariate Analysis on Disease Prevalence and Control

Niue

ABSTRACT

In this project of disease prevalence and controls study, it handles with a data set describing the prevalence of 28 kinds of common or dreadful diseases and controls between 100 regions nationwide. Most of the regions are located around metropolitans or east and west coast.

The methodology used in this report lies in the field of Multivariate Analysis, since no response variables are included in this project. Implementation of some representative graphical displays, then carried out Principal Component Analysis for dimension reduction, factor analysis, compared and chose the most appropriate cluster analysis method, canonical study between disease prevalence and controls, and applied Multidimensional Scaling to the data after factor analysis.

The cluster analysis here discussed K-means with 2-means model and 9-means model selected by different standard. With complete linkage as hierarchical method to make order clustering. The similarity between each groups are displayed by shades of colors of points on the map. 2-means clustering has better interpretability, Cities with low healthcare insurance, high disease prevalence, low prevalence of disease control methods are one cluster, while high healthcare coverage, low disease prevalence and high prevalence of disease controls are another cluster. As for 9-means clustering, it is more precise for clustering and examine the similarities between each group.

Factor analysis is carried out before multidimensional scaling, which detects only 3 factors out of 27 measures of disease in this study. Then a 3-D multidimensional scaling plot is constructed, which similarities between each city are more direct to be displayed

Keywords

Multivariate Analysis, Factor Analysis, Principal Component Analysis, Factor Analysis, Multidimensional Scaling, Cluster Analysis, Data Visualization, Canonical Analysis

1. INTRODUCTION

Healthcare and disease prevention has been heated topics, since more and more individuals and organizations came to realize the necessity of saving more lives. It would be beneficial for the public health organizations to have a general understanding of what kind of cities are likely to have high prevalence in each disease, to always protect the citizens from their needs.

This report mainly focus on the process of sensing the similarity and dissimilarity intuitively, then use justified process to divide the cities into several groups with similar characteristics, which could be a guide for public health departments to enhance their understanding of different regions. And, it could be an open source material for the citizens to be aware of the disease that are more likely to occur to them to take actions in advance.

But it is not always the case that the city that has the larger population has the lowest prevalence of every disease among all the others, which is always considered to be an indication of more advanced medical technology. In contrary, with the increase of material life in big city or metropolitans, it is likely that disease like obesity turns out to be more severe than other regions. In this report, we are going to focus on that as well.

The data here is provided by Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion. By removing non-existence value, and get the raw data. It also contains census data like population, geolocation, place FIPS code, and FIPS code with tract. Thus could be utilize into spatial display, to get a clearer understanding of the location and population of the city, and distinguish between the severities of different diseases.

.

2. GRAPHICAL DISPLAYS

The data after preprocessing in this project have 100 rows and 32 columns, with 27 measures of common and dreadful diseases and controls, and 4 spatial measures. All 27 measures have the numeric range between 0 and 100, which stand for the probability of how likely the city has the prevalence of one specific disease or prevention calculated by Centers for Disease Control and Prevention, they are all variables of evaluating the health environment of a specific region. It is meaningfully reported as the number of cases as a fraction of the total population at risk at that specific regions[1].

Table1. List of Dataset

Column Name	Description
StateAbbr	state abbreviation
placename	city name
placeFIPS	city FIPS code
tractFIPS	tract FIPS code
Place_TractID	combined city and census tract FIPS code
ACCESS2 (X1)	lack of health insurance for adults aged 18-64, 2016
ARTHRITIS (X2)	prevalence of arthritis for adults >=18
BINGE (X3)	prevalence of binge drinking for adults >=18
BPHIGH (X4)	prevalence of high blood pressure for adults >=18
BPMED (X5)	prevalence of taking medicine of blood pressure for adults >=18
CANCER (X6)	prevalence of cancer for adults >=18
CASTHMA (X7)	prevalence of current asthma for adults >=18
CHD (X8)	prevalence of coronary heart disease for adults >=18
CHECKUP (X9)	prevalence of routine checkup for adults >=18
CHOLSCREEN (X10)	prevalence of cholesterol screening for adults >=18
COLON_SCREEN (X11)	prevalence of fecal occult for adults >=18
COPD (X12)	prevalence of chronic obstructive pulmonary for adults >=18
COREM (X13)	prevalence of older adult on a core set of clinical preventive services like Flue shot, PPV shot, cancer scanning

COREW (X14)	prevalence of older adult on a core set of clinical preventive services like Mammogram
CSMOKING (X15)	prevalence of current smoking for adults >=18
DENTAL (X16)	prevalence of visits to dentists for adults >=18
DIABETES (X17)	prevalence of diagnosed diabetes for adults >=18
HIGHCHOL (X18)	prevalence of high cholesterol for adults >=18
KIDNEY (X19)	prevalence of kidney disease for adults >=18
LPA (X20)	prevalence of no leisure time physical activity for adults >=18
MAMMOUSE (X21)	prevalence of mammography for woman aged 50-74 years
MHLTH (X22)	prevalence of mental health not good for adults >=18
OBESITY (X23)	prevalence of obesity for adults >=18
PAPTEST (X24)	prevalence of papanicolaou smear for adults >=18
PHLTH (X25)	prevalence of physical health not good for adults >=18
SLEEP (X26)	prevalence of less sleep for adults >=18
STROKE (X27)	prevalence of stroke for adults >=18
TEETHLOST (X28)	prevalence of all teeth lost for adults >=18
Geolocation	lattitude, longitude of census tract centroid

In the latter part of canonical analysis, the variables above could be divided into two subgroups, the first group contains 10 variables, ACCESS2, BPMED, CHOLSCREEN, COLON_SCREEN, CHECKUP, COREM, COREW, DENTAL, LPA, MAMMOUSE which depicts the prevention approaches of disease control, and the others is another subgroup of common and dreadful diseases prevalence.

The scatterplot matrix below shows the first eight variables, from which we could sense that population have no strong correlation with other variables, and is not related to any disease control and prevalence for the other 7 variables below. Thus for the latter analysis, I decide to discard the data of population for most circumstances. It is only introduced after the Cluster Analysis, as a visualization technique to show how much the city's population are. To provide a direct image of how large the city is.

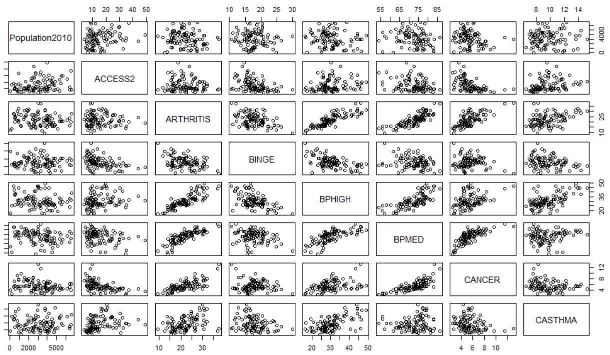


Fig 1. Scatterplot Matrix for First 10 Measures of Diseases

For some of the diseases, the prevalence is highly correlated with other diseases. For instance, the prevalence of high blood pressure (BPHIGH) is highly positively related to taking medicine of blood pressure (BPMED), and arthritis and negatively related to binge drinking. But for some of the other diseases, like current asthma, it is more independent to other diseases, and thus the scatterplot shows no strong trend between other variables.

