Pertanika J. Sci. & Technol. 31 (2): 655 - 670 (2023)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Modified K-Means Clustering for Demand-Weighted Locations: A Thailand's Convenience Store Franchise - Case Study

Chartchai Leenawong* and Thanrada Chaikajonwat

Department of Mathematics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Lat Krabang, Bangkok, Thailand

ABSTRACT

This research applies and modifies K-means clustering analysis from Data Mining to solving the location problem. First, a case study of Thailand's convenience store franchise in locating distribution centers (DCs) is conducted. Then, the final centroids are served at suggested DC locations. Besides the typical distance, Euclidean, used in K-means, Manhattan, and Chebyshev, is also experimented with. Moreover, due to the stores' different demands, a modification of the centroid calculation is needed to reflect the center-of-gravity effects. For the proposed centroid calculation, the above three distance metrics incorporating the demands as weights give rise to another three approaches and are thus named Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev, respectively. Besides the optimal locations, the effectiveness of these six clustering approaches is measured by the expected total distribution cost from DCs to their served stores and the expected Davies–Bouldin index (DBI). Concurrently, the efficiency is measured by the expected number of iterations to the final clusters. All these six clustering approaches are then implemented in the case study of locating eight DCs to distribute to 260 convenience stores in Eastern Thailand. The results show that though all approaches yield locations in close proximity, the Weighted Chebyshev is the most effective one having both the lowest expected distribution cost and lowest expected DBI. In contrast, Euclidean is the most efficient approach, with the lowest expected number of iterations to the final clusters, followed by Weighted Chebyshev. Therefore, the DC locations from Weighted Chebyshev could, ultimately, be chosen for this Thailand's convenience store franchise.

Keywords: Centroid calculation, clustering, Davies–Bouldin index, demand-based, distance metrics, distribution center, K-means, location problem

ARTICLE INFO

Article history:
Received: 03 Febraury 2022
Accepted: 18 July 2022
Published: 06 March 2023

DOI: https://doi.org/10.47836/pjst.31.2.02

E-mail addresses:
chartchai.le@kmitl.ac.th (Chartchai Leenawong)
63605011@kmitl.ac.th (Thanrada Chaikajonwat)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

In Thailand, convenience stores are available on almost every corner (Wang, 2018). New franchises and new stores are emerging regularly, especially in tourist and populated areas. The Eastern part of Thailand is one of the well-known tourist attractions among local and foreign tourists, thanks partly to its terrific location next to the Gulf of Thailand (Ministry of Foreign Affairs, 2017; Surawattananon et al., 2021). Among those popular destinations are Pattaya, Koh Samet, and Koh Kut. Therefore, it is natural for those convenience store franchises to open more branches. Logistics management plays a crucial role in both short-run and long-run plans for franchises to stay competitive. One long-run logistical decision is determining where to locate distribution centers (DCs) (Langley et al., 2020).

This study investigates a case of locating DCs to distribute products to 260 franchised convenience stores in Eastern Thailand. Since Eastern Thailand is comprised of seven provinces: Chachoengsao, Chonburi, Rayong, Chanthaburi, Trat, Prachinburi, and Sa-Kaeo, plus one special governed city, Pattaya, the convenience store franchise of interest chooses to have eight DCs to be located. The objective is to minimize transportation or distribution costs from and to those eight DCs. Figure 1 shows the map of Thailand and the 260 locations of the convenience stores for this study respectively.

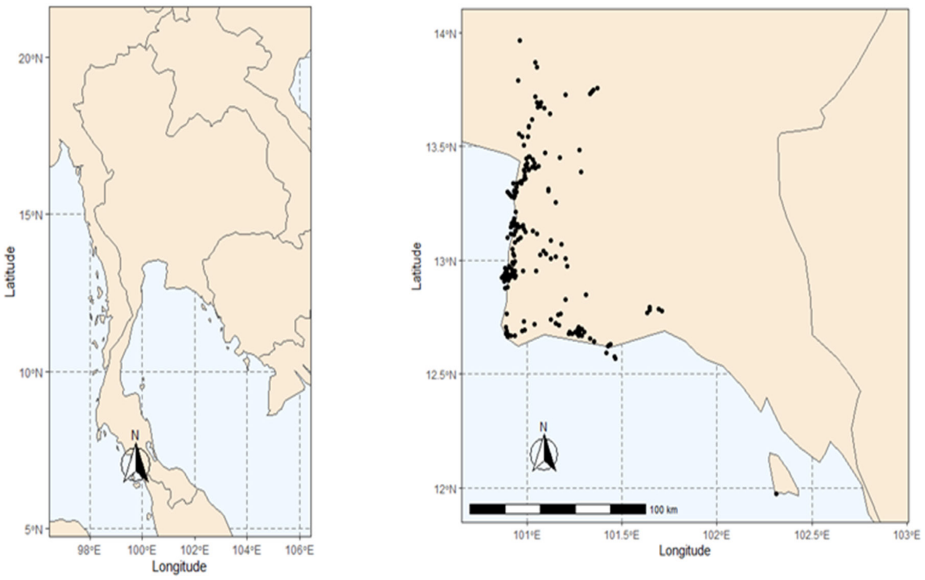

Figure 1. Thailand map and the convenience store locations in Eastern Thailand

There are numerous ways to solve the location problem, for instance, optimization models, the p-center/p-median algorithm, and the grid technique. However, in this research,

as the first contribution of this work, K-means clustering analysis from Data Mining is adapted to find the appropriate locations. The centroid of each final cluster serves as each DC's location. Also, as the second contribution, the typical distance metric, i.e., Euclidean, used in the K-means clustering, is replaced with other distance metrics, namely, Manhattan and Chebyshev. These three-distance metrics constitute the first three clustering approaches to the experiment.

