25S856142: Analysis and Modeling

Final Paper

Analyzing homicide clustering in Washington D.C.: A multi-methodological approach

---

Author:

Max Schneeberger (11903678)

Instructors:

Prof. Gudrun Wallentin

Dr. Christoph Traun

INTRODUCTION

Over the past twenty years, crime mitigation and the geography of crime gained increased interest by criminologists and spatial scientists, counteracting the practice of just concentrating on the offender's attributes (Helbich & Leitner, 2017; Vandeviver & Bernasco, 2017). Already in the year 1989, it was observed that over 50% of police service requests in Minneapolis (United States) were clustered at only 3.3% of all addresses and intersections (Sherman et al., 1989). Consequently, to improve and optimize the temporal and spatial allocation of law enforcement officers, it is necessary to analyze the spatial distribution of crime with respect to clustering and dispersion and to assess what causes these patterns (Dewinter et al., 2020).

One of the most common methods for analyzing the spatial pattern of point data is average nearest neighbor analysis (ANNa) (Colosimo et al., 2021; Melyantono et al., 2021; Thompson et al., 2022). This algorithm provides a global (overall distribution of points in a given study area) assessment whether point locations are either clustered, dispersed or randomly distributed (Colosimo et al., 2021). Clustering or dispersion can manifest at different distances respectively spatial scales however, which ANNa does not consider. For example, wildfire locations may appear as clustered at smaller scales in fire prone regions but dispersed on country scale. One widely recognized algorithm that assesses spatial patterns at various distances is Ripley's K-function (Jannat & Al-Amin, 2024; Wooditch & Weisburd, 2016). Ripley's K-function quantifies the anticipated number of points located within a specified distance from any given observed point. The presence of clustering can be evaluated by comparing the observed Ripley's K-function value to its expected value under the assumption of complete spatial randomness (Self et al., 2023).

With the statistically significant existence of a clustered distribution of crimes, it becomes essential to identify the exact locations of these clusters and to examine the factors associated with their occurrence. One widely used method to detect the locations of crime hotspots is the Getis-Ord Gi* spatial statistic, which calculates statistically significant hot- or cold spots for given input features (Debata et al., 2024; He et al., 2022; Mondal et al., 2022). To explore the factors potentially associated with the emergence of these hotspots, artificial intelligence (AI) methodologies have shown promise in uncovering complex, often non-linear relationships between dependent and explanatory variables (Kwon et al., 2021; Reza et al., 2025).

This research hypothesizes a clustered distribution of homicide point data from 2024 in Washington D.C., influenced by, respectively associated with sociodemographic variables such as population density, income, race, and age distribution. The aim is to investigate which variables have the biggest relationship with the occurrence of homicide hotspots. For initially proving a clustered distribution of homicide locations, Average Nearest Neighbor Analysis and Multi-Distance Spatial Cluster Analysis (Ripley's k-function) will be conducted. Then, the exact location of hotspots will be identified via applying the Getis-Ord Gi* spatial statistic. Finally, an artificial neural network will be trained on labelled training data to classify new potential hotspots based on the given explanatory variables. The results of the model output will also reveal which sociodemographic metrics have the biggest impact on model decision.

METHODS

To assess the overall spatial distribution of homicide locations across the entire study area in Washington, D.C., in 2024, an average nearest neighbor analysis (ANNa) was conducted. ANNa averages all nearest neighbor distances (the distance from a point to its closest neighboring point) in a predefined study area and divides the result by the average nearest neighbor distance of a hypothetical random point distribution. The result is the average nearest neighbor distance index (ANN index). If the ANN index is smaller than 1, then the observed points tend to a clustered distribution. Conversely, if the ANN index is bigger than 1, the observed points tend to a dispersed distribution. The closer the value is to 1, the more likely it is that the points follow complete spatial randomness. To assess the statistical validity of the ANN index value, a z-score and a p-value (derived from the z-score) are calculated additionally.

