ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1181

Predictive space-time model for optimal allocation of security agents in Lima, Perú

Cuya Nizama, Eduardo Andre; Atoche Diaz, Wilmer Jhonny
eduardo.cuya@pucp.edu.pe; watoche@pucp.edu.pe
Department of Industrial Engineering, Pontificia Universidad Católica del Perú Lima, Perú

***Abstract*—Nowadays, one of the most critical problems in Peru is crime and citizen insecurity. Part of this problem is due to a precarious assignment of security agents in geographic strategical zones at the right time. Furthermore, there is a lack of knowledge by the part of community members about certain existing patterns in crime. A large amount of data with geographic information, type of crime, time and other more meaningful variables is available to use. For this reason, the present research project is about a methodology using optimization models, machine learning and forecasting techniques to reduce these crime rates through the development of clusters that will allow us to identify the areas with the greatest confluence of crimes. A predictive spatial time model fed with data in real time will map the areas of high risk at a certain future point in the time. This model will be linked to an optimization model for assigning security agents to optimize and minimize the frequency of crimes and the existing citizen insecurity**

***Keywords*—Data Mining, Forecasting, Clustering, Geographic Information Systems, Assignment Optimization, Simulation, Public Security.**

I. INTRODUCTION

In the first half of 2024, 27.7% of the urban population aged 15 and over reported being victims of a crime, representing an increase of 0.6 percentage points compared to the same period in 2023 [1]. Citizen insecurity remains one of the main concerns affecting the quality of life in Lima and Callao. The primary crime issues identified by the population include street theft, drug addiction, and housebreaking [2].

According to the United Nations Development Programme (UNDP), Peru continues to face significant challenges in public security, with theft being a predominant crime [3]. Furthermore, a considerable gap exists between crime victimization rates and the number of crimes reported to the police. This discrepancy is largely attributed to a lack of trust in law enforcement and dissatisfaction with authorities' response times [3], [4]. As of 2024, 68% of Peruvians still express distrust toward the police, and 54% believe the government is not taking effective measures to improve citizen security [1]. Consequently, a substantial proportion of urban crimes remain unreported [4].

Recent studies also highlight that crime data analysis can play a crucial role in shaping policies aimed at improving citizen security and addressing the factors that contribute to crime [5]. This research aims to develop a predictive model for crime patterns, facilitating the optimal allocation of security resources to mitigate crime rates and improve public perception of safety.

II. THEORETICAL FRAMEWORK

A. *Data Mining*

Data mining is a powerful technique with great potential to help researchers focus on the most important crime information [6]. The technique of data mining has three main objectives: description, prediction and prescription, mainly divided into two main categories, supervised and unsupervised learning [7] Data mining application can be categorized in crime control and criminal neutralization; for crime control the data analysis results are taken into consideration in order to prevent further crimes and for criminal neutralization historical data can be used for seizing and capturing criminals. In order to find crime patterns data mining clustering algorithms are used; these algorithms manage to identify similar data sets that differentiates from the rest of the data. According to Nath [8]. Clustering is the best tool used to find data patterns over any other because crime data changes considerably over time and most of the crimes are not solved.

B. *Operation research*

In synthesis, operations research can be defined as a scientific approach at decision taking starting from the systemic analysis of deterministic problems, for which all the necessary information for their solution is known for certain or stochastic problems for which the necessary information is not known and shows a probabilistic behavior [9], [10]. According to Winston the main steps of the operations research process are as follows:

- Formulating the Problem
- Constructing a Model to Represent the System under study
- Deriving Solution from the Model
- Testing the Model and the Solution Derived from it
- Establishing Controls over the Solution
- Implementation of the Solution

C. *Patrol deploying models*

Starting from the 60's decade a great number of researches has been published about assignation and deployment of police forces in order to prevent crime[11]; however, there was a decrease in the number of related research in the 90's until the September 11 attacks in 2011 that once again raised the interest in this type of research[11]. One of the most famous patrol deployments for crime prevention model [12]

III. METHODOLOGY

This section presents a framework for the optimal allocation of the security agents consisting of 4 stages. In the first stage, crime data is collected. The variables are type of crime, place (latitude, longitude) and time where it was committed. This data is obtained through scraping of a crime statistics web- page from Peru. After that, an exploratory data analysis (EDA) of the data is accomplished to find patterns or clusters of crime points defined through time. Once the EDA is finished, the forecast stage is announced. The aim of the forecast is to predict what will be the points with the greatest crime and which type of crime is the most probably to happen in that point. Next, the optimization model is run to find the optimal points for police patrol allocation in such a way that the distance between those and the crime spots is minimized. Finally, the model is validated through simulation. The space-time model developed should be updated constantly with the data acquired by the police agents, punishing the errors of the prediction and finding new patterns in crime.

