Computers in Human Behavior 127 (2022) 107014

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Data driven smart policing: A novel road distance-based *k*-median model for optimal substation placement

Abinta Mehmood Mir, Ali Hassan, Asma Khalid, Zohair Raza Hassan, Faisal Kamiran, Agha Ali Raza, Saeed-Ul Hassan *, Mudassir Shabbir

Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

ARTICLE INFO

Keywords:
Crime analysis
Optimal location
Geometric clustering

ABSTRACT

In the context of smart city research, finding patterns in crime data to explore trends in crime and to locate the presence of crime has been an exciting research field. Our aim in this paper is optimizing the location of police substations within the jurisdiction of a police station so that law enforcement agencies could work efficiently. These substations are placed near by road in order to immediately respond to nearby crimes. We attempt to optimize the average minimum distance to a crime from its nearest substation. This distance is found by the underlying road network. We are locating these substations in an infinitely large set of points in the given region. Case study has been done on real-world crime data from Lahore, Pakistan to show the efficiency of these methods. We also explored what are trends in crime data with respect to years, seasons and day/night and where to place substations accordingly. We evaluated the placement of substation by our model through comparing the average time taken to report to a crime against random substations placed within the same area. Last but not the least, the analytical insights provided by this study be useful for policy making and smart city research for the inclusion of sustainable cities and communities.

1. Introduction

The prevalence of crime is a pressing global issue, and crime rate highly correlates to the socio-economic status of individuals in a society. Experts study trends in crime data to learn how to better counteract the presence of crime. Finding patterns in crimes has been a popular field of study for decades (Jeffery, 1959). More recently, there has been an influx of crime analysis via the rich tool set provided by the fields of data mining and analytic (Chen et al., 2004; Hassani et al., 2016). In the context of smart city research, the ideas of policy making and strategy planning are brought into the spotlight (Visvizi & Lytras, 2018; Lytras et al., 2020). Research studies also present theories on how decision-making of governments officials are affected by commissioned study (Carroll)and how to boost the urban sustainability through smart decision making (Visvizi et al., 2019). A practical toolkit for regulators is also introduced in the smart city search domain to effectively handle the needs of citizens (Lytras et al., 2019). The goal of this analysis is to construct effective strategies and data driven smart policies for better crime prevention.

This work focuses on optimizing the location of police substations within the jurisdiction of a police station, allowing for law enforcement to operate more efficiently. These substations – often found in countries in the Indian subcontinent (Kumar, 2012), where they are referred to as "chowkis" – allow people to report crimes, and police officers to dispatch correspondents to any nearby crime from multiple locations within a region. In this vein, we attempt to optimize the average minimum distance to a crime from its nearest substation. We also explore a modified version of this objective, based on the type of crime taking place.

Choosing locations based on optimizing a cost function has been researched in a number of manners, in theoretical (Hsu & Nemhauser, 1979; Jain et al., 2003; Williamson & Shmoys, 2011) and practical (Current & Weber, 1994; Singh, 2008; Kalcsics et al., 2000) contexts. facility location algorithms have been proposed demonstrating better approximation factors (Guha & Khuller, 1999; Hochbaum, 1984). These are known as Facility Location, and Optimal-Location Query problems. Generally, in these settings, one is given: *(1)* the facility set, which corresponds to possible locations of the facilities we can open, *(2)* the client set, which corresponds to locations of clients, and *(3)* an objective function dependent on which facilities are chosen, and which client each facility caters to. Typically, as is the case in this work, the objective

* Corresponding author.
E-mail address: saeed-ul-hassan@itu.edu.pk (S.-U. Hassan).

https://doi.org/10.1016/j.chb.2021.107014
Received 17 December 2020; Received in revised form 4 August 2021; Accepted 6 September 2021
Available online 16 September 2021

function is based on the distance between a facility and client. The required output is the subset of facilities that optimizes the given objective function. To illustrate, an example is provided in Fig. 1. We discuss these problems further in Section 2.

Although location optimization is well-explored, the existing works are not directly applicable to the context we investigate in this paper. In particular, our problem setting differs significantly due to the following reasons:

1. This is, to our knowledge, the first work that looks at the problem in the context of crime response. Some notable examples include looking at facility location from a supply chain context (Kalcsics et al., 2000), or catering to customers (Yilmaz et al., 2017).
2. The problem is set in a geodesic space with finite diameter. The distance between a substation and crime is governed by the given region's underlying road network. In comparison, with the exception of some works (e.g. (Chen et al., 2014; Chen et al., 2015; Xiao et al., 2011)), progress towards location optimization in a practical setting has focused on Euclidean metrics/spaces.
3. Most works assume the set of possible facility locations is finite in size (there are some exceptions (Wong et al., 2009; Zhang et al., 2006)). In contrast, we are locating substations in an infinitely large set of points in the given region. This allows proposing solutions for given regions with lesser information. We also assume that the cost of opening a substation at any one of these locations is equal – note that this is not always the case in other contexts.

We perform an extensive case study on real-world crime data provided in the Lahore Crime Dataset. We predicted the placement of optimal substations within the jurisdiction of a police station for our case study. We also explore crime trends with respect to years, seasons and day/night, and discuss changing substation placements. Finally, we evaluate the substations placed by our model by comparing the average time taken to report to a crime against random substations placed within the same area.

The paper is organised as follows. We discuss the related work in Section 2. The preliminaries and problem definition are discussed in Section 3. We propose methodologies to solve the problem in Section 4, and perform a case study in Section 5 to showcase the applicability. Finally, we conclude in Section 7.

2. Related work

In this section we present a brief literature review on works related to choosing optimal facility locations.

2.1. Crime analysis and policing optimization

Criminology has been an area of active research for centuries (Jeffery, 1959). The advent of data mining and machine learning techniques has allowed the community to better automate the analysis of crime (Chen et al., 2004; Hassani et al., 2016), and scale to data of larger proportions (Estivill-Castro & Lee, 2001). Hassani et al. (Hassani et al., 2016) provide a review of common approaches to crime analysis using these techniques. They identify five data mining techniques used within the literature: *(1)* entity extraction, which entails extracting information from sources such as witness reports or news stories, allowing for the analysis of any patterns found within the data; *(2)* cluster analysis, which has been used to find crime "hotspots", and track the model the behaviour of criminals; *(3)* association rule, which has been used to link crimes committed by the same people; *(4)* classification, which has been used for various tasks including detecting suspicious emails, exposing lies, and uncovering credit card fraud; and *(5)* social network analysis, which has been used to detect criminal accounts and analyze characteristics of terrorist organizations.

