Youth, Race and Crimes in the United States in 1995

by Bodi Hu and Jinni Li

Modern Statistical Computing

David Rossell and Martin Wiegand

Pompeu Fabra University

March 23, 2023

1. Introduction

It is a reality that crime-free societies are almost impossible to achieve. However, if we aspire to reduce crime rates in a particular country or society, it is crucial to understand the root causes and contexts in which they occur. By doing so, we can unveil the underlying reasons behind these unlawful acts that happen daily. Knowing the causes of crime is extremely important as it enables us to improve as a modern society.

Primarily, this information provides us with crucial knowledge to prevent future crimes from happening. For example, the relationship between race and crime rate is important because it can help to identify and understand disparities in crime rates across different racial and ethnic groups. These disparities are often due to historical factors such as racism and discrimination, as well as broader social and economic factors such as poverty and lack of access to education and employment opportunities. Policymakers can use this knowledge to develop effective strategies to tackle these problems.

In addition to understanding disparities in crime rates across different racial and ethnic groups, it is also important to examine the relationship between age and crime rate. In particular, investigating whether young age is linked to crime cases can provide insights into potential risk factors and preventative measures. By gaining a deeper understanding of the underlying factors that contribute to criminal behavior, policymakers and communities can develop targeted strategies and interventions to address them.

The goal of this report is to find whether or not there exists a significant relationship between young-age people and criminal cases. The dataset used for analysis is based on criminal cases in the United States in 1995, formed by 2018 districts. It offers demographic and economic variables, including 18 crime variables that are available to be predicted.

Through estimating different count data regression models, we can observe how the negative binomial regression model seems to be more appropriate in order to fit our data. Moreover, we can see variables related to race, education, parental civil status, and demographic variables are significant as expected. One interesting result is that the percentage of white people seems to have a negative relation with the number of violent crimes (in contrast to other races), but other factors such as low educational level, high population, advanced age, and divorces seem to have a positive relation with crimes. The next sections will provide an extensive interpretation of the results.

2. Data

The dataset used for analysis comes from the Machine Learning repository of the University of California Irvine. As mentioned previously, it is formed by 2018 observations which are the different districts around the United States. The original dataset provided 125 independent variables, including some identifying variables such as community name and state abbreviation. 18 crime-related variables are available to be predicted. The source advises only taking one crime variable as a dependent variable, and we decided to choose violent crimes per population. As for our selection of independent variables, we have chosen a total of 17 variables:

- racePctblack: percentage of population that is african american
- racePctWhite: percentage of population that is caucasian
- racePctAsian: percentage of population that is of asian heritage
- racePctHisp: percentage of population that is of hispanic heritage
- agePct12t21: percentage of population that is 12-21 in age
- agePct12t29: percentage of population that is 12-29 in age
- agePct16t24: percentage of population that is 16-24 in age
- agePct65up: percentage of population that is 65 and over in age
- medIncome: median household income
- PctPopUnderPov: percentage of people under the poverty level
- PctNotHSGrad: percentage of people 25 and over that are not high school graduates
- PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed
- TotalPctDiv: percentage of population who are divorced
- PctKidsBornNeverMar: percentage of kids born to never married
- RentMedian: rental housing - median rent
- PctForeignBorn: percent of people foreign born

We have chosen these variables, thinking about what type of people usually commit crimes. As an example, the situation of people with low household income or people under the poverty level may lead them to commit crimes in order to obtain financial gain. The percentage of high school graduates can also tell us that people who have some level of education are more or less likely to commit unlawful acts. Race is also included as percentages of the population that are either African, Asian, Caucasian or Hispanic. Many other demographic variables, such as foreign-born people and unemployment are included because we believe that they are factors that affect crimes. Something important to highlight is that we have included three age variables for the youth population, but later on, we will see that only one variable stays due to correlation. Before everything, we had to do some R data wrangling to select our variables and round the decimals.