A. Data collection

For our scope of study, the district of San Miguel, Lima and surroundings areas are taken as the analysis environment. The election of this city is due to an increase of the crime rate in 2019 at the surroundings of the main university in the district. The district of San Miguel is also known for being one of the districts with the highest investments to reduce crime rate in Lima at present. Based on the report of INEI [1], [15], the crime trends at this district are obtained. The type of data used for the model is structured, due to its ease of analysis and compatibility between models. The source for extracting data is the INEI Integrated crime statistics and citizen security system (DataCRIM) page, which contains the reporting of crimes by type of crime in recent years (see Figure 1).

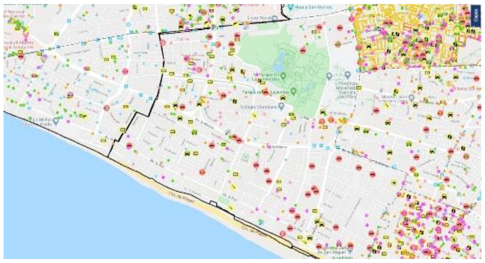


Fig. 1. San Miguel, Crime Map 2019, DATACRIM

Data extraction is not possible directly. Therefore, the scraping technique is used to extract the layer of symbols that are marked on the map for each type of crime. After extracting all the layers, a computer vision algorithm based on Mean shifted cross correlation (1) was used to extract the relative location of each symbol. This algorithm detected the map symbols with an accuracy of 97%.

$$R(x,y)=\frac{\sum_{x',y'}\left(T'(x',y')\cdot I'(x+x',y+y')\right)}{\sqrt{\sum_{x',y'}T'(x',y')^2\cdot\sum_{x',y'}I'(x+x',y+y')^2}} \quad (1)$$

IV. EXPLORATORY DATA ANALYSIS (EDA)

Having crime database ready, an initial analysis is carried out to gain insight into the structure of the database and to extract insights from it (see Fig. 2).

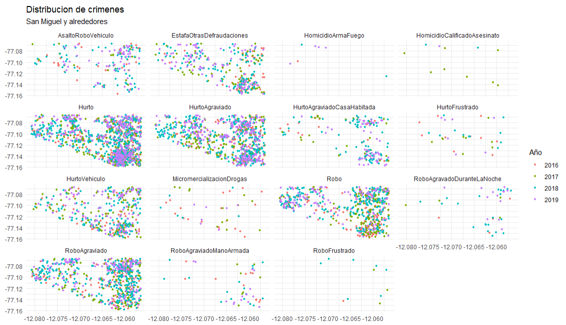


Fig. 2. Crime Distribution by type and year

In order to inspect the distribution of crime points, a quadrant plotting is done to see if the dots appear to be random, uniform, or clustered. As we see in Fig. 3, some of the quadrants have defined crime trends.

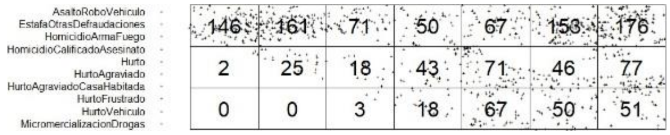


Fig. 3. Crime Distribution by quadrant in 2019

As Dr. Emily Burchfield says [16], Kernel density estimation (KDE) is a methodology used to estimate the density function of a random variable. For each type of crime (e.g. Theft), a subfunction is drawn to measure the effect of crime in space. After that, all of these sub-functions are added up to create an indicative surface of the density of the crime events in the area of analysis. For this study, the Isotropic Gaussian Kernel is chosen to estimate the KDE (see Fig. 4). The cores are generally smoothed in such a way that the bandwidth is the standard deviation of the smoothing kernel.