As for the third contribution of this work, due to the nature of the location problem application, the clustering algorithm needs to be modified to fit this problem better. Accordingly, the centroid calculation during each clustering iteration is adjusted to incorporate the center of gravity's impact. That is done by taking the unequal demands required at the served stores as different weights. As a result, three additional clustering approaches with demand-weighted centroid calculation are proposed and named Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev, respectively.

Altogether, all these six clustering approaches experiment with, and their results are compared, considered both effectiveness and efficiency. The effectiveness is measured by the expected total distribution cost and the expected Davies–Bouldin index (DBI). In contrast, the efficiency is measured by the expected number of iterations to the final clusters.

LITERATURE REVIEW

The facility location problem, or the location problem, refers to how and where to place facilities in a logistics network to minimize total transportation costs from and to those facilities. Four underlying assumptions of the problem are the following: customers assumed to already be at points or on routes, facilities to be located, a space in which customers and facilities are located, and a standard metric that specifies distances or times between customers and facilities. Facilities in the location problem are small relative to the space in which they are located, and interactions between facilities may occur (Farahani & Hekmatfar, 2009).

Facility location decisions are critical to strategic planning for private sectors such as industrial estates, banks, retail facilities, distribution centers (DCs), and public sectors such as hospitals, post offices, and government headquarters. Determining facility locations is one of the broad and long-term decisions influencing numerous operational and logistical decisions. Locating or relocating facilities usually involves huge investments, as it may need to pay enormously for land acquisition and facility construction. Therefore, decision-makers must consider not only every current perspective of the facility but also unforeseen future events that may affect the facility, such as demographics, climate change, and market trend evolution during its lifetime (Farahani & Hekmatfar, 2009).

The location problem was first introduced in 1909 when Alfred Weber considered how to place a single warehouse in such a way as to minimize the total distance between the warehouse and several customers. After that, the location problem was advanced by several other applications. For example, Hakimi (1964) wanted to locate switching centers in a communication network, while Farahani and Hekmatfar (2009) tried to locate police posts along a highway system.

Drezner et al. (2003) studied the best location of a central warehouse to determine the number and the locations of local warehouses. They built simple models that considered inventory and service costs and compared them with those from the traditional model, minimizing the total transportation cost. The models were demonstrated on an example problem with up to 10,000 demand points. Excel Solver solves each model in less than half a second. However, it turned out that the location solutions for all the models were quite different from one another. The conclusion of this research showed that different models led to different locations. Therefore, the decision-maker needed to decide which model was the most appropriate for the situation. In addition, numerical results showed that ignoring inventory costs made the models less accurate.

Yang et al. (2007) investigated the location problem regarding selecting distribution centers from a potential set so that the total relevant cost was minimized under a fuzzy environment. More specifically, the setup cost, turnover cost, and demands of the customers were assumed fuzzy variables. Consequently, a probabilistic-constrained programming model for the problem was designed, and some properties of the model were examined. Tabu search, genetic and fuzzy simulation algorithms were integrated to search for the approximate best solution while satisfying the transportation and assignment constraints of the DCs. The effectiveness and robustness of the hybrid algorithm were tested through a numerical example. As a result, fuzzy chance-constrained programming was constructed as a decision model for the problem. For the convenience of model solving, some mathematical properties of the model were also obtained.

Dantrakul et al. (2014) applied greedy, p-median, and p-center algorithms to the facility location problem to minimize the sum of the setup and transportation costs. Those two costs were considered a function of the number of opened facilities. The network in this work represented the road transportation system of six provinces in Northern Thailand. The facility location model with bounds for the number of the opened facility was constructed in this work. The performances of the constructed methods were tested using 100 random data sets. Simulation results showed that the method developed from the greedy algorithm was suitable for solving the problem when the setup cost was higher than the transportation cost. In contrast, the p-median-based methods were more efficient for the opposite case when the setup cost was lower.

Sharma and Jalal (2017) developed a new clustering and mixed-integer linear programming-based hybrid approach for solving the facility location problem. The main objective was to utilize the facility by maximizing the number of possible customers to maximize profit. The numerical results showed that the profit started to decrease as the number of clusters increased. If the profit kept decreasing, it indicated that the solution procedure would stop.

