Identifying Crime Micro Hotspots in Seattle through Density Clustering

Edouard Seryozhenkov[1], **Himanshu Naidu**[1], **Joobee Jung**[1], **Soohyun Hwangbo**[2]
[1]MS in Data Science (MSDS), University of Washington
[2]Department of Statistics, University of Washington
edouas@uw.edu, hnaidu36@uw.edu, jbjunguw@uw.edu, shwangbo@uw.edu

1 Introduction

Micro crime places, or micro 'hotspots', are small geographic locations within larger communities that disproportionately contribute to crime. A number of past studies, such as Weisburd et al. [2004] and Khare et al. [2023], have attempted to shed light on how crimes occur in dense hotspots in urban regions. However, these studies tend to focus more on larger scale clusters, and not much has been studied about hotspots at a more granular level.

In this study, we recognize the need to examine the presence of and change in crime micro hotspots in Seattle using updated data and a more granular detection method. Such examination of crime hotspots would provide valuable insights for targeted crime prevention in Seattle and other major cities. However, uncovering micro hotspots presents a unique clustering challenge. Unlike situations with a pre-defined number of clusters, there may be arbitrarily many peaks in crime density at any given time, with an arbitrary amount of "noise" alongside the clusters (crimes that do not belong to a micro hotspot). Furthermore, crime micro hotspots may vary widely in shape, size, and absolute density across geographic regions, and over time.

Our research questions are as follows:

1. Can we identify the crime micro hotspots in Seattle using clustering algorithms, and determine the optimal method and parameter choice for micro hotspot detection?

2. Do crime micro hotspots remain consistent, or do they change in location over time? If so, how much do they change, and what factors may influence these changes?

3. What are the potential factors (geographic identifiers, points of interest) that may be associated with and/or characterize crime micro hotspots?

The first part of our project focuses on clustering to identify crime micro hotspots in Seattle between 2008-2023. In the clustering process, we introduce our assumptions on what qualifies as a 'micro crime hotspot' and track cluster changes over 1-year time windows. Existing literature hypothesizes that regions with low cluster movement may be associated with systemic crime, and may be influenced by discriminatory practices such as redlining. We aim to test this hypothesis by checking for correlations between redlining grade values and cluster movement.

After identifying the crime hotspots and assessing their movement, we then further characterize the clusters by checking if there are certain geographic points of interests that tend to be more present near crime hotspots based on prior literature. Distributional differences in crime type by point of interest type are also explored.

Preprint. Under review.

2 Related Work

2.1 Crime Hotspot Analysis Using Clustering

Past studies have established the theory that crimes in urban regions tend to occur in dense clusters, or hotspots. Using the official crime data of Seattle, Washington over a 14-year period, Weisburd et al. [2004] examined the distribution of crime across street segments in Seattle. They concluded that Seattle crime was not dispersed in occurrence, but was concentrated in micro hotspots, and that historical changes in Seattle crime rate were driven by changes at this scale.

More recent studies like Khare et al. [2023] have identified urban crime hotspots through clustering models. The authors analyzed the movement of these crime hotspots in each district of Philadelphia, over time. The authors associate crime hotspot movement with the systemic nature of the associated crime, and measure the cluster movement by introducing a novel distance metric (called Non-Systemic Index) to measure between sets of cluster representatives from different years (the lower the hotspot movement, the more systemic the crime).

2.2 Association of Geographic Attributes and Crime Hotspots

The idea of characterizing crime micro hotspots using point of interest information is grounded in the concept of "criminology of place'.' In terms of points of interest, Seaman and Linz [2014] shows that the presence of certain businesses or facilities may be associated with more concentrated criminal activity (ex. liquor stores). Engstad [1973] highlighted correlations between businesses like hotels and shopping centers and elevated crime rates, particularly auto and bar crimes. Engstad [1973] suggested a potential hotspot effect, with areas containing these businesses experiencing higher crime rates. Chamberlain [2014] showed reducing the distance to the nearest liquor retailer by one mile leads to an increase of crime rates in Seattle, notably affecting violent and drug-related crimes.

3 Data

3.1 Seattle Crime Dataset

The main dataset used for this project is a database of geo-located Seattle crimes that ocurred between January 2008 and December 2023, freely available on the Seattle Open Data Portal here: Seattle Police Department (SPD) Crime Data, from 2008-Present. This database contains 1.12 million rows of crime reports and consists of 17 columns that include the approximate time of crime occurrence, type of crime, blurred geographical location, and victimization category. Of this data, we intend to primarily focus on the spatio-temporal location and type of the crime for clustering.

