Investigating Factors that Affect Rates of Crime in Milwaukee

Marielle Billig

Marquette University

Milwaukee, WI 53233, USA

[marielle.billig@marquette.edu](mailto:marielle.billig@marquette.edu)

James Reilly

Marquette University

Milwaukee, WI 53233, USA

[james.reilly@marquette.edu](mailto:james.reilly@marquette.edu)

Katy Weathington

Marquette University

Milwaukee, WI 53233, USA

[katy.weathington@marquette.edu](mailto:katy.weathington@marquette.edu)

**Github**: <https://github.com/mbillig/MKECrimeModel>

# Abstract

This paper is a preliminary foray into determining if and how various community descriptors can relate to crime rates in the area. The ultimate goal is to develop an interactive tool that allows the user to explore how different factors and parameters can affect the model. The purpose of this iteration is initial analysis and descriptive statistics.

# Author Keywords

Crime; Milwaukee; Regression; Police Districts; Aldermanic Districts; Parks; Communities;

# ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

# Introduction

Crime is a common concern in cities and the public consciousness. National, state, and local governments allocate a significant amount of resources dedicated to funding law enforcement. Many pre-existing biases regarding crime hinder an effective utilization and distribution of police resources. These biases manifest as ignoring certain factors or assigning arbitrary importance to others. We will examine a number of variables beyond just population, race, and poverty level and their ability to predict crime rates in Milwaukee based on criminology research.

# Literature Review

*Crime Data Mining*

Data mining and data science techniques such as clustering, classification, and entity extraction are used to learn about crime [1]. Geospatial analyses have been used to attempt an optimal redrawing of police districts in Buffalo, NY, though the expressed purposes from Buffalo Police Department was the reduction of officer workload and emergency response time rather than reduce over policing [2]. Criminologists struggle to effectively use out of the box clustering packages which take into account only the location and existence of a crime due to uncertainty about the number of clusters to choose and the significance of said cluster [3].

*Crime Indicators*

High amounts of police patrolling can reduce crime rates as much as 13% and decrease public disorder by even more [4]. We need to identify auxiliary indicators of crime, so that assignment of resources results from a data-driven decision regarding where crime is most likely to happen,

Taking advantage of the extensive research to determine what environmental factors contribute or correlate to crime rates may improve the models. A case study from Los Angeles suggests not only a correlation between the presence of liquor stores and crime, but also a possible causal relationship [5]. Casinos appear to have a mixed effect on crime rates [6]. The presence of vegetation is shown to correlate with a reduction in several forms of crime, but specifically not theft [7]. Income inequality, as measured by the GINI index, positively correlates with crime rates [8]. High levels of voter turnout have a negative correlation with crime rate [9]. Lower percent completion of high school also correlates to higher crime [10]. Using additional quantitative measurements about an area could improve the models of crime.

# Method

## Data Acquisition

The data for this project is publicly accessible through the city of Milwaukee’s crime data portal. The crime data is a subset of the data used in previous research. As a byproduct of the previous research, this crime data organized in 120 files: two files per month for a 5-year period.

The dataset including the data about each Police District was compiled manually because the City of Milwaukee’s interface required GUI inputs and outputted a PDF of the resulting data. While it would have been possible to automate this process, considering there were only seven police districts, it was faster to manually parse the data and create an ideally formatted csv.

## Data Cleaning

Data Cleaning was a surprising large endeavor considering that the data was already very processed. However, the crime data was formatted in .js files that were designed to be easily read and used by javascript web apps. This meant that a significant amount of parsing was required to be able to represent the crime data in python as a pandas DataFrame. Anticipating future needs, special cleaning was applied to the address fields to remove unnecessary characters and enable fuzzy matching. Dates and times were also converted from string to python datetime dates. To reduce the amount of time to load and initialize all the variables in python, a library called pickle was used to save the variables so that they could be loaded directly instead of recalculated in the future.

After both sets of data had been loaded, they had to be merged. In this initial stage, a simple count of all the crimes in each district was calculated. However, future work will likely include breaking this down further by year, month, and or crime type.

## Generating Descriptive Plots

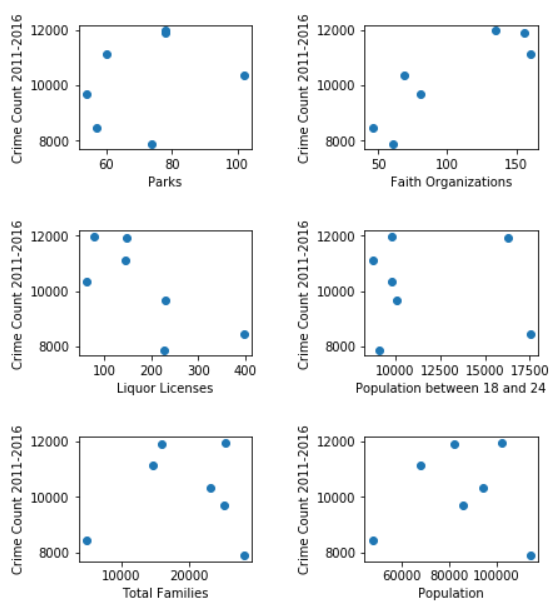
After cleaning the data, generating descriptive statistics was relatively simple. The matplotlib.pyplot library was used to generate simple scatter plots to help gain an initial understanding of the relationships between the different fields.