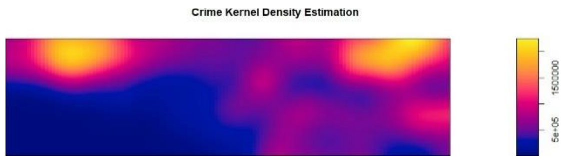


Fig. 4. Kernel Density Estimation Crime San Miguel 2019

Visualization is important, but it is not enough. It is necessary to determine if data is randomly distributed across space or not. To determine this, statistical tests were carried out. These tests have the null hypothesis that the crime points are distributed in space as a complete spatial randomness (CSR). By testing randomness null, this study aims to determine if point is significantly clustered in space.

One approach to test CSR is the Kolmogorov-Smirnov test, which is widely used to evaluate the goodness of fit of a sample from an unknown population to some hypothetical model of a specific population.

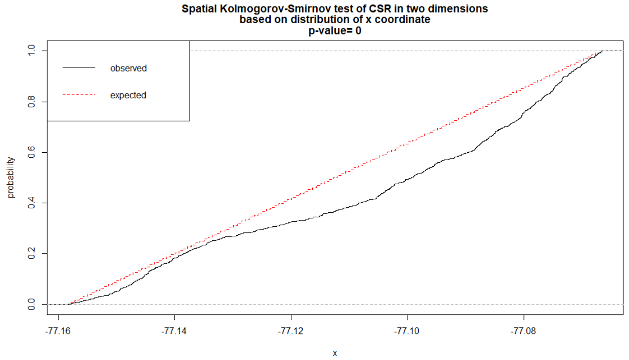


Fig. 5. Kolgmorov-Smirnov Test

In both cases (see Fig 5, 6), the p-value from the test is near zero, which suggests that the crime data was not obtained from a CSR population distribution. We can also see this in our graphs: the observed and expected cumulative distributions are quite different.

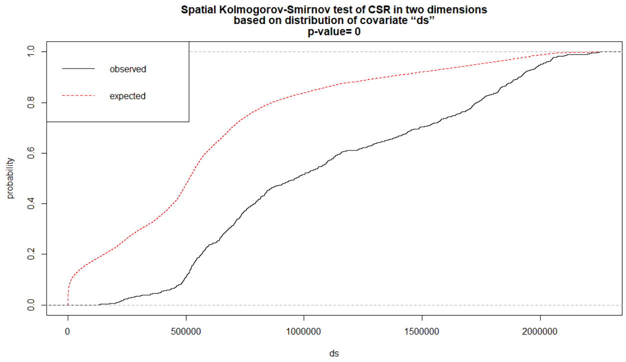


Fig. 6. Spatial Kolgmorov-Smirnov Test

V. CLUSTERING

First, to determine whether the spatial distribution of each type of crime behaves randomly or follows a pattern, a complete spatial randomness (CSR) test is required. This test does not identify the specific spatial patterns present in the data or the nature of their non-randomness. However, it is useful for assessing whether crime points exhibit a regular or clustered distribution. This knowledge is essential for selecting the appropriate algorithm to develop a clustering model for the data.

To evaluate CSR, several techniques are available. However, the primary statistical methods rely on a Point Pattern Process (PPP) to compare the observed spatial distribution of points with an empirical model of distances between points [16]. In this study, the G, F, and K functions are used as statistical measures to assess spatial dependence.

First, the G function, also known as the Nearest Neighbor Distance function, is used to test spatial dependence by measuring the distance from each point to its nearest neighbor. As shown in Fig. 7, the blue line represents the theoretical estimation of a homogeneous Poisson Point Process, while the black, green, and red lines correspond to different edge effect corrections (Kaplan-Meier, border, and Hanisch, respectively) for the empirical distribution of crime points. Analyzing the graph, the blue line is consistently lower than the others, indicating that the crime points are closer together than would be expected under a Poisson Point Process [16]. In other words, this suggests a clustered spatial pattern.