3. Exploratory Analysis

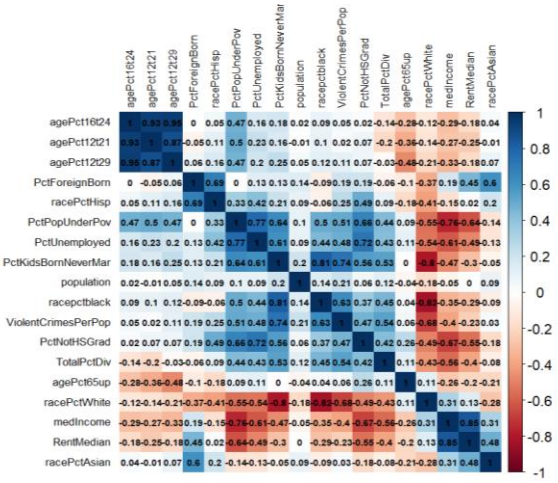

Figure 1. Correlation plot between all predictors

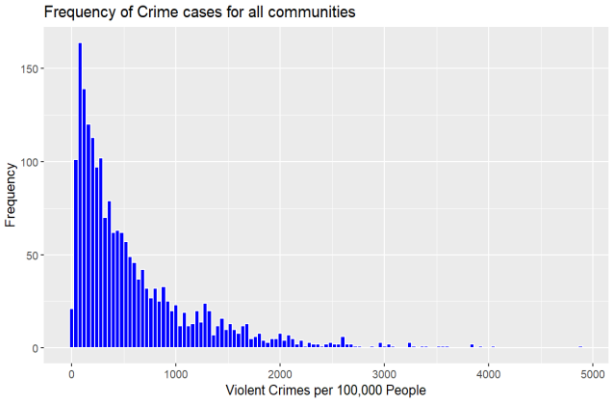

Figure 2. Frequency of Crime cases

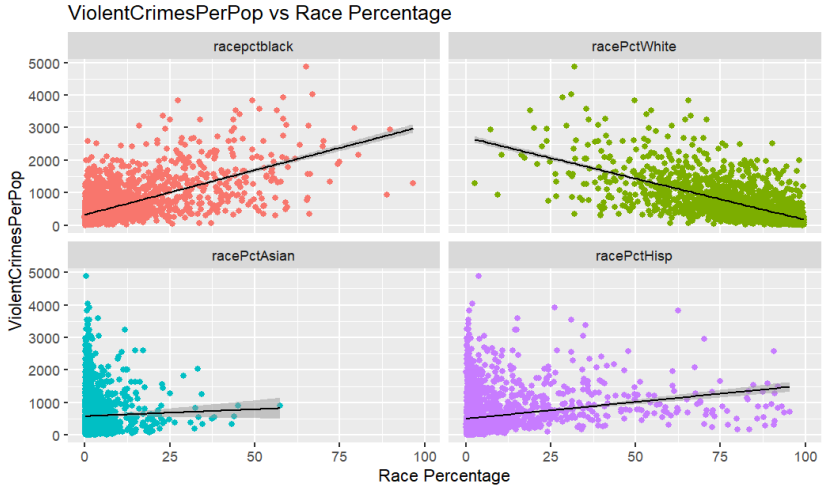

Figure 3. % Race and Crime cases

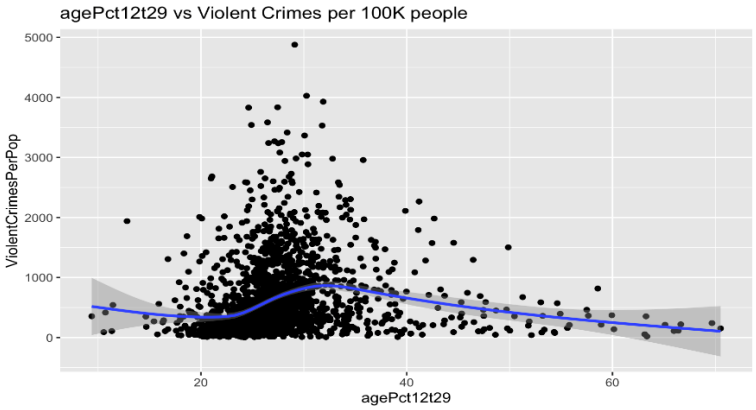

Figure 4. % Aged 12-29 and crimes cases

Figure 1 displays a heatmap illustrating the correlation between all variables except for community names and state notations. Notably, age variables 12 to 21, 12 to 29, and 16 to 24 exhibit strong correlations, which may result in multicollinearity issues. To mitigate this, we will use the age variable from 12 to 29.