$ANN = \frac{\overline{D}_O}{\overline{D}_E}$ where: $\overline{D}_o = \frac{\sum_{i=1}^{n} d_i}{n}$ and $\overline{D}_E = \frac{0.5}{\sqrt{n/A}}$

$z = \frac{\overline{D}_O - \overline{D}_E}{SE}$ where: $SE = \frac{0.26136}{\sqrt{n^2/A}}$

d_i = distance between a feature i and its closest neighbor
n = total number of features
A = area of interest

As mentioned in the introduction, the distribution of spatial point patterns is dependent on the scale of analysis. In order to investigate the arrangement of the crime locations at different analysis scales, a multi-distance spatial cluster analysis via a derivative of the Ripley's K Function, namely the L(d) transformation was conducted. Given our study area A (Washington D.C.) encompassing n points (homicide locations), the process involves computing pairwise distances between all points. For each point i, the distance to every other point j is identified, and an indicator function $k_{i,j}$ is utilized to count whether point j lies within a specified distance d from point i. If the condition is met, $k_{i,j} = 1$, otherwise $k_{i,j} = 0$. The quantity of point pairs that meet the distance criterion $d_{i,j} \leq d$ are then summed and normalized by including the study area A, the number of points n, and the constant π to correct for spatial scale variations. Taking the square root of the result value balances the variance and aids interpretation.

$$L(d) = \sqrt{\frac{A \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k_{i,j}}{\pi n(n-1)}}$$

Under the assumption of Complete Spatial Randomness (CSR) the anticipated value of $L(d)$ is equal to d. If $L(d)$ is greater than d, there is a higher point concentration within the distance d than would be expected under CSR, indicating a clustered distribution. If $L(d)$ is less than d, there is a lower point concentration within the distance d than would be expected under CSR indicating a dispersed distribution.

With evident statistically significant clustering of homicide locations in the study area, the next step was to locate where the hotspots appear exactly. Therefore, a Getis-Ord Gi* hot spot analysis has been conducted. This method evaluates for each spatial unit (in our case, census block groups in Washington, D.C.) whether it is a statistically significant hot or cold spot. This is achieved by comparing the local sum of homicide incidents (homicide counts normalized per 1000 population for the current spatial unit and its neighboring units within a specified distance) to the expected local sum of homicide counts under complete spatial randomness, which is based on the global mean of all incidents in the dataset. The result is the Gi* statistic, which produces a z-score and a p-value for each location. High or low z-scores in combination with low p-values indicate statistically significant hot or cold spots, respectively.

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}}$$

x_j = number of homicides at feature j

$w_{i,j}$ = spatial weight between feature i and feature j

$\bar{X}$ = global mean of all homicide counts

S = standard deviation of all homicide counts

n = total number of features (census block groups in our case)

Finally, to answer the research question which sociodemographic metrics have the biggest association with whether a census block group is classified as a hotspot or not, a feedforward artificial neural network (ANN), specifically a Multilayer Perceptron (MLP) was trained on labelled training data. The training data consisted of 456 census block groups (80% of the total census block groups) where each observation had the following attributes: *population density per km^2, percentage of population with low income, percentage of population with low to moderate income, percentage of white population, percentage of black population, percentage of population over age 18,* and the binary label *is hotspot (derived from the results of the Getis-Ord Gi* hot spot analysis)* indicating whether the current census block group is a homicide hotspot or not. The network architecture was made up of an input layer (including the six explanatory features) followed by two hidden layers, each containing 4 neurons with ReLU (Rectified Linear Unit) activation functions. This made it possible to capture the partly non-linear relationships between the dependent and explanatory variables and to assess which sociodemographic metric is contributing most to the model's decision. To ensure representativeness of class distributions across different training and test splits, stratified k-fold cross-validation with 5 folds was performed.

ArcGIS Pro 3.4 was used for average nearest neighbor analysis (ANNa), Multi-Distance Spatial Cluster Analysis (Ripley's k-function), and Getis-Ord Gi* hot spot analysis. The artificial neural network was designed with Python and the usage of the libraries TensorFlow/Keras, Scikit-learn, and SHAP. As for the data sources, open government data from the District of

Columbia were employed (Metropolitan Police Department, 2024). Figure 1 shows the entire methodological workflow.

