Analyzing Crime Statistics using Machine Learning techniques for Smart City Applications

Varshini Selvadurai

Poolesville High School

Poolesville, MD 20837

Aaron Gilad Kusne

101 Bureau Drove

Gaithersburg, MD 20899

Abstract

Modern cities and communities are not as effective as they could be with their services. Although a vast amount of data is collected, it isn't being utilized to distribute resources effectively. A smart community is capable of managing its resources and services solely based on data collected by sensors throughout the community. Montgomery County, MD is one of the richest counties in the United States. It is not currently a smart community. However, data collected within the community can help make Montgomery County smarter. This paper presents an analysis of possible predictors of crime based on Montgomery County's recorded crime incident database. The crime data was expected to have daily, weekly, monthly, and seasonal temporal trends, high positive correlations with spatial quantities such as population, average house prices, and urbanity. Additionally, it was hypothesized that a relationship between crime in multiple zip codes could be used to predict future crime frequencies in those zip codes. Using evidence from various data analysis methods and machine learning techniques, it was found that the crime data in Montgomery County has a weekly and daily seasonality is correlated to the level of urbanity, and that crime in certain zip codes can help predict crime in others. This project opens an avenue for further research regarding the crime patterns. In the future, these patterns can be exploited to create a safer county.

Introduction

Technological advancements lead to a safer and smarter world. The Internet of Things is a network of devices such as sensors and software that share information with each other to create a dynamic system that can be used to monitor different properties of smart cities. Smart communities are urban areas that use electronically collected data, ranging from energy usage, traffic to environmental data, to discover, interpret, and analyze the relationships between community properties that can be used to improve city resources and services and predict future patterns of those properties. For instance, electricity usage data can be used to minimize energy usage in homes and large buildings to optimally match the supply and demand of all customers without power failures and reduce cost inefficiency [1]. Traffic data can be used to optimize and regulate traffic flow by determining traffic patterns. If an anomaly, such as a sudden increase in traffic, is detected it could indicate an accident, allowing the city to allocate emergency services to that area.

Montgomery County is one of the most populated and economically well-off counties in Maryland and The United States. Although Montgomery County is not a smart community with sensors to collect and store data, the county has public records, collected by public officials, of data, including crime, that could be used to analyze county statistics. The main goal of this project will be to analyze and visualize crime data to identify similarities, patterns, and predictors of crime. Then through various machine learning techniques, the observed patterns are used to predict future crime patterns. Trends found can help allocate the necessary resources in zip codes with higher crimes rates and prevent crime. With such information, Montgomery County can become a smarter community.

In order to analyze the crime data, its properties such as location, date and time at which crime incidents occurred must be incorporated to find temporal, spatial and spatiotemporal relationships amongst the crime incidents.

Temporal trends are determined by finding relationships between temporal characteristics such as date and time and the crime data. It was assumed that crime is more likely to occur during time periods with higher interactions between people, such as rush hour, weekends, holiday months, and active seasons like summer. These recurring periods of high interactions indicate the possibility of daily, weekly, monthly and seasonal trends in crime in Montgomery County.

Spatial trends concern the relationship between crime and demographics of the location at which incidents occurred. Crime patterns could be strongly influenced some basic demographics such as population, average home prices, population density, or how urban or rural a location is.

By combining the temporal and spatial qualities, trends regarding crime over time for different locations can be assessed. The dataset records crime incidents for around 60 zip codes in Montgomery County. Creating a time series for crime incidents split by zip code incorporate the expected temporal and spatial relationships. Assuming shared temporal and spatial relationships exist as predicted, relationships between zip codes such that the number of crime incidents in one zip code can be used to predict crime incidents in another zip code can be expected.

Procedure

The overall methods used in this project included retrieving the data, preprocessing or wrangling the data, visualizing the crime statistics, and analyzing and exploring the relationships found. The intricacies of the procedures changed based on whether temporal, spatial, or spatiotemporal trends were being analyzed. Additionally, data analysis was done in Python, a interpretive high-level programming language.

Temporal Relationships

Data Retrieval

All datasets used for this project were retrieved from government data collection websites with the exception of the satellite image of Montgomery County which was downloaded from [S1] and house price information which was downloaded from [S2]. The crime data used throughout the project was found on [S3], Montgomery County's data collection website. The downloaded dataset contained information regarding over one hundred thousand recorded crime incidents from July 2016 to June 2018. Similarly, zip code boundary details (as of October 2017) were also retrieved from this website. Population information was gathered from [S4] and temperature information was found on [S5], the national center for environmental information.

Data Wrangling

The project involved the use of various sets of data in many formats. All quantitative datasets (crime, population, house price, temperature) were converted into excel files, while all geographic datasets (zip code boundary, satellite image) were downloaded as shapefiles. Unifying the data types of similar datasets made them easier to manage and apply throughout the project. The quantitative data were loaded into the program using the python pandas library, while geographic data were loaded using the python geopandas library.