3.2 Mapping Inequality (Redlining)

In order to test the association between regions with possible systemic crime, and redlining practices, we utilized the Mapping Inequality dataset from Nelson et al. [2023]. The "Mapping Inequality" dataset by the Digital Scholarship Lab is a significant digital humanities project that visualizes the Home Owners' Loan Corporation (HOLC, redlining) maps, from 1930s. In these maps, areas marked with Grades A or B were deemed 'Safe', while those with Grades C or D were deemed 'Hazardous'.

This data is provided in geojson format, which we utilize by performing a spatial join between the clustered data points and the geojson data, to find the HOLC grades for the given points.

3.3 OpenStreetMap

To further characterize our crime hotspot clusters, we utilized OpenStreetMap (OpenStreetMap contributors [2017]). OpenStreetMap (OSM) is a collaborative project that aims to create a free, editable map of the world. Unlike traditional mapping services that rely on proprietary data and closed systems, OSM allows anyone to contribute geographic information, such as roads, trails, landmarks, and points of interest, enabling construction of crowdsourced geographic data. The Overpass API (Wiki [2023]) offers a flexible interface for accessing OSM's rich geospatial data. We used the Overpass API for retrieval of specific points of interest (eg., schools, hospitals, subway stations, smoke shops, etc.) near micro hotspots.

4 Methods

4.1 Mining the Clusters

We propose that crime micro hotspots are crime density peaks of variable shape and size, but typically no bigger than 500 meters in diameter. In model selection, we therefore required a clustering method that was density-based, flexible, yet also capable of yielding stable, size-limited clusters. This turned out to be a challenge, as locations are blurred in the SPD Crime Dataset. As a result, the crimes in each neighborhood are approximately evenly spaced, which makes setting linkage parameters for clustering algorithms very difficult (either nothing is connected, or everything is). We tested DBSCAN Ester et al. [1996], OPTICS Ankerst et al. [1999], and HDBSCAN McInnes et al. [2017] clustering algorithms as baseline approaches. Model search suggested that HDBSCAN McInnes et al. [2017] could work for our use case, with the help of outlier filtering and cluster post-processing.

4.1.1 Clustering Setup

From the perspective of policing, Seattle is broken down into several different units of geographical analysis—Precincts, sectors, and beats. In EDA, we noticed crime density, along with geographical density varies by neighborhood. We therefore chose to cluster by sector, which typically represent 1-3 adjacent neighborhoods, as this choice offered a good tradeoff between maximizing data per unit while assuring geographical homogeneity. We then implemented a "sliding time window" and clustered the crimes by sector by window.

For our distance metric, we chose *haversine*, or great circle distance, which can be computed between two points from their latitude and longitude. This distance assumes a spherical Earth, and so for our purposes needed correction by the radius of the earth at the latitude of interest (we used the radius of the Earth at the Space Needle as a reference). For a given inner angle a, with latitude longitude points (λ_1, ϕ_1) and (λ_2, ϕ_2), the haversine distance is given by:

$$\begin{aligned} \text{hav}(a) &= \sin^2\left(\frac{a}{2}\right) \\ \text{hav}(\Delta) &= \text{hav}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\text{hav}(\lambda_2 - \lambda_1). \end{aligned}$$

4.1.2 HDBSCAN

We clustered the crime using the HDBSCAN algorithm. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) by McInnes et al. [2017] is a follow-up to the popular DBSCAN clustering algorithm Ester et al. [1996] that extends the model by converting it to a hierarchical clustering algorithm, then extracting a flat output based on cluster stability. The key point in the implementation that made us prefer HDBSCAN is that it automatically searches over its linkage parameter for the most stable output, saving us from the variance problem we mentioned earlier. In broad strokes, the HDBSCAN algorithm works as follows:

1. Set a density threshold `min_samples`. If a point has `min_samples` many crimes in its neighborhood, then it is considered dense, and called a `core point`.
2. Consider the neighborhood radius ε that would be required for each point in the data to be considered a `core point`.
3. For each ε, build the clusters by grouping the `core points` at that ε using single linkage. Use these clusters to build a hierarchy of clusters.
4. Return the most persistent clusters across the ε range.