2.2. Related problems

An avenue of research with motivation similar to ours is the optimization of police patrol routes, as these works also aim to minimize crime response time (Dewinter et al., 2020; Li et al., 2011; Mukhopadhyay et al., 2016). However, this differs from our problem significantly; unlike patrols, the substations we set up are static in location. Thereby, the models in these works are not directly applicable to our problem.

The Facility Location and Optimal Location Query problems are concerned with opening new facilities at optimal locations with respect to the clients they have to cater to. The Facility Location problem and its variations have received significant theoretical interest in the past, such as its `NP-Hardness`, and hardness of approximation (Hsu & Nemhauser, 1979; Jain et al., 2003; Williamson & Shmoys, 2011). In general, the cost of a facility catering to a client may be arbitrary, but metric variants which assume the costs are dependent on distances between facilities and clients have also been shown to be `NP-Hard`. The general version is `Log-APX-Hard`, while the metric version is `NP-Hard` to approximate for a constant factor under 1.463 (Guha & Khuller, 1999). Another variant introduces capacities on each facility, i.e. each facility can only cater to a fixed number of clients (Melkote & Daskin, 2001). Although quite similar, techniques to solve facility location problems can not be applied to the problem in this work; facility location deals with a finite set of facilities and a known set of clients to cater to, however, this is not the case for our problem.

Practically oriented works frame the problem as Optimal Location Query (Du et al., 2005). In these settings, there is typically an existing set

(a) An instance of Facility Location. (b) The optimal solution.

Fig. 1. An example of Facility Location and its solution where the objective function is based on the average Euclidean distance between a client (red circle) and its nearest facility (gray box). The chosen facilities have been darkened in the solution.

of facilities that already cater to the customer, but the aim is to open a new facility to further optimize an objective, for instance, Yilmaz et al. (Yilmaz et al., 2017) point out two well studied objectives: *(1)* maximizing the number of clients attracted by a facility, and *(2)* minimizing the average distance between each client and the nearest facility. Clients may also be weighted to signify greater importance. Reverse Nearest Neighbor (Korn & Muthukrishnan, 2000) and Bichromatic Reverse Nearest Neighbor (Wong et al., 2009) are two measures that are well-studied in this context. These measures are based on the number of clients closest to the new facility (Yilmaz et al., 2017).

Unlike typical facility location setups, some works do consider a possibly infinite set of regions where a facility can be opened (Wong et al., 2009; Zhang et al., 2006). Most work in this area is focused towards opening a single new facility, but there are works that discuss opening multiple facilities under the same conditions (Chen et al., 2019). In this work, we aim to open k facilities, where k is a parameter provided as input.

Most works are assumed to exist in a Euclidean setting. Works on road distance assume an underlying road network that dictates the distance between two points (Chen et al., 2014, 2015; Xiao et al., 2011). Note that our work is significantly different than those mentioned due to the fact that "client" set of crimes is unknown to us. Yilmaz et al. are the first to address this issue and explore multiple strategies based on the amount of information one may have when setting up a new facility (Yilmaz et al., 2017). However, they only consider opening one new facility in their approach, and their work assumes a Euclidean setting.

We employed k-medians clustering in our methodology. k-medians is preferred to the typical k-means algorithm as the clusters formed by k-medians are more compact than k-means. k-medians clustering approach works on minimizing sum of distances approach instead of sum of the squared distances. The criterion function formed in this way is better with k-median and is used in facility location applications (Bradley et al., 1997).

3. Preliminaries

In this section we define the terminology and notation used throughout the paper, and formally introduce our problem.

3.1. Notation

Let R be a be a set of points representing a geodesic space (where $R \in \mathbb{R}^2$ and each point x in R is represented by its latitude and longitude). Let $R^* \subset G$ be the points where substations could be potentially set up. Since substations are typically placed on roads, we assume that R^* is the set of points that map to a road in the real world. We assume that we are given a distance function, $d : R \times R \rightarrow \mathbb{R}$, that computes the road distance, between points x and y, where $x, y \in R$, denoted by $d(x, y)$, based on the real-world road network. The road distance between two points is illustrated in Fig. 2.

Let $\mathbb{R}$ be a set of points representing a geodesic space. Let $R^* \subset \mathbb{R}$ be the points where substations may be set up. Since substations are typically set on roads, we assume that R^* is the set of points that map to a road in the real world. We assume that we are given a distance function, $d : R \times R \rightarrow \mathbb{R}$, that computes the road distance, denoted $d(x, y)$, from x to y for $x, y \in R$, based on the real-world road network in R. The road distance between two points is illustrated in Fig. 2.

Let $\mathcal{C} \subset R$ be the multiset of crimes in an area. We refer to $F \subset R$ as a set of points corresponding to substations, where $|F| = k$. Let $f_c = \arg\min_{f \in F} d(f, c)$ be the nearest substation to a crime $c \in \mathcal{C}$.

3.2. Problem formulation

Given an integer $k \in \mathbb{Z}^+$, and a region R, we aim to open k substations such that we minimize the average distance from each crime to its nearest substation:

$$\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} d(c, f_c)$$

where $\mathcal{C}$ is an unknown set of crimes. Note that when $\mathcal{C}$ is known and d is the Euclidean distance between two points, our problem is equivalent to the formulation for k-means clustering. The k-means clustering is known to be `NP-Hard` in the general case (Vattani, 2010), but in practice is efficiently solvable for small values of k. We assume that k is provided as input by the user, as it can be realistically assumed that the number of substations that can be opened is dependent on the resources available to the governing police station.