For better display, I consider using the correlation heatmap in Fig 2, maps the rounded correlations (1-digit) into different shades of colors. In this report, darker red stands for high positive correlation, and deeper blue stands for high negative correlation.

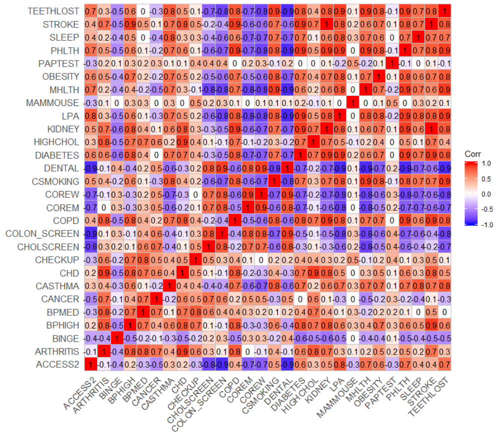


Fig 2. Heatmap / Correlation Matrix of Disease

From the plot, it is more clear to extract basic understanding of relationships between these variables prevalence. For instance, for a city with high frequency of visits to dentists (DENTAL), it should be always true that it has great health insurance for a wide range (low ACCESS2 value). The city of Columbus has the highest value of visits to dentists, while ACCESS2 is only 3.9, which means that health insurance is very prevalent. Some of the diseases are occur along the same time for patients, thus could lead to high correlation under this circumstances. For later research, some data reduction methods are carried out to solve this problem.

In order to detect some outliers, as well as describe the distribution of data. The below is an example of using boxplot for the first two components after Principal Component Analysis. The inner solid ellipse contains about half of the points. And there's only on outlier, which is Los Angeles, located just on the edge of outside circle.

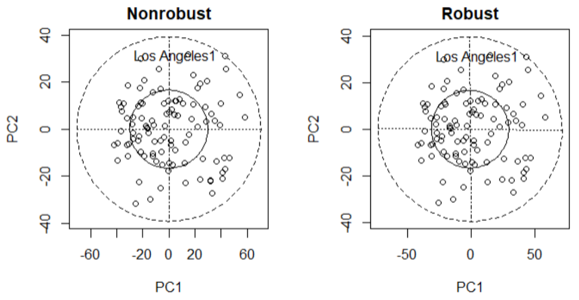


Fig 3. Nonrobust and Robust Bivariate Boxplots

The Starplots below matches with the numeric values of each variable, and draws a shape for each sample. The only concern here is that Starplots could be hard to focus on when the number of variables increase. Thus I plot the starplots of only five dimensions, ACCESS2, ARTHRITIS, BINGE, BPHIGH, and BPMED. The first 10 cities diverse a lot in healthcare environment, while Milwaukee is a city with severe prevalence of these problems, for San Francisco, it only has minor binge drinking (BINGE) and taking medicines of blood pressure (BPMED). The larger the area is, the more problems it needs to be refined in fields of disease control.

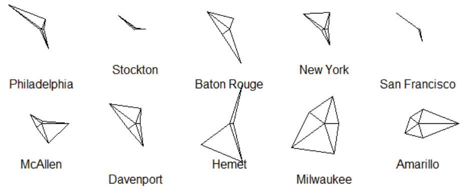


Fig 3. Star plots for first Ten Cities

There is also another way of drawing possible plots for each city. The Radar plot here provides a better illustration.

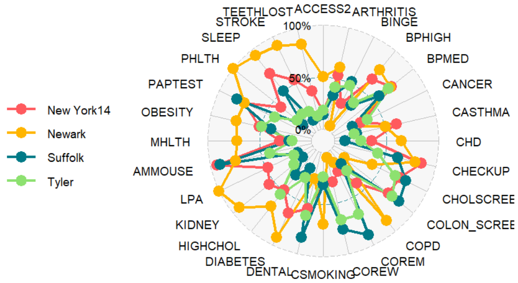


Fig 4. Radar plot for last four Cities

Since the Radar plot here uses more dimensions in a well-performed manner, it also draws the polylines for each city. The city of Newark here has the most severe healthcare environment, most of the items take up the highest value among the four. Both of the Star plots and Radar plots are useful tools for finding similar patterns of each sample. It is a guide for making cluster analysis, where similar shapes of individuals might have been from a group with similar characteristics.

For each regions of the dataset, their geolocation could be extract online. Thus, the subsequent spatial data visualizations are carried out in order to identify whether similar healthcare conditions cities are aggregated on map^2.

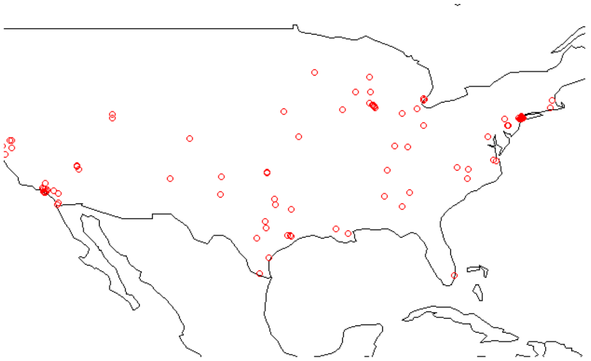


Fig 5. Geolocation of 100 regions

The graph above displays the location of 100 regions and cities used in this project. The dense patterns of points occur at the west and east coast, while with other points dispersed uniformly in the Midwest and the Northeast.

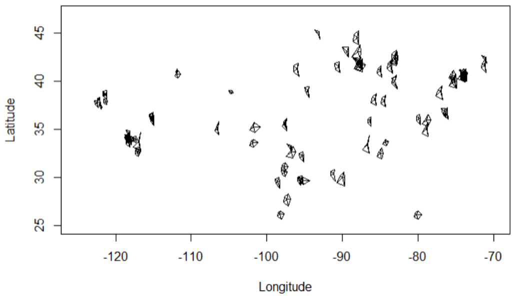


Fig 6. Star Plots Arranged by Relative Geolocation

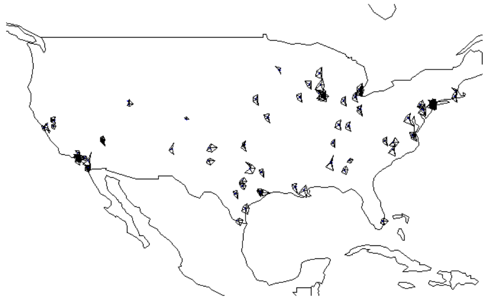


Fig 7. Star Plots Arranged by Precise Geolocation on Map

Figure 5 and six are Star plots of 100 regions taking consideration of geolocation. Some of the patterns show amazing similarity from the outer shape. The stars in the Midwestern area are very similar to each other. While the shape of the stars implies high healthcare insurance coverage (low ACCESS2), and relatively low prevalence of high blood pressure (BPHIGH). With the increase of latitude, the general health condition is more pleasant, since the total area of the stars decrease, while the prevalence of binge drinking (BINGE) increases. It is likely to happen, since more wine and beer factories are located in the Midwestern area, and are consumed more often than the southern region.

