Crime Behavior Analysis Final Report

Lauryn Davis, Jessica Malinowski, L Dettling
School of Computing
Grand Valley State University
Allendale MI

***Abstract*—Analyzing crime levels can present as a difficult challenge due to the size and dimensionality of the crime data within Portland. In this paper, we analyze crimes of Portland, Oregon from 2012 - 2017 using "calls-for-service" data provided by the Portland Police Bureau. We aim to analyze where and when the most crimes occur. Initial multivariate visualizations will be conducted to study the frequencies among variables. Interactive geospatial website development will provide a more descriptive density interpretation of where the crimes are occurring. Finally, An implementation of a popular statistical model named ARIMA (Autoregressive Integrated Moving Average) is presented. Inspection of the relationship between density plot findings and time series prediction with ARIMA will be noted. It is found that Disorder and Non-Criminal/Admin crimes tend to be the most frequent across time. Finally, there was a statistically significant relationship between the year a crime was committed and the type of call group associated with it.**

I. Introduction and Related Work

Crime analysis has become an increasingly popular study for Portland, Oregon in the past few years, with several studies indicating a change in crime rates impacted by factors like race, income, and location within Portland [1] [2] [3]. For the benefit of advancement in data science in tangent with the complexity of crimes and justice, the Portland Police Bureau initiated the Real-Time Crime Forecasting Challenge Posting for students across the United States, and provided the "calls-for-service" data [4]. The initial use of the data within the competition was for students to develop algorithms to study crime-forecasting.

We decided to analyze "calls-for-service" data provided by the Portland Police Bureau. The National Institution of Justice (NIJ) partnered with this bureau in 2017 with the goal of trying to better inform the federal government with crime forecasts and insights. With this being said, the domain for this project is centered around criminal justice. Research questions and motivations for this project are summarized as follows: where do the most crimes occur, when do these crimes occur, and is there a relationship to be made between years of crime data. In doing this, we believe that the federal and state governments would be able to have a clear understanding of various logistic questions. These logistic questions include knowing the level of patrol needed in a certain area and potential investments or technologies that should be utilized to combat this crime.

A brief insight into our approach includes incorporating our knowledge of data mining, multivariate statistics, and geospatial techniques in predicting various logistical questions centered around criminal justice. Multivariate techniques such as correspondence analysis aided our research in geospatial analysis. Geospatial interactive mapping was achieved to understand what call groups were most populated within a certain area. We implemented a series of ARIMA (Autoregressive Integrated Moving Average) models to try to predict the patterns seen within the various call-groups from 2016 - 2017.

The rest of this paper is structured as follows: Section II contains instructions for users that wish to view our interactive dashboard. Our methodology is covered in Section III. We then discuss our formal results in Section IV. Section V contains a brief discussion of results and conclusions. The final Section VI discusses shortcomings, limitations, and future work.

II. Download and Implementation Instructions

Updated code can be found within the GitHub Repository [5]. Key files within this repository include the "Multivariate.RMD", Shiny App, and "Models.ipynd". This includes the visualizations and ARIMA models respectively. Finally, the file labeled "Shiny Tutorial.mp4" within the "Geo Code" folder of our repository has a video tutorial of the Shiny app.

Currently, efforts are being made to publish the Shiny app. Since this is still in progress, however, a video tutorial was conduced. This video will contain an overview of the app and ways that it can be manipulated. Some of the key takeaways from the video are the fact that a user can click on a year from 2012 - 2017 and see the density plot for various call groups. This can be toggled to view only certain call groups. In addition to this, an income to poverty ratio layer is read over each of the years data. This allows for comparison as to where the crimes are occurring in consideration of an economic indicator. A user can zoom in and out of the map that is utilized within Shiny. If they wish to revert back to the full map, a button can be pressed to do so. If the user wanted to learn more about the project or contact the creator, they can click on specific tabs within the website to do this.

III. Methodology

The "calls-for-service data" provided by the Portland Police Bureau consists of five zip-folders, each with several files including excel, shape, and dbf files. Data for years 2016 and 2017 were split up into portions of the year, and into the respective excel, shape, and dbf files. Our data was collected by downloading each zip-folder, and extracting all but our shape files into our GitHub Repository. Unfortunately, our shape files were too large to directly upload into our repository. To solve this challenge, we integrated LFS (Large File Storage)

tracking, so that we could upload our larger files (see [6] for more information).