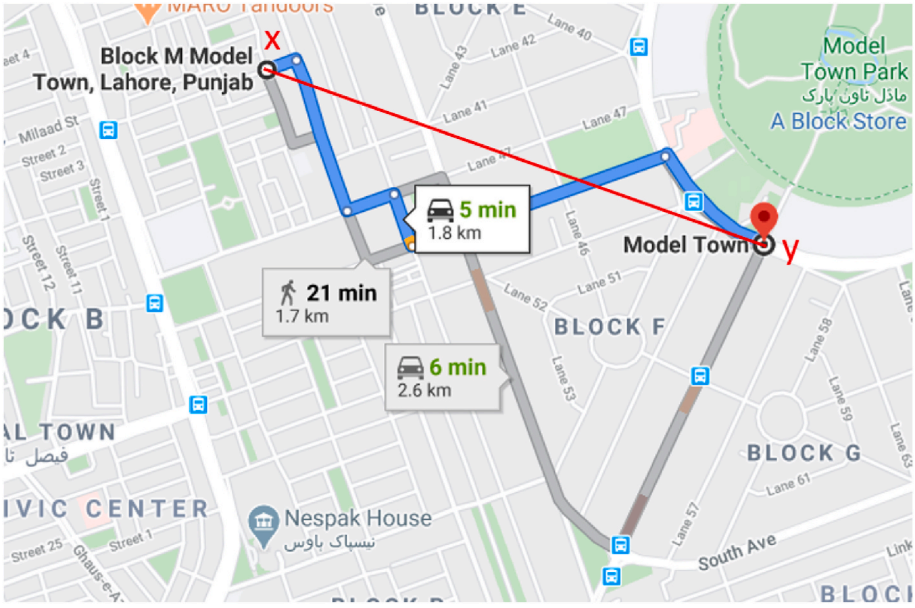


Fig. 2. An example image illustrating the difference between road distance and straight line distance between two points x and y on google map. The blue line shows shortest road distance between points x and y. The red line is for straight line distance between two points x and y. Based on this assumption, the distance function would be computed on the real-world road network.

4. Substation positioning

In this section we discuss our approach to find an optimal set of substation locations F to minimize the response time to crimes in $\mathcal{C}$.

We assume that we are given a finite set of points $C \subset \mathcal{C}$, such that the distribution of crimes in C is similar to that in $\mathcal{C}$. We thereby assume that optimizing substation placement to respond to crimes in C should provide a solution that is (nearly) optimal for $\mathcal{C}$. In this vein, we aim to find the set of substations that minimize the following objective function:

$$\frac{1}{|C|}\sum_{c \in C} d(c, f_c)$$

We employed k-medians clustering on C, as described in Algorithm 1, to obtain our set of locations F. We then projected all locations in F to the nearest point the corresponds to a road to obtain our final set of substation coordinates.

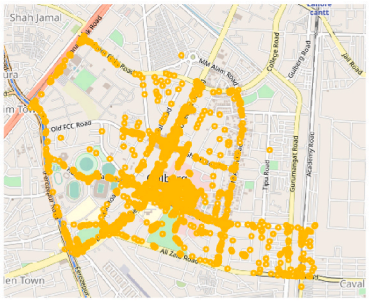
Algorithm 1
Computing F for a given C

Input: Set of points C
Output: F, set of locations for substations
1 Compute pairwise distances for given set of points
2 Randomly initialize k centroids, F
3 Identify nearest points in C for each centroid in F
4 Recompute centroids in F by taking median of data points nearest to it
5 Repeat 3–4 until there is no change in F
6 Return F

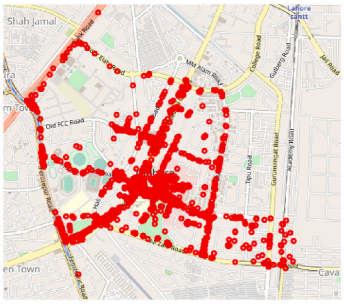
5. Case study

In this section we discuss our analysis on a real-world crime dataset, in the context of optimal substation placement prediction.

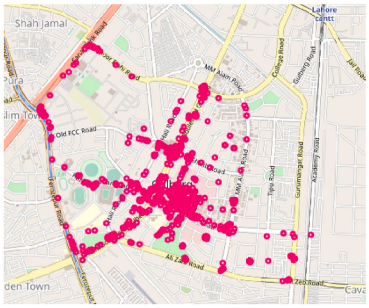
Implementation Python 3.3 and the `pyclustering` library (Novikov, 2019) were used to implement k-medians. For computing road distance between two points we used the `herepy` library. The `overpy` library was used to find the nearest road to a given coordinate.



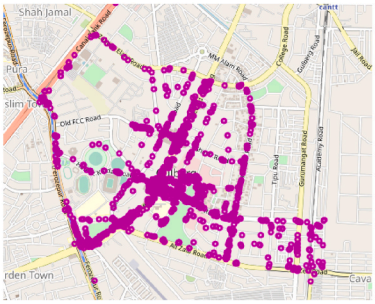
(a) Visualization of crimes during day time



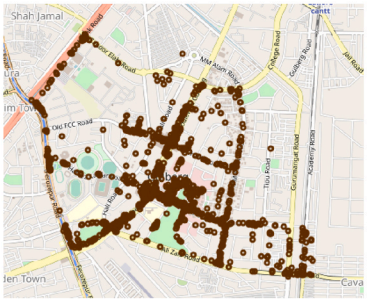
(b) Visualization of crimes during night time



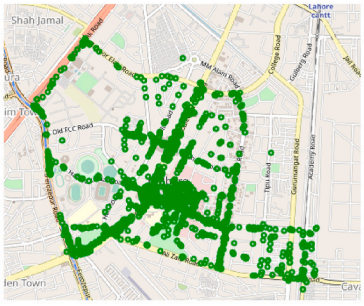
(c) Visualization of crimes occuring in 2014



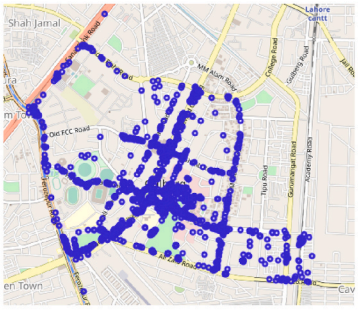
(d) Visualization of crimes occuring in 2015



(e) Visualization of crimes occuring in 2016



(f) Visualization of crimes during summer season



(g) Visualization of crimes during winter season

Fig. 3. Visualization of crime points for seven sub-datasets. For differentiation colors are assigned to each sub-dataset respectively. These data splits would be used to study the optimal substation placement with respect to different criterion.

5.1. Lahore Crime dataset

We perform our analysis and experiments on the Lahore Crime dataset. The dataset has information for a total of 220,719 crimes reported in police stations from 2014 to 2016. The dataset has various attributes for each report, including crime type, geodesic coordinates, and the police station the crime was reported in.

