Spring 2019: Statistical Analysis & Learning

Predicting Cleared Murders in the US

Lucas Arbulu (lwa275)

Julia Di Fiore (jld4865)

Timothy Michalak (tjm4224)

Vineet Raman (vmr452)

May 8, 2019

Table of Contents

[Abstract 2](#_Toc7603700)

[Introduction 3](#_Toc7603701)

[Research Question 3](#_Toc7603702)

[Data Sources 3](#_Toc7603703)

[Literature Review 5](#_Toc7603704)

[Types of Studies 5](#_Toc7603705)

[Methods 5](#_Toc7603706)

[Key Findings 5](#_Toc7603707)

[Methodology 6](#_Toc7603708)

[Data 6](#_Toc7603709)

[Results 7](#_Toc7603710)

[Subtitle 7](#_Toc7603711)

[Subtitle 7](#_Toc7603712)

[Discussion 8](#_Toc7603713)

[Conclusions 9](#_Toc7603714)

[Appendix 10](#_Toc7603715)

[References 11](#_Toc7603716)

# Abstract

1. the purpose of the study (the central question);

2. a brief statement of what was done (Methods);

3. a brief statement of what was found (Results);

4. a brief statement of what was concluded (Discussion, in part).

# Introduction

### Research Question

In the United States, homicide clearance rates vary among jurisdictions. A clearance rate is the ratio of solved crimes to unsolved ones. Richmond, VA, for example, clears approximately 74% of homicides, while Chicago clears fewer than 30%. Cities with low clearance rates pose a greater risk to innocent citizens while also denying victims and their families justice.

This study seeks to investigate quantitative and qualitative characteristics associated with whether a murder is cleared by using classification methods of statistical learning. Among these predictors are the percentage of racial minorities in a county, victim characteristics, such as race, and murder weapon used. If Hispanic victims in counties that have a population with a high percentage of white individuals are going unsolved, then we would expect there to be a relationship between race and whether or not a case is solved.

### Data Sources

**General Considerations**

Our focus is on the years 2012 – 2017. We chose this period to limit the scope of the project, and to allow for data aggregation between our two sources.

**Murder Data (National)**

The Murder Accountability Project (MAP), under the Freedom of Information Act, compiles two datasets that are housed by the Federal Bureau of Investigation: the Uniform Crime Report (UCR) from 1965 to present, and the Supplementary Homicide Report (SHR) from 1976 to present. The FBI does not collect the raw data for the UCR, rather law enforcement agencies are responsible for reporting information to the FBI on a monthly basis. The SHR, a subset of the UCR, contains more detailed information about each homicide. MAP has obtained more than 27,000 homicide cases not reported to the Justice Department, meaning this source is the most accurate and complete source of data on homicides in the U.S.

This dataset includes parameters such as whether a murder is solved or not, victim sex, race, age, and ethnicity, weapon used, location of murder (by county), agency jurisdiction (such as Municipal Police, County Police, etc.), and situation (combinations of single or multiple victims and offenders).

**Census Data (National)**

Murder data from the Murder Accountability Project was aggregated with data from the Census Bureau, specifically, the American Community Survey, which is conducted annually. Our focus is on the data collected between the years 2012 and 2017, which is the most recent five-year period available. The collected ACS data was collected at the county level from 49 US states. We decided to leave out Alaska due to having Census areas and bureaus instead of counties. We used this data to identify various parameters about each county, specifically, education rate, which is defined by percentage of population who completed high school, poverty rate, which is the percentage of the population that falls below the federal poverty level, and race/ethnicity, as defined by the percentage of the population that fall within the following categories: Hispanic, multiracial, other race, pacific islander, Asian, native American, African American, and white.

# Literature Review

### Types of Studies

In our literature review, we looked at a variety of studies spanning different time periods, locations, scope, and focus. We looked at six studies in total; two focused on single cities (Boston and Los Angeles), while the others spanned multiple cities; one study went as far back as the 1970s, some were longitudinal studies that spanned several decades, while still others were looking at an intervention between two shorter periods of three years and one year. One study focused specifically on murders of Hispanic individuals. In reviewing each of these studies, several themes emerged.

### Methods

The outcome variable fell into two main categories: clearance rate (regression) and whether or not an individual murder was cleared (classification). The study that looked at the clearance rate primarily used multiple linear regression as their model (Keel, Jarvis, and Muirhead), while those who looked at whether a murder was cleared used logistic regression (Lee; Braga & Desseault). The remaining studies utilized a combination of bivariate regression analysis (Wellford & Cronin, and Davies), Cox regression analysis (Roberts & Lyons), and finally a cross-sectional, time-series linear model analysis using feasible generalized least squares (Davies).

