Getting Insights Using Minneapolis Police Data Analysis

Kolby Johnson*, Nitish Reddy†, Brennan Rhoadarmer‡, and Fateh Sandhu§
Department of Computer Science and Engineering, University of Nebraska - Lincoln
Lincoln, NE 68508
Email: *kolbyjohn@huskers.unl.edu, †npinninti2@huskers.unl.edu, ‡brhoadarmer@huskers.unl.edu, §fatehkaran@huskers.unl.edu

***Abstract*—Insights on the features for which 911 calls or other emergencies would impact the various understanding of the Minneapolis response to the people for which the government can issue active AI responses. As per the design we have estimated the different sets of the datasets that were accessed through the govt websites opendata.minneapolismn.gov. Now each of the datasets is described based on the Police stop data, Use of Force data and shots fired data. From each such scenario, we have implemented a case model for analyzing the graphical aspects of each data set depending upon the requirement. Since the data will be represented with cluster information on the feature of the spatial model and its implementation using the DBSCAN algorithm.**

I. Introduction

In today's world, criminals are becoming more technologically sophisticated with their criminal activities, and one challenge faced by intelligence and law enforcement agencies is the difficulty in analyzing large volumes of data involved in criminal and terrorist activities. As a result, agencies must learn techniques to catch criminals and stay ahead in the never-ending race between criminals and law enforcement. As data mining applies to collecting or mining information from vast volumes of data, it is seen here on a high volume crime dataset, and the knowledge obtained from data mining techniques is helpful and supports police forces. Clustering is a data mining technique that clusters a collection of objects in such a way that objects in the same category are more identical than objects in other categories, and it involves different algorithms that vary greatly in their notions of what makes a cluster and how to efficiently locate them. In this article, a spatial cluster scan is implemented to retrieve a valuable knowledge from a large crime dataset and to classify the data, assisting police in identifying and analyzing crime trends in order to prevent future incidents of related occurrences and provide information to deter crime.

II. Related Work

In the research and review of criminology, data mining may be divided into two categories: crime prevention and crime containment De Bruin et al. proposed a system for crime rates focused on a recent distance metric for evaluating and clustering all people based on their profiles[1]. Manish Gupta et al. highlight current e-governance programs utilized by Indian police and suggest an open query-based framework as a crime detection platform to aid police in their operations. He proposed an interface for extracting valuable knowledge from the National Crime Record Bureau's (NCRB) massive crime archive and locating crime hot points utilizing crime data mining methods such as clustering. The suggested interface's efficacy has been shown using Indian crime reports[2]. Nazarene Mohamad Ali et al. explore the implementation of the Visual Interactive Malaysia Crime News Retrieval Framework (i-JEN), including the methodology, user tests, system design, and potential plans. Their main goals were to construct crime-based events, investigate the use of crime-based events in improving classification and clustering, develop an interactive crime news retrieval system, visualize crime news effectively and interactively, integrate them into a usable and robust system, and evaluate the usability and system performance[3]. Sutapat Thiprungsri looks into how cluster analysis is used in the accounting sector, specifically difference identification in audits. The aim of his research is to look at how clustering technologies can be used to simplify fraud filtering during audits. When analyzing community life insurance cases, he used cluster analysis to aid auditors to concentrate their efforts. A. Malathi et al. look at how missing meaning and clustering algorithms can be used in a data mining method to forecast crime trends and speed up the investigation phase. To predict crime rates, Malathi[4]. An et al used a clustering/classify dependent model. The city crime data from the Police Department were analyzed using data analysis techniques. The information gleaned from this data mining may be useful in a variety of ways.

For the next several years, it may be employed to reduce or even deter violence. Malathi and Dr. S. Santhosh Baboo A study aimed to create a crime detection method for the Indian context utilizing various data processing tools that could aid law enforcement departments in effectively handling criminal investigations. The proposed tool allows organizations to disinfect, characterize, and interpret crime data in a simple and cost-effective manner in order to detect actionable patterns and trends. Kadhim B. Swadi Al-Janabi proposes a method for analyzing and detecting crime and criminal data utilizing Decision Tree Algorithms for data classification and the Simple K Means algorithm for data clustering[5]. The paper is intended to assist experts in recognizing patterns and developments, forecasting, identifying associations and plausible causes,

