CITADEL | CITADEL | Securities

TheData Open

West Coast Regional Datathon Spring 2023

Actionable Insights on Philadelphia Crime Hot-Spots: Clustering and Statistical Analysis to Inform Future Crime Legislation

Team 1

Aditya Bora, Ishan Khare, Tarun Martheswaran, and Rahul Thomas

Saturday, March 11, 2023

I. Topic question

Philadelphia, one of the most populated cities in the United States, has struggled with high crime rates for decades. In recent years, this problem continues to be exacerbated as Philadelphia's residents, community leaders, and law enforcement officials struggle to address the root causes of the problem and make the city safer for all. In this work, we aim to more deeply understand crime in Philadelphia and offer novel insights for crime mitigation within the city. In particular, we aim to understand and develop an answer to the following questions:

- **What actionable insights can we make to help Philadelphia policy makers create targeted crime reduction legislation?**
- **What can we learn about the underlying nature of crime in Philadelphia from geotemporal analysis?**

II. Executive Summary

IIa. Problem Statement and Background

Philadelphia is one of the most populated cities in the United States, with a rich history and a diverse population. However, like many large cities, Philadelphia has struggled with high crime rates for decades. It has long been known for its high crime rates, particularly violent crimes such as homicide and aggravated assault. In recent years the city continues to experience higher crime rates compared to many other large cities in the United States. According to the Philadelphia Police Department, in 2021 there were 499 homicides, which is the highest number in the city in 30 years.

Over the past 10 years, Philadelphia has implemented various policing strategies aimed at reducing crime rates in hotspot districts. These include community policing, which emphasizes collaboration and communication between the police and community members, as well as hot-spot policing, which focuses on targeting areas with high levels of criminal activity. Despite these efforts, crime rates in some districts of Philadelphia remain stubbornly high: policing these districts requires a nuanced understanding of the local community, its needs, and its priorities.

Indeed, one of the key challenges in addressing crime in Philadelphia is the issue of policing districts. Philadelphia is divided into 21 police districts, each of which has its own unique characteristics and challenges. Some districts, such as Center City and University City, are densely populated and experience a high volume of tourism and commerce, while others are more residential and may be marked by poverty and disinvestment. Any crime prevention strategy that neglects these differences will ultimately be ineffective.

Furthermore, it is critical that community leaders understand the policing alone is very unlikely to reduce crime [12]. Especially given the recent events regarding police brutality, is it paramount that policy makers and community leaders use different methods to try and reduce crime. For instance, tools such as educational investments and infrastructure investments have proven to be effective ways of solving root causes of crime. These among other techniques should be heavily considered and appropriately implemented when creating crime mitigation strategies [9] [3].

Our project aims to explore and utilize the temporal and geospatial trends in Philadelphia's crime rates in order to better inform the policing, infrastructure, and educational decisions within particular districts. Ultimately, we seek to provide insight into the complex and multifaceted nature of crime and policing in Philadelphia. By examining the data collected regarding previous crimes, demographics, systemic discriminatory practices like redlining, among other factors, we formed a deeper understanding of the issues exacerbating crime rates in Philadelphia.

Ultimately, we discovered that crime can be classified based on location and temporal features which can assist policy makers in creating more effective targeted legislation.

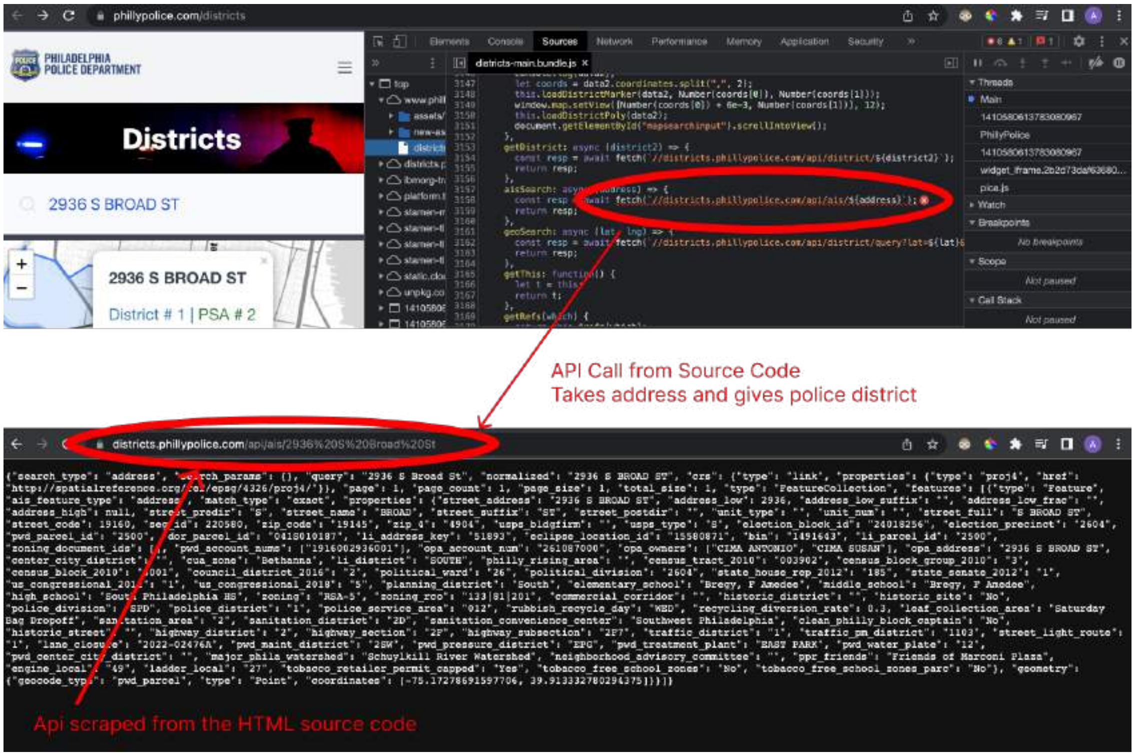
III. Technical Exposition

IIIa. Data Cleaning and Feature Engineering

We extensively use 'crimes.csv', a comprehensive table of data of crimes reported in Philadelphia from 2012 to 2022. The particularly important features are the time of dispatch, geospatial data represented in terms of latitude and longitude, the street address of the crime, and the type of crime committed. *The main feature we use is the geospatial data, in terms of latitude and longitude.*

Data cleaning of 'crimes.csv' is kept relatively simple, as the main data we wish to use is the geospatial information of crimes. First, we delete rows with 'NA' listed under the location. Second, we remove rows with invalid dispatch times and dates, and street addresses. These invalid entries account for 0.7% of all rows. In addition, we remove data points belonging to districts 4 and 23 as there is only one datapoint for each of the districts.