Bubble plots are more preferable to examine the size and degree of each variables. The below graphs are bubble plots for the prevalence of cancer and obesity. This time, I plotted the size of the bubbles with the population data, and shade of color as prevalence of disease. The deeper the color, the more popular it is.

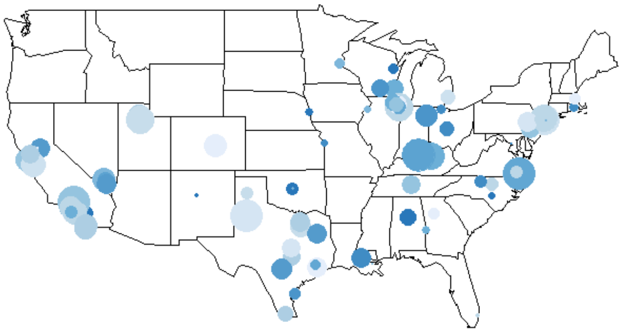


Fig 8. Bubble Plot for Prevalence of Cancer

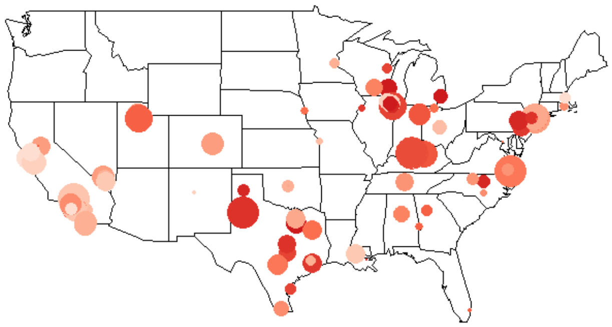


Fig 9. Bubble plot for Prevalence of Obesity

Figure 8 and figure 9 depicts two different shades of disease. The ones having more prevalence of cancer turns out to be less prevalent in obesity. In a lot of regions, it acts as if it is on the reverse side. For Cancer, it middle and smaller range of cities are more likely to suffer from it. For obesity, the west coast has obviously less prevalence of obesity, while the Southern and Midwestern are more likely to suffer from obesity. This is also likely to happen, taking considerations of the weather conditions are so different among them. While it is more preventing for people in Midwestern and Southern regions to work out when climate is tough.

In the later section, I am focusing on how to divide these cities into different groups by their different prevalence, and to dig out the most influential factors and hidden relationships.

3. GOALS

3.1 Describe the dataset with principal component analysis

Since there are many variables in my dataset, I will then use Principal Component Analysis to reduce the dimension of my dataset and find the principal components to better explain the data. I proposed some of the questions to be answered during the process.

- How many principal components should be used to describe the dataset? What are those principal components?
- How does each principal component describe the information in the dataset?
- How do the scores of all principal components for all the cities and regions distributed?

Principal components provide a concise way of describing data by providing components that minimize the variance. With a more succinct summary of the dataset, it will be easier to compare the measurements on the disease prevalence and control across different samples.

3.2 Group the cities into clusters according to measures of disease prevalence and control

In the second step I would like to group the cities by clustering according to their measures on disease prevalence and control. I hope to answer the following questions:

- How are cities clustered in each method based on the measurements in their standard of living?
- What is the difference of clustering by using diversified methods?
- Which is the clustering method that I maintain?
- How are different clusters related to each other?

Through answering these questions, I would like to investigate which cities are similar to each other or which cities are different from each other. This would finally picture on the measures of disease prevalence and control around the world.

3.3 Find out deciding factors out of 27 measures

Very similar to PCA, but factor analysis strives to explain correlations among multiple outcomes as the result of several factors. It also involves data reduction, and represent the set of variables by a smaller number. Thus, I made several questions:

- How many factors should be used to describe the dataset?
- What are those factors?
- What's the point of deciding the factors?

In order to answer those questions, I would choose the multiple factors from the dataset, and then use the dimensional for later study of Multidimensional scaling.

3.4 Provide visual representation of proximities among objects

Since in this problem, I am also interested in the distances between various cities, while comparing it to the actual relative distance on map to see if there are any similarities between them, the questions that I made are:

- What's the dimension of Multidimensional Scaling?
- What is the method chosen?
- How similar are the MDS plot with plot of geolocation?

The dimensions of multidimensional scaling could come from ones after factor analysis. In this question, I visualize the distribution of cities after MDS plot, in comparison with geolocation plot.

3.5 Find possible relationships between common disease prevalence and control approaches

The dataset could be divided into two parts: one with only 10 disease control variables, such as regular examinations, screening, denoted by X. The other are 18 variables of common and dreadful disease prevalence. I want to figure out the potential relationships between these two groups of data. The questions going to be answered are:

- What is the correlation between two groups of data?
- What is each groups' canonical variates from two groups of data?
- What is the trend of variation between two canonical variates and how are they related to each other?

By answering these questions, one is able to describe the association between those disease control methods and other 18 diseases through the canonical variates that best explain the variability both within and between sets.

4. MAIN RESULT

4.1 Principal Component Analysis

In consideration of keeping the original percentage of each vriables, I chose not to scale the data when carrying out Principal Component Analysis.

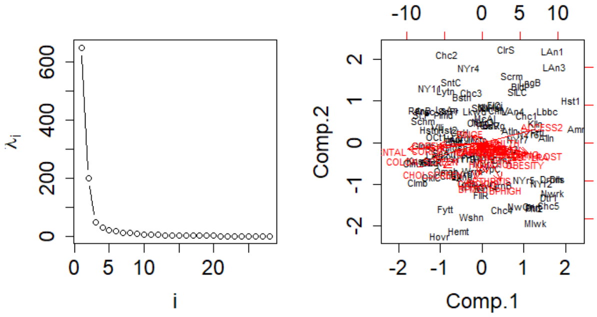
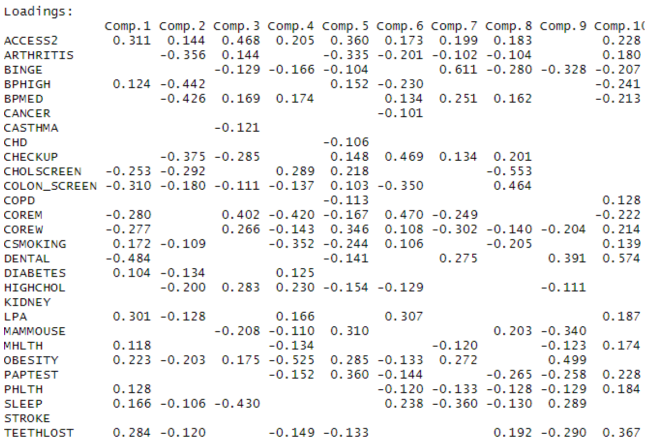