mapping crime networks, and identifying potential offenders. Aravindan Mahendiran et al. mine crime data sets using a variety of methods to find evidence that is obscured from human experience. We view the trends found by our algorithms neatly and intuitively using state-of-the-art simulation tools, allowing law enforcement agencies to channel their efforts appropriately. Sutapat Thiprungsri investigates the feasibility of auditing using clustering technologies[6]. Continuous audits may benefit greatly from automating fraud filtering. The aim of their research is to see whether cluster analysis can be used as an additional and novel anomaly detection strategy in the wire transfer method. K. Zakir Hussain et al. used data mining and a simulation model to try to catch years of human knowledge in computer simulations[7].

III. Datasets

For this project, three datasets were utilized: Minneapolis Police Stop Data[8], Minneapolis Police Use of Force Data**??**, and Minneapolis Shots Fired Data**??**. The first dataset, Minneapolis Police Stop Data, or Stop Data, has 175,771 entries within it, dating from October 31, 2016 until present. Each entry has 18 attributes, a summary of which can be seen in Table I. The second dataset, Minneapolis Police Use of Force Data, or Use of Force, has 32,383 entries within it, dating from January 2, 2008 until present. Each entry has 28 attributes, summarized in Table II. The third dataset, Minneapolis Shots Fired Data, or Shots Fired, has 69,287 entries within it dating from March 3, 2007 until present. Each entry has 14 attributes, summarized in Table III.

IV. Objectives

As this project was rather complex, with multiple possible analysis to preform or ways to approach the data, we decided to split our goals up into four objectives. The first objective was the ”trivial” data preprocesing. It only sounds simple and trivial but requires a number of decisions of how to decide if data is valid and this is the objective that propagates to all the others, so if this objective is slacked on then future work is more difficult.

For objective two, we wanted to do some association analysis over the nominal data from the police stop and crime data. We wanted to determine if there were extenuating factors that would more often than not lead to stops, force, or shots fired. We specifically wanted to focus on the time around the BLM protests in May, 2020, and potentially determine if there was a difference in how race was associated with other data before and after those events.

For objective 3, we wanted to be able to determine if there was a section of the city that was being focused on more, or if there were certain elements of the datasets that were strongly tied to a physical location within the city. Therefore, we decided to utilize DBScan in order to analyze the spatial data within the datasets.

Finally, for objective 4, we wanted to expand upon objective 3, and add in time as another dimension which would allow for some potentially more interesting discoveries. We wanted to see if there was a physical anchor for an element of a dataset that was also tied to time, which would make it difficult to extinguish within only a spatial analysis. Therefore, we decided to build off of objective 3 and continue to perform DBScan over the databases while providing the time data in this new analysis.

These were our four objectives, outlined in a summary below:

Objective 1: Preprocess the datasets for easier manipulation later.

Objective 2: Perform association analysis over the nominal data from the stop and crime data.

Objective 3: Perform spatial clustering analysis over the datasets, focusing over different attributes to determine if such clustering is viable.

Objective 4: Build off of objective 3 and add time into the mix to perform spatio-temporal clustering analysis.

For objectives 3 and 4, we also wanted to output some sort of visualization as that easily pairs with their output and more easily allows for understanding and analysis.

V. Data Pre-processing

Each objective requires its own pre-processing but there were some irrelevant attributes in all the datasets which would not be required in any of the objectives. First, we removed the irrelevant attributes and chose to keep only the ones that would be required later. Then, there were various records with null values in some attributes, we could not replace them with average values as the attributes were nominal. For the attributes that were boolean and null we filled in those values with false or no in this case. We decided to remove the records with null values from the datasets so that the null values do not pose a problem in any of the objectives. There were some records that had no null values except they had 0 for latitude and longitude, since our clustering relies on those attributes we removed all records that had 0.

VI. Approach

A. Data Preprocessing

The data preprocessing was done using Pandas dataframes and Microsoft Excel. The CSV files were first converted into dataframes using Jupyter Notebooks and then once the general preprocessing i.e removing null records and irrelevant attributes was done, they were converted back into CSV files to be used for other objectives.