Visualization

The crime data were expected to have a daily, weekly, monthly, and seasonal trends. In order to visualize daily and weekly trends, the data were grouped by date and hour and the number of crimes per hour or day was plotted. In order to clearly see trends throughout a day or week, crime data over multiple random two week periods were initially looked at. In order to visualize monthly and seasonal trends, the data were grouped together by date. The frequencies were plotted for the year of 2017 along with the average temperature data per day. The temperature dataset included the daily high and low temperature. These values were averaged and compared with the crime data as seasonal trends would depend on temperature changes. After apparent trends were identified, they were quantified in order to verify their existence.

Relationship Analysis

In order to quantify the suggested seasonalities, standard seasonality identifiers such as periodograms and autocorrelation were used. Periodograms use the Fourier transformation to decompose a time series into its equivalent sine and cosine components. The periodogram plots the spectrum density value at each frequency. High spikes in the plot indicate the most prevalent frequencies in the time series, which help determine the seasonality. Autocorrelation plots determine randomness by comparing a time series to itself at varying time lags. If a seasonality exists, then the autocorrelation plot will peak and fall according to a constant period. This period indicates the trend at which the autocorrelation plot suggest a high correlation, meaning there is a seasonality determine by the said period.

Spatial Relationships

Data Retrieval & Data Wrangling

Data were collected and processed as mentioned with the temporal relationships.

Visualization

The effects of various demographics should be considered when analyzing crime data as there is a possibility that these characteristics play a role in increasing crime. Since spatial trends rely on zip codes, choropleth maps of Montgomery County, split by zip codes, were used. Choropleth maps for each demographic were compared to the choropleth map of the crime count (based on incidents that occurred in 2017) to determine which demographics should be further investigated.

The level of how urban or rural a place is was not retrieved from a dataset, instead, it was calculated by processing a satellite image of the county and determining the gray ratio of each zip code. The satellite image was processed using MATLAB, a matrix based programming language, and the values of each pixel was collected. The pixel values were then categorized based on which zipcode they fall in and the gray ratio was calculated using the following formula:

$$\frac{Urban}{Rural} = Gray\ Ratio = \frac{\Sigma\ Pixel\ values\ \geq 100}{\Sigma\ Pixel\ values\ < 100}$$

The gray ratio values were then used as the representative of how urban or rural a zip code is.

Relationship Analysis

Demographics that showed a relative relationship with the crime frequency were plotted against the crime frequency. Using linear regression, a line was fit to the data points, and Pearson's r coefficient was calculated after outliers were removed from the data. Demographics with high r values were said to be useful in determining the level of crime occurrences in zip codes.

Spatiotemporal Relationships

Data Retrieval & Data Wrangling

Data were collected and processed as mentioned with the temporal relationships.

Visualization/Relationship Analysis

The spatiotemporal relationship analyzed was the possibility of using the time series of crime frequencies in one zip code to predict crime frequencies in another. When given two zip codes, such as the ones represented in Figure 1a, it is difficult to determine whether one is similar to the other. Therefore, in order to verify the possibility of such a relationship existing, the Granger causality was used. The Granger causality is a statistical test that is used to determine whether one time series can be used to improve forecasting of another. The Granger test will be run between every pair of zip codes in Montgomery County. If the p-value is less than 0.05, then that zip codes will be used. Additionally, only zip codes with total crime over 100 will be used because zip codes with less crime have no evident time-related patterns that can be used for prediction as shown in Figure 1b.

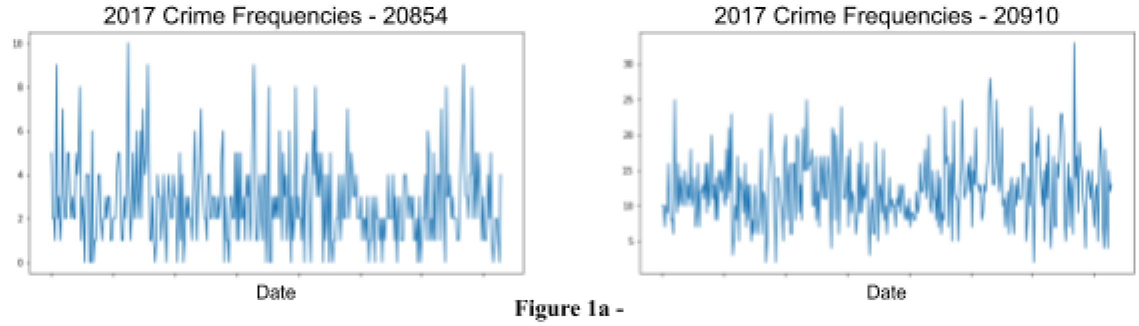


Figure 1a -
The human eye is incapable of finding similarities between these two time series simply by looking at them. This creates the need for the granger test.

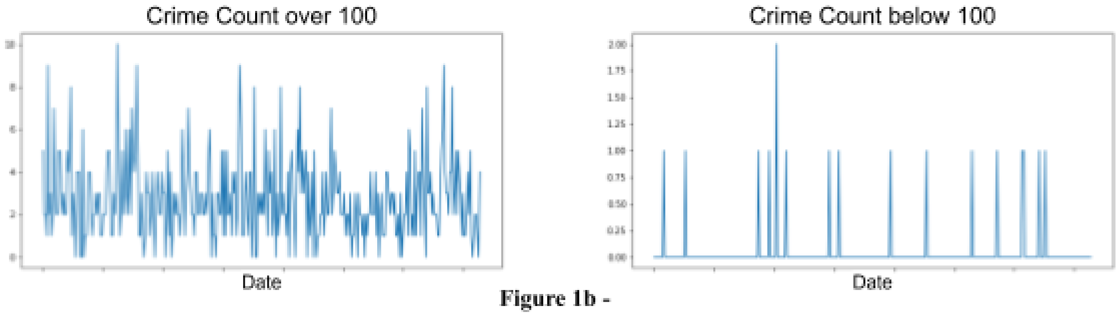


Figure 1b -
If data from the zipcode with a crime count below 100 is used, most of the predictions will be at or near 0, making the models unfairly advantageous.