4.1.3 Parameterization and Raw Output

To tailor HDBSCAN for our problem, we had to adjust the single linkage parameters, density thresholds, and maximum values for ε. Due to differences in density and crime count between neighborhoods that caused significant variance in output, we chose the density parameter flexibly, in

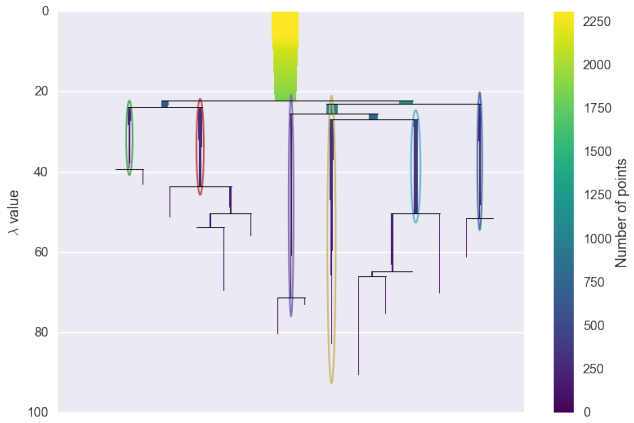
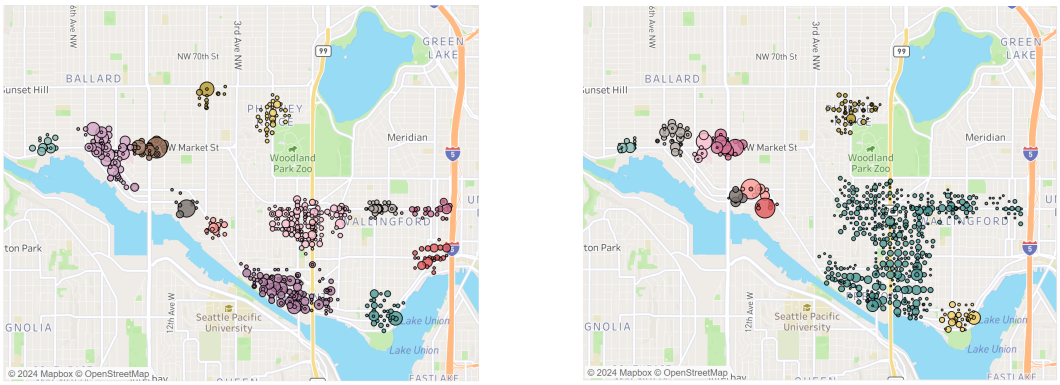


Figure 1: A sample HDBSCAN dendrogram, showing cluster persistence by ε (called λ in the figure)

terms of of a percentage α of the crime in the window. We then performed a grid search over the possible α's and distance thresholds for each sector. Parameter selection was based on the percentage of window crime clustered, the total number of clusters generated, and the size and shape of the clusters, which we visualized in Tableau.

In testing, we found that the best output was achieved when there 5-10 clusters that captured about 30-50% of the crime in the window. We found such output to have the most compact, stable, and nontrivial clusters. Example raw clusters for two windows in Ballard/Fremont are shown in figure 2 . The clusters were compact for most windows, but output was degenerate for a selection of windows/neighborhoods.



(a) Raw clusters in Ballard and Fremont in 2013

(b) Raw clusters (2020)—note the oversized cluster

Figure 2: Though clustering with HDBSCAN was largely successful, there remained windows with oversized clusters, even upon multiple runs. This indicated the need for cluster post-processing.

4.2 Post-Processing the Clusters

The objective of post-processing was to eliminate outliers within each cluster and reduce the radius to 250 meters from each centroid, refining clusters for crime cluster characterization by points of interest using OSM. This radius was chosen for its practicality, aligning with the average 5-minute walking distance for 250 meters, defining micro-hotspots conveniently traversable on foot. It also served as a threshold for querying the OSM API, ensuring retrieved data corresponded accurately to refined clusters. The Haversine distance was used to calculate the distance from data points to centroids.