Figure 2 presents a histogram that provides insights into the distribution of violent crimes per 100,000 people in the dataset. The majority of observations fall between 0 and 1000 cases of violent crimes per 100,000 people, with only a few values exceeding 1000 cases. Additionally, the distribution is skewed to the right, indicating a higher concentration of observations with lower values of violent crimes per 100,000 people. The presence of a long tail of higher values suggests the existence of some districts with particularly high rates of violent crime.

Figure 3 comprises four scatterplots that display the relationship between each race and crime cases, as well as the percentage of each race in the population. These scatterplots reveal that most districts are inhabited by people of white ethnicity, and the Asian population has the smallest percentage in the population. Furthermore, the majority of districts have less than 3000 crime cases. Each race exhibits a positive relationship with its race percentage, except for the white ethnicity.

Figure 4 showcases a scatterplot that depicts the non-linear relationship between the percentage of people aged 12 to 29 and criminal cases. The plot suggests that most crime cases occur when the districts have less than 40% of young people.

Additionally, we have plotted for residuals and fitted values between each predictor and the dependent variables using a linear model to check for linearity. All of the residual plots show non-linear patterns except for the percentage of the Asian population. All of these residual plots are created using the base function in R, and since there are 17 of them, it will not be included in the report, but the plots can be reproduced with the function available in the quarto document, or you can see them in dashboard files. We also found out that 221 observations did not have data for the dependent variable, which were omitted in the regression analysis.

4. Method

Our aim is to use the Poisson Model to estimate the effects of the mentioned predictors on the number of violent crime cases per 100k population in United States communities during the 1990s. We begin by calculating the Poisson estimates and the dispersion parameter to check the model assumption.

$$V(y_i) = \mu_i \; ; \frac{Residual\ Deviation}{Degree\ of\ freedom\ of\ Residuals} = 1$$

Otherwise, we will have an over-dispersion problem. To address this issue, we have planned two models: the Quasi-Poisson Model and the Negative Binomial model.

The Quasi-Poisson Model assumes that the mean and variance of the response variable are proportional to each other but allows for the variance to be larger than the mean. In contrast, the Negative Binomial model assumes that the variance of the response variable is greater than its mean and allows the degree of over-dispersion to vary independently of the mean by estimating an additional dispersion parameter. The Negative Binomial Model is preferred when the degree of over-dispersion is large and the Quasi-Poisson model is insufficient.

Figure 5. Coefficient plot Poisson vs. Quasi-Poisson vs. Negative Binomial Model

To determine which model is appropriate, we perform residual analysis to ensure that the residual variance is constant, which is an indicator of over-dispersion. Once the over-dispersion issue is resolved, we add interactions between the population percentage of non-High-school diplomas (PctNotHSGrad) and the percentage of kids whose parents never got married (PctKidsBornNeverMar) to our model. We hypothesize that children born to unmarried parents may receive less education, which could influence the impact of PctNotHSGrad on violent crime.

Finally, we conduct an ANOVA test to compare the model without the interaction to the model with the interaction and choose the one that better fits our analysis to interpret the coefficients.