Machine Learning

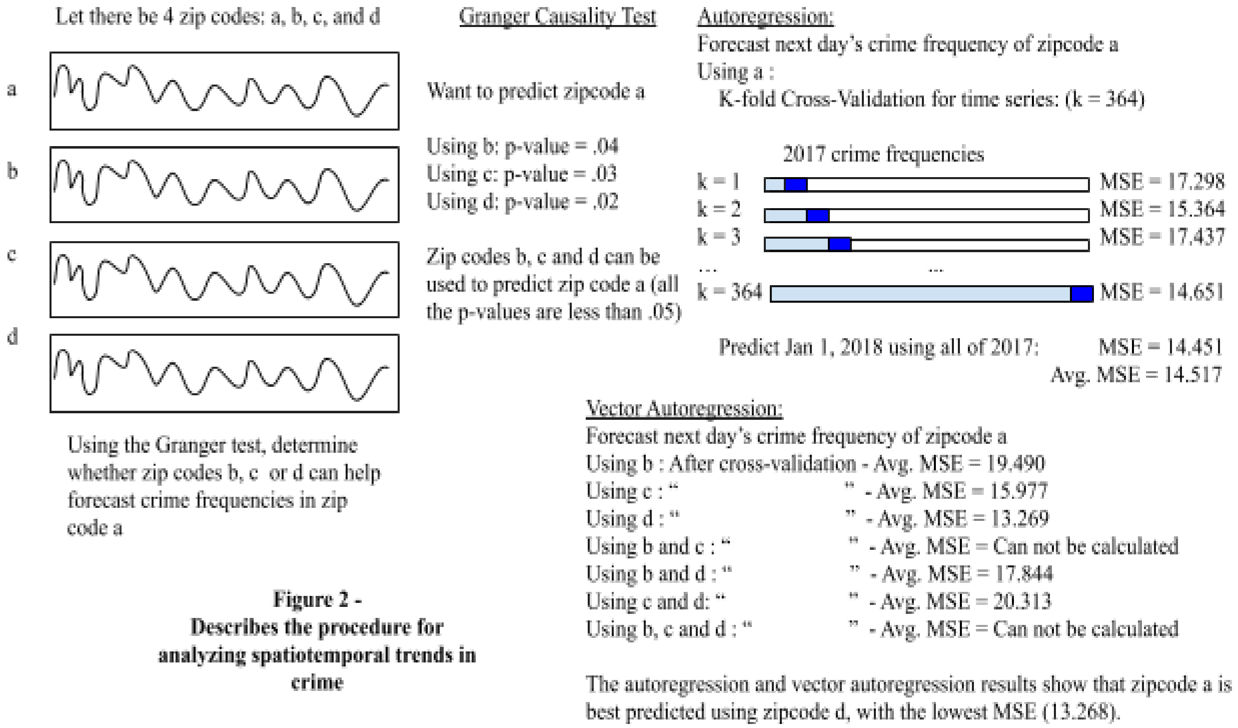

Let there be 4 zip codes: a, b, c, and d
a
b
c
d
Using the Granger test, determine whether zip codes b, c or d can help forecast crime frequencies in zip code a
Granger Causality Test
Want to predict zipcode a
Using b: p-value = .04
Using c: p-value = .03
Using d: p-value = .02
Zip codes b, c and d can be used to predict zip code a (all the p-values are less than .05)
Autoregression:
Forecast next day's crime frequency of zipcode a
Using a :
K-fold Cross-Validation for time series: (k = 364)
2017 crime frequencies
k = 1
MSE = 17.298
k = 2
MSE = 15.364
k = 3
MSE = 17.437
...
...
k = 364
MSE = 14.651
Predict Jan 1, 2018 using all of 2017:
MSE = 14.451
Avg. MSE = 14.517
Vector Autoregression:
Forecast next day's crime frequency of zipcode a
Using b : After cross-validation - Avg. MSE = 19.490
Using c : "   " - Avg. MSE = 15.977
Using d : "   " - Avg. MSE = 13.269
Using b and c : "   " - Avg. MSE = Can not be calculated
Using b and d : "   " - Avg. MSE = 17.844
Using c and d: "   " - Avg. MSE = 20.313
Using b, c and d : "   " - Avg. MSE = Can not be calculated
The autoregression and vector autoregression results show that zipcode a is best predicted using zipcode d, with the lowest MSE (13.268).