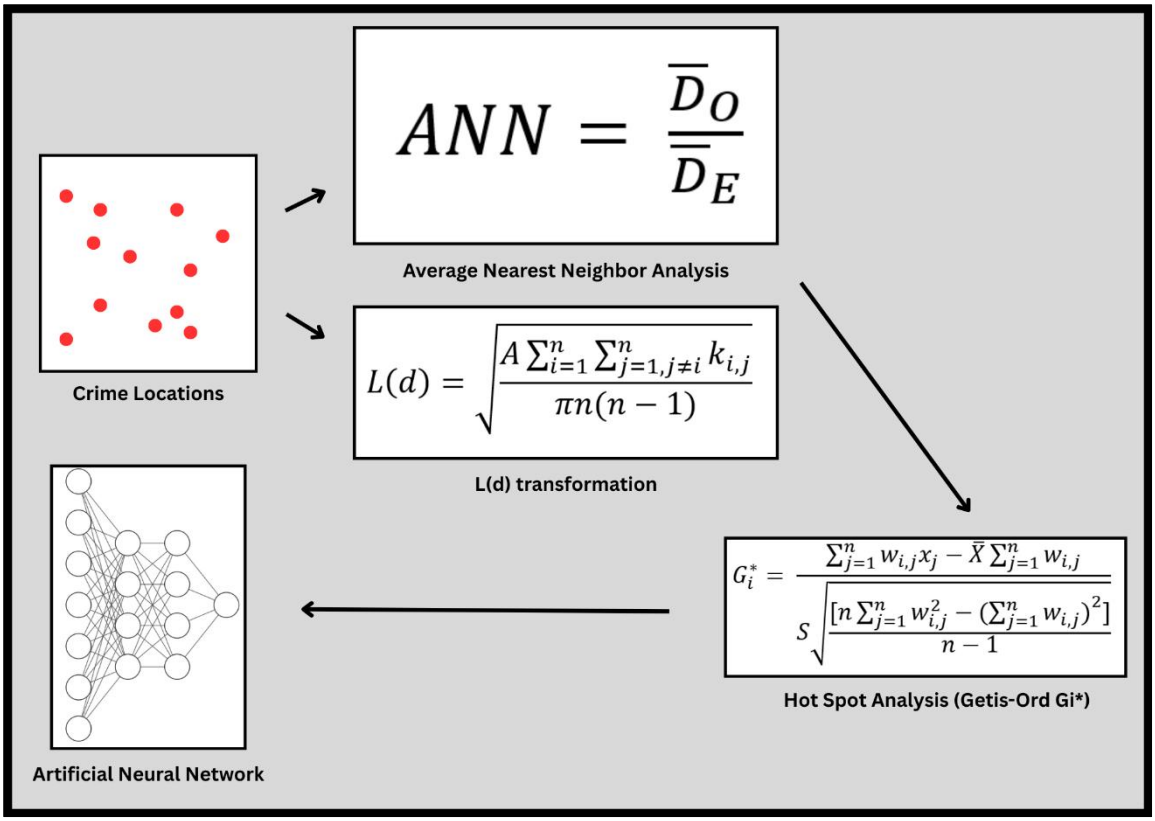

Figure 1: Methodological Workflow Diagram
Note. Created by the author.

RESULTS

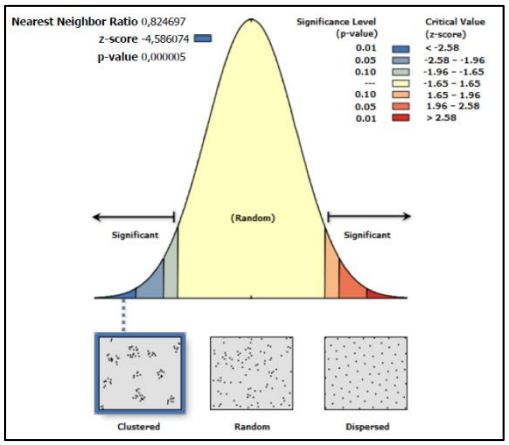

Figure 2: ANN Results

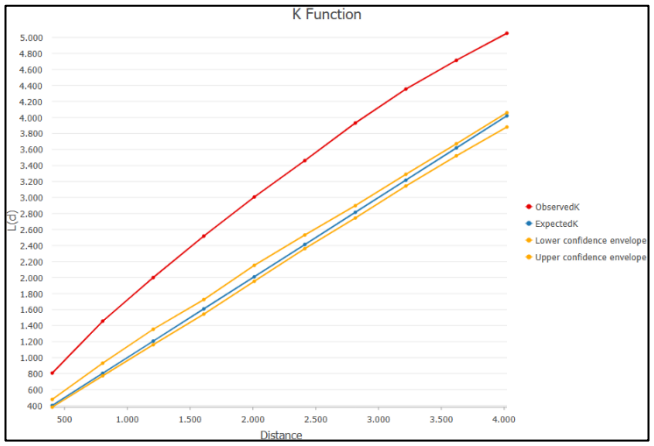

Figure 3: Ripley's K-function (L(d) transformation) Results

The Average Nearest Neighbor Analysis (see Figure 2) resulted in a Nearest Neighbor Ratio of 0.82, a z-score of -4.59 and a p-value of 0.000005. This indicates that homicides follow a statistically significant clustered pattern with less than 1% likelihood that it could be the result of random chance. The results of the Ripley's K-function (see Figure 3) show that ObservedK is higher than ExpectedK throughout the entire distance range which indicates a clustered pattern of homicide locations regardless of the scale of analysis for the given study area.