B. Association Analysis

We used Weka Explorer for the association analysis part. We were initially going to use the Python Apyori library but the input for the algorithm required us to convert the entire

Attribute Name	Type	Notes
masterIncidentNumber	String	Unique identifier for incident
responseDate	Date and Time	Given in format YYYY/mm/dd HH:mm:ss+00
reason	String	Options: Citizen/911, Moving Violation, Investigative, or Equipment Violation
problem	String	Options: Traffic Law Enforcement, Suspicious Person, Suspicious Vehicle, Attempt Pick-Up, Curfew Violations, or Truancy
callDisposition	String	Listing the disposition of the 911 caller
citationIssued	Boolean	Yes or No was a citation issued
personSearch	Boolean	Was the person searched
vehicleSearch	Boolean	Was the vehicle searched
preRace	String	Race assumed before stop. Options: Unknown, Black, White, Native American, East African, Latino, Other , and Asian
race	String	Race determined after stop. Options: Unknown, Black, White, Native American, East African, Latino, Other , and Asian
gender	String	Gender of the stopped person. Options: Male, Female, Unknown, Gender Non-Comforming
lat	Float	Latitude of the stop
long	Float	Longitude of the stop
x	Float	Meaning unknown
y	Float	Meaning unknown
policePrecinct	Int	Numbered 1-5
neighborhood	String	String representation of the neighborhood
lastUpdateDate	Date and Time	Given in format YYYY/mm/dd HH:mm:ss+00

TABLE I
MINNEAPOLIS POLICE STOP DATA ATTRIBUTES

dataset into a list of lists and the algorithm that we wrote for the conversion was not viable as the datasets were really big. Then to use the datasets for association analysis with Weka, we converted all the useful attributes into nominal attributes, input minimum support bound as 0.1 and maximum support bound as 1.0 so it automatically chose the best support for the analysis. The metric chosen for rules was confidence and minimum confidence was set to 0.7

C. Spatial Clustering Analysis

As per the design characteristics which we have procedure to implement the different set of the features that have to be clustered using the directions and its functional analysis on the clustering algorithm where python based solution becomes easier more applicable.

Design Procedure:

1) Initiate the data set from the datasets
2) Features and its labelling for each data cluster formation based on the attribute established.
3) We improvise Python tools such as anaconda to provide the specific libraries ensuring the different set of attributes are acquired from the CSV file.
4) Each attribute are labelled with label encoder to initiate the design model for implementing the different structures on each data fitting scenario for the different locations observed for each set of the region considered.
5) Perform DBSCAN algorithm for the data set ensuring the different attributes observed on the cluster formed.
6) Plot the data to visualize the different crime data attribute using pyplots and scatter plots.
7) Repeat the 1-6 steps for each database chosen.

Our design aims to provide a clustering feature to represent the design aspects on the basis of police force data. We have implemented a data cleaning process to initiate the design acquisition based on the Numeric responses of the columns where each set of the data attribute is Case-id, X, Y , Problem and Neighborhood. Each set of the clustering model is represented with the features accepted with the design, and its attributes are clustered with correct response based on the data fitting.

The data labeling is modelled with "label_encoder.fit_transform" utilizing the different string values to corresponding numeric response. For each required column with Dataset_table are applied to visualize as a cluster feature.

Finally we apply DBSCAN algorithm to initiate the data fitting of the Set of columns and row which are implemented. For case of 1000 we achieve Figures 1 through 5.

From figure 1 we have represented the design for Clustering model on basis of n= 1000 and n = 10000.

For N= 1000, We have only considered with [CaseNumber X Y Problem EventAge] attributes ensuring the clusters represented based on the above four.

Hence for the selection of column 911 call would change representation of the clusters either for X in Is911 call and Y in Is911call, as seen in Figure 2 and Figure 5.