Figure 2 -
Describes the procedure for analyzing spatiotemporal trends in crime

To forecast the crime frequencies of a zip code, it is possible to use previous data from the zip code to forecast itself using a machine learning technique called autoregression. A similar technique called vector autoregression can be used to forecast a time series using one or more time series along with itself. Both the autoregressive and vector autoregressive models were created using the 2017 data and tested on the 2018 data. The models were used to predict the first day of crime frequencies in 2018. Then the mean squared error (MSE) was calculated using the predicted frequencies and the actual frequencies. However, to ensure these models are consistent and can be used for prediction, a k-fold cross-validation testing was completed on the 2017 data using both the autoregressive and vector-autoregressive models. Using a k value of 364, the data was split into 364 parts, each part with 1 data point. For the first iteration, the first split (Jan 1, 2017's data) was used to predict the next day (Jan 2, 2017). The second iteration used the first 2 data splits (Jan 1 and 2, 2017) to predict the third day (Jan 3, 2017), and so on. For each iteration the MSE of the prediction was calculated and at the end the average of the cross-validation values and the original MSE (for 2018 data) was calculated. The model, AR or VARs, that produced the lowest mean squared error was deemed the best model. This procedure is outlined in Figure 2.

Results

<u>Temporal Relationships</u>

Visualization

For the daily and weekly trend visualization, three random weeks from the year of 2017 were used: January 15 - 28, March 12 - 25, and May 7 - 20. Figure 3 shows the hourly frequency of crime for March 12 - 25. The gridlines mark the start of a new day in the two week period. The time series shows a consistent rise and fall in crime (Figure 3.A) throughout the day suggesting a daily trend. Additionally, the spike in crime near the end of the week (Figure 3.B) suggests a seasonal trend. Similar trends were identified when looking at the January 15 - 18 and May 7 - 20 data. However, these trends have only been visually identified and must be quantified.

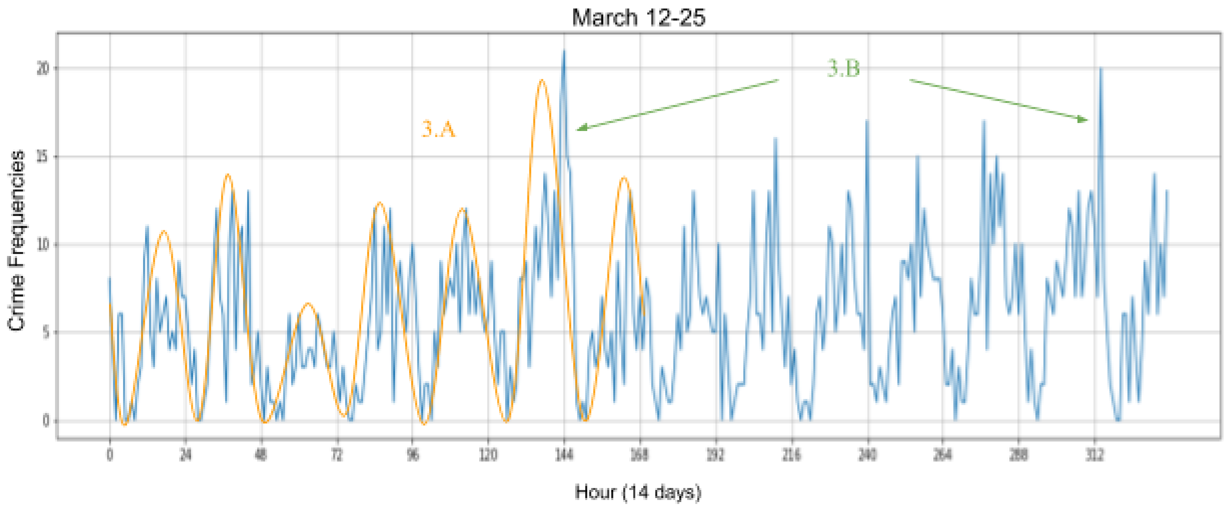


Figure 3-
Using crime frequencies from March 12-25, 2017 to identify daily and weekly trends in crime