Figure 4 shows the result of the Getis-Ord Gi* hot spot analysis which visually supports the clustered pattern of homicide hotspots.

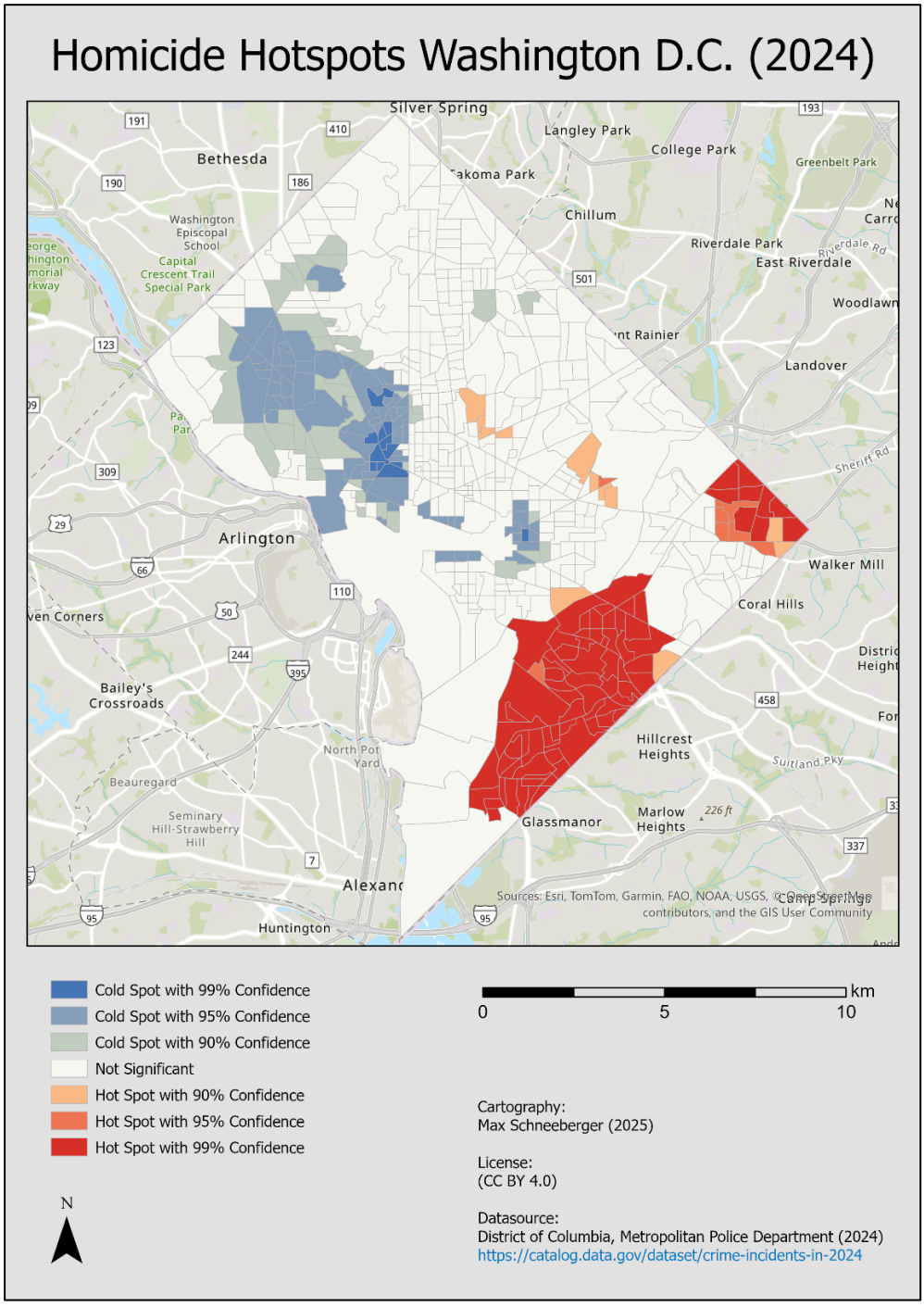

Figure 4: Getis-Ord Gi Hot Spot Analysis Results*

After training the Multilayer Perceptron with five folds of training data, the average performance evaluation metrics for unseen data are as follows: **Average Accuracy: 0.909 | Average Precision: 0.702 | Average Recall: 0.792 | Average F1 Score: 0.742**