After our files were uploaded, we created Python files to read in our excel data, and save into a large data frame using the Pandas library (see [7] for more information on Pandas). After our excel data had been extracted and saved, we decided to read in our shape files for potential geospatial analysis. Unfortunately, we encountered extreme errors with portability and productivity using the GeoPandas library (see [8] for more information on GeoPandas). Because members of our team had R experience, we decided to transition to using R for our geospatial development and analysis. It was imperative that we derived meaningful geospatial and multivariate analysis' that could aid in how we setup our ARIMA model.

As mentioned in Section I, various contingency tables were created to map the distribution of call groups and crime categories across time. With various packages in R Studio, balloon plot visualizations were created. Following this, correspondence and multiple correspondence analysis were computed in R Studio. This was done by utilizing the R package titled, "FactoMineR" (see [9] for more information).

After multivariate analysis was produced, efforts shifted to validating if what we were seeing was true within a geospatial representation of the data. An interactive website application called "Shiny for R Studio" allowed for efficient implementation of this goal (see [10] for more information on Shiny). Shape files were read in with the "Sf" package (see [11] for more information on sf). A popular way of mapping data is done through a package called "Leaflet" (see [12] for more information on Leaflet). Coordinate systems within the shape files were transformed into CRS 4326. Additional geometry files were read in pertaining to the income to poverty ratio within that years census data (see [13] for more information on Tidy Census).

Finally, each year's data is plotted within separate events that are triggered when a user selects a year. The output is a density plot layer on top of the income to poverty ratio indicator. Results will be mentioned and instructions for how to access and view manipulation of the website are provided.

The final facet of our methodology, as mentioned, was the implementation of ARIMA models. This is noted to have exemplary results on capturing linear trends within time series data [14]. As mentioned within lecture notes, ARIMA is a simple linear equation for stationary time series data. If there is dependence among values, it's often beneficial to employ ARIMA. The noteworthy parameters within ARIMA include p, d, and q. p is the number of Auto-Regressive terms, d is the number of nonseasonal differences, and q is the moving average, or lagged forecast errors [14].

The implementations were measured with rolling predictions (processes each future time-step one at a time and each prediction is fed back as input for the next prediction) [15], and evaluated with RMSE (Root Mean Squared Error) scores.

Pseudo code for our implementation of ARIMA can be found in Algorithm 1.

ARIMA Model

```
model = ARIMA(train_dataDISORDER['count'], order=(2, 2,
↪  11))
results_ARIMA = model.fit()
yhat = results_ARIMA.forecast(steps=24)

plt.plot(DISORDER['count'], label="Actual Data")
plt.plot(yhat, color="red", label="Forecasted Data")
plt.plot(results_ARIMA.fittedvalues, color="red",
↪  label="Fitted ARIMA")

# Calculating the RMSE (Root Mean Squared Error)
# rmse = mean_squared_error(test_dataDISORDER['count'],
↪  yhat, squared=False)
# plt.title(f"RMSE: {rmse:.4f}")
plt.legend()
plt.show()
```

Fig. 1: ARIMA Implementation

Separate ARIMA'S were conducted for each call group. The ARIMA model for disorder will be highlighted within Section IV. Although, RMSE scores for each model will be provided within Section IV, we chose to highlight disorder due to the fact that the multivariate distribution plots highlighted this crime as being on average the most frequent.

IV. Results and Discussion

A. Multivariate Analysis

The results that were obtained from the initial multivariate analysis were beneficial for how we decided to proceed in our study. For example, the Figures 2 and 3 were created within R Studio.

