Analyzing Crime Statistics for Smart City Applications

Varshini

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 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, but it is not currently a smart community. However, data collected within the community can help make Montgomery County smarter without the use of sensors. This project presents an analysis of possible predictors of crime based on Montgomery County's recorded crime incident database. The crime data was expected to have both daily and weekly temporal trends and high positive correlations with spatial quantities such as 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

Smart Communities

Technological advancements lead to a safer and smarter world. The Internet of Things (IoT) 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 in smart cities. Smart cities are urban areas that use electronically collected data to interpret and analyze similarities in data that can be used to improve city resources, services and predict future patterns. This data can range from energy usage, traffic to environmental data.

An effective community should know how resources should be spread throughout the community. For this reason unconventionally researched data such as crime should be looked at to determine the spread of police stations and officers. Effectively distributing resources will save money and time for the county.

Montgomery County

Montgomery County is the most populated and economically well-off county in Maryland. Although Montgomery County is not a smart community with sensors to collect and store data, the county has public records of data that could be used to analyze county statistics. Using machine learning techniques to analyze the crime data, relationships can be found between crime and zip codes. These relationships can help determine future crime patterns and 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.

[1]

METHODS

- **Retrieve Data** • Identify pertinent data
- **Data Wrangling** • Unify diverse data types
- **Visualization** • Investigate data to identify qualitative trends
- **Relationship Analysis** • Quantify found trends
- **Machine Learning** • Predict future trends

Retrieve Data & Data Wrangling

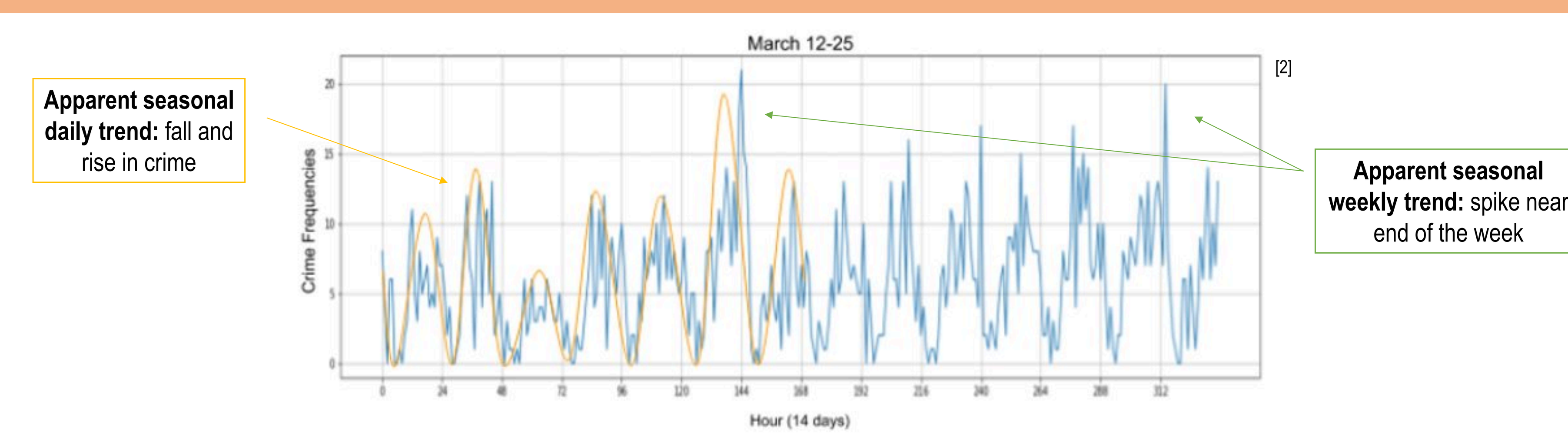
Data	Source	Downloaded As
Crime	Montgomery County Database	Excel
Zip Code Boundary	Montgomery County Database	Shapefile
Temperature	National Center for Environmental Information	Excel
Montgomery County Satellite Image	Google Maps	Shapefile
House Price	Zillow	Excel
Population	Census	Excel