Fig 10. Scree plot and Biplot of PCA



```
Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
ACCESS2       0.311  0.144  0.468  0.205  0.360  0.173  0.199  0.183         0.228
ARTHRITIS           -0.356  0.144        -0.335 -0.201 -0.102 -0.104         0.180
BINGE                      -0.129 -0.166 -0.104         0.611 -0.280 -0.328 -0.207
BPHIGH        0.124 -0.442                0.152 -0.230                      -0.241
BPMED               -0.426  0.169  0.174         0.134  0.251  0.162        -0.213
CANCER                                          -0.101
CASTHMA                    -0.121
CHD                                      -0.106
CHECKUP             -0.375 -0.285         0.148  0.469  0.134  0.201
CHOLSCREEN   -0.253 -0.292         0.289  0.218               -0.553
COLON_SCREEN -0.310 -0.180 -0.111 -0.137  0.103 -0.350         0.464
COPD                                     -0.113                              0.128
COREM        -0.280         0.402 -0.420 -0.167  0.470 -0.249               -0.222
COREW        -0.277         0.266 -0.143  0.346  0.108 -0.302 -0.140 -0.204  0.214
CSMOKING      0.172 -0.109        -0.352 -0.244  0.106        -0.205         0.139
DENTAL       -0.484                      -0.141         0.275         0.391  0.574
DIABETES      0.104 -0.134         0.125
HIGHCHOL            -0.200  0.283  0.230 -0.154 -0.129               -0.111
KIDNEY
LPA           0.301 -0.128         0.166         0.307                       0.187
MAMMOUSE                   -0.208 -0.110  0.310                0.203 -0.340
MHLTH         0.118               -0.134               -0.120        -0.123  0.174
OBESITY       0.223 -0.203  0.175 -0.525  0.285 -0.133  0.272         0.499
PAPTEST                           -0.152  0.360 -0.144        -0.265 -0.258  0.228
PHLTH         0.128                             -0.120 -0.133 -0.128 -0.129  0.184
SLEEP         0.166 -0.106 -0.430                0.238 -0.360 -0.130  0.289
STROKE
TEETHLOST     0.284 -0.120        -0.149 -0.133                0.192 -0.290  0.367
```

Table 2. Parts of PCA Loadings

The above Scree plot indicates that choosing the 3 components are just the case for this problem. From the loadings of Principal Component Analysis, the first component accounts for 0.636 of the variations in the dataset and it summarizes the information of health insurance, binge drinking, cholesterol screening, core set of clinical preventive services, current smoking, visits to dentists, diabetes, no leisure time for physical activity, mental health not good, obesity, physical health not good, less sleep and total teeth loss. While the coefficient of visit to dentists are slightly greater than other variables, indicating that this variable may have larger variance.

From the biplot above, it shows that the cities on the right upper conor has the one of the highest amount of lack of health insurance, as well as less disease prevalence, like some parts of Los Angeles, with low prevalence in diseases (cancer, obesity and etc). While the left lower corner implies low lack of health care insurance, and relatively high prevalence of diseases and control methods, like the city of Hoover from the samples.

The second principal components have 0.196 of variance, which are linear combination of positive lack of health insurance with negative influence of mainly high blood temperature, cancer and routine checkup. The third components have only 0.05 of variance, which are mainly combination of positive impact of lack of health insurance, core set of clinical preventive services, and negative lack of sleep.

Three components takes up to 0.878 of total variance, which implies that three components could explain the dataset. And the three principle components are:

$$\begin{aligned} PC_1 = 0.311X_1 + 0.124X_3 - 0.253X_{10} - 0.310X_{11} - 0.280X_{13} \\ - 0.277X_{14} + 0.172X_{15} - 0.484X_{16} \\ + 0.104X_{17} + 0.301X_{20} + 0.118X_{22} \\ + 0.223X_{23} + 0.128X_{25} + 0.166X_{26} \\ + 0.284X_{28} \end{aligned}$$

$$\begin{aligned} \mathrm{PC}_2 = 0.144\mathrm{X}_1 - 0.356\mathrm{X}_2 - 0.442\mathrm{X}_4 - 0.426\mathrm{X}_5 - 0.375\mathrm{X}_9 \\ - 0.292\mathrm{X}_{10} - 0.180\mathrm{X}_{11} - 0.109\mathrm{X}_{15} \\ - 0.134\mathrm{X}_{17} - 0.200\mathrm{X}_{18} - 0.128\mathrm{X}_{20} \\ - 0.203\mathrm{X}_{23} - 0.106X_{26} - 0.120X_{28} \end{aligned}$$

$$\begin{aligned} \mathrm{PC}_3 = 0.468\mathrm{X}_1 + 0.144\mathrm{X}_2 - 0.129\mathrm{X}_3 + 0.169\mathrm{X}_5 - 0.121\mathrm{X}_7 \\ - 0.283\mathrm{X}_9 - 0.111\mathrm{X}_{11} + 0.402\mathrm{X}_{13} \\ + 0.266\mathrm{X}_{14} + 0.283\mathrm{X}_{18} - 0.208\mathrm{X}_{21} \\ + 0.175\mathrm{X}_{23} - 0.430\mathrm{X}_{26} \end{aligned}$$

Overall, these three principle components showed that most of the variations in the dataset could be accounted as the lack of health insurance, visits to dentists, high blood pressure, cancer, routine checkup, clinical preventive services and lack of sleep. The first component focus on the overall poor health insurance coverage with less disease control approaches. The second component provides more on some daily controls, like regular checkup, blood pressure medicines. The third component add details into more disease controls for elderly and some common diseases.

From the plots with scores on the first component below, we can see that the cities and regions that has lowest visit to dentists, but higher cholesterol screening, no time for physical activity, all teeth loss shows PC1 that is obviously higher than others, the city of Houston, who has one of the highest frequency to dentists, shows a great healthcare conditions among all cities.

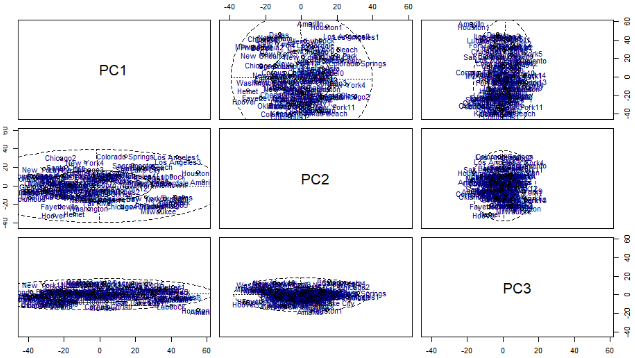


Fig 10. Pairwise plot of Component Scores

Many cities with high PC1 also shows high PC2, like the city of Los Angeles, for it has very less prevalence of different disease, and high prevalence of checkout, and set of examination services.