5. Results Analysis

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           5.067e+00  3.107e-02  163.08   <2e-16 ***
racepctblack          6.105e-03  2.174e-04   28.07   <2e-16 ***
racePctWhite         -7.221e-03  2.109e-04  -34.24   <2e-16 ***
racePctAsian          1.010e-02  3.303e-04   30.59   <2e-16 ***
racePctHisp           6.781e-03  1.439e-04   47.14   <2e-16 ***
population            1.317e-07  2.200e-09   59.89   <2e-16 ***
agePct12t29           5.659e-03  3.330e-04   16.99   <2e-16 ***
agePct65up            2.017e-02  3.759e-04   53.66   <2e-16 ***
medIncome            -1.490e-05  3.301e-07  -45.15   <2e-16 ***
PctPopUnderPov        5.805e-03  2.569e-04   22.59   <2e-16 ***
PctNotHSGrad          2.685e-03  1.664e-04   16.13   <2e-16 ***
PctUnemployed        -1.584e-02  5.652e-04  -28.02   <2e-16 ***
TotalPctDiv           1.100e-01  5.154e-04  213.37   <2e-16 ***
PctKidsBornNeverMar   4.017e-02  4.731e-04   84.90   <2e-16 ***
RentMedian            5.025e-04  1.802e-05   27.89   <2e-16 ***
PctForeignBorn        3.658e-03  1.890e-04   19.36   <2e-16 ***
```

Figure 6. Coefficients of the Poisson Model

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1032501  on 1993  degrees of freedom
## Residual deviance:  373972  on 1978  degrees of freedom
## AIC: 389340
##
## Number of Fisher Scoring iterations: 5
```

Check if there is over-dispersion

```
#residual deviance / degrees of freedom
373972 / 1978
```

```
## [1] 189.0657
```

Figure 7. Dispersion Parameter for Poisson Model

As Figure 6 shows, under the Poisson regression model, all variables are significant at 95% confidence level. However, throughout the dispersion parameter, we can see that this model does suffer a problem of over-dispersion leading us to develop a Quasi-Poisson Model in order to solve it.

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.067e+00  4.551e-01  11.135  < 2e-16 ***
racepctblack          6.105e-03  3.185e-03   1.917 0.055386 .
racePctWhite         -7.221e-03  3.089e-03  -2.338 0.019504 *
racePctAsian          1.010e-02  4.838e-03   2.088 0.036889 *
racePctHisp           6.781e-03  2.107e-03   3.219 0.001309 **
population            1.317e-07  3.221e-08   4.089 4.51e-05 ***
agePct12t29           5.659e-03  4.877e-03   1.160 0.246074
agePct65up            2.017e-02  5.505e-03   3.664 0.000255 ***
medIncome            -1.490e-05  4.834e-06  -3.082 0.002081 **
PctPopUnderPov        5.805e-03  3.763e-03   1.543 0.123096
PctNotHSGrad          2.685e-03  2.438e-03   1.101 0.270888
PctUnemployed        -1.584e-02  8.278e-03  -1.913 0.055841 .
TotalPctDiv           1.100e-01  7.549e-03  14.569  < 2e-16 ***
PctKidsBornNeverMar   4.017e-02  6.929e-03   5.797 7.84e-09 ***
RentMedian            5.025e-04  2.639e-04   1.904 0.057020 .
PctForeignBorn        3.658e-03  2.768e-03   1.322 0.186444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 214.4999)

    Null deviance: 1032501  on 1993  degrees of freedom
Residual deviance:  373972  on 1978  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 8. The Quasi-Poisson Model

