Kevin Brown
Final Research Project
Data Mining Methods

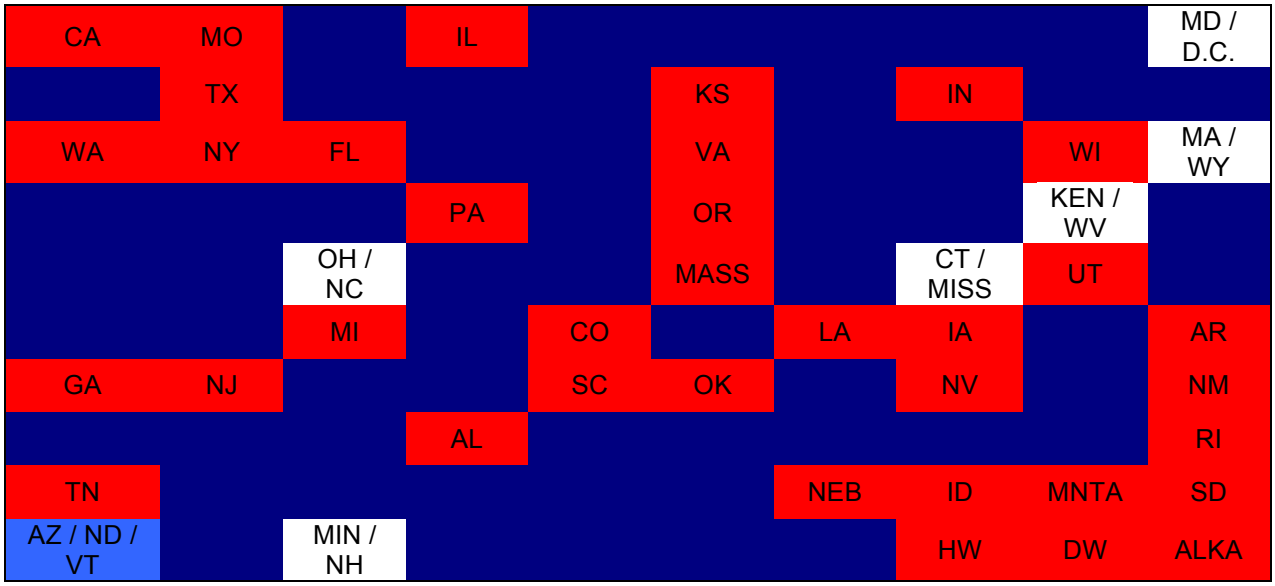
Crime & Violence Statistics of the United Sates As Defined by the Self Organizing Map (SOM)