# Results

Overall there were 71370 records for crimes that were considered. These records were limited to either simple assault arrests or theft arrests and the date range for the data extended between January 1, 2011 and January 1, 2016.

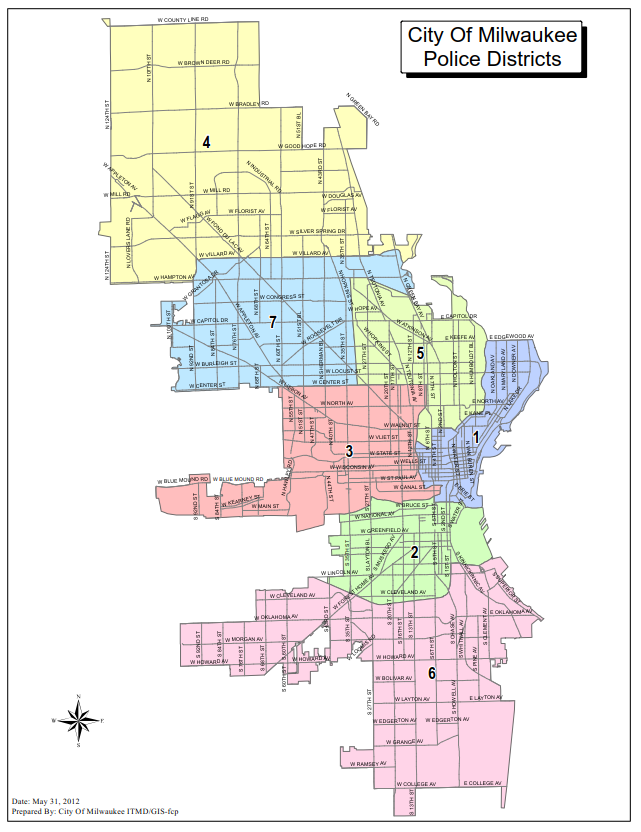
The police district data set was much smaller, with only 7 records, but this will be addressed in the future work section. Relevant fields in the dataset include the population of the district, the area, the distribution of the population ad different ages, the numbers of men and women, the total number of families, the total number of liquor licenses, faith organizations, and parks.

To gain a preliminary understand of the data, the following scatter plots were created to investigate if any of the fields had an obvious relationship with the crime count.



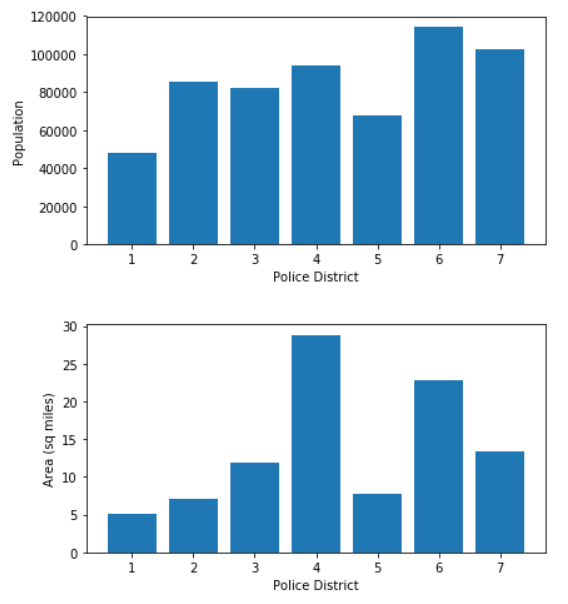
**Figure 1:** Scatter plots showing the relations between several fields and the total count of crimes in each police district

The plots in Figure 1 seem to show that some of the fields appear to have a relationship with the crime count such as faith organizations and liquor licenses. However, a closer look reveals the opposite of what might be expected as Figure 1 shows that an increase in faith organization corresponds to an increase in crime count. This observation led to the realization that the data needed to be normalized by population and area of the police districts since there is a wide variation across the different districts regarding population and area as evidenced the map of the police districts shown below.



**Figure 2:** Police District Map

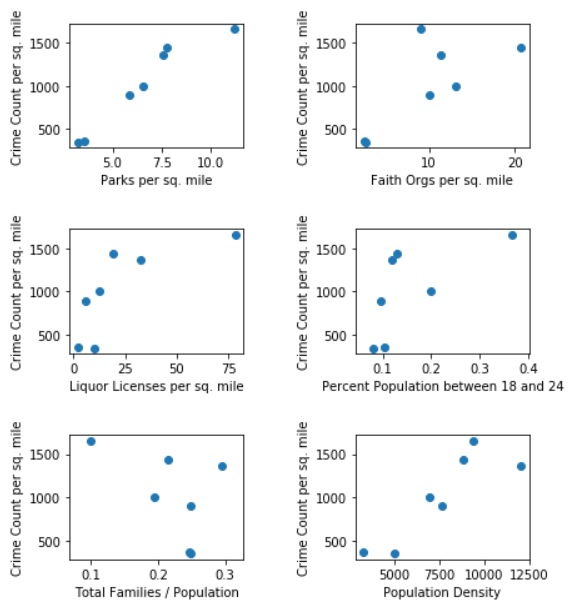
The map clearly shows that some districts, such as District Four, are much larger than others, such as District One.