D. Spatio-Temporal Clustering Analysis

For spatio-temporal clustering analysis, we used the implementation of DBScan found in the sklearn library, while we read the data into pandas dataframes. This made it fairly trivial, as all we had to do was provide the longitude, latitude

Attribute Name	Type	Notes
PoliceUseOfForceID	Int	Unique identifier for interaction
CaseNumber	Text	Case number
ResponseDate	Date and Time	Given in YYYY/mm/dd HH:mm:ss+00 format
Problem	Text	Many possible options
Is911Call	Boolean	Was 911 called for this interaction
PrimaryOffense	Text	Many possible options
SubjectInjury	Boolean	Was the subject injured
ForceReportNumber	Int	Identifier
SubjectRole	Text	Various options
SubjectRoleNumber	Number	Identifier
ForceType	String	Has 11 possible options
ForceTypeAction	String	Various options
Race	String	Race of subject.
Sex	String	Sex of subject. Options: Male, Female, not recorded, unknown
EventAge	Number	Small counting number of something. Maybe minutes
TypeOfResistance	String	Reason force was used
Precinct	String	1-5 or special task force
Neighborhood	String	String representation of the neighborhood
TotalCityCallsForYear	Int	Self explanatory
TotalPrecinctCallsForYear	Int	Self explanatory
TotalNeighborhoodCallsForYear	Int	Self explanatory
CenterGBSID	Float	
CenterLongitude	Float	Longitude of interaction
CenterLatitude	Float	Latitude of interation
CenterX	Float	Unknown
CenterY	Float	Unknown
DateAdded	Date and Time	Date entry added to dataset

TABLE II
MINNEAPOLIS POLICE USE OF FORCE DATA ATTRIBUTES

Attribute Name	Type	Notes
Longitude	Float	
Latitude	Float	
Juisdiction	String	Jurisdiction of incident
Master_Incident_Number	String	Unique ID of incident
Division	String	
Problem	String	Options: Sounds of Shots Fired, ShotSpotter Activation, Shooting, Shooting Report Only
Response$_{D}ate$	Date and Time	Given in YYYY/mm/dd HH:mm:ss+00 format
Time	String	String representation of time from Response_Date
DayOfWeek	String	String representation of day of week
Day	Int	Day in month
Month	Int	Month in year
Year	Int	Year

TABLE III
MINNEAPOLIS POLICE SHOTS FIRED DATA ATTRIBUTES

and time for each entry. In order to provide the time, we used python's datetime library to create a timestamp, or a time since UNIX epoch representation from the given date and time. The issue with this is that the timestamp is given in seconds, so the distance between points is rather large, while the spatial elements are given in longitude and latitude. So over the whole dataset, the spatial elements differ about .15, while the temporal elements differ by 31556926 per year they are separated. This meant that the controlling factor on clustering was time, in that clustering was only being done on elements related by time. So we had to do two things, factor the time by constant to group the temporal elements more similarly to the spatial elements and determine a good ϵ value that would allow good clustering but not cluster everything together. Determining those two values proved most important and very difficult.

The issue was that neither of the values had a good baseline, so changing one, meant changing the other. For the time constant value, we made it so that the separation of the city, i.e. about .15 was equal to the separation of about one week temporally, which meant our time constant was equal to the seconds of 140 days. For our ϵ value we got results for values around 0.005 for the stop data. The issue for the other data is that those datasets had much less data, that was much less dense, so the time constant and ϵ values did not work nearly as well for them.

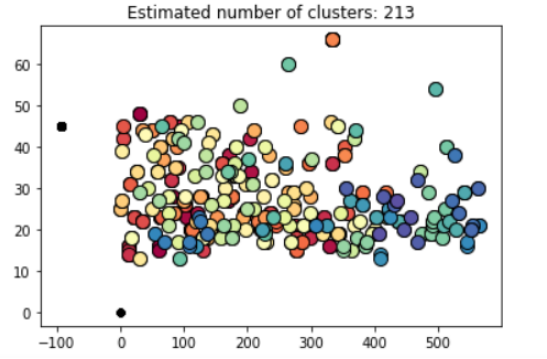

Fig. 1. Number of Clusters for n=1000

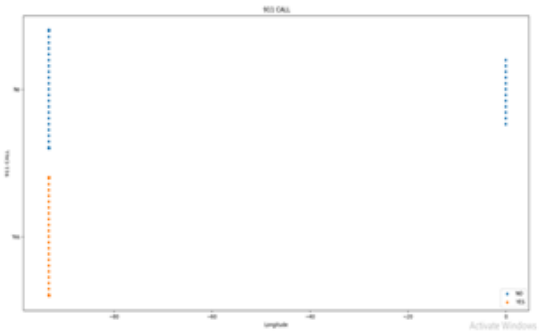

Fig. 2. Is911Call and clustering

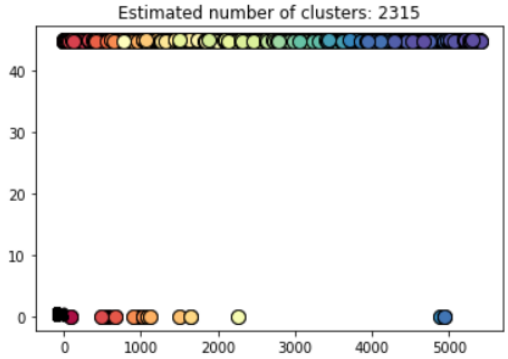

Fig. 3.