### Key Findings

The predictors used within each model varied, however there were many overlapping predictors among studies. The following predictors came up consistently in most or all of the studies:

* Population of the jurisdiction
* Percentage of non-white population
* Demographics of victim (race, ethnicity, age, sex)
* Weapon used

Most studies found a statistically significant impact of predictors on the outcome. Those that looked at race found that a murder was less likely to be solved if the victim was non-white. Those that looked at management and police practices found in jurisdictions where there were better practices (or in the case of the Boston study, when an intervention was introduced that improved practices), the clearance rate was higher (more murders solved).

# Methodology

### Variables Included

**Murder Parameters**

Response Variable:

The objective of our study was to predict whether a murder was solved or unsolved. The response variable was binary: Solved (“Yes”) or unsolved (“No”). This is not to be confused with the clearance rate of a case or a county.

* Solved/Unsolved: Each murder was either solved (“Yes”) or unsolved (“No”)

Predictor Variables:

* Agency Type: n-1 dummy variables for each agency type (“agentype”). We decided to exclude Municipal Sheriff from the model.
  1. Municipal Police Sheriff
  2. Primary State LE
  3. County Police
  4. Special Police
  5. Regional Police
  6. Tribal
* Situation: Each murder had certain characteristics relating to the numbers and types of victims and offenders (“Situation”). We created n-1 dummy variables for each situation. We decided to exclude Single Victim/Single Offender from the model.
  1. Single Victim/Single Offender
  2. Single Victim/Unknown Offender(s)
  3. Multiple Victims/Multiple Offenders
  4. Single Victim/Multiple Offenders
  5. Multiple Victims/Single Offender
  6. Multiple Victims/Unknown Offender(s)
* Sex: Each murder consisted of the sex of the victim(s) and offender(s) (or unknown). We created n-1 dummy variables for each. We decided to exclude male from the model.
  1. Male
  2. Female
  3. Unknown
* Race: Each murder contains the race of the victim (or unknown). We created n-1 dummy variables for each. We decided to exclude Black from the model.
  1. Black
  2. White
  3. Indian or Alaskan Native
  4. Asian of Pacific Islander
  5. Unknown
* Weapon: Each murder observation contains the weapon used. We created n-1 dummy variables for each
  1. Handgun – pistol, revolver, etc.
  2. Firearm, type not states
  3. Knife or cutting instrument
  4. Personal weapons, includes beating
  5. Other or type unknown
  6. Blunt Object – hammer, club, etc.
  7. Rifle
  8. Shotgun
  9. Asphyxiation – includes death by gas
  10. Strangulation – hanging
  11. Drowning
  12. Other gun
  13. Fire
  14. Poison – does not include gas
  15. Narcotics or drugs, sleeping pills
  16. Explosives
  17. Pushed or thrown out window

Before importing the data, we needed to install numerous packages to clean and sort the multiple data sources into a master dataset. Listed below are the various packages installed to clean the data.

|  |  |  |
| --- | --- | --- |
| Package | Purpose | Definition |
| Sqldf | SQL | To perform the SQL to determine which class to assign each column |
| FastDummies | Data Cleanup | Create dummy columns from columns that have categorical variables |
| glmnet | Statistical | Fitting lasso or elastic-net regularization path for logistic regression models |
| Caret | Statistical | Functions for training and plotting models |
| ROSE | Statistical | Deal with binary classification problems for unbalanced data |
| DMwR | Statistical | Datamining |
| Boot | Statistical | Functions and datasets for bootstrapping |
| rpart | Statistical | Recursive portioning for trees |
| randomForest | Statistical | Classification based on a forest of trees using random inputs |
| jtools | Statistical | More efficiently sharing the results of regression analysis |
| sandwich | Statistical | Model-robust standard error estimators |
| ROCR | Graphic | Tool to create performance curves (including ROC) |
| ROCit | Graphic | Performance assessment of Binary Classifier with Visualization (including precision and F-Score) |
| ggstance | Graphic | Provides flipped components |
| rattle | Graphic | Utility functions |
| rpart.plot | Graphic | Plots “rpart” models |
| RColorBrewer | Graphic | Provides color schemes |

After installing these packages, we merged multiple datasets from both MAP and ACS. We imported four datasets from the ACS 2012-2017 estimations (Race, Poverty Rate, County Income, and Education) and one dataset from MAP that indicated whether a murder was cleared or uncleared. Next, we pulled the headers from census files and defined them as strings.

After defining the strings and splitting the data into training and test data, we joined all the subsets to form a master table. However, in the master dataset, we needed to account for missing counties due to irregular naming (Parishes, Boroughs, Independent Cities). These counties were removed from the master dataset and from forming the models. After removing the missing counties, we then formed dummy variables for a number of the columns, including whether a murder was solved, agency type, the year, month, situation, the victims sex and race, and the weapon. These dummy variables were then coerced into factors.