Executive Summary

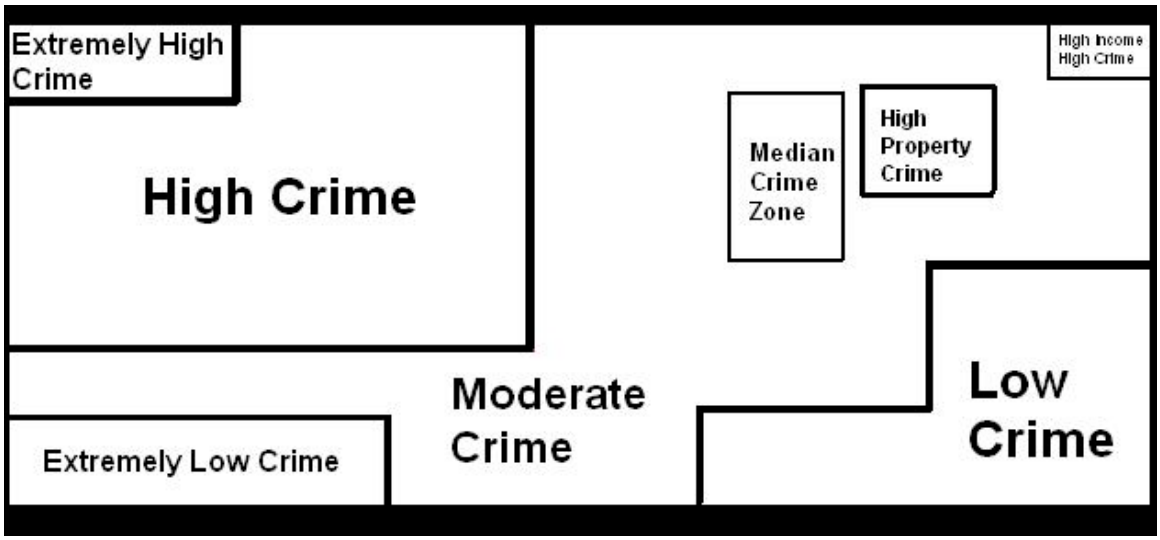
The self organizing map (SOM) is a clustering algorithm that given a data set, creates a two dimensional map of nodes, then trains those nodes with the given data set (3). The SOM places the given data set into the nodes that it created. Each specific set of data is placed into a node that is closest or most similar to the specified data. Each node is then accordingly filled with data that is similar to it; similar nodes will be mapped next to one another (3). Therefore individual pieces of data will be mapped together, creating clusters of 'like' nodes.

In my instance of the self organizing map (SOM) I choose to take statistical data from the U.S. Department of Justice pertaining to crime and violence per U.S. state, the average income per individual in each state, and the population of that state. My hope was that the SOM would map the data in clusters possibly resembling the geography of the U.S. or display geographical regions of the U.S. together, or display clusters with more data specific trends close to one another. For instance states of high crime together, low crime together and medium crime together while factoring in income and population.

My final SOM, including all fifty states of the U.S. and the District of Columbia (Washington, D.C.) looked as follows:



Once I created the map layout from the SOM output, I could conduct a critical analysis of my results. By calculating the minimum, median, and maximum values of each subclass in the dataset, I could begin making comparisons. Also, by observing the dataset as a whole I produced these results:



Each definitive area is mapped off and labeled in my results. Areas are divided into regions of corresponding to the data set and the map, also population and income are factored into these divisions.

Problem Description

The given problem I am trying to address in my research is if I give the SOM a fairly large and complex data set with many attributes, as well as additional attributes not directly pertaining to crime and violence, it will produce a map with significant output and clustering results. The crime and violence dataset alone is complicated with 9 attributes. I complicated the data furthermore by adding income and population. I wish to see whether SOM can create a map with this amount of different data and produce a MAP without consisting of messy or incoherent output.

Analysis Technique

The dataset I have chosen for my research project is crime & violence statistics as composed by the Bureau of Justice Statistics Database, population, and income statistics of every state in the United States including D.C. The data will include the following categories: population, total violent crimes, murder and non-negligent manslaughter, forcible rape, robbery, aggravated assault, total property crimes, burglary, larceny-theft, motor vehicle theft, and the income per individual on average per state.

The algorithm I choose to ideally handle this expansive data set was the self-organizing map algorithm developed by Teuvo Kohonen (1) at Helsinki University of Technology. The SOM, often called the Kohonen Map is a subset of artificial neural networks. The SOM is useful for creating low-dimensional displays of high-dimensional data. In other words a perfect application for what I needed accomplished. SOM consists of a competitive layer of neurons and an input layer (3). Weight vectors are created by the weights of connections from the input layer, to a single neuron or node in the competitive layer (3). The training SOM performs utilizes competitive learning (3). When a training sample is given to the network, its Euclidean distance to all weight vectors is computed (3). While the SOM runs, the program chooses the weight vector with the smallest Euclidean distance from the competitive layer, and trains the said vector to closely resemble the input vector it is closest to (3). This process is repeated over and over again using different input vectors. The result is a map that displays similar nodes clustered together. The formula for Euclidean distance is defined as:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

(5). In this formula, n represents the number of columns of data, p_i is the value of the chosen vector and q_i is the value of the vector that is being compared (5).