Figures 5 to 9 show SHAP (SHapley Additive exPlanations) summary plots for each of the five stratified cross validation folds. The features *percentage of white population,* and *percentage of black population* are constantly among the top two places with regards to model decision influence. The remaining four features have less influence on the model's decision whether a census block group should be characterized as a homicide hotspot or not and fluctuate throughout the folds.

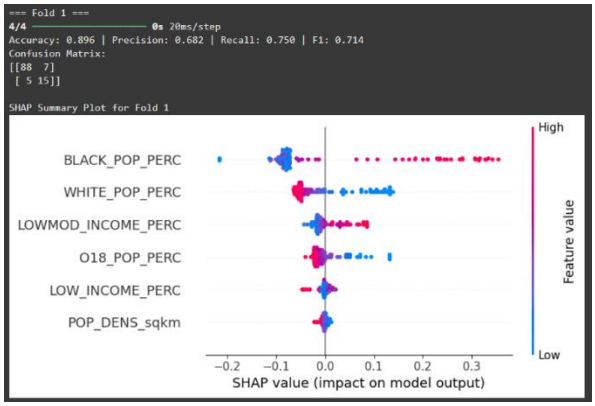

Figure 5: SHAP summary plot (Fold 1)

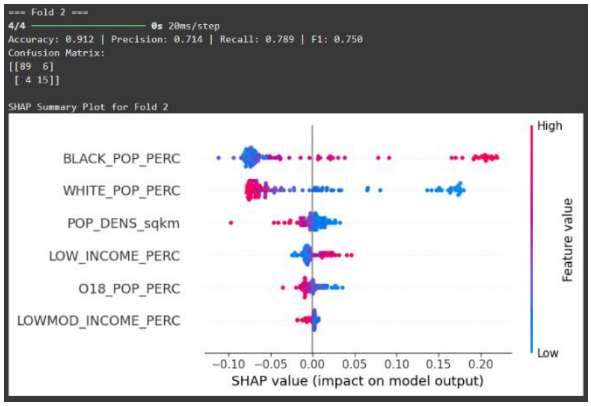

Figure 6: SHAP summary plot (Fold 2)

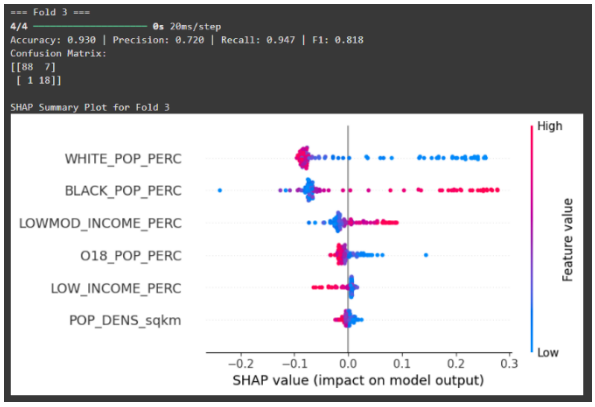

Figure 7: SHAP summary plot (Fold 3)

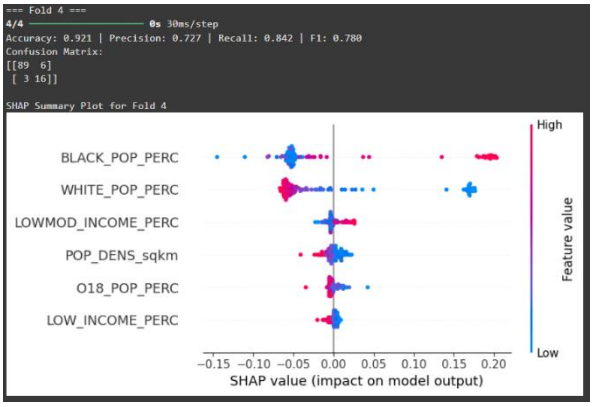

Figure 8: SHAP summary plot (Fold 4)