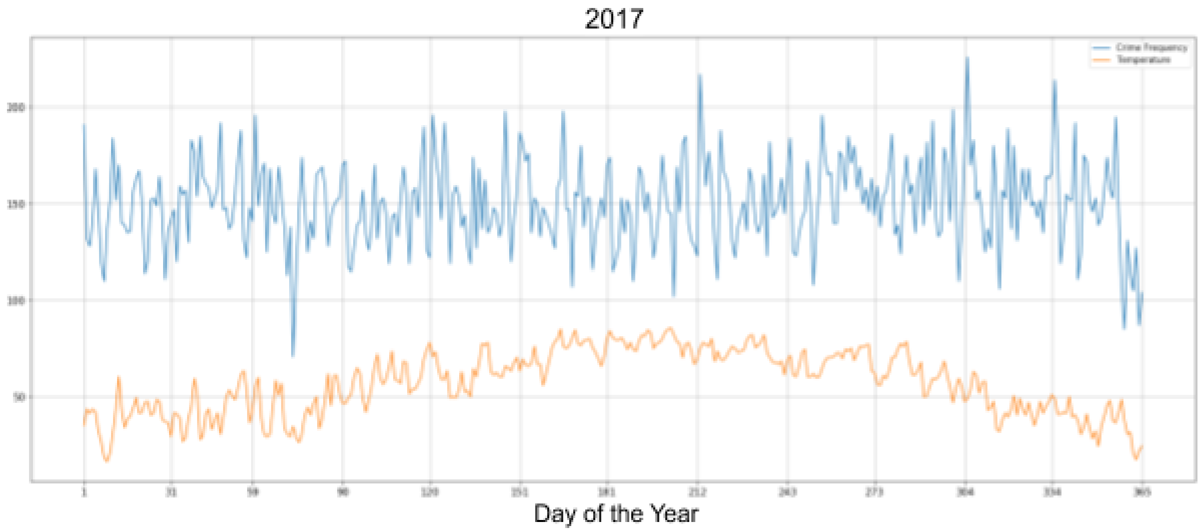


Figure 4 -
Using crime frequencies from 2017 to identify monthly and seasonal trends in crime. The blue line indicates the crime frequency per day throughout 2017 and the orange line represents the average daily temperature for the year of 2017.

Figure 4 shows the daily frequency of crime for the year of 2017 along with the average temperature. The gridlines mark the start of a new month. The time series shows no apparent trends of rises or falls in crime throughout the months. Seasonal trends would depend on the temperature, however, Figure 4 shows no relationship between temperature and crime frequencies. Since neither monthly nor seasonal trends are visually apparent, they will not be further analyzed.

Relationship Analysis

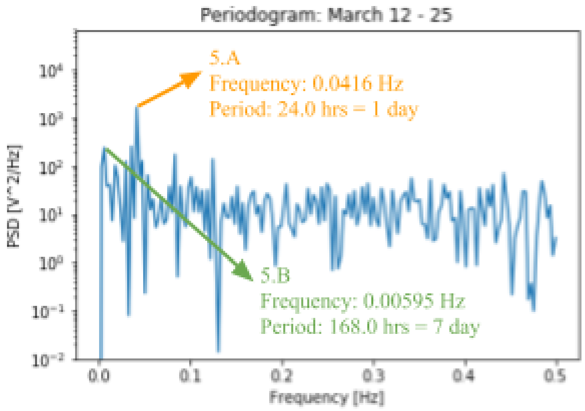


Figure 5-
Periodogram based on March 12 - 25, 2017 crime frequency data to identify daily and weekly seasonality

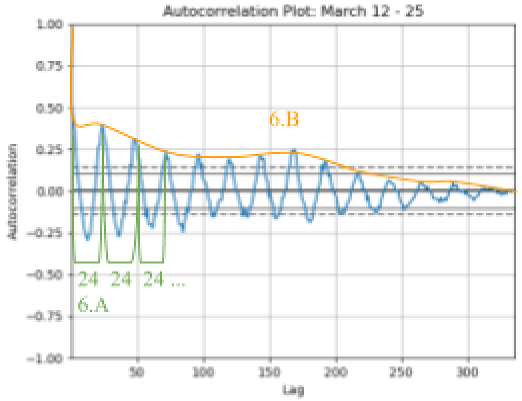


Figure 6 -
Autocorrelation plot based on March 12 - 15, 2017 crime frequency data to identify daily and weekly seasonality

The periodogram of crime data from March 12 - 25 (Figure 3), shown in Figure 5 calls attention to two prevalent frequencies. The first, 5.A, shows the max spectral density at a frequency of 0.0416 Hz or at a period of 24 hours. The second, 5.B, shows a high spectral density at a frequency of 0.00595 Hz or at a period of 168 hours (7 days). The autocorrelation plot in Figure 6 should reinforce the periodograms findings. A peak occurs every 24 lags (6.A), where each lag represents an hour. Generally, the autocorrelation of a time series continuously reduces as the lag increases. However after 7 peaks (6.B), or 7 periods of 24 hours, at the lag of 168, there is a rise in autocorrelation. Similar periodograms and autocorrelation plots were created on the January 15 - 28, May 7 - 20, 2017, and original datasets in order to further prove that the identified patters were not coincidences. Analysis of all the datasets resulted in the confirmation of consistency between other datasets and the March 12-25 dataset.

Spatial Relationships

Visualization

Using Montgomery County zip code boundaries, multiple choropleth maps were created to visualize demographics of the zip codes. Figure 7 features a choropleth map of the frequency of crime per zip code. Each demographic choropleth map was compared to Figure 7 to determine the strength of the possible correlation between the demographic and crime.

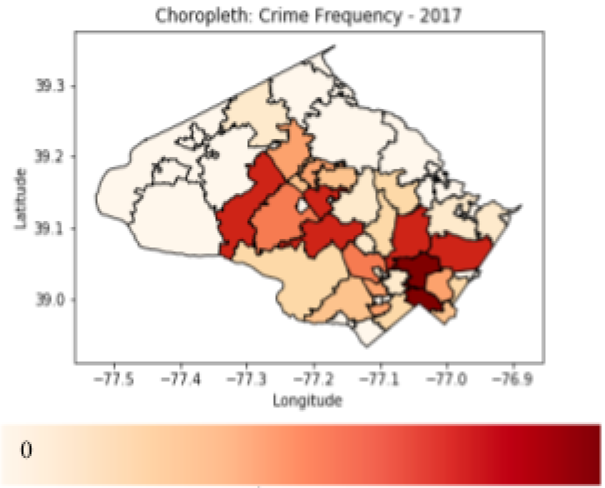


Figure 7 -
Choropleth map used to plot crime frequencies per zip code in 2017. The color spectrum has been split into 10 ranges in which lighter colors represent lower crime frequencies and darker colors represent higher crime frequencies