There are a total of 83 police stations represented in the dataset. For our case study we showcase our results on the areas under the jurisdiction of the Gulberg, Model Town, Johar Town, Naseerabad, Garden Town and Faisal Town (see Fig. 3).

We obtain the geodesic coordinates of each crime within these areas to construct our set of crimes, C. The data was pre-processed and outliers were removed using the z-score technique, giving us total crimes shown in Table 1.

We now discuss constructing sub-datasets by considering differing times, seasons, and years.

5.1.1. Sub-dataset for day and night

We extracted a sub-dataset for day and night to get more insight for optimal substation prediction. Of the 5126 reported crimes, 2325 took place during the daytime, and 2801 during the night. Crimes reported during the day/night are shown in Table 2.

5.1.2. Sub-dataset for summer and winter

We extracted a sub-dataset for the summer and winter seasons for the crimes in our dataset. Summer season in Pakistan runs from March to September and winter season from from October to February. 2940 crimes were reported during the summer, and 2186 during the winter.

Crimes reported during the summer/winter are shown in Table 3.

5.1.3. Sub-datasets w.r.t. years

We extracted three sub-datasets w.r.t. the years 2014, 2015, and 2016. The years sub-dataset detail have been shown in Table 4.

5.2. Results

Recall that in our problem formulation the number of substations is resource dependent. Hence, we arbitrarily set $k = 3$, $k = 5$ and $k = 7$. We employed our methodology on our given dataset, C, and report the objective function:

$$\frac{1}{|C|}\sum_{c\in C} d(c,f_c)$$

The three clusters formed by k-medians are shown with different colors in part (a). Part (b) of Figure represents these clusters on a map. The substation locations were obtained by using the centroid points of each cluster. The substations were then projected to the nearest point that corresponds to a road (see Fig. 7).

The results for a fixed area, Gulberg, Lahore, are illustrated in Figs. 4, 5 and 6. Clustering results for different values of substations k, targeted for each area are shown in Fig. 8, Fig. 9, and Fig. 10. Similarly, Fig. 11 is showing combined clustering results with predicted substations for each of the six targeted areas.

Table 1
Table is showing total number of crime points in each of the targeted area.

Area Name	Number of entries
Gulberg	5126
Garden Town	2810
Faisal Town	3219
Model Town	1623
Naseerabad	3619
Johar Town	3781

Table 2
Table showing number of entries for sub-dataset w.r.t Day and Night time.

Sub-datasets	Entities
Day	2325
Night	2801

Table 3
Table showing number of entries for sub-dataset w.r.t Summers and Winters.

Sub-datasets	Entities	Months
Summer	2940	March, April, May, June, July, August, & September
Winter	2186	October, November, December, January, & February

Table 4
Table showing number of entries for sub-dataset w.r.t Years.

Sub-datasets	Entities
2014	1594
2015	2122
2016	1410

Evaluation After proposing optimal substations F, we evaluated the objective function on the set of crimes C and predicted substations F, results for each area is shown in Table 5.

Baseline comparison To evaluate our methodology we compared the objective value of our model to those of random substations within the Gulberg area for $k = 3$. We picked three random centroid points from C and treat them as substation locations. For each crime point in C we found the nearest substation and evaluated the objective function. This experiment was repeated 100 times, and the results are shown in Table 6.

We observed that our technique for substation placement provided an objective value that was lower than the minimum, average, and maximum value provided by random substations, showcasing the efficacy of our methodology.

5.2.1. Substation placement during day and night time

We showcase in Fig. 12 the affect of splitting according to the day and night. One of the substation is placed similarly in both settings, but the two of them changes locations, showcasing the need for mobile substation placement. The average distance from each crime to its nearest predicted substation was 199 and 180 m for day and night, respectively.

5.2.2. Substation placement for summer and winter

Substation prediction results for crimes reported during summers and winters are shown in Fig. 13.

As in our day/night split, we observe that substations change positions according to the season as well. The average distance from each crime to its nearest predicted substation was 189 m for both summers and winters.

5.2.3. Substation placement w.r.t years

Substation prediction results for crimes reported w.r.t years are shown in Fig. 14.

As shown in Fig. 10 above, predicted substations are changing locations for each year.Therefore, substations should be placed as per changing demand so that average response time should be reduced. The average distance from each crime to its nearest predicted substation was 176, 184 and 297 m for 2014, 2015 and 2016, respectively.

(a) *k*-medians clustering results on 2D plane (b) Visualization of respective clusters on Map

Fig. 4. Clustering results on 2D plane and on Map for the area of Gulberg, Lahore. The value of k is set to 3 which corresponds to three clusters. Clusters are differentiate with different colors.

(a) *k*-medians clustering with predicted centroids for k=3 (b) Visualization of predicted substations on Map

Fig. 5. Illustration of substations predicted for Gulberg Area, Lahore. Three stars in yellow color corresponds to predicted locations in 2D plane. On Map view is showing Map locations for three substations for respective cluster area.

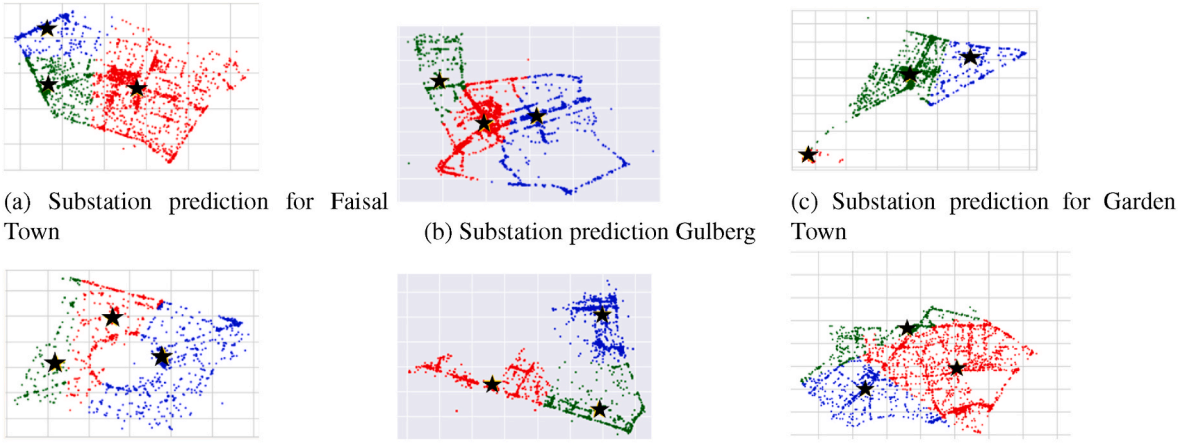
Fig. 6. The location of substations predicted using road distance metric is on road or nearby road as shown in this figure. We can see in zoom-in view that one of the three substations is on the road while rest of the two are also nearby road locations.