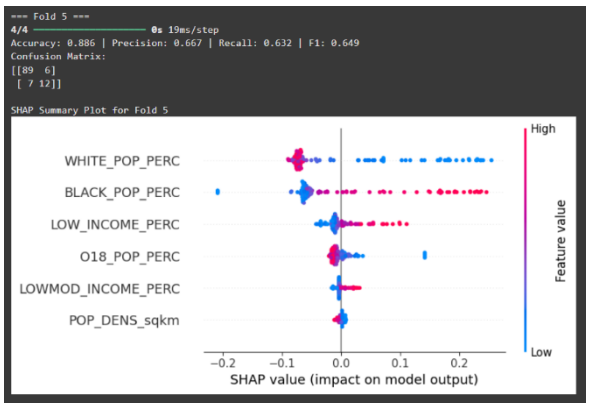

Figure 5:SHAP summary plot (Fold 5)

DISCUSSION

The results of the Average Nearest Neighbor Analysis strongly suggest that homicide locations in Washington D.C. from the year 2024 follow a statistically significant clustered pattern. These findings are consistent with the decades-old assumption that crime is spatially concentrated (Sherman et al., 1989). Also the outcomes of the Ripley's K-function analysis are justifiable, considering the small study area and the fact that variations in clustering and dispersion across different spatial scales are more prominent for larger study areas. The Multilayer Perceptron correctly classified 90.9% of all census block groups. However, given the prevalent class imbalance (only ~17% of observations are labelled as hotspots), this metric alone is not meaningful. Of all the spatial units predicted as hotspots, 70.2% are actual hotspots which shows that the model is not over-predicting hotspots. Among all actual homicide hotspots, the model successfully identifies 79.2% as such, which is solid despite the class imbalance. The F1 Score of 0.742 suggest a reasonable balance between catching true positives and avoiding false alarms. The performance between folds is fluctuating a bit, but overall recall stays relatively high, which is desirable when the main goal is detecting actual hotspots. The SHAP (SHapley Additive exPlanations) summary plots for the respective folds indicate that the variables *percentage of white population,* and *percentage of black population* have the biggest impact on model decision. Census block groups with a high percentage of white population and a low percentage of black population strongly nudge the model to classify this observation as non-hotspot. But SHAP values have to be interpreted with caution – Coupland et al. (2025) point out that while SHAP values accurately report feature importance for the model, they often fail to reveal the underlying causal effects. So, while the raw data used for this study suggests that racial and ethnic composition of census block groups is associated with its likelihood of being classified as a homicide hotspot, Ulmer et al. (2012) emphasize that such patterns largely reflect underlying social, economic, and structural inequalities, as well as systemic biases in the criminal justice system, rather than any inherent racial characteristics. The outcomes of this study serve as an entry point to the current state and patterns of homicide dynamic in Washington D.C. – further research should incorporate historical homicide data to reveal trends and address the fundamental societal, economic and institutional disparities that may reveal the true causation for the development of homicide clusters.

AI STATEMENT

ChatGPT (GPT-4o deep research) was utilized as a supplementary tool to find relevant literature for this paper.

REFERENCES

Colosimo, C., Yon, J. R., Ballesteros, S. R., Walsh, N., Talukder, A., HamIII, P. B., Abuzeid, A. M., & Mentzer, C. J. (2021). Geospatial relationship of trauma and violent crime: An analysis of violent crime and trauma center utilization. *Trauma*, *23*(3), 230–237. https://doi.org/10.1177/1460408620950882

Coupland, H., Scheidwasser, N., Katsiferis, A., Davies, M., Flaxman, S., Hulvej Rod, N., Mishra, S., Bhatt, S., & Unwin, H. J. T. (2025). Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. *BMC Public Health*, *25*(1), 1520. https://doi.org/10.1186/s12889-025-22705-4

Debata, I., Panda, P. S., Karthikeyan, E., Tejas, J., Shruthi, K., Kumar, V. S., & Thirunavukkarasu, D. (2024). Spatial auto-correlation and endemicity pattern analysis of crimes against children in Tamil Nadu from 2017 to 2021. *Journal of Family Medicine and Primary Care*, *13*(6), 2341–2347. https://doi.org/10.4103/jfmpc.jfmpc_1463_23

Dewinter, M., Vandeviver, C., Vander Beken, T., & Witlox, F. (2020). Analysing the Police Patrol Routing Problem: A Review. *ISPRS International Journal of Geo-Information*, *9*(3), Article 3. https://doi.org/10.3390/ijgi9030157