Fig. 2: Call Group Distributions Across Time

These visualizations allowed for an initial gauge of trends that persist across time within the Portland Police Bureau's data. From Figure 2, the highest and most persistent call group is disorder. The size in frequency across time for each call group tends to be stationary with dips in 2017, which makes sense because data was not collected during the second half of 2017. We can get a more representative multivariate picture of what is occurring across time by viewing Figure 3. We

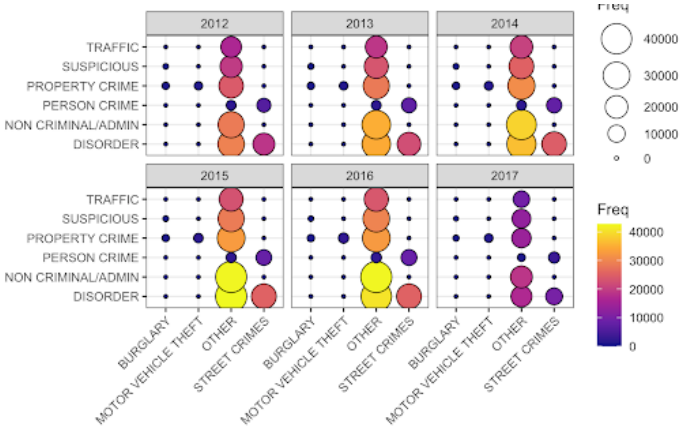

Fig. 3: Multivariate Distributions Across Time

can see that the most frequent other call groups are disorder, noncriminal/admin, and property crime.

After initial distributional visualization, we completed correspondence analysis using the package within R titled, "FactoMineR". This is a statistical technique that uses contingency tables to map relationships among qualitative variables. As shown in Figure 4, we can interpret this by noting that 82.7% of the variability within the data can be explained by the first two dimensions of the correspondence analysis. Further interpretation can be understood by noting that variables farther away from the origin are considered more discriminatory [16]. Therefore, personal crime, non criminal/admin, and traffic crimes seem to be quite discriminatory within the data. We can use a chi-squared independence test to see if there is a significant relationship between years and call groups. The chi-square statistic is equal to 604 and the p-value is $< .0001$. Therefore, we have evidence at the .05 alpha level that there is a statistically significant relationship between the year a crime was committed and the type of call group associated with it.

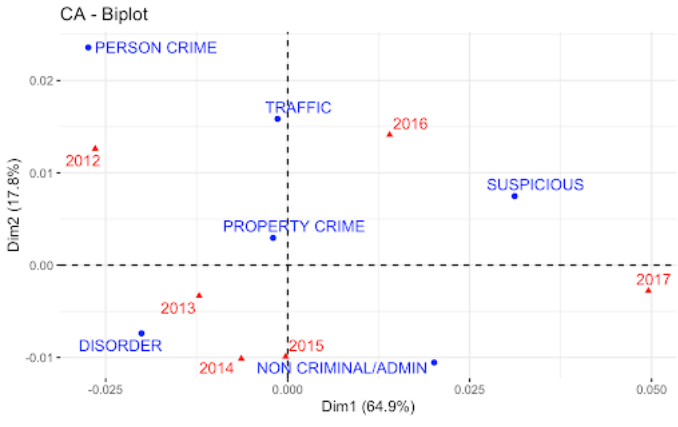

Fig. 4: Correspondence Analysis Between Group Calls and Year

More advanced multivariate techniques can be achieved when viewing multiple correspondence analysis results. From Figure 5, we can represent each variable as a representation of their squared cosine values. These values represent the degree of association between variable categories and a particular axis. If a variable category is represented well within the two dimensions, its cosine squared value will be close to one [17]. Although only 21.9% of the variation within the data is being explained within this plot, it's still interesting to note that the categories street crimes and personal crime have an important contribution to the positive pole of the first dimension, while other and property crime have an important contribution to the negative pole of the first dimension.