Next, we engineer a feature which contains the police district of each crime committed. This information is present in 'investigations.csv', but is missing in 'crimes.csv', where we are only given the PSA, a particular area number within each district. To extract the police district, we use the street address of the crime as follows. At the official district website for Philadelphia police there exists an API that retrieves the police district from a street address. [1] We scrape the API from the source code, as shown in Figure 1, and use it to engineer the police district feature.



```
    let coords = data2.coordinates.split(",", 2);
    this.loadDistrictMarker(data2, Number(coords[0]), Number(coords[1]));
    window.map.setView([Number(coords[0]) + 6e-3, Number(coords[1])], 12);
    this.loadDistrictPoly(data2);
    document.getElementById("mapsearchinput").scrollIntoView();
  },
  getDistrict: async (district2) => {
    const resp = await fetch(`//districts.phillypolice.com/api/district/${district2}`);
    return resp;
  },
  aisSearch: async (address) => {
    const resp = await fetch(`//districts.phillypolice.com/api/ais/${address}`);
    return resp;
  },
  geoSearch: async (lat, lng) => {
    const resp = await fetch(`//districts.phillypolice.com/api/district/query?lat=${lat}&
    return resp;
  },
  getThis: function() {
    let t = this;
    return t;
  },
  getRefs(which) {
```

API Call from Source Code
Takes address and gives police district

districts.phillypolice.com/api/ais/2936%20S%20Broad%20St

```
{"search_type": "address", "search_params": {}, "query": "2936 S Broad St", "normalized": "2936 S BROAD ST", "crs": {"type": "link", "properties": {"type": "proj4", "href":
"http://spatialreference.org/ref/epsg/4326/proj4/"}}, "page": 1, "page_count": 1, "page_size": 1, "total_size": 1, "type": "FeatureCollection", "features": [{"type": "Feature",
"ais_feature_type": "address", "match_type": "exact", "properties": {"street_address": "2936 S BROAD ST", "address_low": 2936, "address_low_suffix": "", "address_low_frac": "",
"address_high": null, "street_predir": "S", "street_name": "BROAD", "street_suffix": "ST", "street_postdir": "", "unit_type": "", "unit_num": "", "street_full": "S BROAD ST",
"street_code": 19160, "seg_id": 220580, "zip_code": "19145", "zip_4": "4904", "usps_bldgfirm": "", "usps_type": "S", "election_block_id": "24018256", "election_precinct": "2604",
"pwd_parcel_id": "2500", "dor_parcel_id": "0418010187", "li_address_key": "51893", "eclipse_location_id": "15580871", "bin": "1491643", "li_parcel_id": "2500",
"zoning_document_ids": [], "pwd_account_nums": ["1916002936001"], "opa_account_num": "261087000", "opa_owners": ["CIMA ANTONIO", "CIMA SUSAN"], "opa_address": "2936 S BROAD ST",
"center_city_district": "", "cua_zone": "Bethanna", "li_district": "SOUTH", "philly_rising_area": "", "census_tract_2010": "003902", "census_block_group_2010": "3",
"census_block_2010": "3001", "council_district_2016": "2", "political_ward": "26", "political_division": "2604", "state_house_rep_2012": "185", "state_senate_2012": "1",
"us_congressional_2012": "1", "us_congressional_2018": "5", "planning_district": "South", "elementary_school": "Bregy, F Amedee", "middle_school": "Bregy, F Amedee",
"high_school": "South Philadelphia HS", "zoning": "RSA-5", "zoning_rco": "133|81|201", "commercial_corridor": "", "historic_district": "", "historic_site": "No",
"police_division": "SPD", "police_district": "1", "police_service_area": "012", "rubbish_recycle_day": "WED", "recycling_diversion_rate": 0.3, "leaf_collection_area": "Saturday
Bag Dropoff", "sanitation_area": "2", "sanitation_district": "2D", "sanitation_convenience_center": "Southwest Philadelphia", "clean_philly_block_captain": "No",
"historic_street": "", "highway_district": "2", "highway_section": "2F", "highway_subsection": "2F7", "traffic_district": "1", "traffic_pm_district": "1103", "street_light_route":
"1", "lane_closure": "2022-02476A", "pwd_maint_district": "2SW", "pwd_pressure_district": "EPG", "pwd_treatment_plant": "EAST PARK", "pwd_water_plate": "12",
"pwd_center_city_district": "", "major_phila_watershed": "Schuylkill River Watershed", "neighborhood_advisory_committee": "", "ppr_friends": "Friends of Marconi Plaza",
"engine_local": "49", "ladder_local": "27", "tobacco_retailer_permit_capped": "Yes", "tobacco_free_school_zones": "No", "tobacco_free_school_zones_parc": "No"}, "geometry":
{"geocode_type": "pwd_parcel", "type": "Point", "coordinates": [-75.17278691597706, 39.913332780294375]}}]}
```

Api scraped from the HTML source code

Figure 1: *Extracting API Information From Website Source for Feature Engineering*

Finally, using the district feature, we computed aggregate counts of crime in each district from 2012 to 2022. We found that in districts 4 and 23, there had only been one crime in 10 years, so we removed these entries. This completes the cleaning and feature engineering for 'crimes.csv'.

IIIB. Exploratory Data Analysis

III.IIIB.A. Understanding Police Districts

In Philadelphia, a police district is a geographic area that is patrolled by officers from the Philadelphia Police Department. There are a total of 21 police districts in the city of Philadelphia.

Each police district covers a specific area of the city and is responsible for responding to emergency calls, investigating crimes, and maintaining public safety in that area. The boundaries of the police districts are based on factors such as population density, crime rates, and geographic features. Each district is staffed by police officers, detectives, and support personnel and is responsible for responding to calls for service, conducting investigations, and engaging with the community.

Given that police resources are allocated within these districts, we begin our analysis by analyzing trends with respect to these district boundaries. We also choose to analyze crime distributions within Philadelphia without the context of the police district boundaries to see if there are larger conclusions/trends which cross district borders.

III.IIIB.B. Temporal Analysis By Police District

To better understand the city of Philadelphia's crime incidence we wish to granularize our analysis. An obvious point of exploration is a location based exploration, specifically by police district. With the upcoming plots we wish to reveal characteristics about the police district level crime rate. Specifically we want to determine the relative frequency of the crime per district. Here we create heatmaps of the crime distribution by police district across the city of Philadelphia over a 9-year time frame with 4-year time increments (Figure 2).

At first glance, we see that there are no changes of greater than 0.5 with respect to the distribution over time within certain police districts. However, we do observe that there are general shifts in distribution within the districts which motivates a more narrowly-focused exploration into individual geographic regions (i.e. districts). This narrow-focus would allow for the discovery of any potential temporal changes that are not otherwise recognizable. In particular, Figure 2 and Figure 3 help us identify and segment out districts where the crime frequency remained consistent over time versus districts where the crime frequency changed. Examining districts individually can help us determine whether there was uniformity in crime frequency within a given district or not. Such exploration may be useful in determining if the general trend for each district casts a shadow upon or hides meaningful insight at a finer level.