Post-processing involved identifying cluster centroids and calculating distances from all points to centroids by year and sector. Mean and standard deviation of these distances provided insight into point spread around each centroid. Outliers exceeding the mean distance by more than two standard

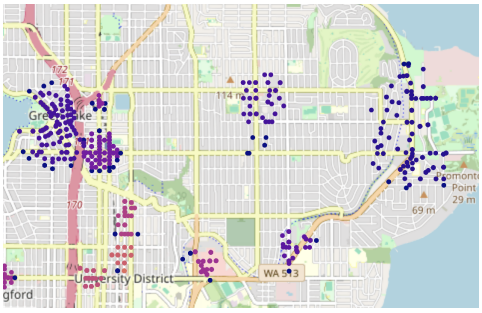
deviations were removed. Afterward, new centroids' latitude and longitude were calculated, and data points were reassigned to the closest centroids within the mean distance plus two standard deviations. Finally, centroids were recalculated with newly assigned cluster points for analysis and results consistency.

Algorithm 1 Post Process Clusters

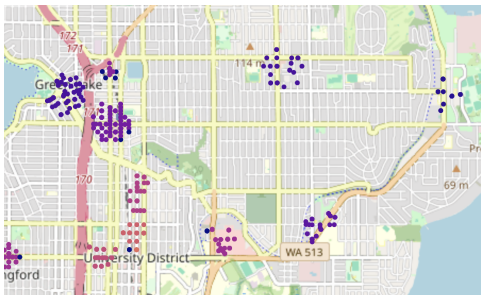
```
procedure POST_PROCESS_CLUSTERS(clustered_df, sectors_to_cluster)
    Parameters: clustered dataframe, sectors to cluster
    for each year from 2008 to 2023 do
        for each sector in sectors_to_cluster do
            Calculate initial centroids.
            Calculate distances to centroids.
            Filter points within mean + 2 standard deviation.
            Recalculate centroids with filtered points.
            Reassign points to closest clusters.
            Update centroids and distances after reassignment.
        end for
    end for
    return result DataFrame
end procedure
```

Figure 3 shows how clusters look in the University District in 2023 after post-processing. Figure 3 (a) shows that even though we reassigned data points to the closest centroids within calculated distance, the large clusters still had larger aggregation than others. Near the University District, there are some dark-blue points near pink and orange clusters, which were filtered out due to their distances from the clusters' centroids.

Once we set the 250-meter radius constraints to the post-processed clusters as in Figure 3 (b), the clusters become consistently similar in size but still reflect density through the number of data points in each cluster.



(a) Post-processed clusters in the University District in 2023



(b) Post-processed clusters with radius 250 meters in the University District in 2023

Figure 3: Post-processed clusters without and with 250-meter Radius Constraints

4.3 Cluster Tracking and Movement Calculation Over Time Windows

The post-processed clusters are used to measure changes in crime hotspot patterns across time-windows for each sector. Existing metrics, like the Non-Systemic Index by Khare et al. [2023], match centroids across windows and sum the distances between matched pairs. However, these metrics overlook key issues:

1. **Pairing Remote Clusters:** Existing methods pair clusters even if they are unreasonably far apart, resulting in distance values that, while relatively high, do not capture the high variability effectively. We propose treating such clusters as separate and assigning a high movement value individually, proportional to their sizes.
2. **Size Differences in Pairs:** Current metrics only consider the distance between centroids, ignoring size differences. A comprehensive measure should include distance, size as well as shape differences.
   As an extreme example, clusters with coinciding centroids but different sizes are still distinct, but existing metrics would treat them as equal.

To address these issues, we propose a nuanced cluster movement calculation procedure.

4.3.1 Cluster Matching

For cluster matching, we follow the overarching idea followed by existing literature, of matching the centroids across windows that are closest to each other. However, to avoid unreasonable matches, we implement a dynamic distance threshold based on cluster sizes.

For each cluster, we calculate the standard deviation of the distances of all data points from the centroid. Clusters are matched if the distance between their centroids is less than twice the standard deviation of either cluster, thus using the standard deviation as a proxy for cluster radius.

We also allow many-to-one and one-to-many matches, recognizing that clusters can split or merge between time windows. This flexible matching accommodates the dynamic nature of cluster evolution over time. Conversely, this matching algorithm would also leave out some clusters unpaired, thus the metric for cluster movement will have to account for this edge case.

4.3.2 Cluster Movement Calculation

To quantify the movement in clusters across time-windows, we use Earth Mover's Distance (EMD) (Rubner et al. [1998]). EMD measures the minimum cost of transforming one distribution into another, providing a robust and versatile metric for our use case. It serves our use case perfectly, as it can handle unmatched clusters, as well as take into account size and distribution differences.