4.2 Cluster Analysis

After gaining an understanding on the different levels of measures of disease prevalence and controls based on Principal Component Analysis, I would like to investigate how cities could be clustered based on the measures.

4.2.1 K-means

The first clustering method that I applied here is by K-means, and I used the within group sum of squares to draw the scree plot to decide the value of k. From the scree plot, the elbow point occurs at k=2, but it seems to be a very subjective decision, then I decided to apply the Average Silhouette Method[3] as in Fig. 11. A high average silhouette width indicates a good clustering. Where the optimal number of clusters k is one that maximizes the average silhouette over a range of possible values for k. Fortunately, two choices of k all implies a 2-means clustering, which the cluster plot shows very separate and equal-sized patterns.

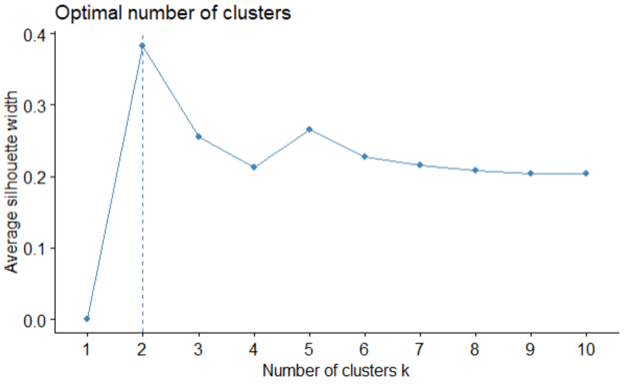


Fig 11. Pairwise plot of Component Scores

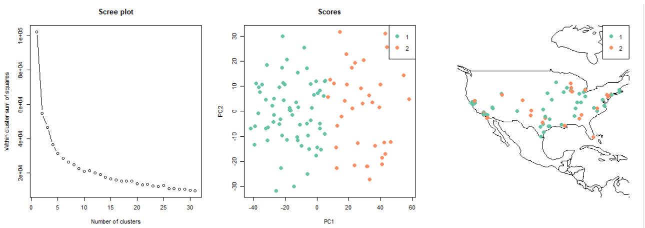


Fig 12. 2-means clustering

Mapping back the 2-clustered data onto the map, we could sense some spatial understanding that the second categories are more likely to be located along the coast, and near the Five Great Lakes.

Table 3. Centers for 2-Means Clustering

	ACCESS2	ARTHRITIS	BINGE	BPHIGH	BPMED	CANCER	CASTHMA	CHD
1	10.67813	21.50625	18.90781	28.49687	73.53750	6.045313	8.901562	5.306250
2	24.76389	22.70556	15.82778	33.82778	70.74167	4.455556	11.088889	6.291667

	CHECKUP	CHOLSCREEN	COLON_SCREEN	COPD	COREM	COREW	CSMOKING
1	70.81562	77.36875	67.20312	5.234375	35.54219	34.11406	14.52500
2	69.41389	65.88333	53.03333	7.544444	22.50278	21.98611	22.18056

	DENTAL	DIABETES	HIGHCHOL	KIDNEY	LPA	MAMMOUSE	MHLTH	OBESITY
1	70.03906	9.059375	33.82031	2.532813	20.43906	79.48594	10.85937	26.27656
2	47.83333	13.672222	35.54722	3.616667	33.48611	79.23889	16.45000	36.30556

	PAPTEST	PHLTH	SLEEP	STROKE	TEETHLOST
1	83.80625	10.31562	34.54219	2.559375	10.65000
2	83.14167	16.28333	41.84444	4.147222	23.75556

The above table shows the centers of two clusters, according to each variables' mean value, we could notice that cluster 1 are cities with better healthcare conditions, which has the lower lack of health insurance, lower prevalence of most diseases, other than binge drinking and cancer. With obviously higher prevalence of disease controls, higher daily checkup, medicines for blood pressure, screening, dental care core preventing services. And cluster 2 contains cities that have relatively more prevalence in diseases and less controls method carried out. However, there are also cities in cluster 1 that tend to have very high lack of healthcare insurance, which is not appropriate to be placed in the first cluster, like in the city of Colorado Springs, though it has relative high level of disease prevalence, but it also has no lack of health insurance, with ACCESS2 value only equals to 19.7, which implies that there is not a lack of health insurance, and thus I considered changing the value of k for better clusters.

Table 4. Categories for first 45 cities

```
 Philadelphia        Stockton      Baton Rouge         New York    San Francisco
            2               2                2                2                2
      McAllen       Davenport            Hemet        Milwaukee         Amarillo
            2               2                2                1                1
    Las Vegas     Los Angeles   Citrus Heights         St. Paul          Chicago
            1               2                2                2                2
   Fall River        Palmdale          Lubbock       Las Vegas1         Chicago1
            2               2                1                2                1
     Chicago2          Layton       Chesapeake      Santa Clara         Glendale
            2               2                2                2                2
    Riverside       Henderson        New York1        New York2        New York3
            2               2                2                1                2
    New York4    Fayetteville     Baldwin Park     Los Angeles1       Fort Wayne
            2               2                1                1                2
        Omaha       New York5        New York6        New York7          Houston
            2               1                1                1                2
     Columbus       New York8        Columbus1           Toledo       Washington
            2               2                2                2                2
```

Fig 12. is a graph of two clusters, the red cluster contains the cities of better health conditions with high healthcare insurance coverage; while the blue cluster are the cities with poorer disease controls and more diseases prevalence.

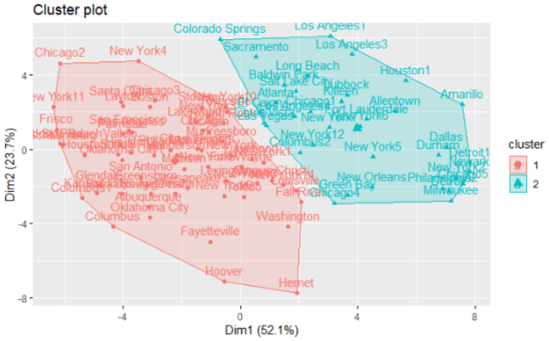


Fig 13. 2-means clustering plot of cities

Reconsidering the number of clusters chosen in this study, trying a couple of more numbers is a sensible way. In the research paper published in 2001[4], the gap statistic has been proposed, based on using Monte Carlo simulations and calculate the intracluster variation for thousand of times, and choose the number k that maximized the gap statistics. Where gap statistic is defined as below:

$$\text{ap}\ \ _n(k) = E_n^* \log(W_k) - \log(W_k)$$

From Fig 13. When the number of k is 9, the gap statistic is maximized. Then I'll use the 9 clusters for later study.