6. Discussion

In this section, we discuss the ramifications, theoretical contributions, limitations and future directions of our work.

Our solution optimizes response time by placing substations that minimize the distance from a crime to its nearest substation. After proposing a data-driven solution based on previous crimes, we showcased the efficacy of our methodology with a case study in Section 5 performed on a real-world crime dataset. Lahore crime dataset has been used for placing substations in a police district to optimize crime response time. In Sections 5.1.1, 5.1.2 and 5.1.3 we exhibit the sub-datasets of different time frames that were extricated from Lahore Crime dataset. In Section 6. We demonstrated the placements of the substations according to minimum average road distance. The position of substations are different for each sub-dataset. Our algorithm ensures that the predicted locations of the optimal sub-stations are either nearby

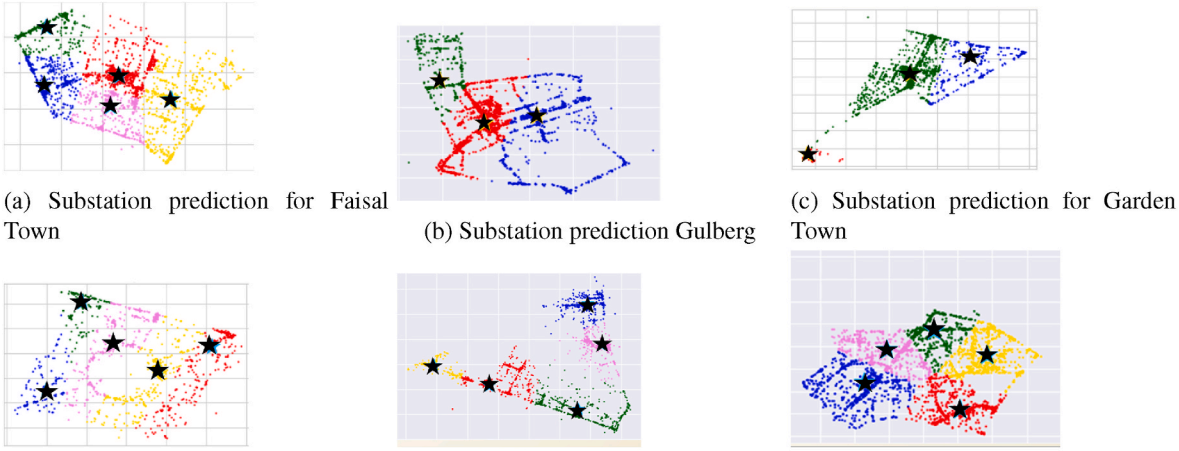
Fig. 7. Visualization of the substations that are projected to the nearest point that corresponds to a road so that finally we could have on-road placement of substations.



(a) Substation prediction for Faisal Town (b) Substation prediction Gulberg (c) Substation prediction for Garden Town

(d) Substation prediction for Model Town (e) Substation prediction for Naseer-abad (f) Substation prediction for Johar Town

Fig. 8. Result of k-medians clustering with predicted centroids as substations for $k = 3$ substations.



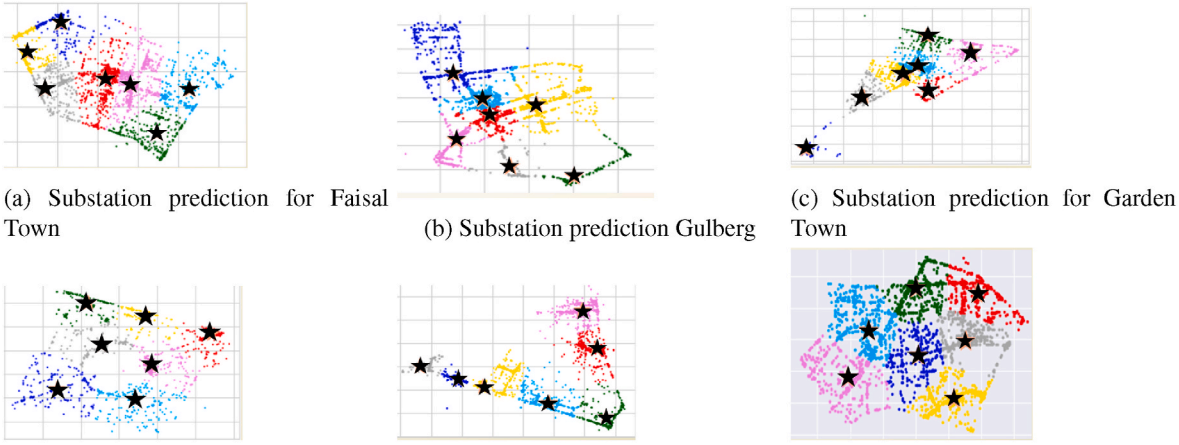
(a) Substation prediction for Faisal Town (b) Substation prediction Gulberg (c) Substation prediction for Garden Town

(d) Substation prediction for Model Town (c) Substation prediction for Naseer-abad (f) Substation prediction for Johar Town

Fig. 9. Result of k-medians clustering with predicted centroids as substations for $k = 5$ substations.

road or on the road instead of randomly predicted on random building or inside of someone's house.

Our proposed system would would allow law enforcement to dispatch correspondents from the nearest substation to a reported crime to respond efficiently (Caplan et al., 2020). We hope that such police presence would diminish the public's fear of crimes, and that quicker crime responses will deter future crimes and create safer communities.



(a) Substation prediction for Faisal Town

(b) Substation prediction Gulberg

(c) Substation prediction for Garden Town

(d) Substation prediction for Model Town

(e) Substation prediction for Naseer-abad

(f) Substation prediction for Johar Town

Fig. 10. Result of k-medians clustering with predicted centroids as substations for $k = 7$ substations.