We first characterize each cluster as a 2-D histogram, capturing the spatial distribution of the points within the cluster, with all the co-ordinates inside Seattle as possible bins. Such a distance metric can also account for unpaired clusters by comparing them with dummy (null) distributions.

EMD is calculated as follows:

Let $C_1 = \{(b_1^1, w_1^1), \ldots, (b_1^n, w_1^n)\}$ and $C_2 = \{(b_2^1, w_2^1), \ldots, (b_2^m, w_2^m)\}$ be the histograms of clusters C_1 and C_2, where b^i represents bin coordinates and w^i represents bins frequencies.

The EMD is given by the optimal solution to the following linear programming problem:

$$\text{EMD}(C_1, C_2) = \min_{\{f_{ij}\}} \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} \cdot d(b_1^i, b_2^j)$$

subject to:

$$\sum_{j=1}^{m} f_{ij} \leq w_1^i \quad \text{for all } i; \sum_{i=1}^{n} f_{ij} \leq w_2^j \quad \text{for all } j; \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} = \min\left(\sum_{i=1}^{n} w_1^i, \sum_{j=1}^{m} w_2^j\right)$$

where f_{ij} represents the flow from bin b_1^i in C_1 to bin b_2^j in C_2, and $d(b_1^i, b_2^j)$ is the haversine distance between bin b_1^i and bin b_2^j.

For each sector S_k, across two consecutive time windows (Y_i, Y_{i+1}), we calculate the average EMD. Let S_k have the following cluster pairs in (Y_i, Y_{i+1}): $\{(C_1^1, C_2^1), (C_1^2, C_2^2) \ldots, (C_1^N, C_2^N)\}$, where we can have dummy clusters to account for unmatched clusters. The average EMD value is given by:

$$EMD_k(Y_i, Y_{i+1}) = \frac{1}{N} \sum_{a=1}^{N} EMD(C_1^a, C_2^a)$$

We finally average the EMD values for a sector across all time windows. Let's say there are I years, thus I-1 consecutive windows. Thus, the final cluster movement value is given by:

$$M(S_k) = \frac{1}{I-1} \sum_{i}^{I-1} EMD_k(Y_{i+s}, Y_{i+1+s})$$

where s is the starting year.

This distance metric thus reflects both the spatial and distributional differences between clusters.

4.4 Associations with Redlining

Based on the work of Khare et al. (2023) Khare et al. [2023], we investigated correlations between the movement in crime hotspots over time and redlining practices. This was achieved by utilizing the Home Owners' Loan Corporation (HOLC) grade distributions associated with historical redlining practices.

Following existing methodologies, we identified sectors with the lowest cluster movement values, indicating systemic crime patterns. For each sector, using the 'Mapping Inequality' dataset, we obtained HOLC grade distributions for crime hotspots of that sector and compared them to the overall Seattle distribution. In line with existing literature, we ran a chi-square test to check how significantly the HOLC grade proportions of that sector differ from that of Seattle.

The chi-square test statistic is calculated as follows: Let's say O_i represents the proportion (proxy for frequency) of the data points inside the clusters of Sector S_k with HOLC grade i, while E_i represents the same proportion for the entire Seattle. The chi-square test statistic is given by:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The p-value of the test statistic is then calculated to check for a significant result.

Despite statistical significance, the chi-square test results did not satisfactorily explain the correlation between HOLC grades and crime hotspot changes (see Section 5 for details). Thus, in order to test the correlation, we propose using an F-test based on a linear regression model, with cluster movement as the response variable and HOLC grade proportions as predictors. The F-test assesses the overall significance of the regression model.

The F-test statistic for a regression model is given by:

$$F = \frac{SSR/p}{SSE/(n-p-1)}$$

where SSR is the sum of squared residuals due to regression, SSE is the sum of squared errors, p is the number of predictors and n is the number of observations.

By applying the F-test and getting the p-value, we can determine if HOLC grade proportions significantly explain the variance in cluster movement values, providing a stronger basis for understanding the correlation between historical redlining and contemporary crime patterns.

5 Results

5.1 Cluster Tracking and Movement Calculation Over Time Windows

The Figure 4 shows the values of cluster movement based on the Earth Mover's Distance, for the 16 sectors in the dataset. We can see significant variation in the movement values.