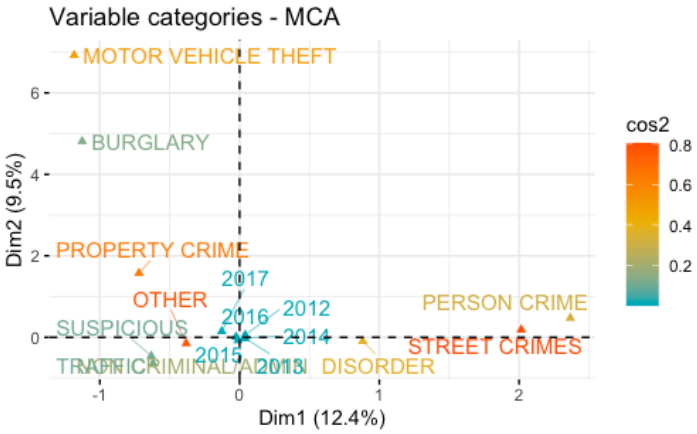

Fig. 5: Multiple Correspondence Analysis Between Category, Group Calls, and Year

B. Interactive Geospatial Analysis in Shiny For R Studio

After the results were obtained from correspondence analysis, density plots were mapped across time within the interactive Shiny web development space. As seen in Figures 6 and 7 the crime categories can be manipulated depending on what the user wishes to see. When toggling the crime category, it is evident that the call group categories of disorder and non-criminal crimes are the most frequent - especially in years 2015 and 2016. This aligns with the results within our multivariate section.

It's also interesting to note that crime is quite sporadic across Oregon, but the densest areas are noted to have a slightly higher income to poverty ratio. Some areas that are quite dark in poverty level, for example, Forest Park, do not have any crime points depicted within them. We suspect this is due to the boundaries of the shape files within the Portland Police Departments jurisdiction. It would be interesting to expand this analysis across other locations to see if further validation could be made pertaining to higher levels of crime associated with poverty.

Overall, the Shiny interactive mapping allowed us to see that crime stayed pretty persistent across time except for the two most frequent call groups. Those seem to increase as the years progress, but as we have stated, we do not have all of the 2017 data. The NHS had only provided about half of the

year, so it would be beneficial to obtain the rest of the data for future studies to see if these patterns persist.

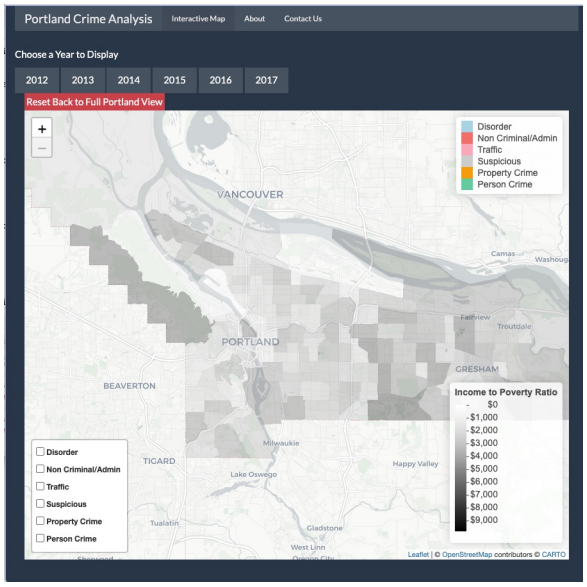

Fig. 6: Income to Poverty Ratio of 2012 Portland, Oregon

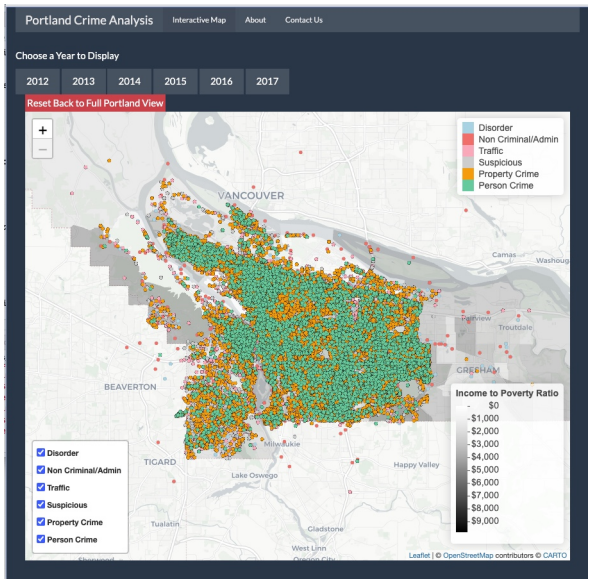

Fig. 7: All Call Groups within Portland, Oregon

C. ARIMA Application

As mentioned, the primary ARIMA model that was focused on was disorder. We first plotted the distribution, in days, of the frequency of disordered crime across time:

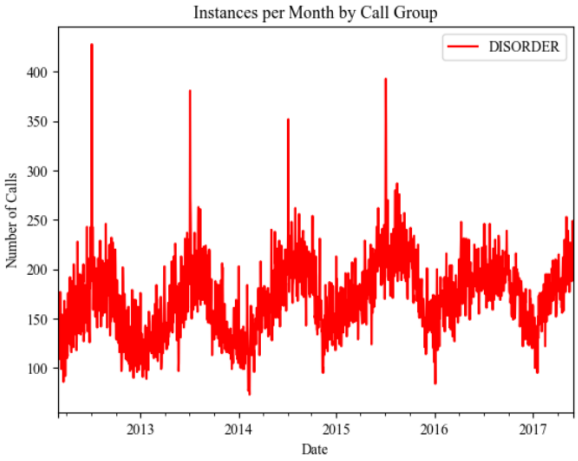

Fig. 8: Disorder Frequency Across Time

As seen within the Figure, there are some significant spikes within the months of July. This is due to the fourth of July holiday. This spike trend stops in 2015. Other significant call group categories that increased as time went on were person, suspicious, admin, and property. After the disorder frequency was seen, we wanted to conduct a time-series prediction using ARIMA. The primary package that was utilized to achieve this was an "Auto-Arima" function called "pmdarima" [18]. The parameters are chosen within this function that are ideal for ARIMA. The parameters that were chosen were 4, 4, and 0 for p, d, q, respectively. The results for this ARIMA can be seen in Figure 9.

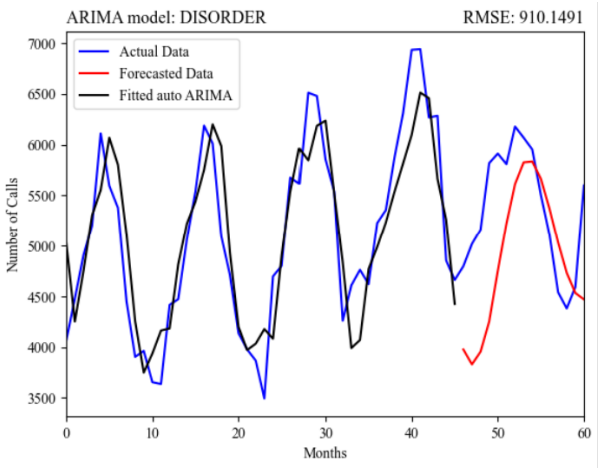

Fig. 9: ARIMA for Disorder

The RMSE for this model is 910.1491. This isn't the best model, but the trend line does seem to follow a similar pattern to the actual data. This model would be interesting to test on additional data within the Portland jurisdiction to see if accuracy improves. The RMSE's for the other call group models had similar results. The best model proved to be when call group was equal to person crime. The RMSE for this model was equal to 131.7.

V. Conclusion

We presented an analysis of the "calls-for-service" data. Initial findings proved that the crime category other had the highest count of crimes. Within the other category, disorder and non-criminal/admin call groups were the most frequent. It was noted that 2017 had dips in frequency due to not having all the data available. Correspondence analysis lead to conclusions of statistically significant patterns relating to the year a crime was committed and the type of call group associated with it.

After multivariate analysis, We created an application for users to view and model different years of data for analysis. We layered our data with a poverty-ratio indicator to gain a better understanding of the data from the 2012-2017 time period. It was found that, on average, areas with a higher poverty rate had slightly denser crime distributions. Crime levels were consistent across time but disorder, person, suspicious, property, and admin grew in frequency as the years went on.

Finally, using this information, we used a series of ARIMA models for time series prediction to see if we could accurately predict different trends. Our ARIMA models were not as accurate as they could have been, but when we set the call group to be person, we achieved our studies minimum RMSE score. In doing this, the Portland Police Department would be able to utilize this model to potentially predict future crime trends. This could be further validated if more data was available. Overall, the disorder category proved to be the most frequent. Also, July fourth spikes were evident through 2015 and should be noted by the Portland Police Department.

