Predicting Crime Rates in the Metro U.S. Based on Crowding and its Implications on the U.S. Prison System



Andy Ninh

Spring 2014

**Abstract**

Crime rate data underwent machine-learning prediction based on population density and other variables. It was predicted that density has a positive affect on crime, which was supported by a dendogram, but unsupported by regression analysis. It was also hypothesized that percentage of nonwhite people in a population positively affects crime rates, which was supported by regression analysis significantly. It is interesting how machine-learning algorithms can predictively cluster this data. U.S. metros are plagued by overcrowding and suffer higher crime rates. This same trend plagues U.S. prisons and maintains a huge social issue.

**Introduction**

A social problem that has received a large amount of attention for many years, and is arguably one of the most important in 2014, is crowding, or overpopulation, and its implied deleterious effects on quality of life. Many studies have been conducted for at least the past 40 years, in all aspects of crowding and its effects, both physical and psychological. With regard to the law, one of the biggest areas seeing concerns with crowding is the U.S. prison system. Overcrowding in prisons causes many problems, which is why it is a serious concern. Once the causes of crowding have been established, researchers can begin to address the problems it causes and deal with them.

In the current study, a dataset comparing population density and crimes in metro cities in the U.S. from 1970 was used to complete a hierarchical clustering analysis in R. As expected, the analytical prediction method accurately clustered cities with similar population densities and relation to crime rates, supporting the hypothesis that crowding will increase the rate of crime. This is incredibly important because over the years, scientists have found confusing results on the topic of crowding, and crowding is an important issue with regard to the public at large, but specifically the prison system. As prisons across the country continue to increase in population, more violent crimes will be committed, making order difficult to maintain. This will completely defeat the purpose of the system. Let this finding be a warning to the justice system.

**Research Problem**

Understanding where crime happens can be a key to understanding why it happens. Models that predict the occurrence of crime by geographical area often use data on the characteristics of the inhabitants (e.g. income, race, family structure) of that area. However, population density has also received considerable attention as it relates to crime. Jane Jacobs, in 1961, contradicted the popular wisdom of city planners with her claim that crowded city streets and sidewalks could be effective deterrents to criminal behavior. A number of national studies tested the relationships between density and crime, with differing results. Some studies found positively correlated relationships between crime and density, while others found the opposite types of relationships. Still, others (e.g. Freedman, 1975) found non-significant relationships between the two variables.[[1]](#footnote-1) So, which is it?

It would seem intuitively that cramming people into a fixed area would not only statistically improve the possibility of increased crime, but also increase the likelihood due to negative effects on the psychology of the situation.

*Hypothesis 1:* Population density, or crowding, will be positively related to crime rate.

Also, high population density usually relates to a high volume of traditionally nonwhite or traditionally poorer neighborhoods. This would lead one to think that metro U.S. cities with high percentages of nonwhite people will have higher crime rates.

*Hypothesis 2:* The percentage of nonwhite persons in a population will positively affect the crime rate.

Although the many factors that might play into this hypothesis are outside the scope of this paper, it still important to assess this factor in the analysis. There is a general understanding, and proven track record, that cities like Detroit have one of the highest crime rates in the country, and it also has a high concentration of nonwhite population. Although this data originates from the 1970 U.S. census, it is likely that population data has not varied significantly for the major cities assessed as most major U.S. metros have maintained their appeal.

**Results + Methods**

The data collected for this study was collected in 1970 by J. Freedman, a famous psychologist who studied the effects of crowding on behavior.[[2]](#footnote-2) It was slightly modified from the original, and is missing some data, but this does not detract from the conclusions. It represents 100 U.S. Metro cities, defined as having a population greater than 250,000, and contained 4 variables: (1) Total Population in 1968 in thousands, (2) Percent of Nonwhite persons in the population in 1960, (3) Density, or Population per square mile in 1968, and (4) Crime rate per 100,000 in 1969. In order to assess the effect of density and crime, this study used an analytical prediction method known as hierarchical clustering in a program called RStudio, which uses a program language called R.

The variables assessed were measured on different scales, so the first step was to standardize the data. In order to do this, the “scale” function was used. The data frame passed through “scale,” along with two vectors. The first is called “center,” which is a vector of values, one for each column of the data frame to be standardized, which was subtracted from every entry in that column. The second, called “scale,” is similar to “center,” but is used to divide the values in each column. These vectors are created with the “apply” function that performs the same operation on each row or column of the data frame. The city names are not useful to portion of the analysis so they were removed. Notice that the “scale” function does not change the order of the rows of the data frame, so it was easy to identify observations using the omitted column from the original data. Refer to the Appendix for output of the Hierarchical Clustering.