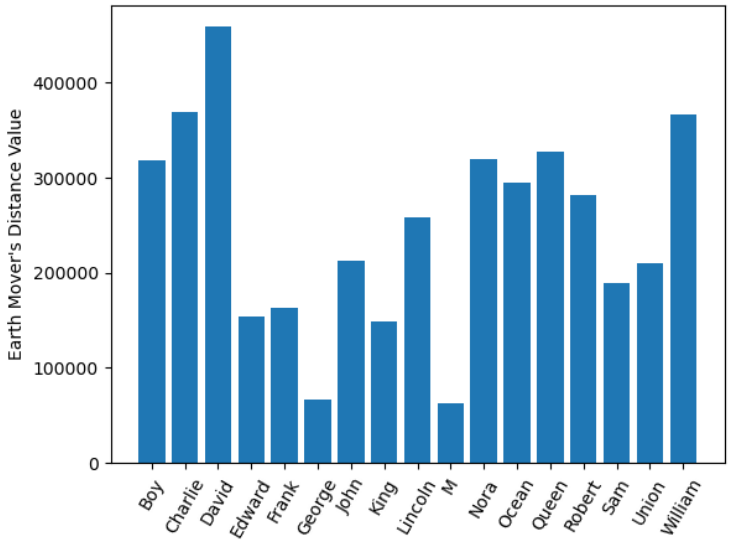


Figure 4: Cluster Movement Values across all Sectors

5.2 Associations with Redlining

The Table 5.2 shows the p-values of the Chi-square tests conducted on some of the sectors with varying cluster movement values.

Sector	Cluster Movement Value (M)	Chi-Square Test Statistic	Chi-Square P-Value
George	66095.868	255.151	$5.031\ 10^{-55}$
King	149287.005	261.672	$1.954\ 10^{-56}$
Edward	153479.967	67.861	$1.224\ 10^{-14}$
Charlie	369246.387	14.387	0.0024
David	458331.558	252.885	$1.554\ 10^{-54}$

Table 1: Chi-Square test results on Correlations between Redlining and Cluster Movement

While these results may indicate an extremely significant correlation between redlining and cluster movement for sectors with low cluster movement (George and King), no conclusions based on them hold valid. This is because we found similar results even for sectors with the highest cluster movement values (Charlie and David). Existing literature failed to report values for such regions, which thus may have given an inaccurate picture of these associations.

On further analysis, it becomes evident that such a test cannot measure correlations between redlining and cluster movement, as it fails to take into account a multitude of other factors in each sector deviates from the entirety of Seattle. Thus, simply comparing the grade distributions of a sector with that of Seattle does not lead to any conclusive results.

The Table 5.2 represents the F-test details on running a regression model, with cluster movement value as the response, and the HOLC grade proportions as the predictors. We see that the F-test does not give a statistically significant result, which would lead one to conclude that the redlining practices may not be correlated with cluster movement. This result conflicts with the notion of existing literature, that regions negatively affected by redlining practices may have lower crime movement.

F-Test Statistic	F-Test P Value	Grade A Coeff	Grade B Coeff	Grade C Coeff	Grade D Coeff
1.996	0.154	26655.056	1072.927	-165.578	-1587.706

Table 2: F-test results on Correlations between Redlining and Cluster Movement

5.3 Crime Cluster Characterization by Points of Interests

After identifying the crime clusters using HDBSCAN and tracking the cluster movement over time windows, we conducted a further analysis of the clusters to understand their characteristics by assessing key geographic points of interests that are in proximity of the clusters. This was based on the prior literature in criminology that the presence of certain businesses or facilities - especially adult businesses like liquor stores, bars, and nightclubs - fall outside the heaviest concentrations of criminal activity.

We first defined a set of geographic 'points of interests' (POIs) that we wanted to examine. This was determined by leads from literature review, as well as our understanding of the recent crime trends in urban areas. OpenStreetMap (OSM) provides a documentation of how their tags are defined, and we conducted verification of the chosen tags to validate that we were retrieving intended geographic points. The points of interests include: *[Bus Stop, Railway Station, School, College, University, Hospital, Alcohol Store, Mall, Supermarket, Park, Bar, Nightclub, Stripclub, Gas Station]*.

In the initial stage of cluster characterization, we retrieved the distribution of these points in the 250-meter radius of the identified crime clusters' centroids. Note that the OSM API did not provide historical records of geographic points, so we limited our cluster characterization to the most recent year: 2023.