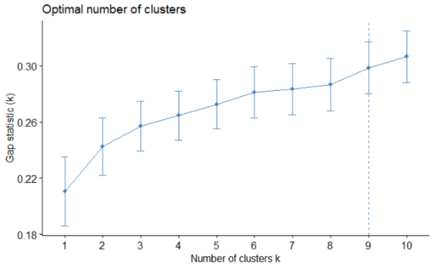


Fig 14. Gap Statistic

4.2.2 Agglomerative clustering

By using hierarchical model in clustering, there are three intercluster dissimilarity measures, which are single linkage, complete linkage, and average linkage. Which could be applied to different circumstances. Single linkage sets the nearest distance between two points as measure, is likely to enhance the chaining effect, and form the clusters of irregular, often thread-like curved shapes. Since in this study, we want the clusters to be more like a shape of ellipsoid rather than curve, single linkage might not be a good choice. While for complete linkage, the similarity of two clusters is the similarity of their most dissimilar members.

From the dendrogram plots in Fig 14. below, the cities are clustered with complete linkage, single linkage and average linkage. Applying the standard of clustering at the nine clusters for each clustering techniques from the previous decisions. One notable result is that, single linkage gives clusters of very unequal size, with one or two regions for some clusters, and is primarily affected by outliers. While for complete linkage, every cluster has relatively similar number of sizes. The result of complete linkage also matches the intuitive that cities like Colorado Springs and Salt Lake City should be clustered into one group, for they have relatively low prevalence of lack of health insurance, while high disease prevalence. In contrast with the average linkage that they belongs to different groups of data, with only three data points in the cluster the same as Colorado Springs. Due to these considerations, it seems like that complete-linkage clustering gives a more favorable result among the three hierarchical clustering methods. Using this method, the clustering of 100 regions and cities is plotted on the nation map.

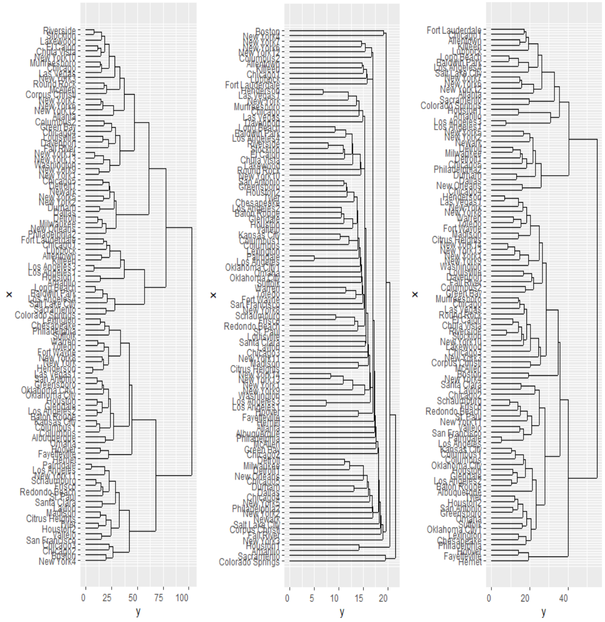


Fig 14. From left to right: complete, single, average linkage

From the graph below, the single linkage plot shows a huge pattern in the middle, while other patterns are scattered around the margins. The size of other clusters are so small to be really bad for clustering, with one or two at the margins. Average linkage performs rather better than single linkage, but the overlaps between different clusters are very severe, and one cluster has only three data points, which also is counterintuitive with the observation. Thus, only the complete linkage displays the nine clusters separately and of equal size between them.

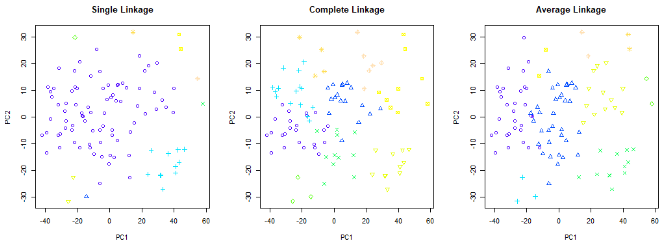
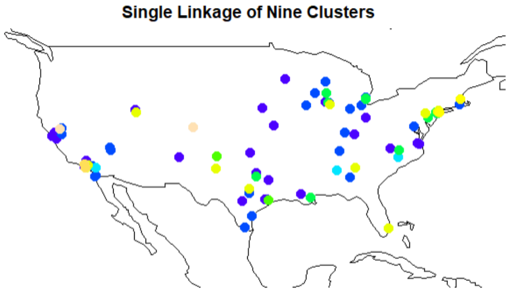
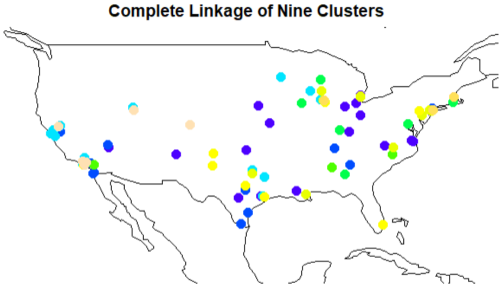


Fig 15. Agglomerate clustering displayed by PC1 and PC2

Then, I maps the data back to the geolocation on the nation map, adding the population of each regions as the size of the circle, here I only focus on the complete linkage from Fig 15. Reason for the colors are that, the similar the color is, the closer two cities are from the dendrogram. The cluster colored yellow and orange that located mostly near the coast and along the lake shore, has the characteristic that are very similar. For instance, Florida has, which is similar to Philadelphia and Killeen, with middle-ranged prevalence of lack of health insurance, high blood pressure and low in prevalence for cancer, heart disease and relatively high prevalence of obesity. The cluster colored in green is mostly located on the east side of states, and most of them are of middle size in population. These cities have the characteristics that has low prevalence of health insurance, but relatively high disease controls, like checkouts and cholesterol scanning, like the city of Hoover. Interestingly, brown color cluster only happens on west side of the states, like Santa Clara, which has low prevalence of lack of health insurance, low heart disease, and high disease controls. These categories distinct with each other a lot. And gives insight into making some conclusion that the coastal area shouldn't worry about the prevalence of health insurance that much, since the prevalence of diseases are still within satisfying range, and disease controls are of high prevalence. But for states in the middle part of the states should not only implement on the healthcare insurance, maybe should look at other factors that lead to the prevalence of disease, and arise citizen's awareness of taking regular medical examinations.


Single Linkage of Nine Clusters


Complete Linkage of Nine Clusters

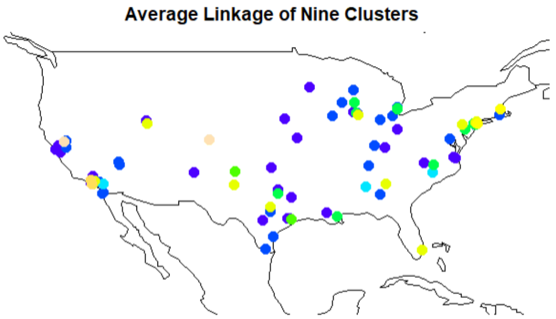


Fig 15. From top to bottom: 9 clusters on nation map

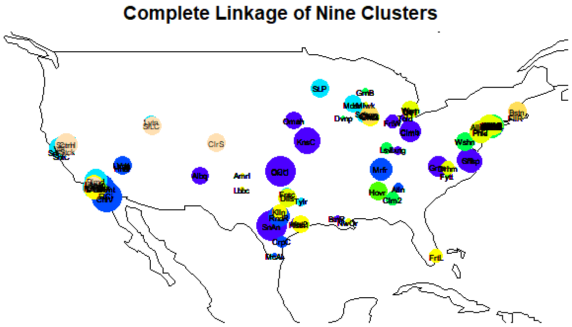


Fig 16. Complete-Linkage clustering on nation map

4.3 Factor Analysis

Since there are so many variables in my dataset, by using factor analysis, I could examine how much multiple factors are useful in this study. And the crucial decision in exploratory factor analysis is how many factors to extract. The nFactors package offer a suite of functions to aid in this decision. It is not an important part, which is only used for decide the dimension for Multidimensional Scaling.