	CaseNumber	EventAge	X	Y	Is911Call
5	151	25	-93.271963	44.981655	0
7	189	35	-93.278819	44.978407	0
10	283	24	-93.271963	44.981655	0
18	269	18	-93.222966	44.897637	0
23	387	22	-93.273141	44.980808	0
...	...	...	...	...	...
9981	5198	38	-93.284894	44.951949	0
9984	5232	29	-93.315888	45.017763	1
9992	5115	21	-93.286823	44.999135	0
9993	5143	16	-93.306962	45.010527	1
9995	5186	49	-93.274317	44.974499	0

Fig. 4. Output table of spatial clustering

For the spatio-temporal clustering analysis overall, we had two major implementations, a dynamic one and a static one. The dynamic implementation reads in the data within a certain range and then steps through the data to the final date, doing DBScan over each range, resulting in multiple sets of clusters, and a png output for each step made. All of the steps were then compiled together and made into a gif. The static implementation did a single DBScan over the entire date range, resulting in one set of clusters and a single output png

VII. Evaluation

Evaluation for each of our objectives was different. For preprocessing, we simply had to see how much preprocessing we had to do for the other objectives. The more we had to do the less effective our preprocessing was. For association analysis the rules were generated by Weka, so we evaluated the rules, based on their confidence. For spatial clustering analysis, we could visually observe the results and use that to determine if what we had returned was meaningful. We used a similar method for determining if we got meaningful results for spatio-temporal clustering analysis.

VIII. Results and Analysis

When looking at the effects of BLM Protests we found out there were different results for stop and force data. We noticed that with the stop data half the amount of black people were stopped, with 40% of stops being black before BLM and only 24% after. When looking at police use of force the amount of black people was the same as before BLM with both being about 60%. For splitting the data we did post BLM as anything after May 25, 2020 and preBLM was the same number of records with being closer to May 25, 2020.

We ran the Apriori algorithm on the data sets but were not able to generate informative rules for all the attributes, so be looked for rules between sets of 2 attributes to help us get further insight.

When looking at association analysis for stop data we tried to focus on the race attribute. The most important rule generated was that a race of white associated with them not having their vehicle searched with a confidence of 95%. Another rule was that a race of black associated that they were searched (the personSearched attribute was highly skewed so we altered the data set to have an equal number of yes/no) with a confidence of 75%. We found out that that if race is unknown then they were reported as a suspicious person or vehicle. This shows that race is still very much a factor in when police stop and search someone. These results can be seen in Figures 9 and 10.

When looking at association analysis for shots fired data Weka did not generate informative rules that had a high confidence. This was unfortunate because we were trying to