The following figure shows the distribution of the specified points of interests, found within 250m radius of Seattle crime clusters' centroids:

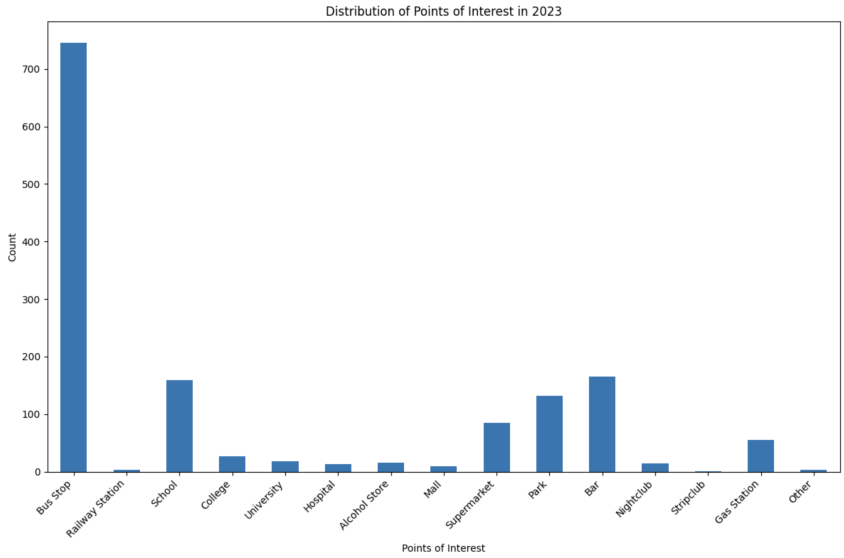
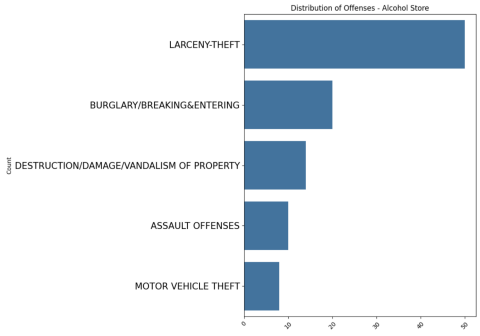


Figure 5: Distribution of Points of Interests Around Crime Hotspots in 2023

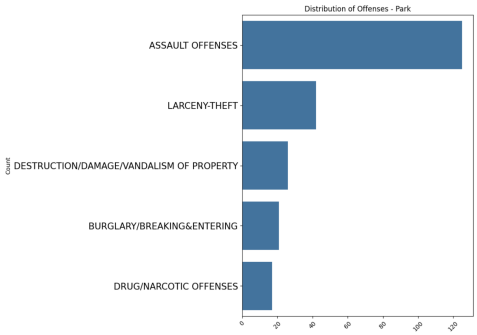
From the above figure, we see that facilities and businesses like bus stops, schools, bars, parks and supermarkets are among the most frequently found geographic points in within crime clusters in Seattle. Since this is a cumulative count of the points of interests, we do not make an argument that having certain points of interests is causally associated with an instance of crime cluster, and restrict our analysis to understanding the characteristics of our final clusters.

Secondly, we proceeded to examine the characteristics of crimes that tend to happen in the most significant clusters around certain POIs, by looking at the distribution of crime occurrences by types. This revealed an interesting difference in crime types of clusters around different POIs: for example, clusters around alcohol/liquor stores were characterized by dominantly high instances of Larceny-Theft, followed by Assault Offenses and Robbery. On the other hand, clusters around parks had the greatest number of Assault Offenses, followed by Larceny-Theft, Property Destruction, and

Drug/Narcotic Offenses. This aligns with the insight from Chamberlain [2014], in that liquor retailers have a positive effect on crime rates, with persistent impacts on violent crimes and shoplifting.



(a) Distribution of Crime Types for the Hotspot Around Alcohol Stores



(b) Distribution of Crime Types for the Hotspot Around Parks

Figure 6

Such characterization of crime types for notable crime clusters around different POIs has significance in helping identify the recent changes in crime trends associated with geographic locations. One example is the increased reports of assaults and violence around hospitals as evidenced by the recent article from Papadakos and Glatter [2023], which is also seen in our cluster analysis: clusters around hospitals tended to have higher cases of Assault Offenses, Larceny-Theft, and Destruction of Property. Identifying micro clusters and their key geographic POIs could therefore inform us of the context behind dense crime clusters, which could be applied to actionable changes in policing allocation and crime prevention measures.