VI. Limitations and Future Work

There were several restrictions and shortcomings with this project. Limitations with larger data sets can slow performance within analysis techniques. Using provided libraries might not always be portable. For example, the RDGAL library in R is now archived, and setting the R environment to use RGDAL requires end users to drop to a previous version of R, which can cause compatibility issues (see [19] for more information on RGDAL). There are also possible restrictions with the size of the provided data - having over one-million rows of excel data requires a machine with updated hardware, which may not be accessible for all users.

In relation to this project, one deterrent to complex analysis was the fact that we decided to switch primary implementation languages from Python to R three weeks before the deadline of this project. This slowed progress, but the benefit of R's visualization yielded intricate and organized results that we couldn't achieve in Python with our visualizations. R also offered more stability with reading and plotting shape files, while Python's GeoPandas struggled to read and write to shape files, a crucial part of our study.

Future work can be done to extend our Shiny application. Additions like aggregating and grouping specific crimes together, the option to add additional data, and implementing data models within the app could improve the usability and productivity of our provided Shiny app. Extensions of this work could be applied to different cities or states to expand the scope of this analysis. Furthermore, applications could be ported to a high-performance cluster for better compatibility across users.

References

[1] M. E. Cahill, "Geographies of urban crime: an intraurban study of crime in nashville, tn, portland, or and tucson, az," Ph.D. dissertation, University of Arizona, 2004.

[2] C. E. Kubrin and E. A. Stewart, "Predicting who reoffends: The neglected role of neighborhood context in recidivism studies," *Criminology*, vol. 44, no. 1, pp. 165–197, 2006.

[3] G. Stewart and K. Henning, "Portland, oregon," *Problem-Oriented Policing: Successful Case Studies*, 2020.

[4] N. I. of Justice, "Real-time crime forecasting challenge posting." [Online]. Available: https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge-posting#data

[5] "Cis 635 repo," 2023. [Online]. Available: https://github.com/GVSU-CIS635/gvsu-cis635-term-project-crime-team-lauryn-jess-l

[6] GitHub, "Git large file storage." [Online]. Available: https://git-lfs.com/

[7] "Pandas documentation," 2023. [Online]. Available: https://pandas.pydata.org/docs/

[8] "Geopandas documentation," 2023. [Online]. Available: https://geopandas.org/en/stable/

[9] S. L. J. M. Francois Husson, Julie Josse, "Factominer: Multivariate exploratory data analysis and data mining," 2023. [Online]. Available: https://cran.r-project.org/web/packages/FactoMineR/index.html

[10] "Shiny," 2023. [Online]. Available: https://www.rstudio.com/products/shiny/

[11] "sf: Simple features for r," 2023. [Online]. Available: https://cran.r-project.org/web/packages/sf/index.html

[12] "Leaflet: Create interactive web maps with the javascript 'leaflet' library," 2023. [Online]. Available: https://cran.r-project.org/web/packages/leaflet/index.html

[13] M. H. a. K. E. c. Author: Kyle Walker [aut, cre], "tidycensus: Load us census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames." [Online]. Available: https://cran.r-project.org/web/packages/tidycensus/index.html

[14] Y. Zhuang, "Forecasting air passenger traffic," 2023. [Online]. Available: https://yong-zhuang.github.io/gvsu-cis635/air-passenger-forecast.html

[15] ——, "Assignment 4: Exploring neural network architectures in time series prediction," 2023. [Online]. Available: https://yong-zhuang.github.io/gvsu-cis635/assignment4.html

[16] A. Kassambara, "Statistical tools for high-throughput data analysis," *Statistical tools for high-throughput data analysis*, 2017. [Online]. Available: www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/.

[17] ——, "Mca - multiple correspondence analysis in r: Essentials," *Statistical tools for high-throughput data analysis*, 2017. [Online]. Available: www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/.

[18] T. G. Smith, "Tips to using auto_arima," 2023. [Online]. Available: https://alkaline-ml.com/pmdarima/tips_and_tricks.html

[19] "Rdgal download and documentation," 2023. [Online]. Available: https://cran.r-project.org/web/packages/rgdal/index.html