The SOM consisted of 3 different executable files that required input from 2 separate data files, and a batch file to run the 3 executables. It was my task to create a 'run' batch file by inserting the command lines to run each program and the given parameters necessary to run without error. In each command line I had to choose the size of map I wanted to create, the data files I wanted to read into the algorithm, and the name of the file that would store the output after the algorithm ran its course. In order to format the data files the SOM would be reading in I first needed to format my data in a manner that was appropriate for the SOM to handle. First, I needed to create a data file that was space delimited and listed the number of attributes first in a line by itself. Second, I had to create another data file with the same parameters but this time the rows of data were labeled by state, and the second file had to have a different name from the first. Now, I was able to run the SOM executables since, my data files were created and all corresponding batch files had been updated to load in the newly created data files.

Once the SOM output its first set of results, I had to display them into an excel file to visually represent the map. However not all states were output on the first run of the SOM software. To handle this problem I then had to go back and edit my data files removing all states that had been mapped from the file, leaving those that hadn't been mapped untouched. The original data file with the unlabeled data set also remained static. I ran the SOM again and repeated this process until all states were mapped and visually represented within my color coded excel map.

After my map had been completed I depicted each state mapped by itself in red, each state with a collision (two states mapped into the same node) in white, and each state with more than one collision in bright blue. All neutral space where no nodes were mapped I depicted in dark blue. Once my visual representation was complete I could begin recording what states had clustered together, which ones were isolated, and possible trends amongst clusters. After I had a list of the clustered states I could begin analyzing the original data set to possibly observe why this clustering pattern had occurred.

Due to the large size of my given data set, I had to formulate a method for comparison that would make it feasible for me to make an analysis. I did this by sorting each individual attribute at a time in ascending order using excel. When the given set of data was in order, I recorded the maximum, minimum, and median values of that set, and the state that corresponded to that value. I repeated this method for each individual attribute of the data set. This gave me 11 sets of max., min., and median values with states listed next to each value. This led me to take each of the states that were clustered together and group them along with their data in excel so that I may draw comparisons between the different groups. I related state within the groups to where they fell in relation to the minimum, maximum and median values for each specific attribute in the cumulative data set. I noted where each state and cluster fell the said max/min scale and what attributes were unique to that cluster. After much comparison and observation using these rules as a guideline I was able to create my second diagram displaying what the clusters and groups were and why they were located in that area.