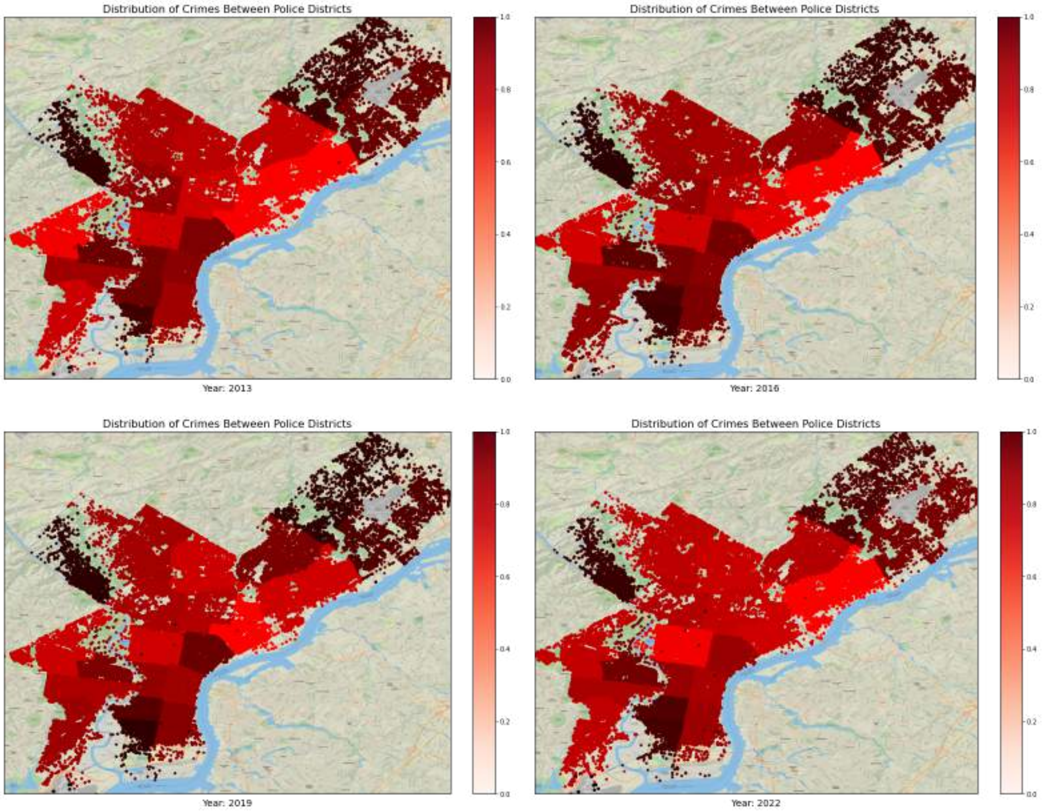

Distribution of Crimes Between Police Districts
Year: 2013
Distribution of Crimes Between Police Districts
Year: 2016
Distribution of Crimes Between Police Districts
Year: 2019
Distribution of Crimes Between Police Districts
Year: 2022
1.0
0.8
0.6
0.4
0.2
0.0

Figure 2: *Heat map of crime percentages for each Philadelphia police district from 2013 - 2019. Dark red regions indicate high crime frequency whereas light red regions indicate lower crime frequency.*

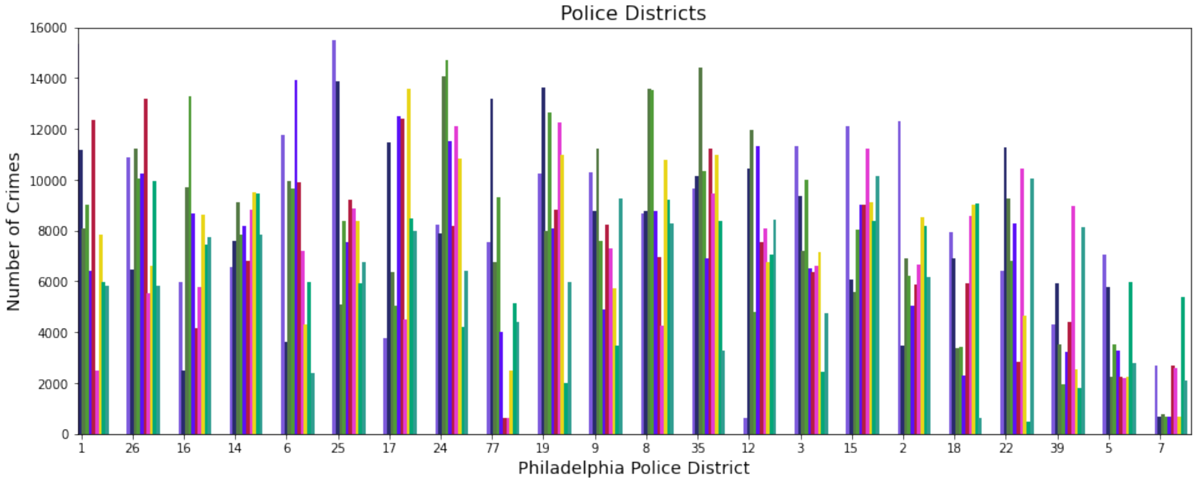


Figure 3: *For each Philadelphia police district, we plot the crime frequencies for every year from 2013 - 2019.*

III.IIIB.C. DBSCAN Clustering

From the previous heat map visualizations we infer that we should *cluster* crimes to extrapolate geospatial hotspots with similar levels of crime. One common clustering algorithm is *k-means clustering*, where we group all points into k clusters of close points. This is done by minimizing an objective function, based on sums of distances squared of points to the centroids of their clusters. However, there are a few downsides to this algorithm:

- First, running k-means requires prior knowledge of the number of clusters, which is inconvenient for a large of clusters.
- Furthermore k-means has difficulty clustering data which has pockets of different sizes and densities, which is something we observe in our heat map visualizations.
- Finally, k-means performs poorly when a large number of outliers are present.

Given these obstacles, we instead opt for a clustering algorithm called **DBSCAN (density based spatial clustering)** [6]. This is a non-parametric algorithm, which groups together points which are packed together; it does not required knowledge of the cluster count beforehand. It is also one of the most commonly-used clustering algorithms, and has been proven to be particularly well-suting in dealing with latitudes and longitudes. Thus, it is well-suited to the task of clustering crimes based on geospatial proximity. Another advantage of DBSCAN is its ability to separate high density areas from low density areas. As seen in the heat map, crime percentages range from 0.4% to over 1%, so we can infer DBSCAN will perform well in this regard.

The main idea behind DBSCAN is that a point belongs to a cluster if it is close to *many* points from that cluster. There are two parameters used in DBSCAN:

$$\epsilon: \text{ minimum distance away for a new point to be added to a cluster,}$$

$$p_{min}: \text{ minimum number of data points to define a cluster.}$$

To choose ϵ, we use KNN (k nearest neighbors). By plotting the geospatial distances from one data point to the others, sorted in ascending order, we form an "elbow" curve. This curve has a sharp change in concavity, which is where we choose ϵ. The value of p_{min} is traditionally chosen to be twice the number of features, so in this case, we set it to $2 \cdot 2 = 4$.