- Quantitative Data → Excel
- Geometric Data → Shapefile
- Uploaded to Python

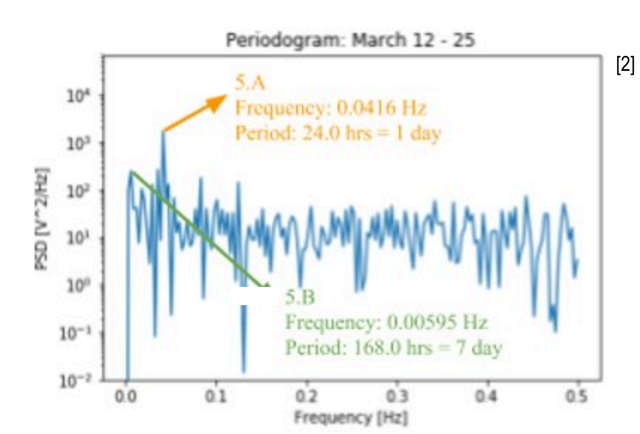
Visualization & Relationship Analysis

Temporal Relationships

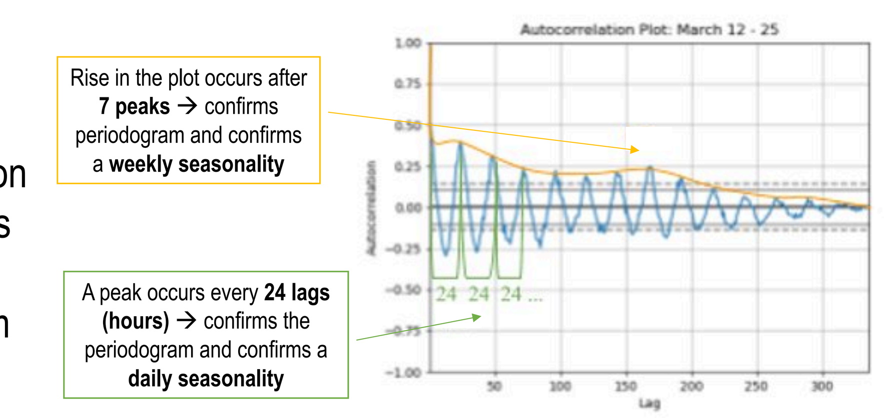
Interested in: Crime trends over time
Expected: Daily & Weekly trends
Confirmed: Daily and Weekly seasonality



The periodogram shows a the highest peak at 24 hours, meaning it is the more prevalent period. The next highest period is at 168 hours confirming a period of 7 days.

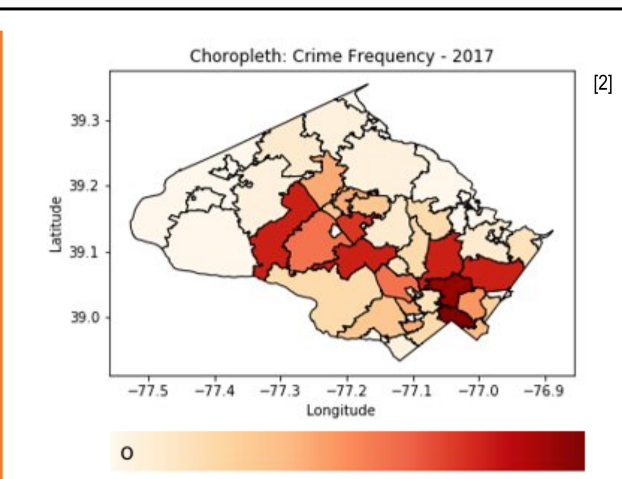


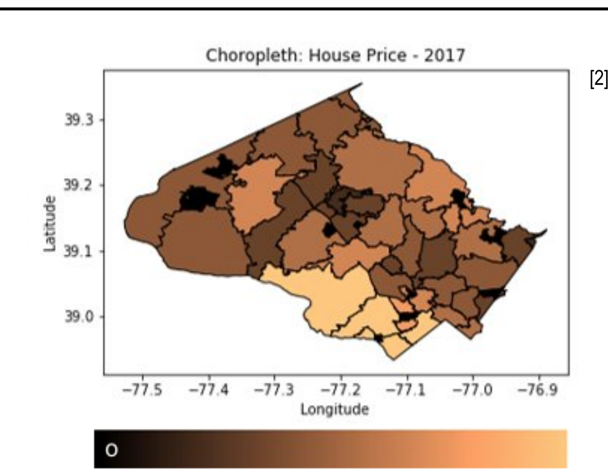
The autocorrelation plot confirms the periodogram

