Research Plan: Analyzing Crime Statistics using Machine Learning techniques for Smart City Applications

Varshini Selvadurai

Poolesville High School

Poolesville, MD 20837

Aaron Gilad Kusne

National Institute of Standards and Technology

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 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.

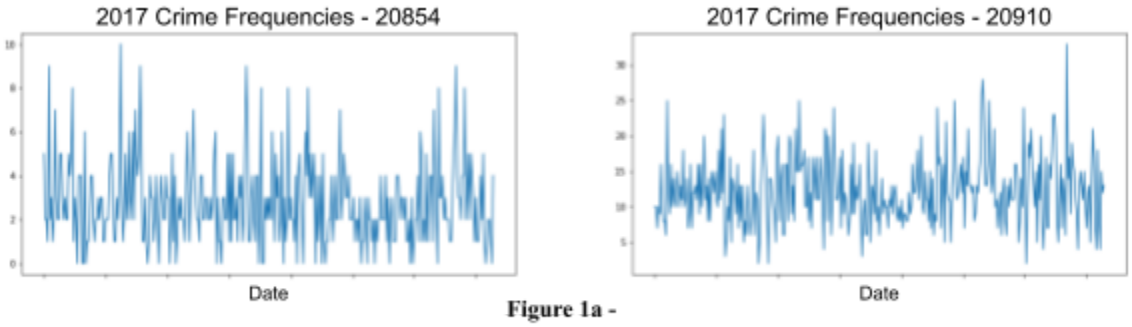

Figure 1a -
The human eye is incapable of finding similarities between these two time series simply by looking at them. This creates the need for the granger test.

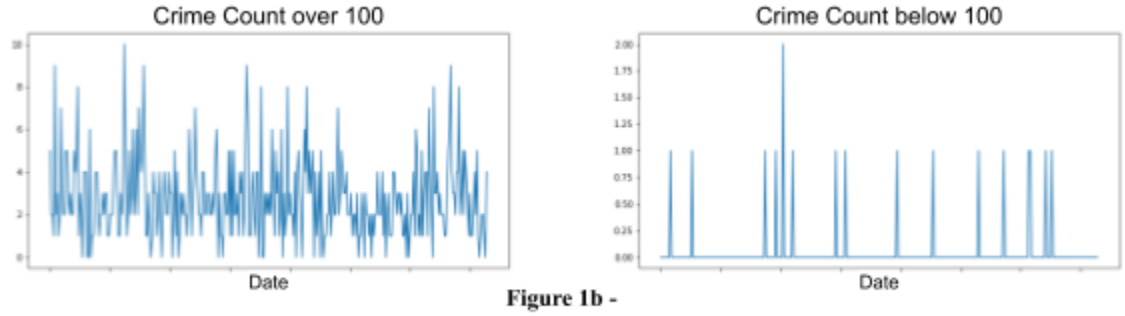

Figure 1b -
If data from the zipcode with a crime count below 100 is used, most of the predictions will be at or near 0, making the models unfairly advantageous.

Machine Learning

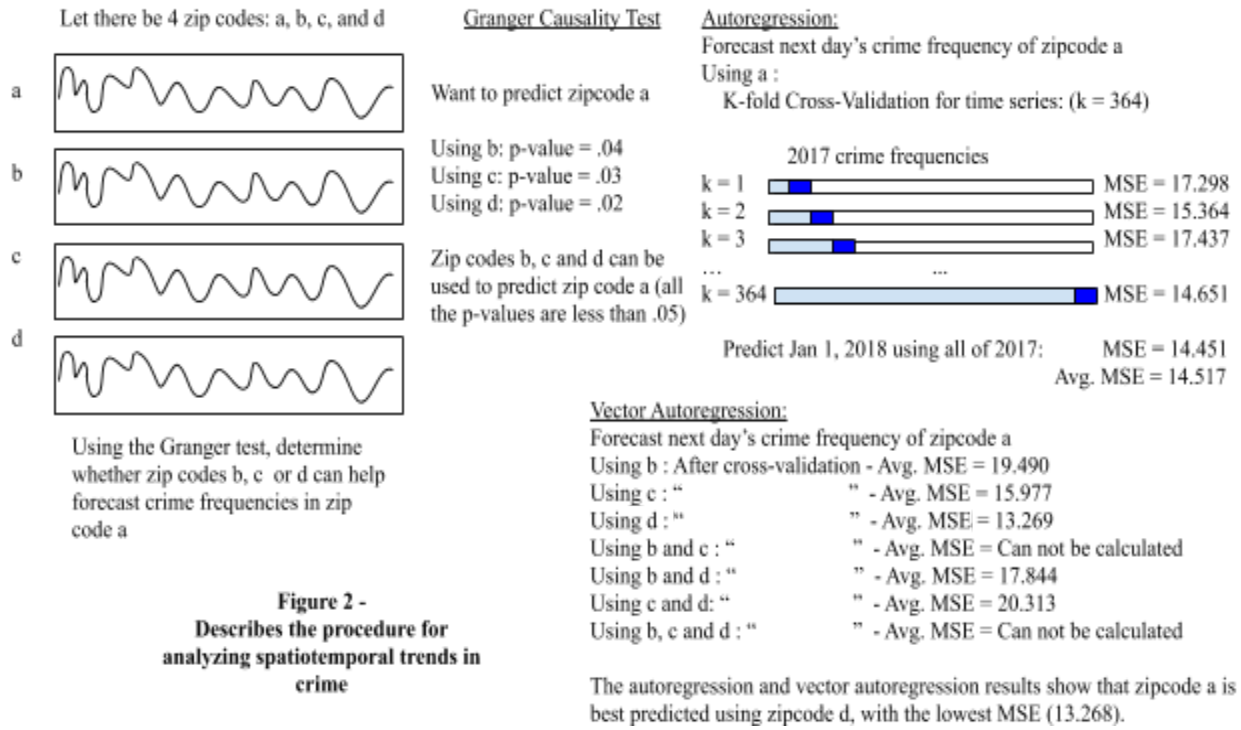

Figure 2 -
Describes the procedure for analyzing spatiotemporal trends in crime

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 the VARs, that

produced the lowest mean squared error was deemed the best model. This procedure is outlined in Figure 2.

Expected Results

Temporal Relationships: I expect the prominence of both a weekly and daily trends.

Spatial Relationships: We expect to see that crime has a strong correlation with both house price and how urban or rural the zip code is.

Spatiotemporal Relationships: Although we expect to see zip codes result in possible zip codes for prediction, we aren't sure that the results will be conclusive due to the lack of data.

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/