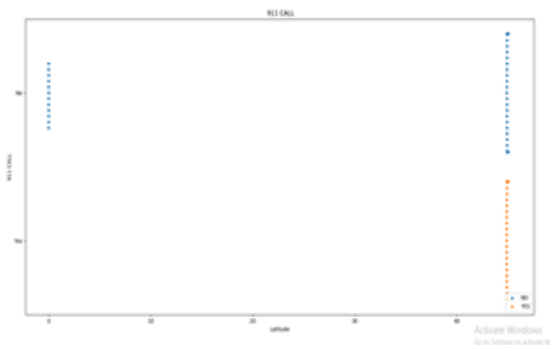

Fig. 5. Is911Call Clustering

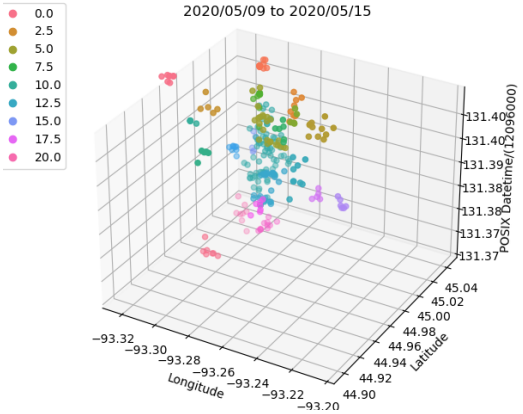

Fig. 6. Example clustered output gotten from data

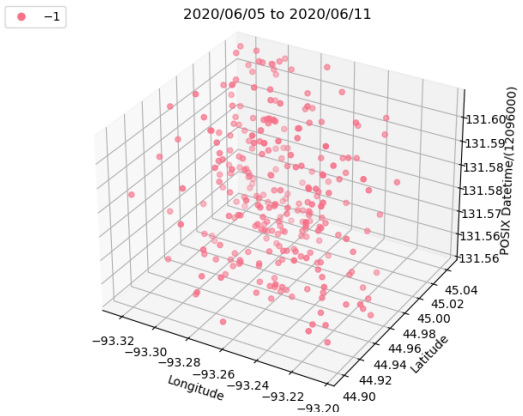

Fig. 7. Example clustered output gotten from data

see if there was a certain precinct that associated with violent crime.

When looking at association analysis for use of force data we tried to focus on the race attribute, but there were no rules generated with a high confidence that associated race. When looking at the rules generated some of the most informative were related to the sex and if the subject was injured. One of the rules generated was that if an officer used a chemical irritant it most likely didn't lead to subject injury with a confidence of 92%. We found that if the subject was injured by the officer they were male with a confidence of 93%. Another of the rules generated stated that subjects who fled on foot were male with a confidence of 97%. We found that if an officer used body weight to pin it was a male subject with a confidence of 87%. Rules generated with sex should be taken with a grain of salt since our data set was skewed with 86% of the records being male. In regards to the neighborhood attribute we found that the Downtown West neighborhood didn't have a 911 call when the police used force with a confidence of 82%. This could be because this neighborhood is over policed, or it has a large amount of crime. These results can be seen in Figures 11, 12, 13, 14, 15.

When looking at the spatio-temporal clustering analysis we got some interesting results around the time of the BLM protests. In Figure 6, you can see an expected result when there are clusters found within the data. Each different cluster has a different color which is explained by the key. We only graph the data that is in the clusters in order to make it easier to see. The interesting thing that came up, was that for this analysis we went dynamically from 2020/05/01 to 2020/07/01, and when we hit 2020/06/05, approximatley 10 days after when we determined the BLM protests started, we had no clusters within the stop data, as seen in Figure 7. The key clearly shows everything is labeled as -1, or noise, as output from the DBScan. When everything is noise, we output everything. This section of having no clusters goes on until 2020/06/17, lasting almost 2 weeks. After those 12 days, we once again get clusters.

This phenomenon is difficult to understand, though we propose a few reasons for it. The first, is that it is simply due to poorly chosen ϵ and time constant values for the data. While possible, this does not explain how or why we were getting clusters before and after this. The better explanation is the reason we were looking in that area in the first place, the BLM protests, spread the out across the city, making clustering much harder, where before and after stops seem to be happening in largely the upper corner of the plots.

We also reproduced these results while filtering based on race, but the gap in clustering stayed. We did some more interesting clusters in that information, as seen in Figure 8. Here we see that there is a center column of a cluster that continues for over a week. The only time this cluster stops is when we have no clusters from 2020/06/05 until 2020/06/17.

IX. Future Work

While we got some interesting results from this project there is some work that could be done in the future. The spatio-temporal analysis could be looked at to improve the time constant and ϵ values used within it, and varying those for each data set would likely provide better results. Then we could also do some analysis over weather data. We had initially planned to use some, but then none of our objectives utilized it so we cut that out. In the future we could reintegrate the weather data to receive more interesting results.