Once DBSCAN is finished running, it classifies all points into three levels of "density".

- A *core point* is one where at least p_{min} points within distance ϵ.
- A *border point* is one that lies at most ϵ away from a core point, but less than p_{min} points within distance ϵ.
- An *outlier (noise)* is anything that is not a core or border point.

The clusters we obtain will consist of the core points: these represent crime hotspots, areas where the geospatial distribution of crimes for a particular year is particularly dense.

iii.iiib.d. Clustering Within Philadelphia Over Time

Our preliminary motivation for this section is to analyze crime within Philadelphia without the context of the police borders. The reason for this is to better understand trends in crime that cross the boundaries. Initially, we plotted a heatmap of crime frequency over all of Philadelphia for the years 2012-2022, as displayed in Figure 4. We can visually observe that there seem to be pockets of higher crime frequency, which motives the need for clustering. Clustering can help identify spatial patterns and groupings in geographic data, allowing for more efficient analysis and decision-making.

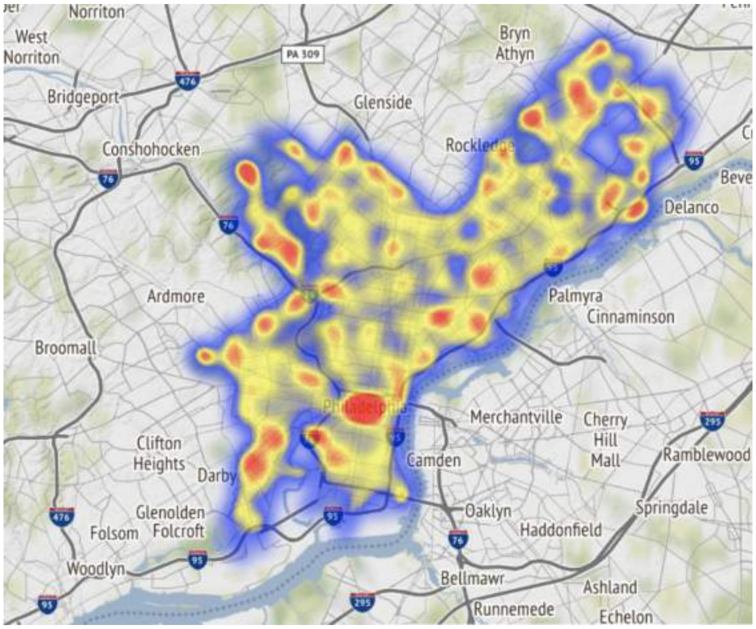


Figure 4: *This figure shows a heatmap of crime frequency over all of Philadelphia for the years 2012-2022.*

Here, we have used DBSCAN clustering over a 9-year time frame. Above, we have selected clustering graphs for every 3 years, starting at 2013 and ending at 2022. Observe that the the locations of particular clusters change significantly over time. Also, notably, the number of clusters has decreased over time. Furthermore, when we overlay one such cluster plot with the district lines, we observe that clusters overlap between different districts. This is actually a reason to consider *clustering within each district, rather than on all Philadelphia:* this overlap suggests an extra level of complexity that cannot be resolved by solely district or solely proximity of latitude and longitude, so we are motivated to use a combination of both.

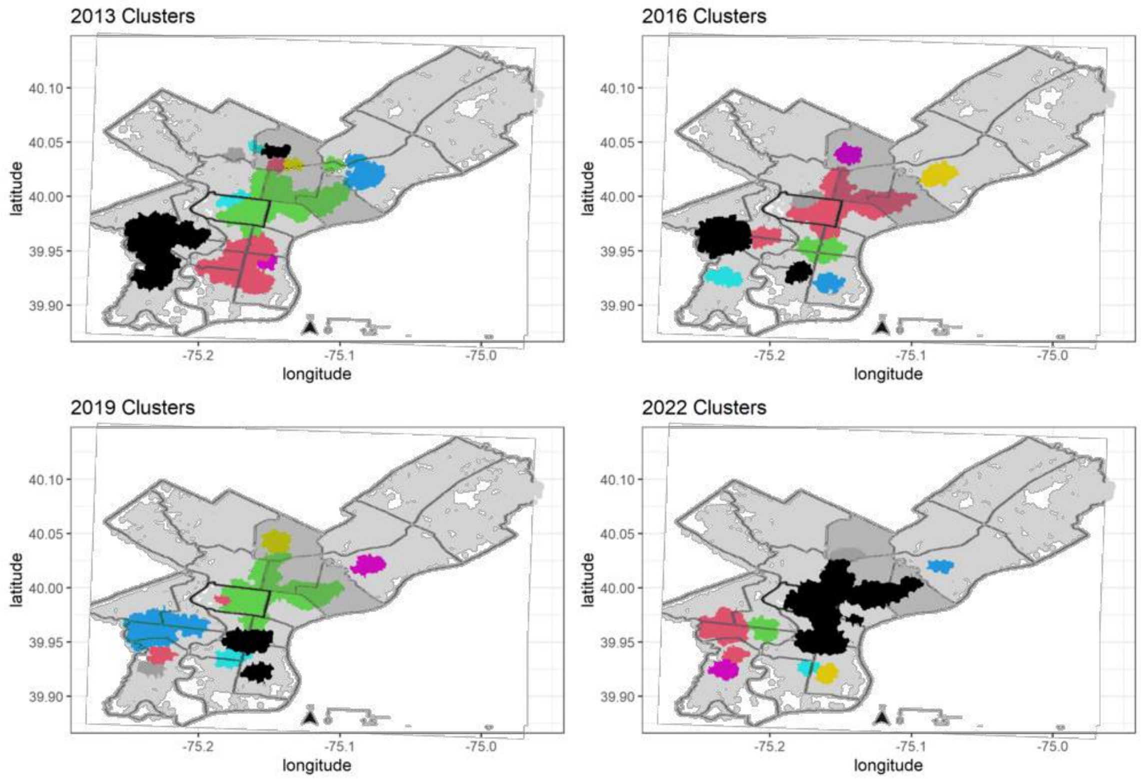


Figure 5: *DBSCAN's Clustering over 9-year time frame*

Now, there are a few other reasons that we want to run DBSCAN clustering within each district. First, from our qualitative observations in Figure 5, the density is different within each district, so it would make sense to represent this non-uniformity through district-focused clustering. Second, aggregate clustering gives at most 13 clusters in all cases, which is significantly less than the number of actual police districts: so, we might gain better insight into geospatial trends of crime by specifying the algorithm runs within each district. Finally, with regards to politics, we would expect districts to separate their police officers and resources: so any sort of policing strategy *should not operate based on clusters which lie in multiple districts.*

III.IIIB.E. Hopkins statistic testing

Before we run clustering on each district, we introduce a statistical testing method which evaluates the success of clustering. The Hopkins statistic provides a measure of how "uniform" a collection $\mathcal{P}$ of N points is. [8] When the Hopkins value is close to 0, the data is largely uniform, and when it is close to 1, the data is highly clustered.