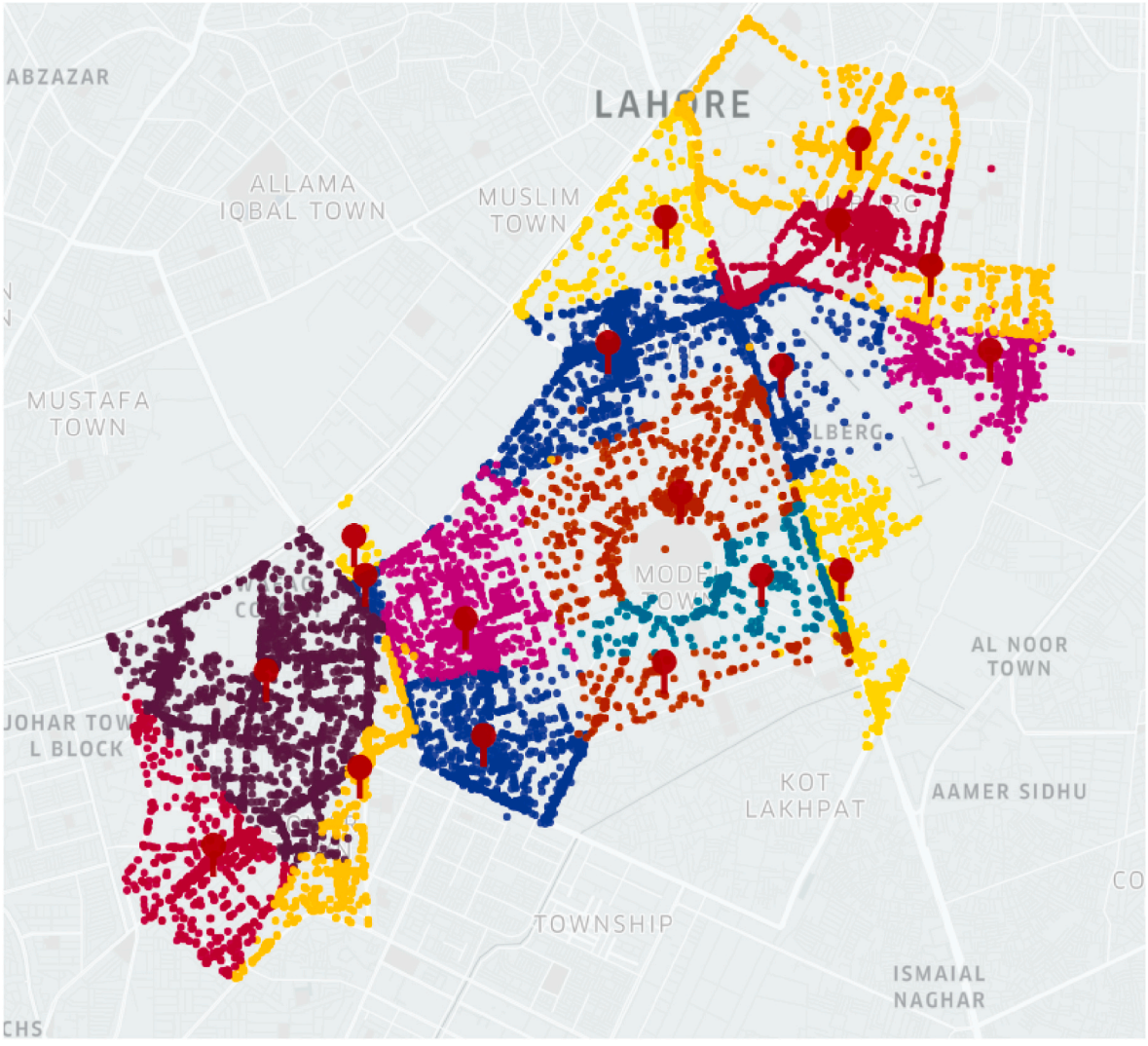


Fig. 11. Combined result for six targeted areas. Different colors are representing different clusters formed for $k = 3$ substations. Substations are represented as red pins on Map.

Table 5
Table is showing evaluation of objective function value in meters for different values of k substations in each of the targeted area.

Objective Function Value	$k = 3$	$k = 5$	$k = 7$
Gulberg	189	153	139
Garden Town	222	157	83
Faisal Town	314	294	240
Model Town	272	257	151
Naseerabad	319	257	206
Johar Town	376	239	170

Table 6
Evaluation of methodology is done by choosing three random substations within Gulberg area. For each crime point, nearest substation is find out and objective function value is evaluated and stored in this table.

Objective Function Value for $k = 3$	Average	Maximum	Minimum
distance (meters)	315	590	212

6.1. Theoretical contribution

The proposed research work mainly contributes in predicting suitable sub-station locations near road for security workforce distribution in order to immediately respond to nearby crimes. Our research work follows a data driven model in predicting optimal sub-stations locations. We proposed that multiple sub-stations within a particular area of a city are required to be established in order to distribute security workforce, since a single police station is not enough to immediately respond to crime happenings within that particular area. Our methodology is using the *k*-medians algorithm with specialized user-defined road distance matrix instead of using default Euclidian distance matrix.

6.2. Implications for practice

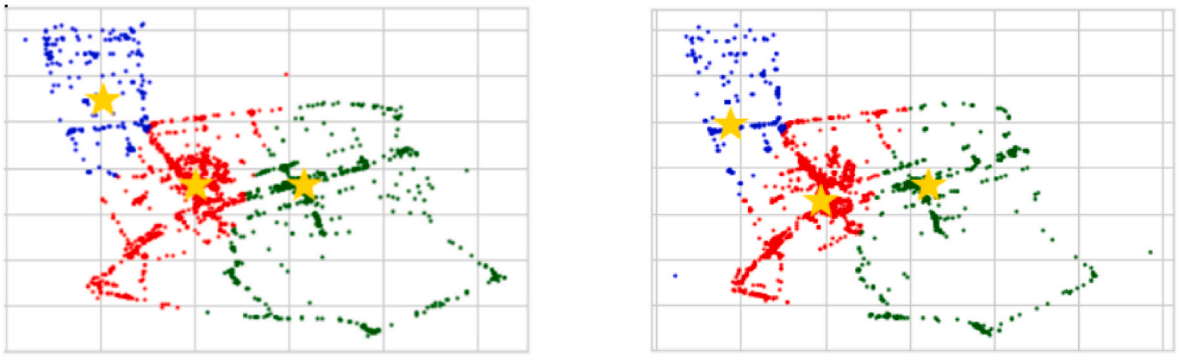
Police stations are facing obstacles in reducing crime due to the optimal distribution of the workforce. Security force resource distribution in an area is a new challenge; a city has different areas and each area region can be extended, but large areas have no sufficient police forces to distribute. Without substations in an area, it would be difficult to report a crime in an area for people, and police cannot deter crimes and criminals effectively, as one police station cannot be sufficient for the immediate based response.