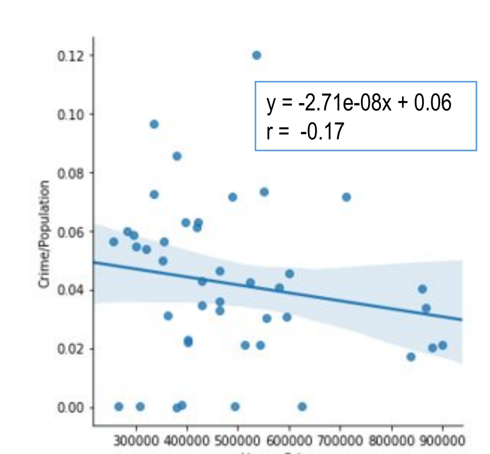


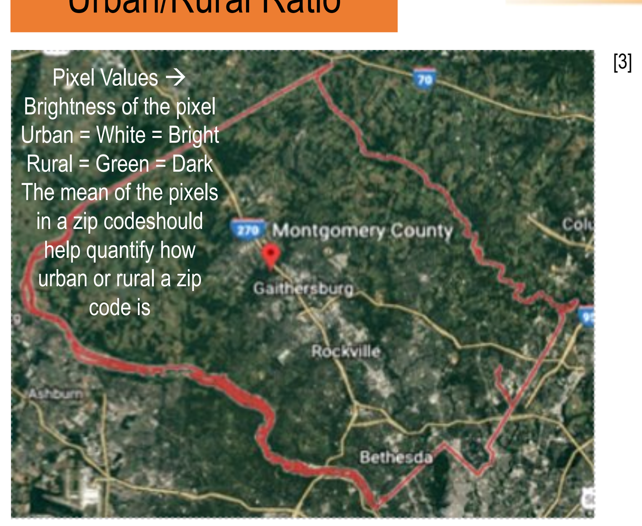
Spatial Relationships

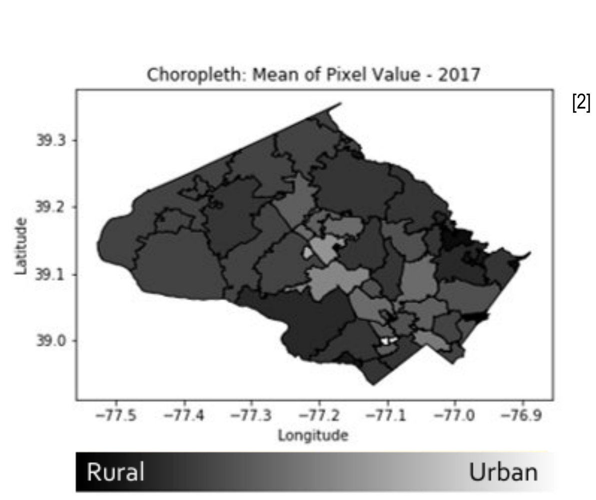
Interested in: Crime trends based on demographics of different zip codes
Expected: Urban/Rural Ratio, Avg. House Price
Confirmed: Urban/Rural Ratio

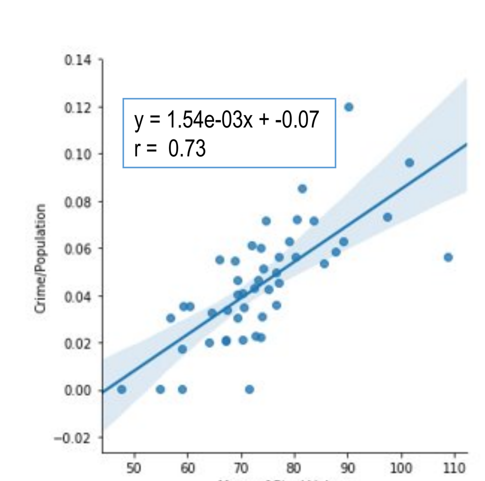












The r-values for Crime Frequency/Population verses House Price and Mean Pixel Value shows that crime has no correlation with house price and a strong correlation with the level or urbanity.