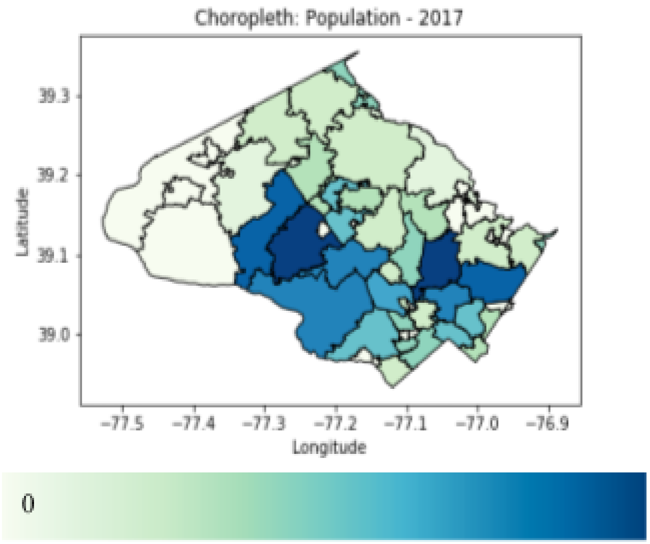


Figure 8 -
Choropleth map used to plot population per zip code in 2017. The color spectrum has been split into 10 ranges in which lighter colors represent lower populations and darker colors represent higher poulations

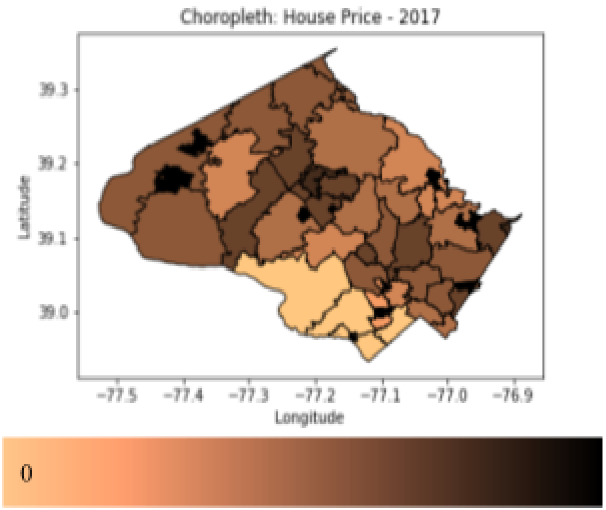


Figure 9 -
Choropleth map used to plot average house price per zip code in 2017. The color spectrum has been split into 10 ranges in which lighter colors represent lower house prices and darker colors represent higher house prices

Figure 8 features a choropleth map of population per zip code. Comparing Figure 7 and 8 shows a possible relationship between crime and population since zip codes are color-coded similarly (based on each color scale).

When comparing average house prices per zip code in the year of 2017, Figure 9, to the crime frequencies, Figure 7, it can be noticed that zip codes are not colored similarly. However,