We compute the Hopkins statistic as follows. We first select $n << N$, which will be our subcollection size for the testing. Then, we choose a random subset $\mathcal{Q}$ of n points from $\mathcal{P}$. Next, we randomly and uniformly generate a set $\overline{\mathcal{Q}}$ of n random points. Finally, as our data is

2-dimensional, the Hopkins statistics for these random sets $\mathcal{Q}, \overline{\mathcal{Q}}$ are given by

$$H(\mathcal{Q}, \overline{\mathcal{Q}}) = \frac{\sum_{q \in \mathcal{Q}} \min_{p \neq q \in \mathcal{P}} d(p,q)^2}{\sum_{q \in \mathcal{Q}} \min_{p \neq q \in \mathcal{P}} d(p,q)^2 + \sum_{\overline{q} \in \overline{\mathcal{Q}}} \min_{p \in \mathcal{P}} d(p,\overline{q})^2}. \tag{1}$$

Here, $d(p,q)$ is the distance between p and q.

III.IIIB.F. Clustering within districts over time

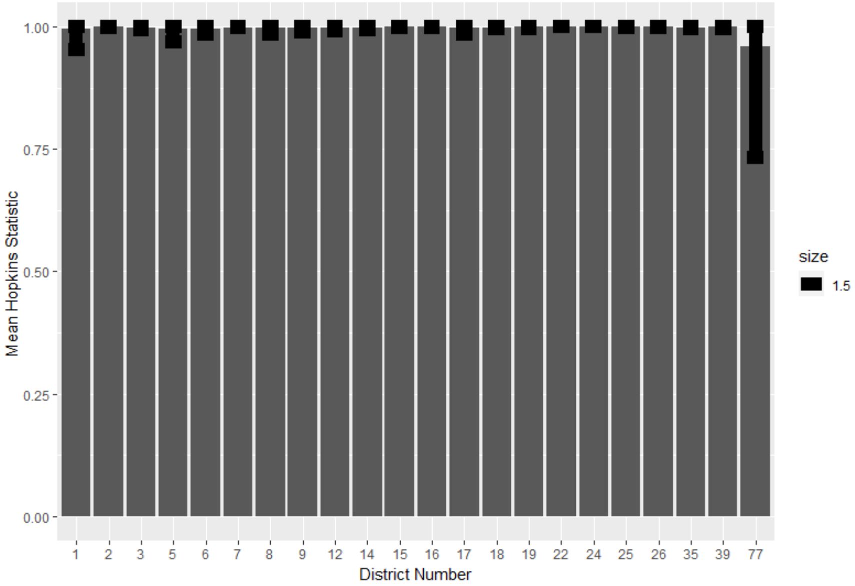


Figure 6: *Mean Hopkins Statistic Per District*

The results of the Hopkins test, as displayed in Figure 6 show that within each district the data is highly cluster-able. This justifies the use of DBScan to visualize these clusters and determine their geographic displacements over time which is critical as explained in methodology.

IIIC. Methodology

The Hopkins test showed us that the crime distribution complexity could be well-modeled within each of the police districts. Given this result, we propose the following methodology to help classify the type of crime occurring in a district in order to implement prevention and mitigation measures which are most likely to be successful.

The overarching goal is to be able to create a model which helps policy makers develop the **most targeted interventions to help prevent crime**.

IIID. Defining "systemic" vs "non-systemic" crime

After visual analysis of the DBScan output on the 21 police districts, a clear theme could be discerned: certain districts would have clusters which don't change geographically over time (Figure 8) whereas other districts do (Figure 7). We believe that differentiating these districts can be extremely useful for policy makers to implement targeted solutions to try and reduce crime.

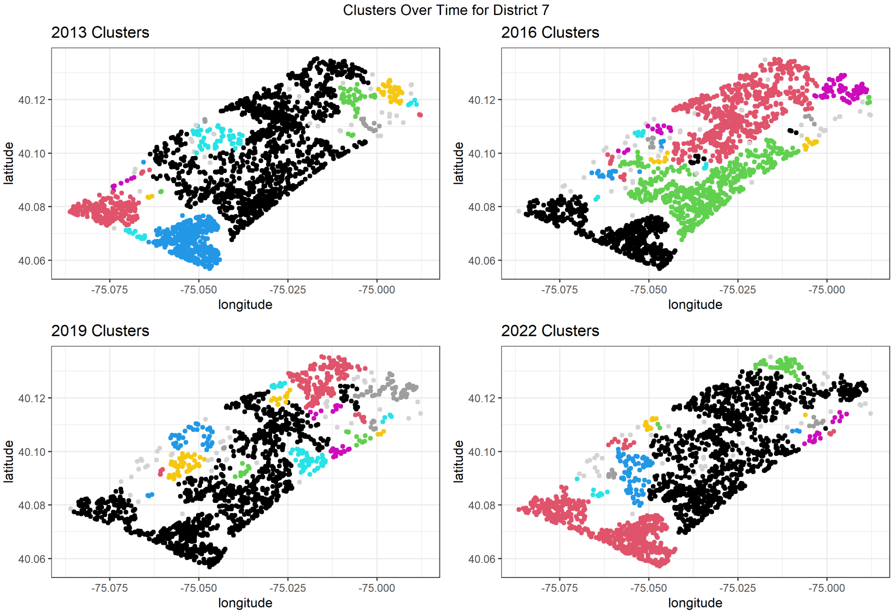


Figure 7: *District 7: Clusters can be seen to change geographic location over time*

Specifically, we propose a method for classifying districts within two categories, systemic and non-systemic, which we define below. This classification allows for targeted implementation of interventions measures.

III.IIID.A. High Level Overview Of Systemic Crime

We have defined systemic crime to be a district in which the crime clusters have not changed in geography over time (statistical procedure for this determination is below). Given this, we suggest that interventions which address the root causes of crime will be more effective than increased policing. For instance, a tool like education and infrastructure investments would be more fruitful interventions in these areas. We attempt to discern factors which drive the "systemic-ness" of the crime such as comparing against historic spatial racism which refers to the use of geographic or spatial factors to perpetuate discriminatory practices.