Chen (2019) studied the location problem of DCs based on the Baumer Walvar model using Jiaji Logistics as a case study. This research aimed to optimize the total DC costs, consisting of four cost components, namely, the transportation cost from the factories to DCs, and from DCs to the customers, the DCs' fixed costs, and the DCs'change fee. The whole cargo of Jiaji logistics was transported from five factories (Chongqing, Chengdu, Xi'an, Zhengzhou, and Lanzhou) to four customers (Guangzhou, Shanghai, Hangzhou, and Tianjin). The company wanted to select the optimal five DCs out of the predetermined eight DCs (Wuhan, Nanchang, Guiyang, Changsha, Shijiazhuang, Beijing, and Nanjing). The economies of scale were also taken into account. The results showed that the minimum cost was 7,301,620 yuan, and the optimal locations of DCs were Nanchang, Nanjing, Guiyang, Changsha, and Shijiazhuang.

As for previous work on the K-means clustering, algorithms, distance metrics, and performance measurement are of our interest and are presented as follows.

Singh et al. (2013) compared the K-means clustering using three different distance metrics: Euclidean, Manhattan, and Minkowski. All the experiments were performed on dummy data. The result showed that Euclidean distance gave the best performance while Manhattan distance yielded the worst.

Sinwar and Kaushik (2014) studied two popular distance metrics, Euclidean and Manhattan, on the simple K-means clustering. They used two real and one synthetic data set, namely, Iris, Diabetes, and BIRCH. The development tool for clustering data items was WEKA, and the numbers of clusters used in this research were 2, 3, 4, 5, 6, and 7. The results showed that the Euclidean method was more efficient than the Manhattan method in terms of the number of iterations performed during centroid calculation.

Gultom et al. (2018) analyzed and compared object clustering from real big data using K-means and K-medoid methods. In both methods, combination testing used three distance metrics: Euclidean, Canberra, and Chebyshev. The sample dataset contained six variables collected from three college classes having 147,679 students at Medan State University. Performance measurement was the Davies-Bouldin index. The results showed that the Chebyshev distance in K-means yielded better results than that in K-medoid in terms of accuracy and quality. On the other hand, the results suggested not to use the Canberra distance in K-means nor K-medoid because the Davies-Bouldin index was undefined.

In the next section, the K-means clustering using three different distance metrics and the proposed demand-weighted approaches is explained.

THE TYPICAL AND PROPOSED CLUSTERING APPROACHES

This section describes the typical K-means clustering along with the proposed modified one in detail. K-means clustering is the most commonly used clustering algorithm and one of the most efficient partitional clustering algorithms. The K-means clustering algorithm's general steps are explained step by step as follows (Gultom et al., 2018; Aggarwal & Reddy, 2014).

Step 1: Determine the number of clusters formed in the dataset, K.
Step 2: Randomly choose K representative points as initial "centroids" of the K clusters.
Step 3: For each point, calculate the distance to each centroid and identify the closest centroid.

Then, assign that point to the cluster.

Step 4: Once all the points are assigned to clusters, update the centroids of all clusters.
Step 5: Repeat step 3 to step 4 until all the points in each cluster do not change. The algorithm stops. The last set of centroids is used as the desired locations.

However, in our application of locating the DCs for a convenience store franchise where the points to be clustered represent the convenience stores and the centroids represent the locations of the DCs serving the stores in the same clusters, it is natural to also take into consideration the different demands at the served stores. Therefore, in our case, the demands are used as weights in computing the updated centroids after the clusters are formed at each iteration.

In the following, the modified K-means clustering algorithm that incorporates the stores' different demands is applied to and explained in our application context. Simultaneously, three distance metrics, namely, Euclidean, Manhattan, and Chebyshev, are experimented with in the algorithm as well. Finally, together with the typical and demand-weighted centroid calculations, six combinations are tried to compete for the best algorithm. The notations used in this article are defined as follows.

K = the number of clusters/centroids/DCs; in our case here, $K = 8$.
N = the total number of convenience stores. Here, $N = 260$.
T_i = the number of convenience stores in cluster i; $i = 1, 2,\ldots, K$.
$X_i = (x_i, y_i)$ refers to the location of centroid i representing DC i, where x_i and y_i are the latitude and longitude of centroid i, $i = 1, 2,\ldots, K$, respectively.
$S_j = (r_j, s_j)$ refers to the location of convenience store j, where r_j and s_j are the latitude and longitude of store j, $j = 1,2,\ldots, N$, respectively.

$S^i_j = (r^i_j, s^i_j)$ refers to the location of store j that is assigned to cluster i.
w_j = the demand at convenience store j

Then, the K-means using each of these three-distance metrics, Euclidean, Manhattan, and Chebyshev, and the modified demand-weighted K-means using each of the above metrics proceed in detail as follows.

Step 1: Random eight initial centroids X_i; i = 1, 2,…, 8, representing eight initial DCs.