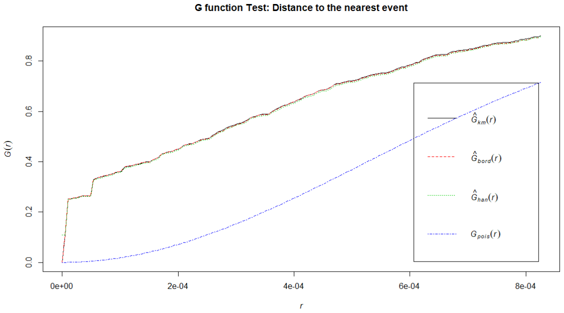


Fig. 7. G Function Test

Second, the F function, also known as the Empty Space Distance function, is used to measure the distance from a fixed reference location to the nearest data point. In contrast to the Nearest Neighbor Distance (G function), a blue empirical line that is greater than the edge effect corrections (Chiu-Stoyan, border, and Kaplan-Meier) suggests that crime points exhibit a clustered spatial pattern [16] (see Fig. 8).

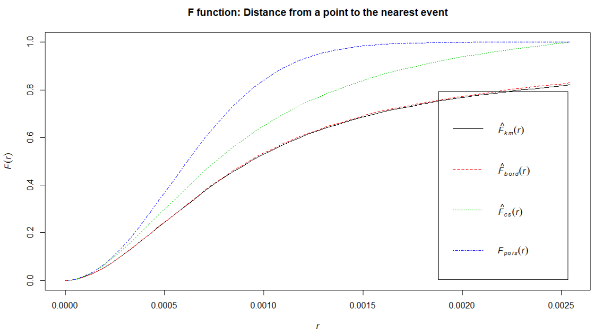


Fig. 8. F Function test

Finally, the K function measures the number of points within a given distance from a reference point. If crime points were distributed according to a homogeneous Poisson Point Process, the expected number of points within a certain distance would remain constant. As shown in Fig. 9, the theoretical blue line is lower than the empirical values obtained using the isotropic, translation, and border estimators. This suggests that the crime data exhibit a clustered spatial pattern.

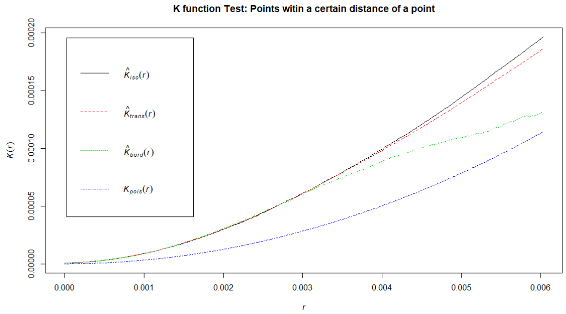


Fig. 9. K Function test

The K, F and G tests confirm that our data have existing clustering patterns. Given the results of the tests, we proceeded to group the spatial crime data for the year 2019 for each type of crime. The procedure for Theft in 2019 is taken as an example, a similar process should be followed for the other types of crime.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is selected as the clustering algorithm due to its ability to identify clusters without requiring a predefined number of clusters, unlike other clustering algorithms, and its robustness to noise. However, a key limitation of DBSCAN is the need to define the epsilon (eps) parameter, which represents the local radius. Since our data fit a Matern Cluster model, we set epsilon (ϵ) equal to the radius of the fitted model (4.585156e−3).

This process is repeated for each type of crime, and the resulting clusters will serve as input for the optimization model to be developed (see Fig. 10).

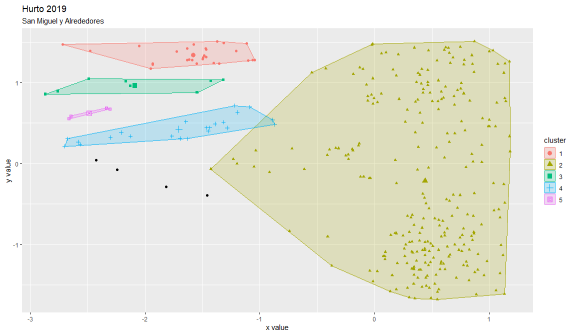


Fig. 10. Clusters Theft 2019

VI. FORECASTING

To forecast the number of crimes in 2020, we split the data into two sets: the training set consists of crimes recorded between 2016 and 2018, while the test set includes crimes from 2019. The design matrix includes the year and the distance from each crime location to the nearest police station as independent variables. Specifically, it contains eight variables representing the distances to the first, second, ..., and eighth nearest police stations. The dependent variable is the type of crime at each location.