**Figure 3:** Bar charts showing the distributions of population and area across the different police districts

Figure 3 reiterates the variety in both area and population of the districts. It also points towards the conclusion that is obvious to most city-dwellers that some parts of the city are more densely populated than others.

As a result of these observations, many new fields were calculated. Population, liquor licenses, faith organizations, and parks were divided by the area of the district, and the demographic breakdowns were divided by the total population. Additionally, instead of comparing the fields to the total number of crimes that had occurred in that district, the fields were compared against the crime count divided by the area (Crime density).



**Figure 4:** Scatter plots showing the relations between several fields and the crimes per square mile in each police district

These plots show that many of the fields seem to have a relationship with increase crime per square mile, but these plots are just preliminary findings and are not yet conclusive.

# Future Work

The main area for future work is developing more records than the 7 police districts. Another way of partitioning the city is into Aldermanic districts. These 15 districts are nearly equivalent in population we have nearly identical information on them as the police districts. Availability of additional per district variables to improve (or have no correlation with) crime prediction will judge the impact of Aldermanic districts.

The primary challenge with including aldermanic districts is that the crime data set does not include the aldermanic district directly. To determine the aldermanic district of a crime, the address string has to be compared to the master property list that contains information on all the properties in Milwaukee (also available from City of Milwaukee website). Understandably, this is a very large file, and this would be an expensive operation even if all the strings were perfect matches. Based on initial exploration, approximately half of crime addresses could be matched exactly to an address in the master property list. The other addresses are currently in the process of being matches using python’s difflib fuzzy matching library. This allows for the addresses that are the most similar to the address in question to be returned. A simple mode of the aldermanic districts of these similar addresses is used for the original address.

Beyond letting the aldermanic district calculation finish running, more work also has to be done regarding creating a model that related the field discussed earlier to the crime densities. It would also be interesting to break down the crime counts/densities to subset my month, time or year, to see if the models that work best overall, work best on these subsets as well and vice versa.

Finally, an interactive d3 application must be developed. This application will show which variables best predict crime by allowing the viewer to actively change which variables are considered and their relative weight.

# Acknowledgements

Special thanks to Dr. Guha for allowing us to branch from the Marquette Computational Social Science’s project. Thanks to the City of Milwaukee for making their crime data public. Thanks to Justin Miller and Griffin Berlstein for making the first steps of getting and cleaning the crime data.

# References

1. Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, and Homa Atabakhsh. 2003. Crime data mining: an overview and case studies. In Proceedings of the 2003 annual national conference on Digital government research (dg.o '03). Digital Government Society of North America 1-5.
2. Steven J. D'Amico, Shoou-Jiun Wang, Rajan Batta, and Christopher M. Rump. 2001. A simulated annealing approach to police district design. (December 2001). Retrieved March 28, 2018 from <https://www.sciencedirect.com/science/article/pii/S0305054801000569>
3. Tony H. Grubesic and Alan T. Murray. 2001. Detecting Hot Spots Using Cluster Analysis and GIS. (December 2001). Retrieved March 29, 2018 from
4. Lawrence W. Sherman and David Weisburd. 2006. General deterrent effects of police patrol in crime “hot spots”: A randomized, controlled trial. (August 2006). Retrieved March 29, 2018 from https://www.tandfonline.com/doi/citedby/10.1080/07418829500096221?scroll=top&needAccess=true
5. Bing-Ru Teh. 20017. Do Liquor Stores Increase Crime and Urban Decay? Evidence from Los Angeles (December 2017). Retrieved March 29, 2018.
6. B.Grant Stitt, Mark Nicholas, and David Giacopassi. 2003. Does the Presence of Casinos Increase Crime? An Examination of Casino and Control Communities. (April 2003). Retrieved March 28, 2018 from <http://journals.sagepub.com/doi/10.1177/0011128702251058>
7. Mary K. Wolfe and Jeremy Mennis. 2012. Does vegetation encourage or suppress urban crime? Evidence from Philadelphia, PA. (September 2012). Retrieved March 28, 2018 from <https://www.sciencedirect.com/science/article/pii/S0169204612002502?via=ihub>
8. Pablo Fajnzylber, Daniel Lederman, and Norman Loayza, "Inequality and Violent Crime," The Journal of Law and Economics 45, no. 1 (April 2002): 1-39. From <https://doi.org/10.1086/338347>
9. Stephen Coleman. 2016. A Test for the Effect of Conformity on Crime Rates Using Voter Turnout. (December 2016). Retrieved March 29, 2018 from <https://www.tandfonline.com/doi/abs/10.1111/j.1533-8525.2002.tb00049.x>
10. Lance Lochner and Enrico Moretti. 2004. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. (March 2004). Retrieved March 29, 2018 from <https://www.aeaweb.org/articles?id=10.1257/000282804322970751>