Step 2: For a fixed convenience store S_j, calculate the distance between the store and each centroid

X_i uses one of the three metrics, i.e., Euclidean, Manhattan, and Chebyshev, according to Equations 1, 2, and 3 (Singh et al., 2013).

$$D_{\text{Euclidean}}(S_j, X_i) = \sqrt{(r_j - x_i)^2 + (s_j - y_i)^2} \quad i = 1, 2,\ldots, 8 \tag{1}$$

$$\text{or } D_{\text{Manhattan}}(S_j, X_i) = |r_j\text{-}x_i| + |s_j\text{-}y_i| \quad i = 1, 2,\ldots, 8 \tag{2}$$

$$\text{or } D_{\text{Chebyshev}}(S_j, X_i) = \max(|r_j\text{-}x_i|, |s_j\text{-}y_i|) \quad i = 1, 2,\ldots, 8 \tag{3}$$

Then, select the centroid i that minimizes the distance from store j. Assign this store S_j to cluster X_i accordingly. Now, S_j becomes S^i_j; that is, store j is grouped in cluster i; in other words, served by centroid or DC i. Repeat this step for all other stores.

Step 3: Calculate the new location of each centroid i, using the typical average of all store locations j in cluster i, as Equation 4.

$$X_i = \left(\frac{\sum_{j=1}^{T_i} r^i_j}{T_i}, \frac{\sum_{j=1}^{T_i} s^i_j}{T_i} \right) \quad \text{for } i = 1, 2,\ldots, 8 \tag{4}$$

On the other hand, the effect of each store's different demand results in the proposed demand-weighted average for computing the new location of each centroid i as Equation 5.

$$X_i = \left(\frac{\sum_{j=1}^{T_i} w_j r^i_j}{\sum_{j=1}^{T_i} w_j}, \frac{\sum_{j=1}^{T_i} w_j s^i_j}{\sum_{j=1}^{T_i} w_j} \right) \quad \text{for } i = 1, 2,\ldots, 8 \tag{5}$$

Step 4: Repeat Steps 2 to 3 until all convenience stores in the final clustering are the same as in the immediate previous clustering.

Step 5: The total distribution cost from DCs to their served stores is calculated, and the Davies–Bouldin index (DBI) is computed to measure the effectiveness. As for the efficiency measurement, the number of iterations to the final clusters is determined.

The details of these measures are given in the next section.

Step 6: Repeat Steps 1 through 5 for 10,000 instances to obtain the expected distribution cost, the expected DBI, and the expected number of iterations to the final clusters, accordingly.

Now that all the algorithm steps have been stated, the effectiveness and efficiency of the six clustering approaches will be measured and compared. These issues will be explained in more detail next.

THE EFFECTIVENESS AND EFFICIENCY MEASUREMENT

The modified demand-weighted K-means algorithm described above employs three different distance metrics and two centroid location calculations. As a result, six different approaches are carried out for each problem instance. The first three approaches are named after the three-distance metrics: Euclidean, Manhattan, and Chebyshev. The other three approaches incorporating the demands as the weights in updating the centroid location calculation are Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev. After the experiments are performed, these six approaches are compared by their effectiveness and efficiency. In terms of effectiveness, the expected total distribution cost and the expected Davies–Bouldin index (DBI) are measured. In contrast, in terms of efficiency, the expected number of iterations to the final clusters is determined for each of the six clustering approaches.

Measurement of Effectiveness: Distribution Cost

For our application, we are most concerned with the overall distribution cost of locating the DCs. Typically, the distribution cost depends on the transportation rate, the shipment weight, and the traveling distance. Let us assume that the transportation rate is \$1 per kilometer per one shipment weight unit. Assume further that the shipment load is the demand at each store S_j served by DC X_i, denoted by l_{ij}. Finally, for the traveling distance between the store and its relevant DC, the Euclidean metric is used in the calculation. Therefore, the distribution cost from DC X_i to store S_j is as in Equation 6.

$$\textit{Distribution cost} = \$1 \times l_{ij} \times D_{\text{Euclidean}}(S_j, X_i) \tag{6}$$

Measurement of Effectiveness: Davies–Bouldin Index (DBI)

The Davies-Bouldin Index (DBI), introduced by David L. Davies and Donald W. Bouldin in 1979, is a metric for evaluating clustering algorithms. It is an internal evaluation scheme in which the evaluation of how well the clustering is performed is based on variables and features that are intrinsic to the dataset. The process of calculating DBI is as follows (Davies & Bouldin, 1979):

Step 1: For each cluster i, calculate the average distance between all stores S_j in the cluster and DC X_i, denoted by A_i, by Equation 7.

$$A_i = \frac{1}{T_i}\sum_{j=1}^{T_i}\left\|S_j^i - X_i\right\| = \frac{1}{T_i}\sum_{j=1}^{T_i}\sqrt{\left(r_j^i - x_i\right)^2 + \left(s_j^i - y_i\right)^2} \; ; i = 1,2,\ldots,8. \tag{7}$$

Step 2: Calculate the distance between DCs X_h and X_i, denoted by M_{hi}, according to Equation 8.

$$M_{hi} = \left\|X_h - X_i\right\| = \sqrt{\left(x_h - x_i\right)^2 + \left(y_h - y_i\right)^2} \tag{8}$$

Step 3: For each pair of DCs X_h and X_i, can calculate using Equation 9

$$R_{h,i} = \frac{A_h + A_i}{M_{h,i}} \tag{9}$$

Then, identify using Equation 10

$$D_i = \max_{h \neq i} R_{h,i} \tag{10}$$

Step 4: Finally, calculate DBI using the following Equation 11.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} D_i \tag{11}$$

Measurement of Efficiency: Number of Iterations to the Final Clusters

To measure efficiency, for each instance, the number of iterations to the final clusters, where all the stores served by the DCs remain unchanged from the previous iteration, is counted. Once the experiment is repeated for 10,000 instances, an average is obtained for each of the six clustering approaches.