Since the task involves predicting a categorical dependent variable (crime type) based on the given independent variables, a multinomial regression model is required. The models selected for this study are K-Nearest Neighbors (K-NN) and a Neural Network (NN).

For tuning the K-NN model, repeated cross-validation is employed. The primary hyperparameter to be optimized is k k, which determines the number of nearest neighbors considered. As shown in Fig. 11, the accuracy function converges at approximately $k = 100$.

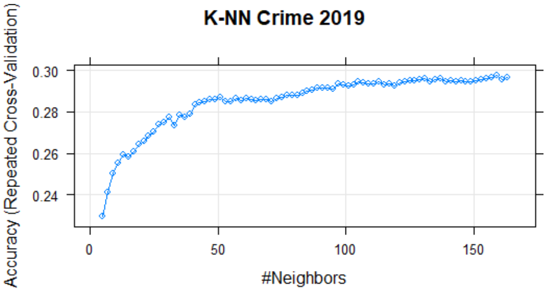
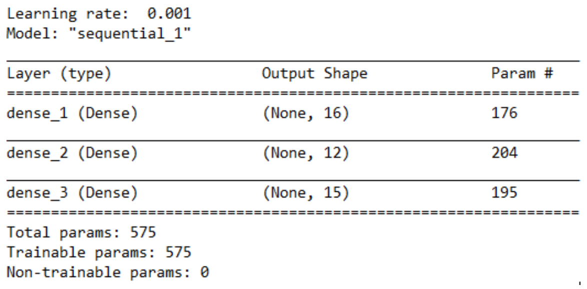


Fig. 11. Hyperparameter tuning K-Nearest Neighbour

The architecture chosen for NN model has 3 layers (see Fig. 12) Since the goal is to have an idea of the probabilities for each classification, the activation function chosen for the output layer is soft-max and the loss of the model is measured by categorical entropy function.

The test accuracy for the K-NN and NN-model is 29.72% and 29.87% respectively. Due to its simplicity the model chosen is K-NN. Even if the accuracy is low. The model is a good start points to analyze the probability distribution of each crime, given a set of characteristics.



```
Learning rate:  0.001
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 16)                176
_________________________________________________________________
dense_2 (Dense)              (None, 12)                204
_________________________________________________________________
dense_3 (Dense)              (None, 15)                195
=================================================================
Total params: 575
Trainable params: 575
Non-trainable params: 0
```

Fig. 12. Neural Network Architecture

VII. OPTIMIZATION MODEL FOR OPTIMAL ALLOCATION

For the development of the optimization model, a simplification of the space through a grid of 40x24 symmetric quadrants will be considered. Therefore, we have 960 nodes, each with its own crime amount in the last 5 years. In this stage, we present a model for security agent deployment optimization to determine the optimal location of police patrols.

The optimization model seeks for an optimal set of nodes that minimize the distance from the neighbors' nodes to its nearest center. This is done by using the p-center algorithm. The p-center model is defined as follow: Given a set of n nodes and the distances between any pair of nodes, the goal of the model is to locate p facilities in such a way to minimize the maximum of the distances from each node to its nearest center. We add a cost factor to this model based on the amount of crime in each node. The mathematical formulation for this model is shown below:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} c_i * d_{ij} * x_{ij} \qquad (2)$$

$$s.t. \sum_{j=1}^{n} X_{ij} = 1 \quad \forall i = 1, 2, \dots, n \qquad (3)$$

$$s.t.\ x_{ij} \leq y_j \quad \forall i, j = 1, 2, \dots, n \qquad (4)$$

$$s.t. \sum_{j=1}^{n} y_j = p \qquad (5)$$

$$x_{ij} \in \{0, 1\} \ \forall\, i, j = 1, 2, \dots, n \qquad (6)$$

$$y_j \in \{0, 1\} \ \forall\, j = 1, 2, \dots, n \qquad (7)$$

In this model, n is the total number of nodes in the area (100). The variable $x_{i\,j}$ is a 0-1 binary variable that represents that the node i should be attended by the police agent in node j; $y_{\,j}$ is a 0-1 binary decision variable that represents if a police agent should be located in node j or not; c_i is the predict amount of crime from the forecast method for the node i; $d_{i\,j}$ is the distance between node i and node j and p is the total number of facilities available.