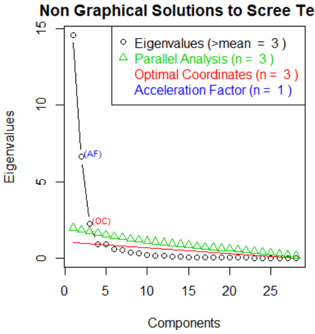


Fig 17. Deciding number of factors

4.4 Multidimensional Scaling

After carrying out the factor analysis from the previous sections, in order to sense the potential similarity from the factors, I used the dimension of 3, as is previously calculated through the factor analysis. One of the main tasks the analyst has is determining the number of dimensions in the MDS model. Each dimension represents a different underlying factor. One of the goals of the MDS analysis is to keep the number of dimensions as small as possible. Thus, 3 is just appropriate for this study.

By using the distance matrix of the dataset, the multidimensional scaling plots of classic and nonmetric displays the same when dimension is chosen to be 3. Firstly, use the 3 dimensional multidimensional scaling, then color them with 2 clusters and 9 clusters separately. From the 2-dimensional plot, the shape of clusters are well-preserved, which also depicts how similarity are between these cities. Though the real distance between New York and Chicago is so far, and they belongs to different regions in the states, some parts of New York and Chicago are very close to each other on the MDS plot, since they are with similar characteristics like a higher lack of healthcare insurance, high disease controls like screening and medicines for blood pressure, while still having suffered from the prevalence of relative diseases, like heart disease, smoking and etc.

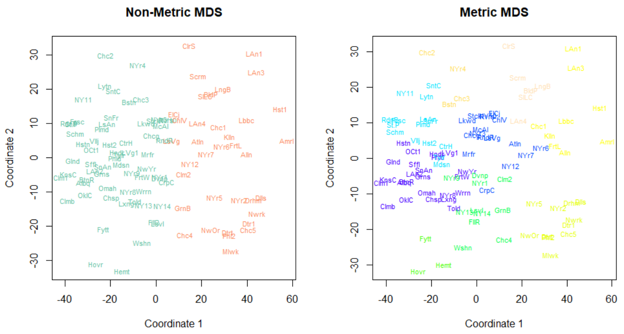


Fig 18. Non-Metric and Classic MDS plot for 2 and 9 clusters

Whether there is any similarity of first two MDS coordinates and geolocation are made by drawing two plots to decide. One with latitude as x-axis and longitude as y-axis, the others use first two MDS coordinates. While these two plots could not overlaps no matter how to rotate, which means that distance on the geological map is not correlated with the relationships through the disease prevalence and controls data.

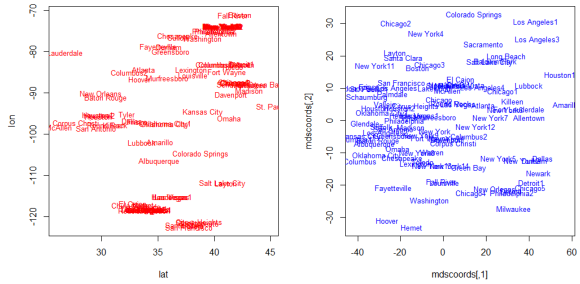


Fig 19. Comparisons with geolocation

The 3-D Multidimensional scaling cubic shows a more comprehensive way of how these samples are related to each other. By rotating the cubic, one could find out the relationship without losing information when projecting on a 2-D plane.

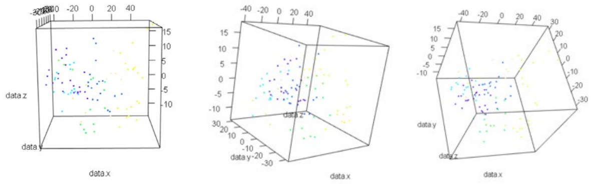
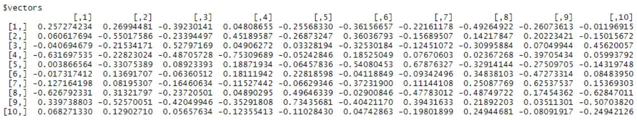


Fig 20. 3-D Multidimensional scaling

4.5 Canonical Analysis

To further study the relationship between the prevalence of disease and disease control examinations and approaches, I divided the data into the two subgroups in the previous steps. The first subset with disease control methods is denoted as X, while the other is Y. And then, calculate the canonical correlations and find variables. Since the data of each variable are all from the same unit, which are all percentages. Thus, I decided not to take the scale, and keep the original data to protect its representative.

Table X. Coefficients of series of U canonical variables



```
$vectors
              [,1]        [,2]        [,3]        [,4]        [,5]        [,6]        [,7]        [,8]        [,9]       [,10]
 [1,]  0.257274234  0.26994481 -0.39230141  0.04808655 -0.25568330 -0.36156657 -0.22161178 -0.49264922 -0.26073613 -0.01196915
 [2,]  0.060617694 -0.55017586 -0.23394497  0.45189587 -0.26873247  0.36036793 -0.15689507  0.14217847  0.20223421 -0.15015672
 [3,] -0.040694679 -0.21534171  0.52797169  0.04906272  0.03328194 -0.32530184 -0.12451072 -0.30995884  0.07049944  0.45620057
 [4,] -0.631697535 -0.22823024 -0.48705728 -0.75309689 -0.05242846  0.18525049  0.07670603  0.02367268 -0.39705434  0.05993792
 [5,]  0.003866564 -0.33075389  0.08923393  0.18871934 -0.06457836 -0.54080453  0.67876327 -0.32914144 -0.27509705 -0.14319748
 [6,] -0.017317412  0.13691707 -0.06360512  0.18111942  0.22818598 -0.04118849 -0.09342496  0.34838103 -0.47273314  0.08483950
 [7,] -0.127164198  0.08195307 -0.16460634 -0.11527442 -0.06629346 -0.37231900  0.11144108  0.25087769  0.62537537  0.15369303
 [8,] -0.626792331  0.31321797 -0.23720501  0.04890295  0.49646339 -0.02900846 -0.47783012 -0.48749722  0.17454362 -0.62847011
 [9,]  0.339738803 -0.52570051 -0.42049946 -0.35291808  0.73435681 -0.40421170  0.39431633  0.21892203  0.03511301 -0.50703820
[10,]  0.068271330  0.12902710  0.05657634 -0.12355413 -0.11028430  0.04742863 -0.19801899  0.24944681 -0.08091917 -0.24942126
```

The canonical correlations between, being the square roots of the eigenvalues, which shows strong correlations here, are 0.9927, 0.9832, 0.9509, 0.8722, 0.7990, 0.7000, 0.6599, 0.5589, 0.3441 and 0.3283.