THE EXPERIMENTS, THE RESULTS, AND THE DISCUSSION

This section presents the experiments, their results, and the discussion. First, all the previously mentioned six different clustering approaches, resulting from a combination of three different distance metrics and two calculation methods for centroid locations, are experimented with for our location problem. More precisely, Euclidean, Manhattan, and Chebyshev, together with the other three demand-weighted approaches, are Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev, are applied to find the optimal eight DC locations for distributing goods to 260 convenience stores in Eastern Thailand.

The experiments conducted in this study use a total of 10,000 different instances. For comparison purposes, each instance randomizes new initial centroids, and these same initial centroids are then used in all six approaches. After the 10,000 instances are carried out for each approach, the effectiveness and efficiency measurement expectations are calculated over these 10,000 instances.

The optimal solutions obtained from these six clustering approaches are first tabulated, followed by their efficiency and effectiveness results reported in tabular and graphical presentation. In addition, a discussion of all the results is provided.

Also, note that all the experiments in this research are run on Intel® Core™ i5-1035G4 with 8 GB of DDR4 memory. The programs are coded in R-programming on RStudio version 1.3.1093.

Optimal Solution Results: The Locations of Eight Centroids or DCs

All eight optimal centroids or DC locations are obtained after implementing all six clustering approaches (Table 1). They all yield the optimal locations nearby, which are not easy to differentiate. Therefore, measurement of the effectiveness and efficiency of the six clustering approaches is needed for comparison purposes.

Table 1

Optimal eight centroids from six different clustering approaches

Clustering Approach	Centroid 1	Centroid 2	Centroid 3	Centroid 4
Euclidean	(13.358,100.988)	(13.017,101.132)	(13.867,101.004)	(12.700,101.341)
Weighted Euclidean	(13.365,100.989)	(13.018, 101.135)	(13.876,101.009)	(12.697,101.337)
Manhattan	(12.380,101.933)	(12.794,101.164)	(13.797,101.208)	(13.152,101.045)
Weighted Manhattan	(12.487,101.842)	(12.786,101.171)	(13.799,101.208)	(13.156,101.042)
Chebyshev	(13.357,100.991)	(13.024,101.126)	(13.867,101.004)	(12.699,101.347)
Weighted Chebyshev	(13.355,100.990)	(13.024,101.130)	(13.876,101.009)	(12.697,101.337)

Clustering Approach	Centroid 5	Centroid 6	Centroid 7	Centroid 8
Euclidean	(12.879,100.912)	(13.624,101.131)	(11.972,102.312)	(13.131,100.950)
Weighted Euclidean	(12.875,100.912)	(13.642,101.151)	(11.972,102.312)	(13.129,100.949)
Manhattan	(13.030,101.060)	(13.507,101.108)	(12.692,100.929)	(12.906,100.928)
Weighted Manhattan	(13.019,101.068)	(13.604,101.075)	(12.691,100.931)	(12.906,100.930)
Chebyshev	(12.876,100.916)	(13.625,101.126)	(11.972,102.312)	(13.131,100.947)
Weighted Chebyshev	(12.875,100.912)	(13.631,101.132)	(11.972,102.312)	(13.130,100.946)

Effectiveness Results: The Expected Distribution Cost

For the effectiveness measurement, the first indicator, the distribution cost from each approach, is calculated (Equation 6) in the previous section and then reported and visualized (Figure 2). The expectation is averaged over 10,000 instances for each clustering approach.

Weighted Chebyshev and Chebyshev produce the first two lowest expected distribution costs of $1,559.66 and $1,564.61, respectively. On the contrary, Weighted Manhattan and Manhattan generate the worst two expected distribution costs of $6,805.71 and $6,650.62,

respectively. Note that the expected distribution costs of these worst two are also far from those of the remaining approaches.