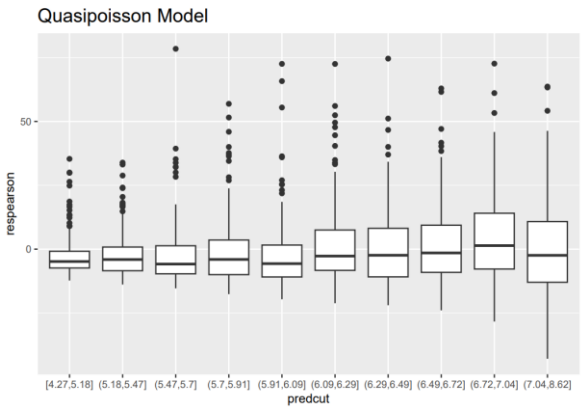

Figure 9. Residual Analysis of Quasi-Poisson Model

With this model, the problem of over-dispersion is solved with the new dispersion rate, which is 214.5 (Figure 8). What is more, some of the variables which were significant with the Poisson model, such as "racepctblack", "PctPopUnderPov, etc., are now no longer relevant since their P-value are higher than 0.5. Then, in order to check the constant residual variance assumption, we divided the predicted samples of the model into 10 intervals to plot a boxplot with Pearson residuals. Figure 9 shows a heteroskedastic distribution of residuals: we can see that the variance of residuals is increasing over the predicted values. Consequently, we moved to our last solution, the Negative Binomial Model.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.476e+00  5.083e-01   8.805  < 2e-16 ***
racepctblack      9.655e-03  4.346e-03   2.222  0.02630 *
racePctWhite     -7.829e-03  4.205e-03  -1.862  0.06261 .
racePctAsian      1.192e-02  6.333e-03   1.883  0.05976 .
racePctHisp       8.075e-03  2.616e-03   3.087  0.00203 **
population        2.207e-07  7.282e-08   3.031  0.00244 **
agePct12t29       9.332e-03  4.623e-03   2.018  0.04354 *
agePct65up        2.233e-02  5.104e-03   4.376 1.21e-05 ***
medIncome        -9.499e-06  3.594e-06  -2.643  0.00822 **
PctPopUnderPov    1.507e-04  4.183e-03   0.036  0.97127
PctNotHSGrad      2.848e-03  2.612e-03   1.090  0.27567
PctUnemployed     1.219e-02  9.895e-03   1.232  0.21779
TotalPctDiv       1.235e-01  7.453e-03  16.576  < 2e-16 ***
PctKidsBornNeverMar  5.106e-02  1.070e-02   4.774 1.81e-06 ***
RentMedian        4.183e-04  2.283e-04   1.832  0.06690 .
PctForeignBorn    2.163e-03  3.737e-03   0.579  0.56273
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2.4803) family taken to be 1)

    Null deviance: 5156.3  on 1993  degrees of freedom
Residual deviance: 2132.0  on 1978  degrees of freedom
```

Figure 10. NB Model without interactions

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       4.753e+00  5.067e-01   9.380  < 2e-16 ***
racepctblack                      4.437e-03  4.406e-03   1.007 0.313961
racePctWhite                     -1.174e-02  4.221e-03  -2.782 0.005399 **
racePctAsian                      2.732e-03  6.439e-03   0.424 0.671432
racePctHisp                       5.576e-03  2.628e-03   2.122 0.033857 *
population                        1.472e-07  7.240e-08   2.033 0.042020 *
agePct12t29                       3.961e-03  4.661e-03   0.850 0.395391
agePct65up                        1.718e-02  5.114e-03   3.360 0.000779 ***
medIncome                        -6.320e-06  3.586e-06  -1.762 0.077995 .
PctPopUnderPov                    3.220e-03  4.172e-03   0.772 0.440245
PctNotHSGrad                      1.398e-02  3.002e-03   4.657 3.21e-06 ***
PctUnemployed                     1.091e-02  9.794e-03   1.114 0.265396
TotalPctDiv                       1.143e-01  7.518e-03  15.198  < 2e-16 ***
PctKidsBornNeverMar               1.653e-01  2.001e-02   8.259  < 2e-16 ***
RentMedian                        3.738e-04  2.261e-04   1.654 0.098202 .
PctForeignBorn                    6.438e-03  3.749e-03   1.717 0.085900 .
PctNotHSGrad:PctKidsBornNeverMar -3.388e-03  4.900e-04  -6.915 4.69e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2.5328) family taken to be 1)

    Null deviance: 5264.2  on 1993  degrees of freedom
Residual deviance: 2129.7  on 1977  degrees of freedom
AIC: 27478
```

Figure 11. NB Model with interactions

Finally, with the new model, we arrived at solving the Dispersion problem as the dispersion parameter now is similar to 1 (1.07), and we can observe that the variance of residuals is more or less constant despite some atypical values.