With our proposed system, the substations will be placed in an area for optimal distribution of the workforce. People can report a crime on time and police can report and deter the crime immediately.

6.2.1. Data-driven policy formulation

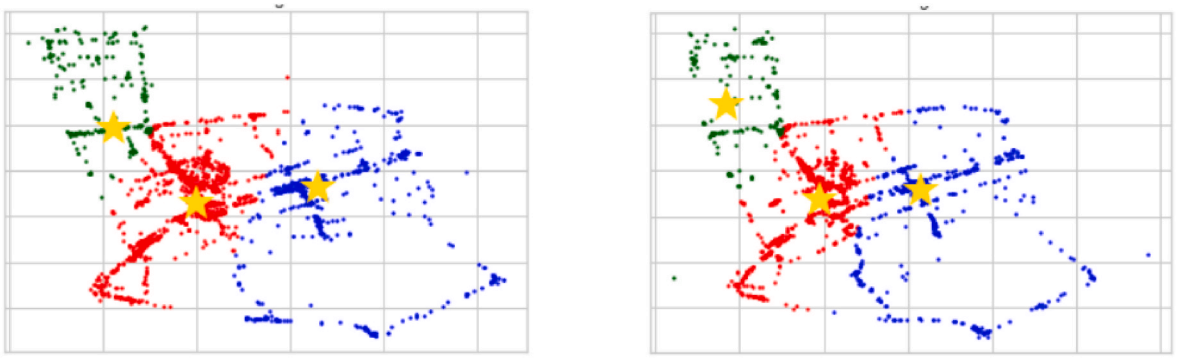
To predict substations in an area we need a relevant crime dataset. Using this data we can make better policies. The crime dataset will aid us to discover the number of substations and it will help us in equal distribution of the workforce. It is required to stream the crime dataset so the systems can pre-process the data and pass it to our model to predict the optimal substations. People can report a new crime to the closest substation easily. The crime dataset will benefit law enforcement in all operational aspects. It will help from actual policing to safer community development.

Data will even showdown that crime is increasing or decreasing in a specific area. This data is even important in creating an accurate budget for resource allocation. More law enforcement initiatives will be created to decrease crime. This can help criminal justice professionals to understand whether their initiatives are successful or not. The crime rate



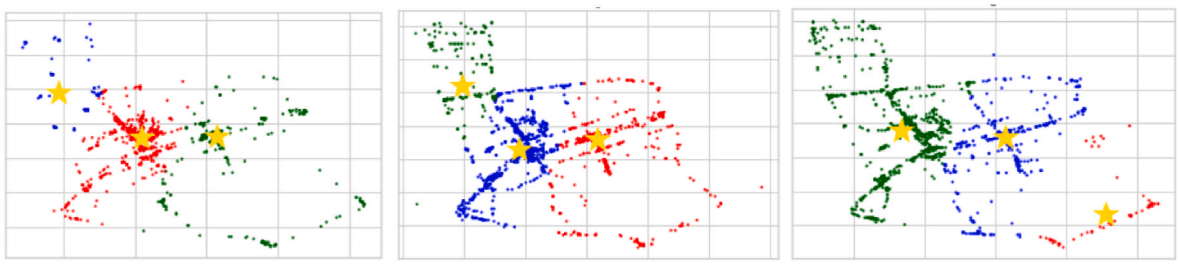
(a) Substation prediction during day time (b) Substation prediction during night time

Fig. 12. Result of *k*-medians clustering with predicted centroids as substations for sub-dataset of day and night time split.



(a) Substation prediction during summers (b) Substation prediction during winters

Fig. 13. Result of *k*-medians clustering with predicted centroids as substations for sub-dataset of summers and winters.



(a) Substation prediction for 2014 (b) Substation prediction for 2015 (c) Substation prediction for 2016

Fig. 14. Result of *k*-medians clustering with predicted centroids as substations for sub-dataset w.r.t years.

will be decreased due to the equal distribution of the security forces. With this solution, people will have timely and effective police services. The data-driven policy must be formalized already so we can easily place sub-stations.

6.3. Limitations and future research directions

Although the proposed model showed promising results, it experiences some limitations. As our placement of substations are data independent. We need crime dataset for equal distribution of security force. As crime dataset has very sensitive nature due to its accessibility to all. Its importance depends on strict data security enforcement. We anticipate continued improvement. A virtuous process will emerge in the best case scenario and we will witness improved availability of crime dataset. We are using the randomly generated points from the dataset to act as centroids. We will find a better technique of randomly generated points for the evaluation of our process. The proposed research work can be extended for traffic workforce distribution in order to predict optimal sub-areas where traffic is most likely to be congested, with many advancements in security forces and their programs many mobile police forces are also introduced, but the main issue is again the distribution of security workforces. The practical implication of our proposed work would be to monitor crime sensitive areas and adjust their security workforce distribution accordingly in order to respond immediately.

7. Concluding remarks

In this work we explored the problem of placing substations to efficiently respond to crimes in an area. We discussed multiple strategies one could follow using the geographical data (region, roads), and possibly old crime data. We performed a case study on real-life crime data from Lahore, Pakistan to show the efficacy of these methods. We showed this main result.

Future work includes making these algorithms more efficient, as time complexity was not the focus of this work. An interesting avenue to explore would be to use deep learning and computer vision to predict optimal locations based on satellite imagery of a region.

Credit author statement

Abinta Mehmood Mir: Conceptualization, Methodology, Investigation, Writing – original draft preparation & editing. Ali Hassan: Methodology, Writing – original draft preparation & editing. Asma Khalid: Validation, Software, revision. Zohair Raza Hassan: Investigation and writing-review & editing Mudassir Shabbir: Supervision, Data curation, Methodology, Writing – original draft preparation & editing. Faisal Kamiran: Investigation and writing-review & editing. Agha Ali Raza: Supervision, Investigation, Validation. Saeed-Ul Hassan: Supervision, Data curation, Methodology, Writing – original draft preparation & editing.