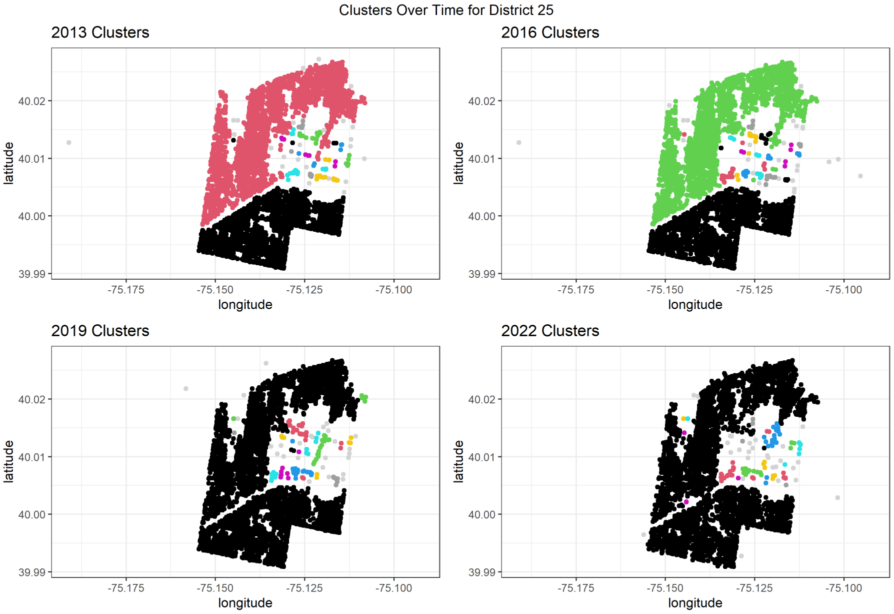


Figure 8: *District 25: Clusters can be seen to stay in the same geographic location over time*

III.IIID.B. High Level Overview Of Non-Systemic Crime

We define a Non-systemic crime district to be one in which the clusters of crime changed geographic locations over time. A real world example of a non-systemic crime district could be one in which crime spiked as a result of recent civil unrest. In a situation such as this, data suggests that increased policing would be more effective in decreasing the rate of crime given that the crime is "non-systemic" in nature. [10] [11]

Now, we introduce our methodology for differentiating clusters within districts as being systemic vs non-systemic.

IIIE. Designing a Systemic Index: Procedure for defining a District with Systemic Crime vs Non-Systemic Crime

We want to differentiate systemic crime, where crime hotspots' geographic locations have not changed over time, from non-systemic crime, where they have changed significantly over time. In this sense, how 'systemic' the crime is corresponds to how much variation there is in the locations of clusters over the years. As it is quite complicated to determine variations in clusters, it would be prudent to instead represent these clusters by their *cluster representative*, the average of their core points in DBSCAN. Formally, for a cluster $\mathcal{C}$, with $\mathcal{C}_{core}$ being the set of *core points* in $\mathcal{C}$, the representative is

$$\tilde{x} = \underset{x \in \mathcal{C}_{core}}{\text{avg}} \; x. \tag{2}$$

Then, for each district, we have temporal data of these collections of *cluster representatives*, and want to determine whether these sets change geospatially over time. The first step is to define a distance between two sets of points.

Suppose we have two sets S_1, S_2 of points in the 2D plane, and we want to see how 'far' apart they are. The traditional idea would be to use a permutation method. Here, we randomly permute the order of S_2, pair off points in S_1 and S_2, and get the sum of the distances between paired points. We average these sums over a large number of random permutations, and this will give a sort of distance metric on sets of points.

One major disadvantage of this method is its bias in classifying areas with multiple clusters as systemic crime, because the sum of distances will be larger. Also, it is not clear which points to pair off if the sizes do not match up. Furthermore, this metric is fairly inaccurate when two sets of points are close together: in fact, for two equal sets of *sparse* points, the computed random permutation metric would be quite large, but the actual distance should be zero.

Thus, we propose a novel metric on two sets of points S_1, S_2. This is defined as follows, where d represents the usual Euclidean metric between two points on a plane, with latitude and longitude being the coordinates:

$$d_{set}(S_1, S_2) = \underset{p_1 \in S_1}{\text{avg}} \min_{p_2 \in S_2} d(p_1, p_2) + \underset{p_2 \in S_2}{\text{avg}} \min_{p_1 \in S_1} d(p_2, p_1). \tag{3}$$

Essentially, for each point in S_1, we find the closest distance to a point in S_2, and then average all these minimal distances; we do the same process from each point in S_2 to S_1; and then finally sum these two averages. This avoids the bias of the random permutation metric, because it takes averages rather than sums. Furthermore, this handles the case where $|S_1| \neq |S_2|$ well, as it computes the average for each of the two sets and adds them. Finally, if two sets of points are close together, because we are taking an average of *minimum distances*, then the metric will return a small value, which is fairly accurate.

With the above metric, we define the **non-systemic index** $\mathcal{NSI}(D)$ for each district D. Suppose we have an array of n sets of points for the district D, say $[S_1, \ldots, S_n]$, where S_i is the set of cluster representatives for year i. Our goal is to quantify the overall physical proximity of these sets. Indeed, by using the distance metric defined in 3, we can form an $n \times n$ *distance matrix* $\mathcal{M}(D)$ for these sets, with entries

$$(\mathcal{M}(D))_{i,j} = d_{set}(S_i, S_j), \quad \forall\ 1 \leq i, j \leq n. \tag{4}$$

Then, the $\mathcal{NSI}$ is computed as the Frobenius norm (2-norm) of this matrix:

$$\mathcal{NSI}(D) = \|\mathcal{M}\|_2. \tag{5}$$

Note that a higher $\mathcal{NSI}$ indicates that the distance matrix is larger in magnitude, meaning the cluster representatives physically move more over time, and thus the crime hotspots change significantly in that district. That is, a high $\mathcal{NSI}$ indicates that the district is **non-systemic**.

From the bar graph in Figure 9, we see that it will be interesting to study the underlying factors for the more systemic and non-systemic districts. Based on the $\mathcal{NSI}$, we will use **District 25 and District 19 in our analysis of systemic districts**, as they have extremely low scores. Also, we will study **District 7 and District 8 in our non-systemic analysis**, as they have the highest $\mathcal{NSI}$ values.

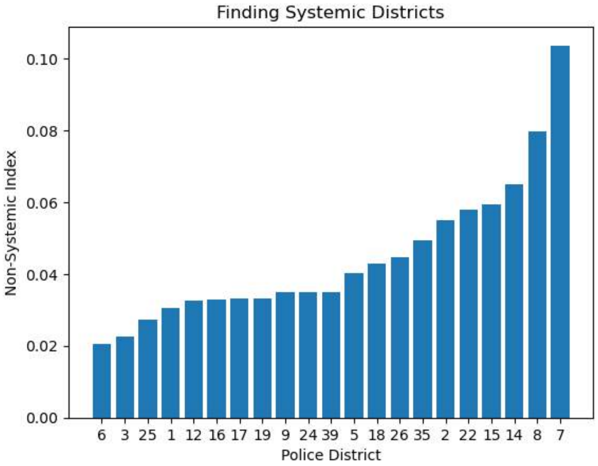


Figure 9: *Police districts are ranked in ascending order by non-systemic indices over the years 2012-2022.*

iiif. Systemic Analysis

We have defined systemic crime as crime in which clusters do not significantly change in geographic position over time. The goal is to now determine what underlying factor has caused the crime to remain in the same location. We propose and hypothesize that one of the major causes for this crime is attributed to systemic racism and the historical redlining that occurred in the city.

iii.iiif.a. What is Redlining?