Next, we needed to coerce certain variables as continuous variables, such as education rate (percentage of county with at least a high school degree), poverty rate, and county racial make-up.

Within our data, the proportion of solved to unsolved cases is disproportionate, in that the number of solved cases far exceeds the number of unsolved ones. To account for this disproportionality, we used a resampling technique called SMOTE. SMOTE balances the cases for each class in classification problems. This method works similarly to bootstrapping, except instead of randomly sampling existing observations, it artificially generates synthetic observations from the minority class using the nearest neighbors of cases within that class. The majority class is under-sampled to balance as well.

*Logistic regression*

The first model run on the aggregated data was a logistic regression predicting whether a murder case was solved or unsolved on all of the predictors. The logistic regressions were run on both the balanced data and the regular training data. With the training logits, we generated probabilities with the training data and balanced data and then empty tables with the unsolved outcome (which we labeled as “No”). Outputs of the logistic models that had probabilities greater than 0.5 were labeled as “Yes”, meaning that the model predicted that the case was solved. Due to concerns of multi-dimensionality, we constructed a Lasso Model using 10 K-Fold, which is the industry standards.

To test the accuracy of the logistic models (both original and lasso models), we decided to use four metrics: MSE, Precision, Recall, and F measure. Our justification for the four metrics was to account for the unbalanced data. Only using one measure to test the accuracy of the model may not give on a robust answer to the performance of the model. We used ROC curves as an additional tool to diagnose the performance of the model.

Precision: Actually Solved Cases/(Actual Solved Cases + Unsolved Cases categorized as Solved)

Recall: Actual Solved Cases/(Actual Solved Cases + Solved Cases categorized as Unsolved)

*Decision Trees*

Our second method used in the project was a decision tree, with solved as the main predictor. We decided to use the balanced data on the tree-based method to account for more information about the solved cases.

# Results

### Logistic Regression

For the 10-Fold CV logistic regression with the regular data, we found the following confusion matrix. The horizontal axis represents the whether the actual base was solved (Yes) or unsolved (No). The vertical axis indicates if the model predicted a case to be solved or unsolved. For the K-Fold with the regular training data, test MSE for the model is .7214. The precision and recall rates for the model are 0.742 and 0.92 respectively.

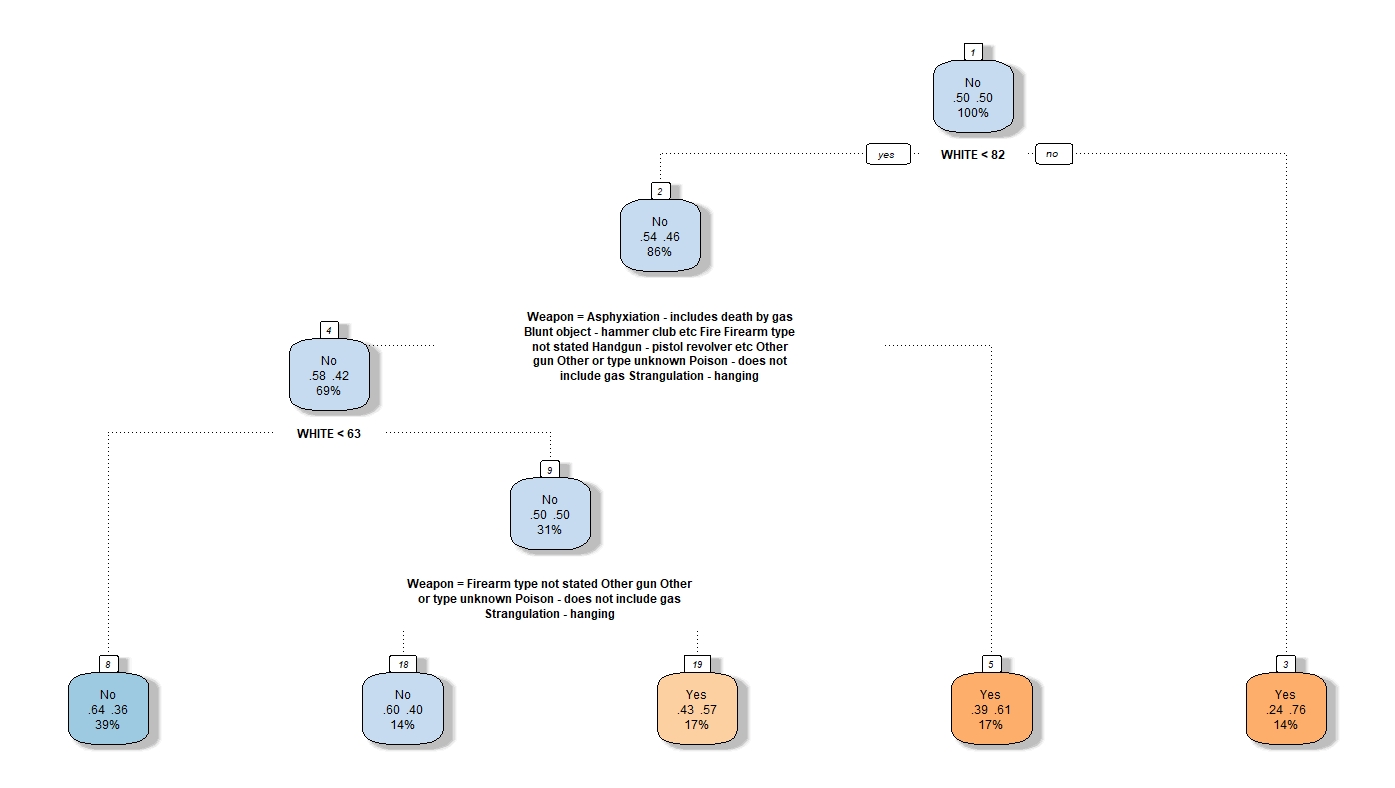
|  |  |  |
| --- | --- | --- |
| 10-Fold | Yes | No |
| Yes | 15736 | 5467 |
| No | 1364 | 1948 |