Acknowledgements

This work was supported by the grant received to establish the Crime Investigation and Prevention Lab, associated with the National Center in Big Data and Cloud Computing, funded by the Higher Education Commission (HEC) of Pakistan.

References

Bradley, P. S., Mangasarian, O. L., & Street, W. N. (1997). Clustering via concave minimization. In *Advances in neural information processing systems* (pp. 368–374).

Caplan, J. M., Kennedy, L. W., Piza, E. L., & Barnum, J. D. (2020). Using vulnerability and exposure to improve robbery prediction and target area selection. *Applied Spatial Analysis and Policy, 13*(1), 113–136.

J. Carroll, One study, four cities: Information impact in neighborhood economic development, Transforming government: People, process and policy.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. *IEEE Computer, 37*(4), 50–56.

Chen, Z., Liu, Y., Wong, R. C.-W., Xiong, J., Mai, G., & Long, C. (2014). Efficient algorithms for optimal location queries in road networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 123–134).

Chen, Z., Liu, Y., Wong, R. C.-W., Xiong, J., Mai, G., & Long, C. (2015). Optimal location queries in road networks. *ACM Transactions on Database Systems, 40*(3), 1–41.

Chen, F., Qi, J., Lin, H., Gao, Y., & Lu, D. (2019). Goal: A clustering-based method for the group optimal location problem. *Knowledge and Information Systems, 61*(2), 873–903.

Current, J., & Weber, C. (1994). Application of facility location modeling constructs to vendor selection problems. *European Journal of Operational Research, 76*(3), 387–392.

Dewinter, M., Vandeviver, C., Beken, T. V., & Witlox, F. (2020). Analysing the police patrol routing problem: A review. *ISPRS International Journal of Geo-Information, 9*(3), 157.

Du, Y., Zhang, D., & Xia, T. (2005). The optimal-location query. In *International symposium on spatial and temporal databases* (pp. 163–180). Springer.

Estivill-Castro, V., & Lee, I. (2001). Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *Proc. Of the 6th international conference on geocomputation, citeseer* (pp. 24–26).

Guha, S., & Khuller, S. (1999). Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms, 31*(1), 228–248.

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining, 9*(3), 139–154.

Hochbaum, D. S. (1984). When are np-hard location problems easy? *Annals of Operations Research, 1*(3), 201–214.

Hsu, W.-L., & Nemhauser, G. L. (1979). Easy and hard bottleneck location problems. *Discrete Applied Mathematics, 1*(3), 209–215.

Jain, K., Mahdian, M., Markakis, E., Saberi, A., & Vazirani, V. V. (2003). Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *Journal of the ACM, 50*(6), 795–824.

Jeffery, C. R. (1959). Pioneers in criminology: The historical development of criminology. *The Journal of Criminal law, Criminology, and Police Science, 50*(1), 3–19.

Kalcsics, J., Melo, T., Nickel, S., & Schmid-Lutz, V. (2000). Facility location decisions in supply chain management. In *Operations research proceedings 1999* (pp. 467–472). Springer.

Korn, F., & Muthukrishnan, S. (2000). Influence sets based on reverse nearest neighbor queries. *ACM Sigmod Record, 29*(2), 201–212.

Kumar, B. (2012). Role of information and communication technology in indian police. *Gian Jyoti E-Journal, 1*(2), 13–21.

Li, L., Jiang, Z., Duan, N., Dong, W., Hu, K., & Sun, W. (2011). Police patrol service optimization based on the spatial pattern of hotspots. In *Proceedings of 2011 IEEE international conference on service operations, logistics and informatics* (pp. 45–50). IEEE.

Lytras, M. D., Visvizi, A., Chopdar, P. K., Sarirete, A., & Alhalabi, W. (2020). Information management in smart cities: Turning end users' views into multi-item scale development, validation, and policy-making recommendations. *International Journal of Information Management*, 102146.

Lytras, M. D., Visvizi, A., & Sarirete, A. (2019). Clustering smart city services: Perceptions, expectations, responses. *Sustainability, 11*(6), 1669.

Melkote, S., & Daskin, M. S. (2001). Capacitated facility location/network design problems. *European Journal of Operational Research, 129*(3), 481–495.

Mukhopadhyay, A., Zhang, C., Vorobeychik, Y., Tambe, M., Pence, K., & Speer, P. (2016). Optimal allocation of police patrol resources using a continuous-time crime model. In *International conference on decision and game theory for security* (pp. 139–158). Springer.

Novikov, A. (2019). Pyclustering: Data mining library. *Journal of Open Source Software, 4*(36), 1230.

Singh, K. N. (2008). The uncapacitated facility location problem: some applications in scheduling and routing. *International Journal of Operational Research, 5*(1), 36–43.

Visvizi, A., & Lytras, M. D. (2018). Rescaling and refocusing smart cities research: from mega cities to smart villages. *Journal of Science and Technology Policy Management, 9*(2), 134–145.

Visvizi, A., & Lytras, M. D. (2019). Chapter 1 - smart cities research and debate: What is in there? In A. Visvizi, & M. D. Lytras (Eds.), *Smart cities: Issues and challenges* (pp. 1–14). Elsevier. https://doi.org/10.1016/B978-0-12-816639-0.00001-6. http://www.sciencedirect.com/science/article/pii/B9780128166390000016.

Vattani A., (2010) The hardness of k-means clustering in the plane, manuscript, accessible at http://cseweb.ucsd.edu/avattani/paper/kmeans_hardness.pdf.

Williamson, D. P., & Shmoys, D. B. (2011). *The design of approximation algorithms*. Cambridge university press.

Wong, R. C.-W., Özsu, M. T., Yu, P. S., Fu, A. W.-C., & Liu, L. (2009). Efficient method for maximizing bichromatic reverse nearest neighbor. *Proceedings of the VLDB Endowment, 2*(1), 1126–1137.

Xiao, X., Yao, B., & Li, F. (2011). Optimal location queries in road network databases. In *2011 IEEE 27th international conference on data engineering* (pp. 804–815). IEEE.

Yilmaz, E., Elbasi, S., & Ferhatosmanoglu, H. (2017). Predicting optimal facility location without customer locations. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2121–2130).

Zhang, D., Du, Y., Xia, T., & Tao, Y. (2006). Progressive computation of the min-dist optimal-location query. In *Proceedings of the 32nd international conference on Very large data bases* (pp. 643–654). VLDB Endowment.