6 Discussion and Limitations

There are several challenges and limitations that we faced through the model implementation and analysis steps. First, the clustering output tended to be highly variable, even after tuning.

Also, it is important to note that assessing the degree and significance of cluster movement depends upon confidence in the raw materials—the clusters themselves. In terms of checking the association between redlining practices and cluster movement, it was difficult to fully incorporate potential confounders into the statistical tests. Nonetheless, we believe that our cluster movement analysis advances existing research in the right step, by providing a more reliable measure for cluster movement, as well as an alternative way to test associations between discriminatory practices and crime patterns.

Lastly, there were limitations in our analysis of Seattle's crime clusters through geographic points of interests, as we were restricted to exploratory analysis and were not able to establish causal relationship between the presence of specific POIs and crime density.

7 Conclusion

In this project, we implemented HDBSCAN to identify crime micro hotspots (or clusters) in Seattle, enforcing the constraint that crime micro hotspots are highly dense with peaks of variable shape and size. We analyzed the extent to which the clusters changed over time, from 2008 to 2023, across all sectors. We proposed a novel procedure for calculating cluster movement, utilizing a flexible cluster matching system and the Earth Mover's distance for the final movement calculation. Contrary to existing research, running an F-test between sector-based cluster movement and redlining grade proportions did not reveal a statistically significant correlation. Additionally, we performed a qualitative analysis of the influence of certain points of interest (such as liquor stores, parks, etc.) on crime patterns. We believe this project paves the way for intriguing analyses on crime and better equips researchers with the appropriate metrics and analytical techniques.

References

David Weisburd, Shawn Bushway, Cynthia Lum, and Sue-Ming Yang. Trajectories of crime at places: A longitudinal study of street segments in the city of seattle*. *Criminology*, 42(2):283–322, 2004. doi: https://doi.org/10.1111/j.1745-9125.2004.tb00521.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9125.2004.tb00521.x.

Ishan S. Khare, Tarun K. Martheswaran, Rahul K. Thomas, and Aditya Bora. Actionable Insights on Philadelphia Crime Hot-Spots: Clustering and Statistical Analysis to Inform Future Crime Legislation. *arXiv e-prints*, art. arXiv:2306.15987, June 2023. doi: 10.48550/arXiv.2306.15987.

Christopher Seaman and Daniel Linz. Are adult businesses crime hotspots? comparing adult businesses to other locations in three cities. *Journal of Criminology*, 2014, 03 2014. doi: 10.1155/2014/783461.

Peter Engstad. *Environmental opportunities and the ecology of crime*. éditeur non identifié, 1973.

Andrew Chamberlain. Urban crime and spatial proximity to liquor: Evidence from a quasi-experiment in seattle. *Urban Economics & Regional Studies eJournal*, 2014. URL https://api.semanticscholar.org/CorpusID:156964502.

Robert K. Nelson, LaDale Winling, and et al. Mapping inequality: Redlining in new deal america, 2023. URL https://dsl.richmond.edu/panorama/redlining. Accessed: 30-May-2024.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

OpenStreetMap Wiki. Overpass api — openstreetmap wiki,, 2023. URL https://wiki.openstreetmap.org/w/index.php?title=Overpass_API&oldid=2624841. [Online; accessed 3-May-2024].

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, jun 1999. ISSN 0163-5808. doi: 10.1145/304181.304187. URL https://doi.org/10.1145/304181.304187.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205. URL https://doi.org/10.21105/joss.00205.

Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998. doi: 10.1109/ICCV.1998.710701.

Robert Papadakos and Peter Glatter. The epidemic of violence in american hospitals, 2023. URL https://time.com/6337450/the-epidemic-of-violence-in-american-hospitals/.

8 Individual Contributions

The following outlines the individual member contributions. Every member had contributions in almost every aspect of the project. Having said that, certain individuals took the lead in some specific tasks, which is what has been outlined below.

Edouard Seryozhenkov: Exploratory Data Analysis, Clustering (Formulation, Testing and Code)

Himanshu Naidu: Cluster Tracking and Movement (Formulation, Code), Testing on Redlining and Cluster Movement (Formulation, Code), OpenStreetMap API (Research)

Joobee Jung: Cluster Post-Processing (Code), Crime Cluster Characterization (Code), Cluster Plotting, OpenStreetMap API (Code)

Soohyun Hwangbo: Report Writeup, Literature Review, Crime Cluster Characterization (Formulation, Analysis)