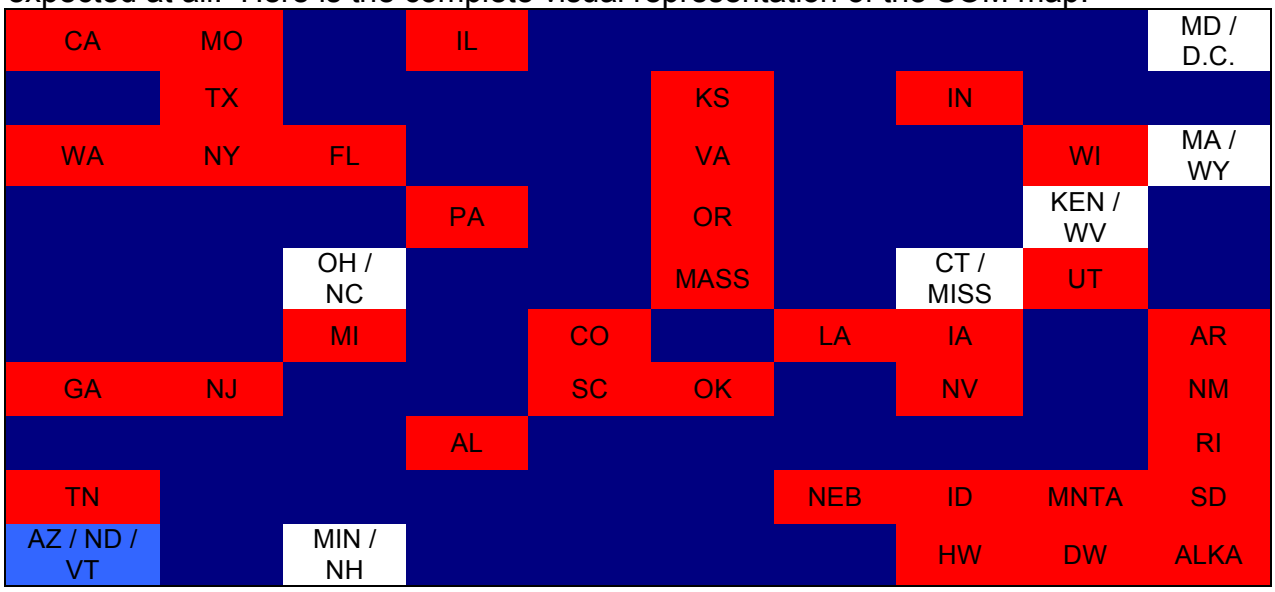
Assumptions

-The SOM will be able to create a simple, easy to read map with a relatively complex dataset
-The SOM may depict parallels with crime and geography
-Some noise may occur due to the complexity of the dataset
-There are multiple sets of data that make up crime dataset as a whole
- Population and annual income will be added to the dataset to possibly see a correlation with crime rates.
-Areas of high crime, low crime, moderate crime, and types of crime will be mapped together.
-Such a large dataset may be too complex to analyze the mapped results; in this case a few attributes may have to be removed.

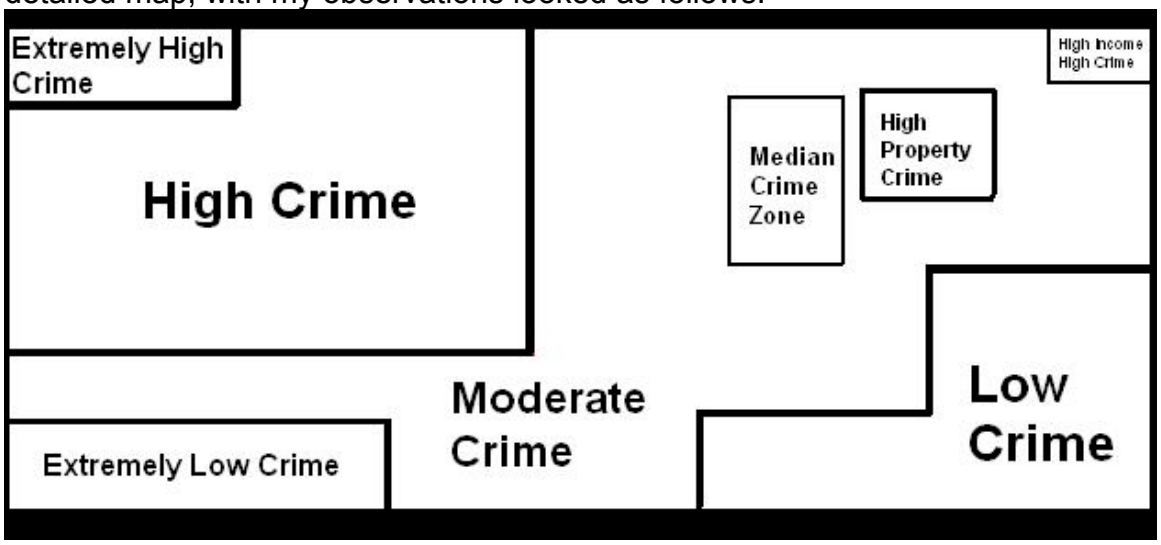
Results

Having a dataset consisting of 50 states, each with eleven different attributes, I was surprised that SOM was able to find likenesses in each node to organize the data. However, I was pleasantly surprised with the results. Ideally,