The coefficients for 14 of the variables in the original logistic regression were set to zero. Some of the removed coefficients included the proportion of the county that was African American, Asian American Victims, and murders committed by handguns. For K-Folded Lasso, the confusion matrix is provided below. The test MSE for the model is 0.627. The precision and recall rates for the lasso model are 0.796 and 0.63 respectively. As illustrated by the confusion matrix, the lasso model lowered the amount of incidences of unsolved cases categorized as solved cases. However, there was a relatively large increase in solved cases categorized as unsolved.

|  |  |  |
| --- | --- | --- |
| K-Fold Lasso | Yes | No |
| Yes | 10766 | 2751 |
| No | 6370 | 4628 |

### Decision Trees

For the unpruned classification tree, the tree was in regard to the proportion of the county that was white. If a county was more than 82% white, then the model predicts that the murder was solved. Counties with a population less than 82% white, had Asphyxiation as the next predictor. If the cause of death was Asphyxiation, then the model predicts the murder as solved. However, if the murder weapon was not Asphyxiation, the next used predictor was if the percentage of the county white was greater or less than 63%. If the proportion of the county “white” was less than 63%, the model predicts the murder to be unsolved. However, if the percentage white is greater than 63%, the final predictor is if the murder weapon was an unknown firearm or unknown poison. If a murder occurred with that type of weapon, the tree categorizes the murder as solved. Murders not committed with that type of weapon, the model predicts the model unsolved. Overall, the tree predicts 48% of all murders to be solved.



# Discussion

* See if variables of importance or trends line up with the literature
  + Point out if there is conflict and see if there is a recommendation that can be made
* Trade-offs between models
  + Based on our statistical understanding, would we expect logistic regression or trees to perform for this binary classification problem?
  + Does that line up with the results?
* Discuss how balancing data improved the tree performance
* Talk about why we are using precision and recall as performance metrics rather than Test MSE

# Conclusions

# Appendix

# References

Keel, T. G., Jarvis, J. P., & Muirhead, Y. E. (2009). An Exploratory Analysis of Factors Affecting Homicide Investigations: Examining the Dynamics of Murder Clearance Rates. *Homicide Studies*, *13*(1), 50–68. <https://doi.org/10.1177/1088767908326903>

Davies, H. J. (2007). Understanding Variations in Murder Clearance Rates: The Influence of the Political Environment. *Homicide Studies*, *11*(2), 133–150. <https://doi.org/10.1177/1088767907300751>

Keel, T. G., Jarvis, J. P., & Muirhead, Y. E. (2009). An Exploratory Analysis of Factors Affecting Homicide Investigations: Examining the Dynamics of Murder Clearance Rates. *Homicide Studies*, *13*(1), 50–68. <https://doi.org/10.1177/1088767908326903>

Lee, C. (2005). The value of life in death: Multiple regression and event history analyses of homicide clearance in Los Angeles County. *Journal of Criminal Justice*, *33*(6), 527–534. <https://doi.org/10.1016/j.jcrimjus.2005.08.002>

Roberts, A., & Lyons, C. J. (2011). Hispanic Victims and Homicide Clearance by Arrest. *Homicide Studies*, *15*(1), 48–73. <https://doi.org/10.1177/1088767910397278>

Wellford, C., & Cronin, J. (n.d.). An Analysis of Variables Affecting the Clearance of Homicides. Retrieved May 1, 2019, from <http://www.jrsa.org/pubs/reports/Clearance_of_Homicide.html>