The first step of the hierarchical method is to calculate a distance matrix. For a data set with n observations, the distance matrix will have n rows and n columns; the (i,j)th element of the distance matrix is the difference between observation i and observation j. In this experiment, the default of Euclidean distance was used, but it is possible to use other metrics. Then, the “hclust” function was applied to update the distance matrix for “complete” linkage. Using this method, when a cluster is formed, its distance to other objects is computed as the maximum distances between any object in the cluster and the other object. At this point, a graphical display of our output was created. The main graphical tool for looking at a hierarchical cluster solution is known as a dendogram. This is a tree-like display that lists the objects, which are clustered along the x-axis, and the distance at which the cluster was formed along the y-axis. The dendogram from this output is found in the Appendix.

If you choose any height along the y-axis of the dendogram, and move across the dendogram counting the number of lines that you cross, each line represents a group that was identified when objects were joined together into clusters. The branches of the dendogram that spread out below the line represent the observations in that group. For example, if we look at a height of 20, and move across the x-axis at that height, we’ll cross three lines. That defines a three-cluster solution; by following the line down through all its branches, we can see the names of the cities that are included in these three clusters. Since the y-axis represents how closer together observations were when they were merged into clusters, clusters whose branches are very close together (in terms of the heights at which they were merged) probably are not very reliable. But if there is a big difference along the y-axis between the last merged cluster and the currently merged one, that indicates that the clusters formed are probably doing a good job in showing us the structure of the data.

Looking at the dendogram for the crime data, there appear to be about 10 distinct groups. Looking at ten groups is a good place to investigate. The first thing we looked at was how many cities were in each group. We created a vector showing the cluster membership of each observation using the “cutree” function. Simply displaying the group memberships is not that revealing. Instead, we created a table and also combined the various groupings using the “sapply” function. This way, we could see solutions ranging from 2 to 10 clusters. To see which cities are in which clusters, we use subscripting on the vector of city names to choose just the observations from a particular cluster. Since we used all of the observations in the data set to form the distance matrix, the ordering of names in the original data will coincide with the values returned by “cutree.” Also, we grouped all the clusters using “sapply.”

The most obvious and important information observed from the dendogram and the tables is that New York and Jersey City were clustered highly for crime due to crowding. Also, Honolulu was notably higher than a majority of the clusters. Other notable cities include Chicago, Los Angeles, Detroit, Boston, Philadelphia, San Francisco, Washington, Cleveland, and Baltimore. Logically, it makes sense that some of these popular metros with large population densities would have higher crime rates. This appears to support Hypothesis 1. The next step was to aggregate the median and mean values and convert the values in the table back into their original forms to make sense of everything. Looking at that table, something odd sticks out. Honolulu has the highest population density of any city, but it’s cluster ranks third highest as we analyzed the data. This suggests that density is not the only factor that increases crime.

In order to determine the significance of our results, we performed a regression analysis on the results. What this does is compare the variables and determines how they affect each other. This method provides a string of “coefficients” that indicate how each independent variable affects the dependent variable and to what degree. Also, this provides p-values that tell us whether the independent variables are statistically significant in affecting the dependent variable. In other words, it tells us if this places an important role in affecting the outcome of crime data. The regression analysis shows a slightly negative correlation between density and crime, which does not support our hypothesis. However, this result is not statistically significant. Also, percentage of nonwhite people in the population seems to have a strong positive effect on crime, and is statistically significant (p < 0.05). Also, the total population of a city seems to have a slightly positive affect on the crime rate with statistical significance. Both of these points make logical sense, and support Hypothesis 2.

**Discussion**

The hypotheses that drove this study were based on a review of literature, which showed differing or conflicting relationships between density and crime. It was hypothesized that density would positively affect the crime rate, in that the higher the population density, the higher the crime rate. Also, it was hypothesized that the higher the percentage of nonwhite people in the population, the higher the crime rate would be.

The first hypothesis was not supported in a significant way because a regression analysis showed a slightly negative relationship but without statistical significance. This could be because either density is not a statistical predictor of crime rate, or our data is just not full enough to predict a true outcome. I suspect the latter is the case because there is support for the hypothesis in the clustering analysis. This is actually very notable, in that a clustering algorithm combined the cities with the highest densities, which still stands today.[[3]](#footnote-3) Also, these cities are reputed for having higher than average crime rates.