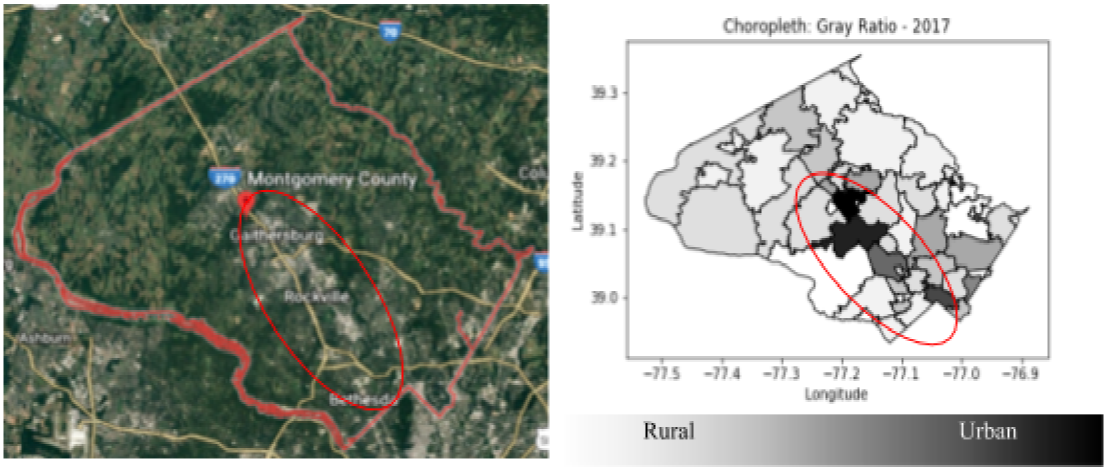


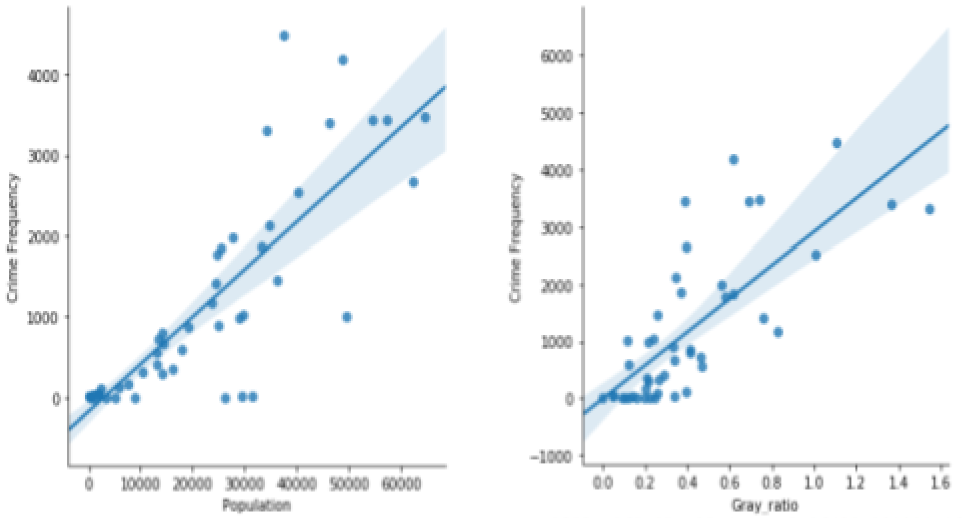
Figure 10 -
On the left is a satellite image of Montgomery County, MD. This image was used when determining the gray ratio values per zip code represented in the choropleth map on the right. Like the other choropleth maps mentioned previously, this color spectrum was split into 10 ranges where lighter colors represent more rural areas and darker colors represent more urban areas. Notice the regions marked by the red indicators. The areas in the satellite image with high concentrations of white (representing cement and urban areas) have gray ratio values in the higher end of the color spectrum.

quantifying this relationship will provide stronger evidence that there is no or little correlation between crime and house prices.

Figure 10 feature the choropleth map representing the gray ratio. Comparing Figure 10 and 7 shows a possible relationship between how urban or rural a place is and crime.

Relationship Analysis

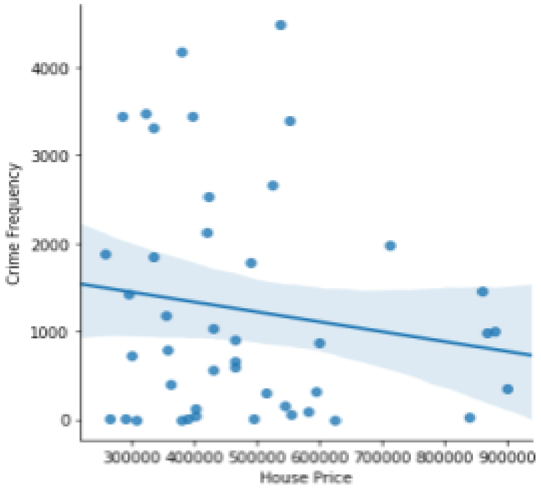
Using linear regression the visualized correlation between, population, house price and the gray ratio and crime frequency were confirmed. Figure 11 shows the relationship between population and crime frequency. The relationship has an r-value of .83, suggesting a strong linear correlation. Although the population may not cause crime it may be used to predict it. In Figure 11, by comparing the gray ratio to the number of crime incidents in every zip code a strong linear correlation with an r-value of .76 was found. This suggests that there may be a difference in crime level in rural versus urban areas.



```
y = 0.05867432182141776x + -175.29830531304515      y = 2923.5369722770274x + -8.364907224331546
r =  0.833722230541378                              r =  0.7561684539736295
```

Figure 11 -
The linear regression on the left compares population (as of the 2010 census) of a zip code against crime frequency of that zipcode. The linear regression on the right compares the calculated gray ratio of a zip code against the crime frequency of that zip code.

Figure 12 shows the lack of a relationship between the average house prices and crime frequency. The r-value -0.15 confirms that no relationship exists between the house prices and frequencies.



```
y = -0.001123282393845748x + 1783.0240899362834
r =  -0.15173644795724817
```

Figure 12 -
The linear regression compares the average house price in a zip code to the crime frequency of that zip code in the year 2017

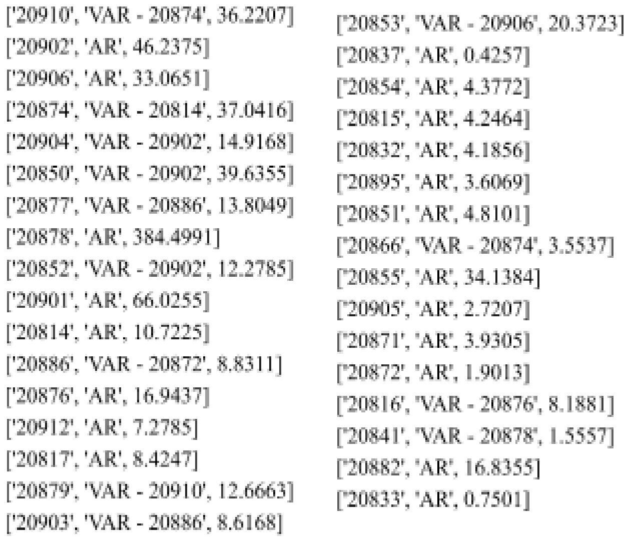
Spatiotemporal Relationships

Visualization/Relationship Analysis

There are 60 zip codes recorded in the crime dataset. After removing the zip codes with under 100 crime incidents for the year of 2017, 33 zip codes remained. The Granger causality test was run on every pair of these 33 zip codes, and of the 1056 pairs, 282 pairs passed the test. The relationship between these zip codes could be exploited to predict their crime frequencies.