Redlining, the practice of denying financial services such as loans or insurance to residents of certain neighborhoods based on their race or ethnicity, has had a significant impact on crime rates in affected communities. Redlining occurred primarily in the United States from the 1930s through the 1960s, although its legacy continues to affect many communities today. One of the ways in which redlining has impacted crime is by exacerbating poverty and economic inequality in redlined neighborhoods. By denying residents access to credit and other financial resources, redlining has made it more difficult for people in these neighborhoods to start businesses, buy homes, and otherwise build wealth. This has led to higher levels of poverty and unemployment, which are both risk factors for criminal behavior [4].

In Figure 10, we overlay the red-lined neighborhood geography on top of our clusters to visually examine whether there is an overlap in the red-lined neighborhoods and the clusters from DBScan. To support our hypothesis, we are correlating the presence of redlining and institutionalized racism. We map each latitude-longitude pair to redlining Home Owner's Loan Corporation (HOLC) grade (A is the best whereas D is the worst)- with Mapping Inequality data from The University of Richmond [2].

We elect to use police districts 19 ($\mathcal{NSI} = 0.0332$) and 25 ($\mathcal{NSI} = 0.0273$), as these are classified as systemic, falling below the median $\mathcal{NSI}$ for districts of ≈ 0.06. For our analyses, we

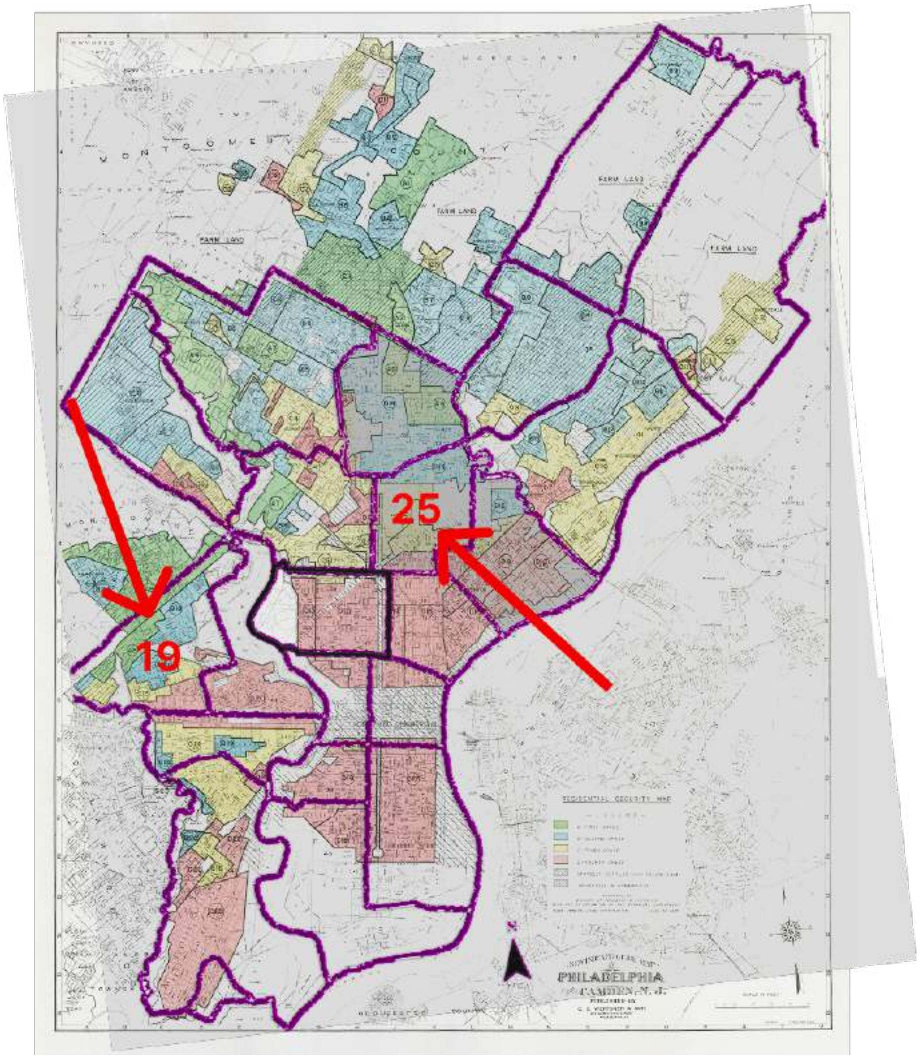


Figure 10: *Redlining Data from 1930s with Current Day Police District Map Overlayed*

re-cluster each district with data across all years, as $\mathcal{NSI}$ established temporal homogeneity. We then compute the breakdown of HOLC grade for each cluster and observe notable differences and inequities. Both police districts are characterized by one far larger cluster and several relatively small clusters. We chose to combine the HOLC grade counts for these smaller clusters. To assure mitigation of introduced bias, we visually observe the HOLC grade breakdown from 10. By cross-referencing these with our earlier made DBSCAN clusters, we see, geographically, that the grades in smaller clusters are primarily homogeneous. We then use Chi Square Tests for Independence to determine if there is a statistically significant association between the two variables.

Both Districts 19 and 25 displayed significant associations with HOLC grade, implying differences between the two cluster groups. It is important to note that District 25 contained no "A" HOLC grade-regions and combined clusters were only in "C" grade regions. However, this still leads us to an important discussion surrounding the lasting effects of redlining.

HOLC Grade	Cluster 1	Other Clusters
A	0.4%	7%
B	20%	39%
C	16%	52%
D	63.6%	2%

Table 1: *Percentage breakdown of HOLC grade by Cluster in District 19. ($\chi^2 = 90.2$, $p = 2.02 \times 10^{-19}$)*

HOLC Grade	Cluster 1	Other Clusters
A	0%	0%
B	17%	0%
C	54%	100%
D	29%	0%

Table 2: *Percentage breakdown of HOLC grade by Cluster in District 25. ($\chi^2 = 59.0$, $p = 1.52 \times 10^{-13}$)*

iiig. non-Systemic Analysis

We base our non-systemic analysis on police districts that displayed a high non-systemic index. This implies low variance in the geographic distribution of crime clusters from year to year. We attempt to determine if there is a present association between the clustered crime regions and the distribution of historical redlining grades.