The canonical variables are: (use two decimals and first two canonical variates as example)

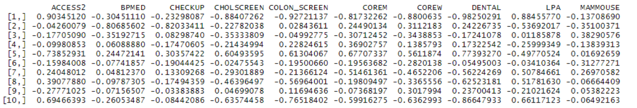
$$\begin{aligned} U1 = 0.26X1 + 0.06X2 - 0.04X3 - 0.63X4 - 0.02X6 - 0.13X7 \\ - 0.63X8 + 0.34X9 + 0.07X10 \end{aligned}$$

$$\begin{aligned} V1 = 0.22Y1 - 0.09Y3 + 0.27Y4 - 0.02Y5 - 0.37Y6 + 0.37Y7 \\ - 0.08Y8 + 0.20Y9 - 0.06Y10 - 0.52Y11 \\ - 0.30Y12 - 0.16Y13 + 0.08Y14 \\ - 0.07Y16 + 0.38Y17 - 0.18Y18 \end{aligned}$$

U1 is mainly a contrast between cholesterol scanning, visits to dentists and lack of health insurance. And could be best described as very severe lack of health insurance, low frequency of cholesterol scanning and visits to dentists. Which is commensurate to our intuitive understanding that lack of health insurance is usually a sign of lack of bad disease controls and preventions, in drastic contrast with receiving regular examinations like screening and dental care. With all these three variable combined together, the disease control could be represented.

Variable V1 is mainly a contrast between arthritis prevalence, chronic obstructive pulmonary, stroke and coronary heart disease, obesity, mental health not good. For V1, distinguishing between each diseases are more subtle, but it could be more intuitively described as contrast between diseases are likely to be caused by manual works, taking in cheap and unhealthy foods (obesity and heart disease for example). And ones that could be possibly caused by less exercise, taking in seafood and alcohol too much and often (stroke for instance).

Table X. Correlations between Canonical and Original Variables (U and X)



```
          ACCESS2       BPMED     CHECKUP  CHOLSCREEN COLON_SCREEN       COREM      COREW      DENTAL         LPA    MAMMOUSE
 [1,]  0.90345120 -0.30451110 -0.23298087 -0.88407262  -0.92721137 -0.81732262 -0.8800635 -0.98250291  0.88455770 -0.13708690
 [2,] -0.04260079 -0.80685602 -0.82033411 -0.22782038   0.02843611  0.24490134  0.3112183  0.24226735 -0.53692017 -0.35100371
 [3,] -0.17705090 -0.35192715  0.08298740 -0.35333809  -0.04992775 -0.30712452 -0.3438853 -0.17241078  0.01185878  0.38290576
 [4,] -0.09980853  0.06088880 -0.17470605 -0.21434994   0.22824615  0.36902757  0.1385793  0.17322542 -0.25999349 -0.13839313
 [5,] -0.73852931  0.24472141  0.30357422  0.60493595   0.61304067  0.67707337  0.5611874  0.77393270 -0.49770524  0.01692659
 [6,] -0.15984008 -0.07741857 -0.19044425 -0.02475543  -0.19500660 -0.19563682 -0.2820138 -0.05495003 -0.03410364 -0.31277271
 [7,]  0.24048012  0.04812370  0.13309268 -0.29301889  -0.21366124 -0.51461361 -0.4652206 -0.56224269  0.50784661  0.26970582
 [8,]  0.39077880 -0.09787305 -0.17494359 -0.46396497  -0.56964001 -0.19809497 -0.3365556 -0.62523181  0.51781630 -0.06664409
 [9,] -0.27771025 -0.07156507 -0.03383883  0.04699078   0.11694636 -0.07368197  0.3017994  0.23700413 -0.21021624  0.05382223
[10,]  0.69466393 -0.26053487 -0.08442086 -0.63574458  -0.76518402 -0.59916275 -0.6362993 -0.86647933  0.66117123 -0.06492163
```

From the table above, for U1, there are great differences, which seem to be best and correspond to what has seen from the data. For example, city of Houston has the highest lack of health insurance problem, low cholesterol scanning, while the chronic obstructive pulmonary (COPD) and stroke is of 7 and 4.3, which are one of the smallest prevalence among all of 100 regions. But it is not always precise, since some of the variables are highly correlated, and thus might cause problems in evaluating.

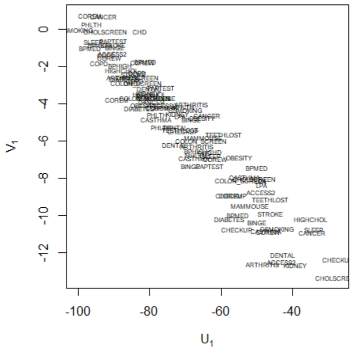


Fig 21. Plot of V1 vs. U1

From the plot above, there is no specific outliers, and U1 and V1 are highly negatively correlated to each other. Which also matches our intuition since when the lack of health insurance is more severe, it is more likely that the individual leads a poor-conditioned life, and thus the probability of getting a heart disease and obesity is higher, while heart disease is lower.

5. CONCLUSION

This study provides a general picture on the different measures of disease control and prevalence of 100 cities and investigate what factors contribute to such differences by using Principal Component Analysis and Factor Analysis. Three components are selected to describe the dataset, with variables mainly from as the lack of health insurance, visits to dentists, high blood pressure, cancer, routine checkup, clinical preventive services and lack of sleep.

Cluster analysis here uses 2-means clustering and complete linkage for clustering, and they display different result, which 2-means simply output cities with pleasant healthcare conditions and cities don't. While 9 clusters from complete linkage precisely identify each category. By displaying them through the nation map, we could make reasonable suggestions for the middle part of the cities to promote not only the health insurance coverage, but also other ways to deal with disease controls. And also to make sure that coastal area doesn't need to focus on promoting health insurance at this stage.

Canonical analysis focus on the relationship between disease prevalence variables and disease controls, which also matches our intuitive understanding of how canonical variables affect the original data.

Hopefully, it could be a study of 100 cities as a start, and could be then implemented towards the global disease prevalence and controls.

6. REFERENCES

[1] Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., & Croft, J. B. (2014). Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor

Surveillance System. American Journal of Epidemiology, 179(8), 1025–1033. doi: 10.1093/aje/kwu018

[2] Longitude and latitude geolocation data from https://www.arcgis.com/home/item.html?id=f7f805eb65eb4ab787a0a3e1116ca7e5

[3] UC Business Analytics R Programming Guide from https://uc-r.github.io/kmeans_clustering

[4] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411–423. doi: 10.1111/1467-9868.00293