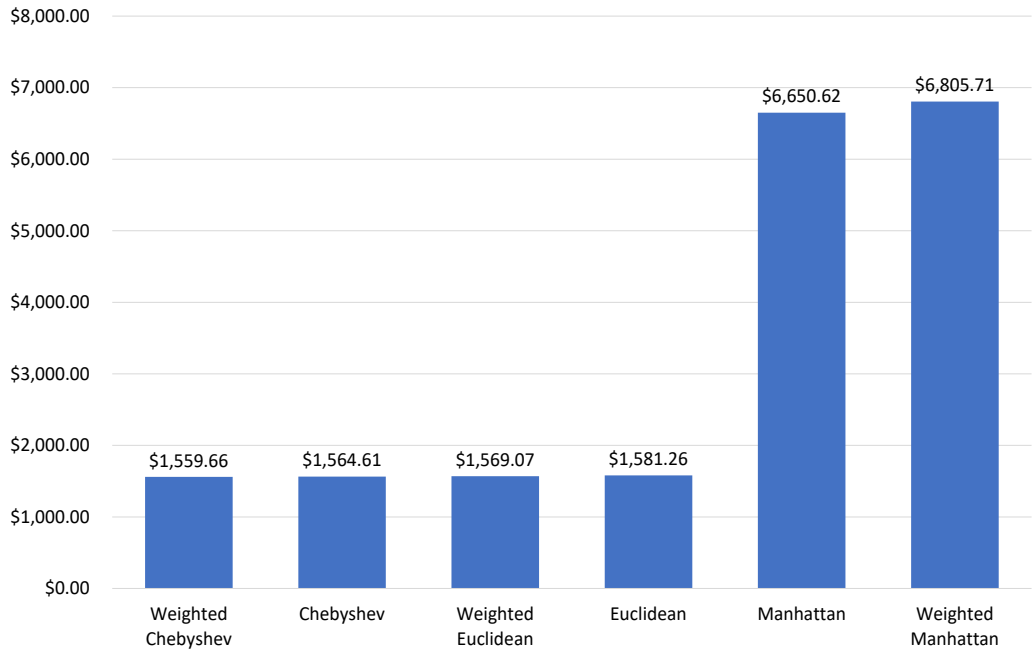

Figure 2. Bar chart of the expected distribution costs over 10,000 instances from each different clustering approach

Effectiveness Results: The Expected DBI

The other indicator of effectiveness is the expected Davies-Bouldin Index (DBI) from the six approaches. They are calculated according to the steps in the previous section and then reported and visualized by bar charts in Figure 3.

The results show that Weighted Chebyshev and Chebyshev yield the best two expected DBIs of 0.6779 and 0.6793, respectively. In contrast, Manhattan and Weighted Manhattan yield the worst two DBIs of 2.1905 and 2.0939, respectively. Similar to the above effectiveness results by the expected distribution costs, the two DBIs of these two worst approaches are far away from those of the remaining approaches even though the worst here, Manhattan, and the second worst, Weighted Manhattan, are interchanged from before.

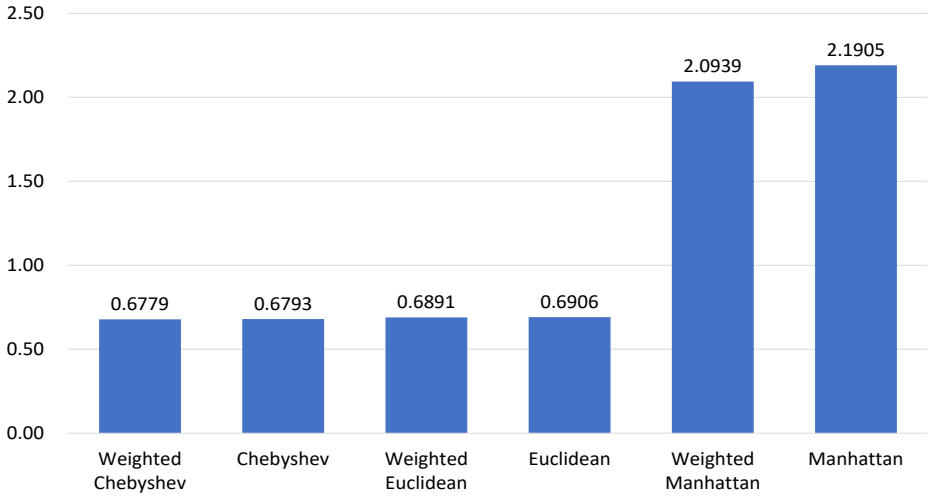

Figure 3. Bar chart of the expected DBI over 10,000 instances from each different clustering approach

Efficiency Results: The Expected Number of Iterations to the Final Clusters

For the efficiency measurement of all six approaches, the expected numbers of iterations to the final clusters are determined and compared. They averaged over 10,000 instances for each clustering approach. Euclidean yields the lowest expected number of iterations at 8.65 (Figure 4). Slightly in the second and third bests are Weighted Chebyshev at 8.88 and Chebyshev at 8.97, while Weighted Manhattan and Manhattan are the worst two with the numbers far away from the rest, that is, 16.01 and 14.84, respectively. Thus, in terms of efficiency, it is fair to say Euclidean, Weighted Chebyshev, and Chebyshev are among the most efficient approaches.