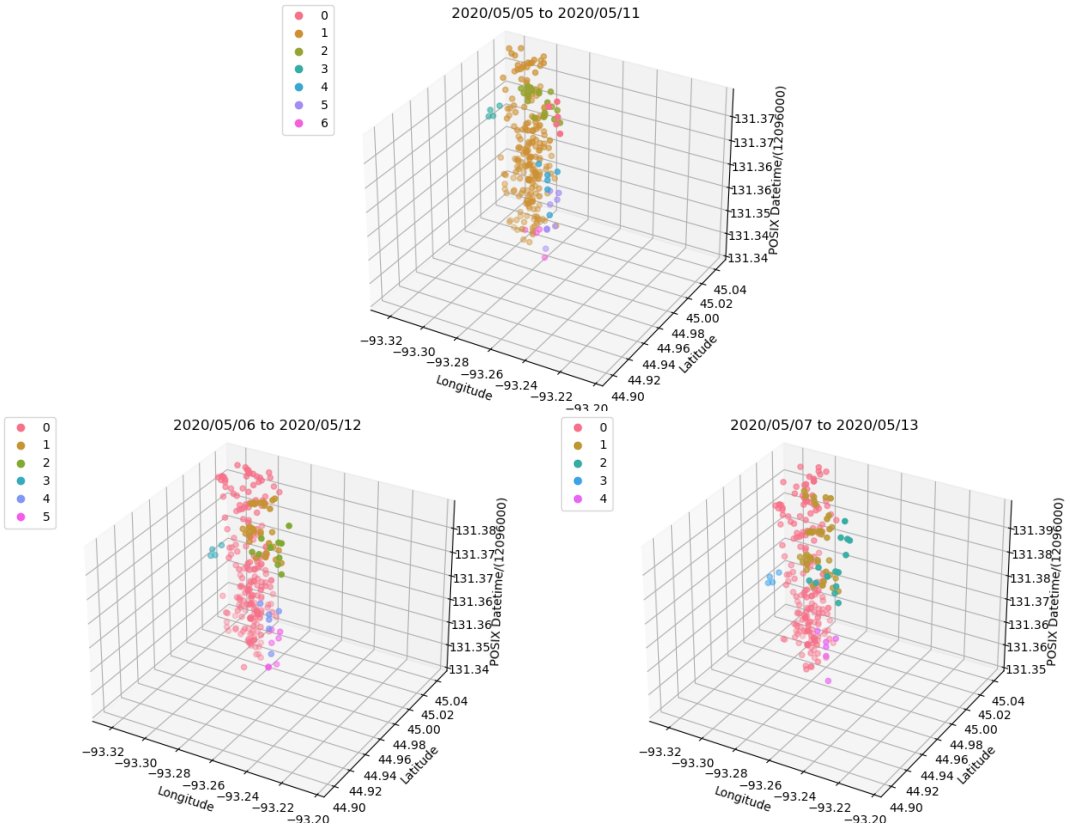

Fig. 8. Stop data filtered to only Race=Black entries

REFERENCES

[1] H.-c. Chen, W. Chung, Y. Qin, M. Chau, J. Xu, G. Wang, R. Zheng, and H. Atabakhsh, "Crime data mining: An overview and case studies," 12 2003.

[2] C. Fan, K. Xiao, B. Xiu, and G. Lv, "A fuzzy clustering algorithm to detect criminals without prior information," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 238–243.

[3] R. Kiani, S. Mahdavi, and A. Keshavarzi, "Analysis and prediction of crimes by clustering and classification," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, 08 2015.

[4] A. Malathi and S. Baboo, "An enhanced algorithm to predict a future crime using data mining," *International Journal of Computer Applications*, vol. 21, 05 2011.

[5] J. Oyelade, I. Isewon, O. Oladipupo, O. Emebo, Z. Omogbadegun, O. Aromolaran, E. Uwoghiren, D. Olaniyan, and O. Olawole, "Data clustering: Algorithms and its applications," in *2019 19th International Conference on Computational Science and Its Applications (ICCSA)*, 2019, pp. 71–81.

[6] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques." [Online]. Available: http://www.stat.cmu.edu/~rnugent/PCMI2016/papers/DocClusterComparison.pdf

[7] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. Wiley-Interscience, 2003.

[8] City of Minneapolis, "Police stop data," https://opendata.minneapolismn.gov/datasets/police-stop-data

[9] ——, "Police use of force data," https://opendata.minneapolismn.gov/datasets/police-use-of-force.

[10] ——, "Shots fired data," https://opendata.minneapolismn.gov/datasets/shots-firede.