Machine Learning

For each zip code, a cross-validation of the autoregression and vector autoregression were used to determine the best model for forecasting.



```
['20910', 'VAR - 20874', 36.2207]
['20902', 'AR', 46.2375]
['20906', 'AR', 33.0651]
['20874', 'VAR - 20814', 37.0416]
['20904', 'VAR - 20902', 14.9168]
['20850', 'VAR - 20902', 39.6355]
['20877', 'VAR - 20886', 13.8049]
['20878', 'AR', 384.4991]
['20852', 'VAR - 20902', 12.2785]
['20901', 'AR', 66.0255]
['20814', 'AR', 10.7225]
['20886', 'VAR - 20872', 8.8311]
['20876', 'AR', 16.9437]
['20912', 'AR', 7.2785]
['20817', 'AR', 8.4247]
['20879', 'VAR - 20910', 12.6663]
['20903', 'VAR - 20886', 8.6168]
['20853', 'VAR - 20906', 20.3723]
['20837', 'AR', 0.4257]
['20854', 'AR', 4.3772]
['20815', 'AR', 4.2464]
['20832', 'AR', 4.1856]
['20895', 'AR', 3.6069]
['20851', 'AR', 4.8101]
['20866', 'VAR - 20874', 3.5537]
['20855', 'AR', 34.1384]
['20905', 'AR', 2.7207]
['20871', 'AR', 3.9305]
['20872', 'AR', 1.9013]
['20816', 'VAR - 20876', 8.1881]
['20841', 'VAR - 20878', 1.5557]
['20882', 'AR', 16.8355]
['20833', 'AR', 0.7501]
```

Figure 13 -
Results of Machine Learning; The results are formed such that the list contains the zip code being forecasted, followed by the model that resulted in a smaller MSE. If VAR resulted in a better model, the zip codes used in prediction are listed.

Figure 13 shows the results of the AR and VAR model comparisons.

Conclusion

Temporal Relationships

Temporal analysis of the crime frequency data confirmed the existence of a daily and weekly trend in crime. The high spectral density at a frequency of 24 hours in the periodogram (Figure 5, 5A) along with the recurring peaks of the autocorrelation plot every 24 hours (Figure 6, 6A) confirm a cyclic fall and rise in crime based on the 24-hour cycle. The spectral density at 168 hours in the periodogram (Figure 5, 5B) and the unusual rise of the autocorrelation plot after 168 hours (Figure 6, 6B) suggest a pattern based on a 7 day cycle. It is important to note that the weekly trend shows a rise in crime near the end of the week (Friday and Saturday). With this knowledge, public safety officials can increase resources and surveillance on certain days and at certain times based on the recorded trends.

Spatial Relationships

Spatial analysis was able to confirm a strong correlation between population and crime frequency and gray ratio and crime frequency. The high Pearson’s R-value found using linear regression suggests that population and gray ratio can be used to determine the level of crime in each zip code. On the other hand the low Pearson’s R-value for the relationship between house price and crime frequencies suggest that there is no relationship. However the information regarding the population and level of urbanity can help redistribute public safety officials to more populated and urban areas for maximized efficiency.

Spatiotemporal Relationships

The spatiotemporal analysis confirms the existence of relationships between crime frequencies in different zip codes. The cross-validated autoregression and vector autoregression show the possibility of predicting crime frequencies in one zip code using past crime data from various zip codes. However, the data used in this analysis is too small to confirm the accuracy of the predictions. Although the data isn’t conclusive, this opens avenues to explore such spatiotemporal trends between zip codes in the count and in a couple of years, the information can be exploited to predict crime in Montgomery County.

Data Sources

[S1] https://www.google.com/maps/

[S2] https://www.zillow.com/research/data/

[S3] https://data.montgomerycountymd.gov/

[S4] https://www.census.gov/

[S5] https://www.ncdc.noaa.gov/

References

[1] Sun, M., Wang, Y., Strbac, G., & Kang, C. (n.d.). *Probabilistic Peak Load Estimation in Smart Cities Using Smart Meter Data*. Retrieved from IEEE website: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8304659

[2] Sun, Y., Song, H., Jara, A. J., & Bie, R. (2016, February). *Internet of Things and Big Data Analytics for Smart and Connected Communities*. Retrieved from iEE Xplore website: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7406686

[3] U.S. Census Bureau (2016, June). *QuickFacts: Montgomery County, Maryland* Retrieved June 10, 2018, from https://www.census.gov/quickfacts/geo/chart/montgomerycountymaryland/PST045216

[4] Woyke, E. (2018, March/April). A smarter smart city. *MIT Technology Review*, *121*(2). Retrieved from https://www.technologyreview.com/s/610249/a-smarter-smart-city/