In 1957, Schmitt found a high correlation between population density, juvenile delinquency, and adult crime in Honolulu during the years 1948-51, with a large number of juvenile delinquents and adult prisoners from overcrowded neighborhoods having multiunit construction and a large ratio of population to residential land.[[4]](#footnote-4) It is interesting that Honolulu had the highest population density of all the major metros, and also had a very high crime rate. The findings today do lend support to that hypothesis.

Also, Hypothesis 2 was supported. There was statistical significance found for the positive relationship between high percentage of nonwhite population and higher crime rates. This makes sense and is important. Consider Detroit, Michigan for example. In 1970, its population was approximately 13.1% nonwhite and had a crime rate of 4,283 per 100,000, and this was strongly indicated on the dendogram and table. Currently, Detroit has a much higher percentage of nonwhite people concentrated in the city, and Detroit is known to be one of the top three most dangerous cities to live in due to high crime. Of course, this is not the only reason, but the example draws support for the hypothesis.

The most important thing to gather from this experiment is that crime can be predicted to some degree using machine-learning methods, based on seemingly important factors like density, or crowding. Also, this has a huge implication on a huge social issue of overcrowded prisons. Prison overcrowding is caused by a variety of issues, such as not enough room in prisons, fluctuating crime rates, changes to laws and improvements to law enforcement tactics. Researchers have determined that some of the causes of prison overcrowding are harsher penalties for criminal activities, changes to laws that make new actions illegal, high recidivism rates and needed improvements to the penal system.[[5]](#footnote-5) Based on our findings, you might suspect that prison facilities in the cities with the highest population densities would be the most overcrowded. However, prisons tend to be larger in these areas. It is probably possible to figure this out, but is outside the scope of the current research. More importantly, overcrowding prisons will have a similar effect as the results indicate. More crime will be committed inside prisons, which defeats the purpose of prison.

A long-term study on prison population density (PPD) in Japan found that it has a direct correlation with prison violence rates (PVR). This study conclusively states that ‘the effect of PPD was significant and positive on PVR, even after controlling for the effects of the proportions of males, age younger than 30 years, less than one-year incarceration, and prisoner/staff ratio.[[6]](#footnote-6)

This problem must be solved for the U.S. prison system to continue to be effective.

A limitation of this research may be inadequate data to provide strong results for Hypothesis 1. This is not a huge issue in this case because there was some support for the hypothesis and it makes logical sense. Bigger or more popular metro cities such as New York are expected to have higher crime rates and higher population densities. Future research in this area could expand the data set, as well as including other possibly significant variables such as average household income, amount of livable property, etc.

**Conclusion**

It is clear that crowding impacts crime and although our experiment did not yield perfectly significant results, it remains a powerful theory. Population density does affect crime rates in U.S. metros. This affects the U.S. prison system in a truly negative way, and unless a solution to the problem is created, it will fail.