MACHINE LEARNING

Spatiotemporal Relationships

Interested In: Combing spatial and temporal qualities to assess trends regarding crime over time for different zip codes

Expected: The number of crime incidents in one zip code can be used to predict crime incidents in another zip code

Granger Causality: A statistical test that is used to determine whether one time series can forecast another

- 60 zip codes → 33 zip codes after removing ones that can't be used
- Conducted the Granger Causality test for every combination of the 33 zip codes

Autoregression(AR): Forecasts data using previous data from the time series

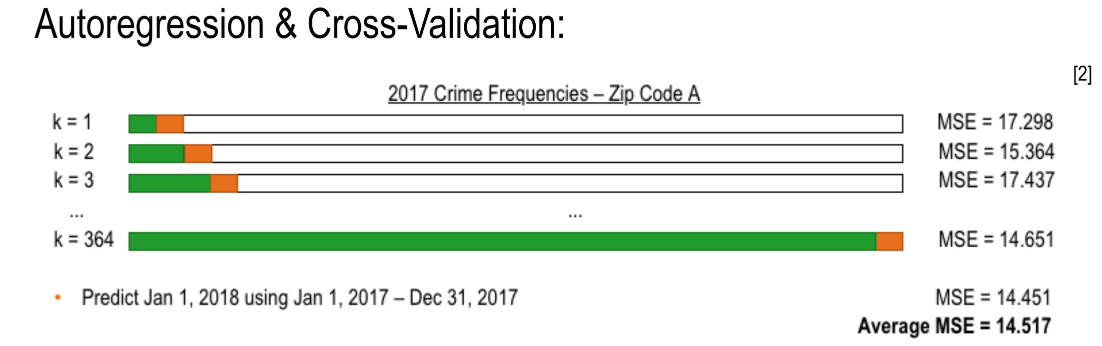
Vector Autoregression(VAR): Forecasts data using previous data from other time series

Cross-Validation: Validates the stability of the first 2 models

Mean Squared Error(MSE): Used to compare the forecasted data with the actual data

Example: For some zip code A, where the Granger Causality passed with a significance of below .05 for zip codes B, C and D

Autoregression & Cross-Validation:



Where in step k = 1 Jan 2, 2017 is being predicted using data from Jan 1, 2017 and in step k =2 Jan 3, 2017 is being predicted using data from Jan 1, 2017 – Jan 2, 2017

Vector Autoregression & Cross-Validation:

After Cross-Validation Using	Average MSE
B	19.490
C	15.977
D	13.269
B &C	Can not be calculated
B &D	17.844
C &D	20.313
B, C &D	Can not be calculated

The best forecaster for zip code A is Vector Autoregression & Cross-Validation using zip code D because it resulted in the least MSE.

This process was conducted on all 33 zip codes, although there were results, nothing is conclusive because of the lack of data

CONCLUSION

	Temporal	Spatial	Spatiotemporal
Predictors	Daily ✓ Weekly ✓	Urban vs. Rural ✓ House Price ✗	There exists a possibility to predict crime
Future Work	Monthly Seasonal Yearly	Population Density Urban vs. Rural: different method of creating the value	Need more data to make a conclusive statement

REFERENCES

[1] Montgomery County Database [2] Self made [3] Google Maps [4] 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 [5] 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 [6] U.S. Census Bureau (2016,June) *QuickFacts: Montgomery County, Maryland* Retrieved June 10, 2018, from https://census.gov/quickfacts/geo/chart/