ANOVA TEST

```
anova(nbfit,nbfit2)

## Likelihood ratio tests of Negative Binomial Models
##
## Response: ViolentCrimesPerPop
##
## 1                                racepctblack + racePctWhite + racePct
## 2 racepctblack + racePctWhite + racePctAsian + racePctHisp + population +
##      theta Resid. df    2 x log-lik.   Test    df LR stat.      Pr(Chi)
## 1 2.480299       1978       -27488.20
## 2 2.532755       1977       -27442.09 1 vs 2     1 46.11199 1.116829e-11
```

Figure 12. ANOVA test. Interaction vs Non interaction

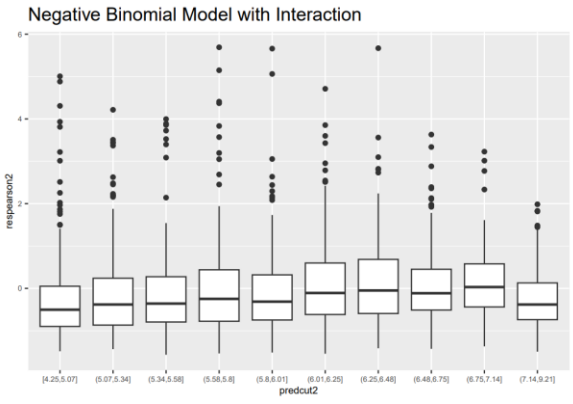

Figure 13. Residual Analysis of NB Model

With the ANOVA test, we arrived at the conclusion that the model with interactions adapts better to our project since the P-values are lower than 0.05. Accordingly, we are going to realize the rest of the analysis with the Negative Binomial Model with Interactions.

```
glm.nb(formula = ViolentCrimesPerPop ~ racepctblack + racePctWhite +
    racePctAsian + racePctHisp + population + agePct12t29 + agePct65up +
    medIncome + PctPopUnderPov + PctNotHSGrad + PctUnemployed +
    TotalPctDiv + PctKidsBornNeverMar + RentMedian + PctForeignBorn +
    PctNotHSGrad:PctKidsBornNeverMar, data = df_com, init.theta = 2.532754841,
    link = log)
```

Figure 13. R command for the Negative Binomial Regression Model

The following figure shows all the significant variables for our model. In the column of "Estimate", we can see that the percentage of white population and the interaction of percentage of non-high-school diploma obtainers with the percentage of kids whose parents never got married are negative apart from their original beta. Then, from all the other variables, the two that have most effect on the outcome are surprisingly one of the interaction variables, percentage of kid born with non-married parents, following we have the percentage of divorced couples and in the third place, non highschool graduates.

	Estimate	2.5 %	97.5 %
(Intercept)	4.753418e+00	3.738869e+00	5.771428e+00
racePctWhite	-1.174268e-02	-2.030926e-02	-3.197565e-03
racePctHisp	5.576438e-03	2.751345e-04	1.096739e-02
population	1.472201e-07	-1.681941e-08	3.598947e-07
agePct65up	1.718254e-02	7.384764e-03	2.709651e-02
PctNotHSGrad	1.397832e-02	8.070025e-03	1.987205e-02
TotalPctDiv	1.142613e-01	9.918400e-02	1.293740e-01
PctKidsBornNeverMar	1.652675e-01	1.257241e-01	2.049314e-01
PctNotHSGrad:PctKidsBornNeverMar	-3.388423e-03	-4.325612e-03	-2.435791e-03

Figure 14. Variable estimated coefficients with significant p-value

6. Conclusion

Upon conducting a comprehensive analysis and making multiple comparisons between models, it is imperative to note that the dispersion of the data is huge, even with the negative binomial model, we cannot say that the variance is 100% constant and there are actually numerous outliers that could affect the estimation. Some discoveries worth highlighting are that the effect of the percentage of people above 65 years of age is greater than that of the percentage of Hispanic people. Regarding the Poisson models with interactions, since the estimates of the interactions themselves are small, it does not provide information by interpreting the interactions. Also, the estimate of percentage of the white population is a negative value as compared to other races. We believe that additional resources or information may be necessary to undertake a more thorough investigation, culminating in a conclusive determination regarding the underlying causes of criminal cases.

References

Dataset:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

U. S. Department of Commerce, Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a (Computer Files).

U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)

U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)

U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)