As shown in Figure 9, the two districts that exhibited the most non-systemic behavior from 2012-2022 were district 7 and district 8. Let us now attempt to understand potential reasons why this observation may be arising. From district map data and online reading, we know that both districts 7 and 8 have large areas of farmland and are more agricultural/rural communities. Let us first turn to district 7, which, on average, has the lowest frequency of crime among all the districts (Figure 3). As defined earlier, being non-systemic means having clusters of crime that change geographic locations over time. In general, given the smaller sample size (total number of crimes committed) in district 7, small changes in the number of crimes committed can have a significant impact on the overall crime rate and consequently the clustering.

Now, we can turn to district 8. Figure 3 shows that district 8 has the highest incidences of crime during the years 2014-2015. It is well known and well-documented that economic conditions can be a causal factor for crime [7]. In particular, rural areas may be more vulnerable to changes in economic conditions, such as a downturn in the agricultural sector. When people lose their jobs or experience financial hardship, they may be more likely to engage in criminal behavior. News articles from 2014-2015 announced that the "Average household farm income in 2015 is projected to be $19,121, down from last year's high of $28,687, according to the USDA"[5]. Thus, the economic pressure in the agricultural communities of districts 7 and 8 may be leading to the non-system patterns we observe. The authors acknowledge that sometimes it can be easy to "explain-away" certain observations so further investigation into districts 7 and 8 would be necessary before making strong claims.

iiih. Interpretation of Results: Targeted Intervention

In our analysis, we have identified that police districts can be classified as either being hosts to systemic or non-systemic crime. This determination is extremely useful for legislators as data shows that different crime mitigation techniques are more effective within systemic vs non-systemic districts. Overall we can say that we have taken a major stride towards achieving our goal of creating actionable insights for effective/targeted crime mitigation strategies.

While our model is capable of analyzing any particular district within the city of Philadelphia, we chose to analyze particular districts due to their more extreme $\mathcal{NSI}$ scores. As a recap, the $\mathcal{NSI}$ score allows us to determine the "systemic-ness" of a particular district. For the following districts, our model along with current event information could result in suggestions similar to the following.

Districts 7 and 8 (Northeast Philadelphia): The high $\mathcal{NSI}$ score indicates these districts as non-systemic. Poverty within this district has also recently been on the rise with neighborhoods such as the Northeast neighborhood of Holmesburg, which went from a 2 percent poverty rate to 19 percent.

Given these two factors, we suggest more reactionary measures. Specifically, a potential policy suggestion would be to increase police presence within this area as well as implementing economic stimulus legislation.

Our analyzed systemic Districts (19,25), both displayed significant statistical associations with HOLC grade. For District 19, we see that 63% of crimes took place in regions historically rated as D. This is clearly vastly different from the 2% in the other combined clusters. This implies that historical and institutionalized bias still plays an important role in determining the geographic distributions of crime. In District 25, the statistical results are not as reliable, as we see all smaller clusters reside in C grade regions. However, the lack of A, and small amount of B graded regions overall in District 25 leads us to a similar conclusion as District 19. In these cases, a push for such interventions as targeted economic development can attract investing and ultimately reduce poverty rates in this area. Another initiative that can combat the historical impacts of redlining, specifically in the context of crime, is community policing. Using such methods as DBSCAN, we have reliably identified crime hotspots within each District. With our research, police forces should formulate plans to increase attention towards these hotspots in the future.

iiii. Limitations

While the Hopkins statistic can be a useful tool in determining the clustering tendency of a data set, there are some assumptions and limitations in its use. These include:

1. *Sensitivity to data distribution:* the Hopkins statistic assumes that the data is uniformly distributed, and it may not work well if the data has a skewed distribution or contains outliers. It is unclear if the clustered data is normally distributed.
2. *The dataset has no noise:* The Hopkins statistic assumes that the dataset does not contain any noise or outliers that could bias the clustering tendency, but in the crime dataset there may

be a large number of outliers in certain districts.

3. *The dataset has a moderate to high number of dimensions:* The Hopkins statistic assumes that the dataset has a moderate to high number of dimensions, as otherwise, it might not be able to distinguish between a clustered and non-clustered dataset. Here, this is potentially problematic, as we only use two dimensions (latitude, longitude). *Nevertheless,* a counterpoint is that because the dataset had many clusters and a complex structure, low dimensionality was not an issue.

Similarly, the χ^2-model has a few assumptions and limitations to consider:

1. *Independence:* The observations should be ideally independent of each other here in the χ^2-test. However, here, the HOLC grade observations may be dependent on each other, based on historical and demographic factors that have affected redlining.
2. *Sample size:* The sample size should be sufficiently large to ensure that the expected frequencies are not too low. This was not an issue in District 19, but was potentially problematic in District 25, where some clusters had zeroes in their percentage breakdown.
3. *Association, not causation:* The χ^2-model can detect an association between two variables, but it cannot determine whether this is causual.

Violation of these assumptions can lead to incorrect results and conclusions. However, violating one or more of these assumptions does not necessarily mean that the Hopkins statistic or χ^2-model cannot be used. Rather, it is generally important to be aware of these assumptions and their potential impact on the results of the analysis.

IV. Conclusion and Future Work

In this paper, we began by exploring crime distributions within the city of Philadelphia and ultimately decided upon temporal and geospatial analysis within set police district boundaries. We then proposed a method for classifying the type of crime within these districts and began to explore the potential underlying causes including factors such as redlining. In order to further this work, we would like to explore different underlying factors and their relationships to the types of crime within the city.

References

[1] Districts: Philadelphia police department. https://www.phillypolice.com/districts/.

[2] Mapping inequality: Redlining in new deal america. https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58.

[3] Public school investment reduces adult crime, study shows | the university record. https://record.umich.edu/articles/public-school-investment-reduces-adult-crime-study-shows/.

[4] Wikipedia Contributors. Redlining. https://en.wikipedia.org/wiki/Redlining, 04 2019.

[5] Christopher Doering. Low commodity prices hit farm income in 2015. https://www.desmoinesregister.com/story/money/agriculture/green-fields/2015/08/25/farm-income-usda/32345389/, Aug 2015.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[7] David Fergusson, Nicola Swain-Campbell, and John Horwood. How does childhood economic disadvantage lead to crime? *Journal of Child Psychology and Psychiatry*, 45:956–966, 07 2004.

[8] BRIAN HOPKINS and J. G. SKELLAM. A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany*, 18(2):213–227, 04 1954.

[9] Daniel Montolio. The effects of local infrastructure investment on crime. *Labour Economics*, 52:210–230, 06 2018.

[10] Siladitya Ray. 35 arrested after protests at atlanta's 'cop city'. https://www.forbes.com/sites/siladityaray/2023/03/06/35-arrested-after-protests-at-atlantas-cop-city/?sh=3132e8da3cbe.

[11] Patrick Sharkey. Perspective | cops prevent violence. but they aren't the only ones who can do it., 06 2020.

[12] Simone Weichselbaum. More cops. is it the answer to fighting crime? https://www.usatoday.com/story/news/investigations/2019/02/13/marshall-project-more-cops-dont-mean-less-crime-experts-say/2818056002/, 02 2019.