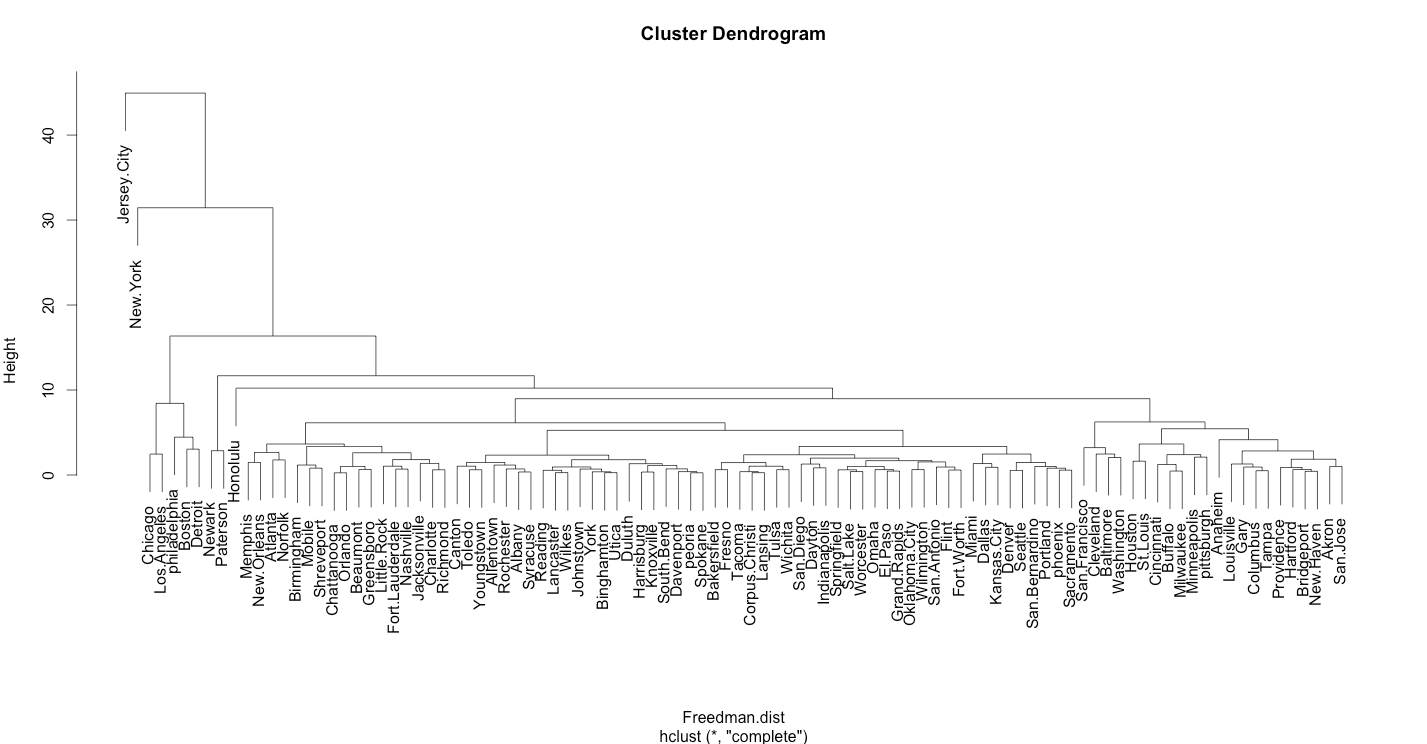
**Appendix**

*R Output for Paper*:

All output from R is beautifully displayed on my GitHub Account. A link to the R Markdown page is found here:

<https://github.com/A-Ninhja/Legal_Analytics_Spring_2014/blob/master/Final_Project/Crowding.md>

*The Dendogram Built from Hierarchical Clustering*



Click [here](https://raw.githubusercontent.com/A-Ninhja/Legal_Analytics_Spring_2014/master/Final_Project/Cluster.png) to view an enlarged version of the dendogram.

*R Output for Regression Analysis*

> reg\_1 <- lm(crime ~ density + nonwhite + population, data=Freedman)

> summary(reg\_1)

Call:

lm(formula = crime ~ density + nonwhite + population, data = Freedman)

Residuals:

Min 1Q Median 3Q Max

-2003.18 -657.16 70.94 603.73 2213.20

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2193.70088 143.04566 15.336 < 2e-16 \*\*\*

density -0.02145 0.06578 -0.326 0.745045

nonwhite 26.03770 8.76746 2.970 0.003764 \*\*

population 0.24495 0.06095 4.019 0.000116 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 877.2 on 96 degrees of freedom

Multiple R-squared: 0.2288, Adjusted R-squared: 0.2047

F-statistic: 9.493 on 3 and 96 DF, p-value: 1.496e-05

1. *See* Brian Christens & Paul W. Speer, *Prediciting Violent Crime Using Urban and Suburban Densities*, Behavior and Social Issues, 14, 113-127 (2005). [↑](#footnote-ref-1)
2. United States (1970), *Statistical Abstract of the United States*, Bureau of the Census. [↑](#footnote-ref-2)
3. *See* *List of United States Cities by Population Density*, Wikipedia, http://en.wikipedia.org/wiki/List\_of\_United\_States\_cities\_by\_population\_density (last visited May 7, 2014). [↑](#footnote-ref-3)
4. *See* Chalsa M. Loo, *Crowding and Behavior*, 93 (1974) [↑](#footnote-ref-4)
5. *See* *Prison Overcrowding is a Growing Concern in the U.S.*, Portland State University, http://online.ccj.pdx.edu/resources/news-article/prison-overcrowding-is-a-growing-concern-in-the-u-s/ (last visited May 7, 2014). [↑](#footnote-ref-5)
6. *Id.* [↑](#footnote-ref-6)