He, Z., Lai, R., Wang, Z., Liu, H., & Deng, M. (2022). Comparative Study of Approaches for Detecting Crime Hotspots with Considering Concentration and Shape Characteristics. *International Journal of Environmental Research and Public Health*, *19*(21), 14350. https://doi.org/10.3390/ijerph192114350

Helbich, M., & Leitner, M. (2017). Frontiers in Spatial and Spatiotemporal Crime Analytics—An Editorial. *ISPRS International Journal of Geo-Information*, *6*(3), Article 3. https://doi.org/10.3390/ijgi6030073

Jannat, R., & and Al-Amin, M. (2024). Spatial statistics for legal process. *Journal of Spatial Science*, *69*(2), 327–347. https://doi.org/10.1080/14498596.2023.2226672

Kwon, E., Jung, S., & Lee, J. (2021). Artificial Neural Network Model Development to Predict Theft Types in Consideration of Environmental Factors. *ISPRS International Journal of Geo-Information*, *10*(2), Article 2. https://doi.org/10.3390/ijgi10020099

Melyantono, S. E., Susetya, H., Widayani, P., Tenaya, I. W. M., & Hartawan, D. H. W. (2021). The rabies distribution pattern on dogs using average nearest neighbor analysis approach in the Karangasem District, Bali, Indonesia, in 2019. *Veterinary World*, *14*(3), 614–624. https://doi.org/10.14202/vetworld.2021.614-624

Metropolitan Police Department. (2024). *Crime Incidents in 2024* [Dataset]. Metropolitan Police Department. https://catalog.data.gov/dataset/crime-incidents-in-2024

Mondal, S., Singh, D., & Kumar, R. (2022). Crime hotspot detection using statistical and geospatial methods: A case study of Pune City, Maharashtra, India. *GeoJournal*, *87*, 5287–5303. https://doi.org/10.1007/s10708-022-10573-z

Reza, M., Bisaria, A., Advaitha, S., Ponnekanti, A., & Arya, A. (2025). CriX: Intersection of Crime, Demographics and Explainable AI. *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, 714–725. https://doi.org/10.5220/0013316200003890

Self, S., Overby, A., Zgodic, A., White, D., McLain, A., & Dyckman, C. (2023). A hypothesis test for detecting distance-specific clustering and dispersion in areal data. *Spatial Statistics*, *55*, 100757. https://doi.org/10.1016/j.spasta.2023.100757

Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place. *Criminology*, *27*(1), 27–56. https://doi.org/10.1111/j.1745-9125.1989.tb00862.x

Thompson, A. E., Walden, J. P., Chase, A. S. Z., Hutson, S. R., Marken, D. B., Cap, B., Fries, E. C., Piedrasanta, M. R. G., Hare, T. S., Iii, S. W. H., Micheletti, G. J., Montgomery, S. M., Munson, J., Richards-Rissetto, H., Shaw-Müller, K., Ardren, T., Awe, J. J., Brown, M. K., Callaghan, M., … Chase, D. Z. (2022). Ancient Lowland Maya neighborhoods: Average Nearest Neighbor analysis and kernel density models, environments, and urban scale. *PLOS ONE*, *17*(11), e0275916. https://doi.org/10.1371/journal.pone.0275916

Ulmer, J. T., Harris, C. T., & Steffensmeier, D. (2012). Racial and Ethnic Disparities in Structural Disadvantage and Crime: White, Black, and Hispanic Comparisons. *Social Science Quarterly*, *93*(3), 799–819. https://doi.org/10.1111/j.1540-6237.2012.00868.x

Vandeviver, C., & Bernasco, W. (2017). The geography of crime and crime control. *Applied Geography*, *86*, 220–225. https://doi.org/10.1016/j.apgeog.2017.08.012

Wooditch, A., & Weisburd, D. (2016). Using Space-Time Analysis to Evaluate Criminal Justice Programs: An Application to Stop-Question-Frisk Practices. *Journal of Quantitative Criminology*, *32*(2), 191–213. https://doi.org/10.1007/s10940-015-9259-4

Ye, X., Wu ,Ling, & and Lee, J. (2017). Accounting for Spatiotemporal Inhomogeneity of Urban Crime in China. *Papers in Applied Geography*, *3*(2), 196–205. https://doi.org/10.1080/23754931.2016.1268969