VIII. MODEL EVALUATION

To evaluate the performance of our proposed model, the current policy of the San Miguel district police department was compared with the results of our proposed optimization model. For the comparison, the K-means algorithm was added as a heuristic proposal to our model. This heuristic is more efficient in computational terms than the Mixed Integer Linear Programming (MILP) algorithm proposed first. However, it does not give us an optimal result. The simulation is carried out generating scenarios with the number of patrols in a range from 1 to 14.

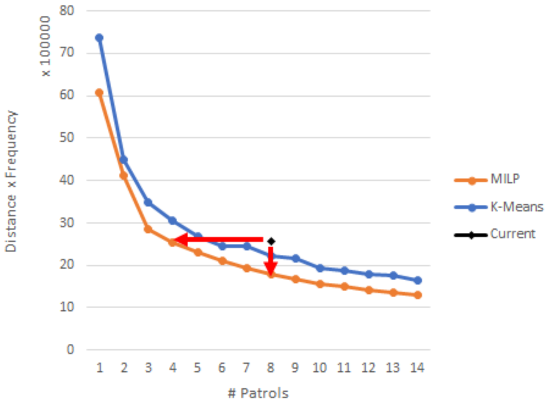


Fig. 13. Trade-off curve: Number of patrols deployed vs Crime rate

There is a trade-off between the number of patrols and the crime rate that is represent by the indicator distance from the patrol to the node of crime and the amount of crime expected in that node (see Fig. 13). Furthermore, the current patrol deployment policy that the police department of San Miguel has is not optimal. There are possibilities of optimizing the current policy both through the use of the optimization model or the proposed heuristic:

- Keep the same number of patrols and follow the locations determined by the MILP algorithm in order to reduce the crime rate. This decision increases the effectiveness of the current policy.

- Reduce the number of patrols to 4, keeping the same crime rate. This decision increases the efficiency in cost of the current policy.
- Reduce the number of patrols to 6 and reduce the crime rate. This decision increase both the effectiveness and efficiency.

IX. DISCUSSION

For the development of this study, the open-source programming language R and Python were used to make scraping, data preprocessing, exploratory data analysis, clustering and forecasting. Keras framework was used to train the neural network. On the other hand, the Gurobi solver (Full Student License) was used to solve the optimization models. The development of this study was carried out on a Windows 10 64-bit computer, Intel Core i7, 2.4GHz, 16GB RAM.

In the realization of this model, several assumptions were made, which are points of improvement for the methodology. First of all, for the distance between nodes, the Euclidean distance was taken. One solution of this would be to introduce the Google Maps API and calculate the real distance and the time it takes to move from one node to another. Furthermore, by having these values, the maximum coverage model could define more precisely its range of coverage for each node. Another assumption that was made was to consider that each node behaves independently, without analyzing if there was any correlation between the criminal behavior for those nodes. On the other hand, given the computational speed of the Gurobi Application Program Interface, it is convenient to analyze the development of a web application that facilitates the work of the police planner.

Future extension of this work could incorporate models of human rationality to improve decision-making of agents. This problem could also be framed as a sequential decision analytics model as crime evolves over time. Factors such as perceived risk of apprehension, result of the deployment of more patrols, can modify the crime patterns previously identified, therefore requiring adjusting the allocation of security agents.

X. CONCLUSIONS

This paper presents a methodology to predict crime using a framework based on a clustering, forecasting and optimization model. This framework is applied to the district of San Miguel and surrounding areas. With our optimization models it is shown that is possible to cover the crime hotspots by deploying the security agents in the correct nodes. This result reveals that is possible to improve urban security and reduce crime rates by forecasting with a space-time model the areas that are potentially riskier and deploying the two stages model to optimally cover with limited resources. In particular, we believe that this methodology will contribute to the body of knowledge on policing. A possible future extension of the current study is to explore the use of other advanced algorithms like Long short-term memory neural networks with crime data segregated by date to improve the forecasting. We believe that this proposed methodology will have a lot of potential applications in other areas such as transportation, warehouse location, marketing. We leave it for future studies.