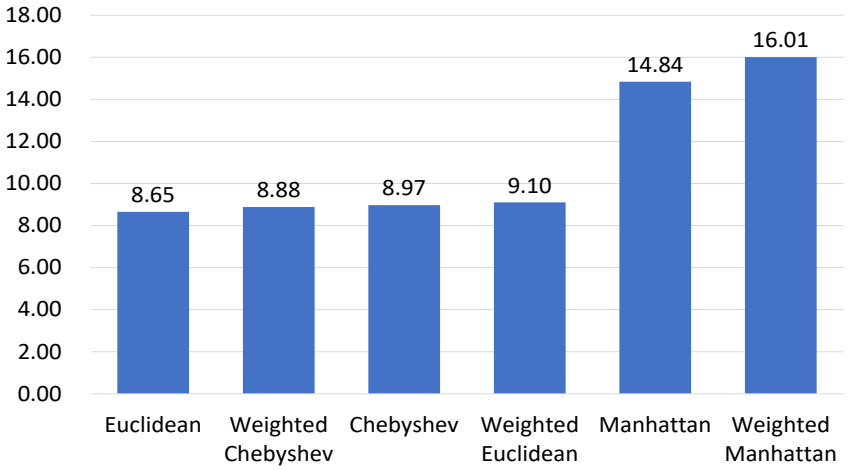

Figure 4. Bar chart of the expected number of iterations to the final clusters over 10,000 instances from each different clustering approach

The Discussion of the Results

The optimal locations obtained from the six clustering approaches are not significantly different (Table 1). Thus, the effectiveness and efficiency measurement can be good indicators for differentiating the six approaches as reported in the previous subsections. Nevertheless, a discussion on the combined results across every approach is needed and hence given here.

Starting with a summary of the effectiveness and efficiency results (Table 2) and obviously, Weighted Chebyshev is most effective either judged by the expected distribution cost or the expected DBI (Figure 5). Moreover, even though Euclidean is the most efficient among the six approaches, the second most efficient, Weighted Chebyshev, is just slightly behind. Hence, Weighted Chebyshev could be the clustering approach that best fits our case study of locating the DCs to serve their convenience stores with different demands.

Table 2

Summary of the effectiveness and efficiency of all six different clustering approaches

Clustering Approach	Effectiveness		Efficiency
	Expected distribution cost	Expected DBI	Expected number of iterations
Weighted Chebyshev	$1,559.66	0.6779	8.88
Chebyshev	$1,564.61	0.6793	8.97
Weighted Euclidean	$1,569.07	0.6891	9.10
Euclidean	$1,581.26	0.6906	8.65
Manhattan	$6,650.62	2.1905	14.84
Weighted Manhattan	$6,805.71	2.0939	16.01

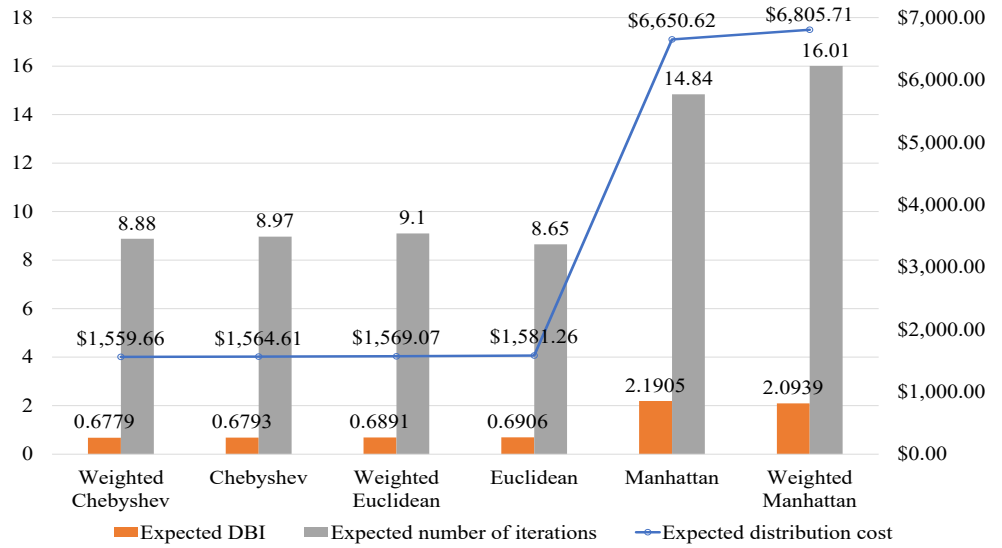

Figure 5. Graphical summary of the effectiveness and efficiency results from all six different clustering approaches

In this section, the results from the six experiments using the three types and three proposed clustering approaches have been reported and discussed combined results. In the next section, a conclusion of this work is first given. Then, suggestions for future research improvement are provided in the end.

CONCLUSION AND SUGGESTIONS

This research examines the location problem with a case study of locating DCs for Thailand's convenience store franchise. The K-means clustering algorithm is adapted so that the centroids in the final iteration can be used as the appropriate locations. In addition to the Euclidian distance typically used in the K-means clustering, two other distance metrics, Manhattan and Chebyshev, are also used.

Furthermore, due to this particular location problem's characteristics of having different demands at the stores and thus different shipment sizes, the locations should be pulled by the center-of-gravity rule. Therefore, modifications to the algorithm are necessary to suit this application better. This research proposes one way of doing so by adjusting the centroid calculation. As a result, the centroid calculation at each iteration is weighted by the stores' different demands. Besides the first three distance metrics, namely, Euclidean, Manhattan, and Chebyshev, another three modified distance metrics are proposed and named Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev, respectively.