I had hoped for the SOM to possibly map states in a possible resemblance to geographic location. In reality, the SOM mapped the sates in ways I had not expected at all. Here is the complete visual representation of the SOM map:



The areas in red were mapped alone with no collisions, white boxes consisted of 1 collision, bright blue consisted of 2 collisions, and dark blue was space containing no states. After carefully analyzing the data using the aforementioned techniques I was able to draw some conclusions. My final detailed map, with my observations looked as follows:



California was mapped in the very upper-left corner. It was mapped in this fashion because it held the maximum value in every attribute in the entire data set. Interestingly enough, Missouri, a much smaller state was mapped close to California. This is due to Missouri having an extremely high crime rate in relation to its population. In the surrounding areas around CA and MO, are other areas of relatively high crime consisting of states such as TX, NY, FL, MI, WA, GA, NJ,

and a few others. What deemed this states to be high crime is that they were all well above the median values for high crime in each attribute, MO and California happened to be standout datasets in this group and therefore mapped as extremely high areas of crime. Interestingly enough, areas of extremely low crime were mapped very close to the areas of high crime, even though areas of low crime are mapped on the other side of the grid. One would assume that areas of extremely low crime would be mapped within the low crime group similar to the extremely high/high crime group. A possible cause of this unique mapping may be the population of these extremely low crime areas in comparison the other low crime areas is in stark contrast.

In the upper right hand corner more unique mapping occurred. Maryland and the District-of-Columbia mapped together, as areas of high income and high crime. Uniquely enough these areas are geographically located right next to one another. Four states Wisconsin, Kentucky, West Virginia, and Utah mapped together as areas of high property crime. Why these mapped together is most likely due to the fact that the states were average in comparison to other states, except had a relatively high property crime rate considering population. Also to be noted, as that in areas of low crime, most of the included states have a low population as well, or are slightly isolated, and often contain no major U.S. cities. States with larger urban areas tended to mapped together in the regions of high crime.

One could possibly draw the conclusion from the Mapping clusters that the closer you cram people together, regardless of climate or geographic location, the more violent crime will occur. Areas of high population typically had higher crime rates than those with low populations. Also noted is that fact that areas of high income does not necessarily mean there will be low crime. In fact, some of the most affluent areas of the U.S. tended to have the highest crime rates per capita. The healthy majority of states found themselves located in the median crime range zone, being violent only on a statistical average level.

.

Issues

Originally, this project was going to display major nations that were members of the UN and their crime and violence statistics factored in with population and economical data. However in the final stages of my research the UN logistic data sets I was using at the time were summarily blocked by the UN for further use. What happened was this, the UNECE statistics department whose data I was using was subsidized by the main UN logistic department website. The UN no longer allows private researchers to access this data without a paid subscription for an outrageous rate. They also blocked any use of datasets previously downloaded from this mirror site. More and more data is being purchased and deemed private, unavailable for public use without paid

subscription. It is rumored that large corporations are buying up large amounts of data for this purpose.

Another issue was the shear size of my dataset. It was hard to draw conclusions from such a large and complex dataset when the majority of data was relatively similar. The most difficult part of this research was formulating an efficient and logical method to compare data and draw conclusions from the map.

Appendices

1 Self Organizing MAP SOM (tools):
mercury.webster.edu/aleshunas/

2 Crime and Violence Data Sets:
http://bjsdata.ojp.usdoj.gov/dataonline/Search/Crime/State/statebystatelist.cfm

3 Self Organizing Map SOM (information):
http://en.wikipedia.org/wiki/Self-organizing_map

4 Population and Income Data Sets:
http://quickfacts.census.gov/qfd/index.html

5 Euclidean Distance:
http://en.wikipedia.org/wiki/Euclidean_distance