ACKNOWLEDGMENT

I would like to express my gratitude to my teacher Atoche Diaz, Wilmer Jhonny who guide me through the development of this study as well as Pontificia Universidad Catolica del Perú who gave me the wonderful opportunity through the Programa Apoyo a la Investigacion (PAIN) to do this project centered on a problem that is a general interest of citizens.

REFERENCES

[1] INEI, Estadísticas de Seguridad Ciudadana, Enero-Junio 2024, Instituto Nacional de Estadística e Informática, 2024. [Online]. Available: https://www.inei.gob.pe/media/MenuRecursivo/boletines/estadisticas-de-seguridad-ciudadana-enero-junio-2024.pdf

[2] Lima Cómo Vamos, Informe de Percepción sobre Calidad de Vida en Lima y Callao, 2024, 2024. [Online]. Available: https://www.limacomovamos.org/reportespercepcion/

[3] United Nations Development Programme, Informe de Desarrollo Humano para América Latina y el Caribe: Seguridad Ciudadana y Desarrollo Sostenible, 2023.

[4] Observatorio de Criminalidad del Ministerio del Interior del Perú, Reporte de Opinión Ciudadana sobre Seguridad, Dic 2023 - May 2024, Ministerio del Interior del Perú, 2024. [Online]. Available: https://observatorio.mininter.gob.pe/sites/default/files/proyecto/archivos/Server%20Reporte%20Opinion%20Ciudadana-Semestre%20movil-dic2023-may2024.pdf

[5] A. Rojas and M. Díaz, "Crime data analysis for public safety policies: A systematic review," Journal of Public Safety Research, vol. 12, no. 4, pp. 212-230, 2023.

[6] Thongtae, P. and Srisuk, S. (2008). An analysis of data mining applications in crime domain. In Computer and information technology workshops.

[7] Vadim, K (2018). Overview of different approaches to solving problems of Data Mining.

[8] Nath, S.(2006). Crime pattern detection using data mining.

[9] Hillier, Frederick and Gerald Lieberman (2015) Introducción a la investigación de operaciones. México, D.F: McGraw-Hill.

[10]Winston, Wayne L. (2005) Investigación de operaciones, algoritmos y aplicaciones. México: International Thompson Editors.

[11]Green, L. and Kolesar, P. (2004) Improving emergency responsiveness with management science. Management Science.

[12]Tongo, C. and Agbadudu, A. B. (2010). A dynamic programming model for optimal distribution of police patrol efforts in Nigeria: a case study of Lagos State. International Journal of Operational Research.

[13]Chaiken, J. and Dormont, P. (1978) A patrol car allocation model: background. Management Science.

[14]Taylor, P. E., and Huxley, S. J. (1989). A break from tradition for the San Francisco police: Patrol officer scheduling using an optimization-based decision support system. Interfaces, 19(1), 4-24

[15]INEI (2018). Informe técnico de estadísticas de seguridad ciudadana Enero-Junio 2018.

[16]Burchfield, D. (2020). Point pattern analysis. Retrieved from 2https://www.emilyburchfield.org/courses/gsa/point_pattern_lab

[17]Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2015), shiny: Web Application Framework for R. R package version 0.11.

[18]Sean J. Taylor, Benjamin Letham (2018) Forecasting at scale. The American Statistician 72(1):37-45

[19]Church, R. and ReVelle, C. (1974), "The maximal covering location problem", Papers of the Regional Science Association, Vol. 32 No. 1, 1pp. 101-118.

[20]Cheung, C.-Y., Yoon, H. T., & Chow, A. H. (2015). Optimization of police facility deployment with a case study in Greater London Area. Journal of Facilities Management, 13(3), 229–243

[21]Curtin, K. M., Hayslett-McCall, K., & Qiu, F. (2007). Determining Optimal Police Patrol Areas with Maximal Covering and Backup Covering Location Models. Networks and Spatial Economics, 10(1), 125–145. doi:10.1007/s11067-007-9035-6