After these, six clustering approaches are experimented on in the case study of locating eight DCs to service 260 convenience stores in Eastern Thailand. The resulting eight DCs' locations show insignificantly different, and thus the effectiveness and efficiency of these approaches play a significant role. In conclusion, the clustering approach best fits this certain problem is the proposed demand-weighted Chebyshev.

Apart from this, several possible ideas for future research are suggested. First, the cost of constructing a DC at each location is usually different, so it should somehow be reflected in the algorithms. The same logic can also be applied to the different transportation rates at different locations.

Also, another way to incorporate the center-of-gravity impact of the stores' different demands is by introducing another attribute into the distance metric Equation. In addition to the latitude and longitude attributes, the store's demand can be treated as another attribute.

Furthermore, other than the three-distance metrics employed here, the distances between the convenience stores and their distribution centers may be figured from the real world based primarily on existing land routes.

Finally, as in the traditional K-means clustering, the complexity of the initialization steps can be viewed as a trade-off to the number of iterations to the final clusters. It is suggested to explore further into this issue to obtain higher algorithm efficiency.

ACKNOWLEDGEMENT

The authors want to thank the anonymous reviewers whose valuable comments and thoughtful suggestions helped enhance the quality of this manuscript. The authors would also like to express sincere thanks to the editors who were always there for us since the very first step of this publication process.

REFERENCES

Aggarwal, C. C., & Reddy, C. K. (Eds.). (2014). *Data clustering algorithms and applications*. CRC Press.

Chen, H. (2019). Location problem of distribution center based on Baumer Walvar model: Taking Jiaji logistics as an example. *Open Journal of Business and Management*, *7*(2), 1042-1052. https://doi.org/10.4236/ojbm.2019.72070

Dantrakul, S., Likasiri, C., & Pongvuthithum, R. (2014). Applied p-median and p-center algorithms for facility location problems. *Expert Systems with Applications*, *41*(8), 3596-3604. https://doi.org/10.1016/j.eswa.2013.11.046

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1*(2), 224-227. https://doi.org/10.1109/TPAMI. 1979.4766909

Drezner, Z., Scott, C., & Song, J. S. (2003). The central warehouse location problem revisited. *IMA Journal of Management Mathematics, 14*(4), 321-336. https://doi.org/10.1093/imaman/ 14.4.321

Farahani, R. Z., & Hekmatfar, M. (Eds.). (2009). *Facility location: Concepts, models, algorithms and case studies.* Springer Science & Business Media.

Gultom, S., Sriadhi, S., Martiano, M., & Simarmata, J. (2018). Comparison analysis of K-means and K-medoid with Euclidean distance algorithm, distance, and Chebyshev distance for big data clustering. In *IOP Conference Series: Materials Science and Engineering* (Vol. 420, No. 1, p. 012092). IOP Publishing. https://doi.org/10.1088/1757-899X/420/1/012092

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research, 12*(3), 450-459. https://doi.org/10.1287/opre.12.3.450

Langley, C. J., Novack, R. A., Gibson, B., & Coyle, J. J. (2020). *Supply chain management: A logistics perspective*. Cengage Learning.

Ministry of Foreign Affairs. (2017). *Tourism industry in Thailand.* Netherlands embassy in Bangkok. https://www.rvo.nl/sites/default/files/2017/06/factsheet-toerisme-in-thailand.pdf

Sharma, A., & Jalal, A. S. (2017). Clustering based hybrid approach for facility location problem. *Management Science Letters*, *7*(12), 577-584. https://doi.org/10.5267/j.msl.2017.8.007

Singh, A., Yadav, A., & Rana, A. (2013). K-means with three different distance metrics. *International Journal of Computer Applications*, *67*(10), 13-17. https://doi.org/10.5120/11430-6785

Sinwar, D., & Kaushik, R. (2014). Study of Euclidean and Manhattan distance metrics using simple K-means clustering. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *2*(5), 270-274.

Surawattananon, N., Reancharoen, T., Prajongkarn, W., Chunanantatham, S., Simakorn, Y., & Gultawatvichai, P. (2021). *Revitalising Thailand's tourism sector*. Bank of Thailand. https://www.bot.or.th/Thai/MonetaryPolicy/EconomicConditions/AAA/250624_WhitepaperVISA.pdf

Wang, M. (2018). *The research of strategy for the 7-eleven convenience store in Thailand* (Masters dissertation). Siam University, Thailand. https://e-research.siam.edu/wp-content/uploads/2019/08/IMBA-2018-IS-The-Research-of-Strategy-for-the-7-Eleven-Convenience-Store_compressed.pdf

Yang, L., Ji, X., Gao, Z., & Li, K. (2007). Logistics distribution centers location problem and algorithm under fuzzy environment. *Journal of Computational and Applied Mathematics*, *208*(2), 303-315. https://doi.org/10.1016/j.cam.2006.09.015