DATA MINING FOR THE DEMAND–BASED LOCATION PROBLEM AND HOLT'S FORECASTING WITH EVENTS

THANRADA CHAIKAJONWAT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN APPLIED MATHEMATICS

DEPARTMENT OF MATHEMATICS SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2023
KMITL-2023-SC-D-001-004

Thesis Title	Data Mining for the Demand–based Location Problem and Holt's Forecasting with Events
Student Name	Miss Thanrada Chaikajonwat
Student ID	63605011
Degree	Doctor of Philosophy (Applied Mathematics)
Department	Mathematics
Year	2023
Thesis Advisor	Assoc. Prof. Dr. Chartchai Leenawong

Abstract

This thesis proposes modified data mining techniques for addressing the location problem and sales forecasting with special events. First, the typical K-means clustering algorithm is applied and also modified to find optimal locations for distribution centers for a Thailand-based convenience store franchise. Three clustering approaches using different distance metrics, namely, Euclidean, Manhattan, Chebyshev, along with their three modified ones that incorporate the demands, are compared for effectiveness and efficiency. The Weighted Chebyshev approach offers the best results in terms of expected distribution cost and the Davies–Bouldin index, while Euclidean is the most efficient. In addition, this research proposes three modified Holt's-based methods to better forecast time-series data affected by events, such as the COVID-19 pandemic, on Thailand's automotive industry. The proposed methods show improved accuracy values in terms of mean absolute percentage error (MAPE) and symmetric mean absolute percentage error (SMAPE), compared to the traditional Holt's method. The Holt's with seasonality and events yields the best MAPE of 8.64% and SMAPE of 8.90%.

Keywords : Data mining, Event component, Holt's method, K-mean clustering, Location problem

Acknowledgements

I extend my heartfelt thanks to all the individuals who have been instrumental in the completion of this thesis. My profound gratitude goes to my advisor, Associate Professor Dr. Chartchai Leenawong, for his unwavering support, guidance, and invaluable insights throughout my research journey. His expertise and encouragement have been pivotal in bringing this thesis to its final form.

I would also like to express my gratitude to Assistant Professor Dr. Sukrawan Mavecha, Associate Professor Dr. Pattrawut Chansangiam, and Assistant Professor Dr. Thawatchai Khumprapussorn, who have provided constructive criticism and invaluable guidance. I am also thankful to Associate Professor Dr. Nisakorn Sangwaranatee, who served as an external expert and offered insightful suggestions. I extend my appreciation to Pornpol Laempetch for his support in R-programming.

I am grateful to the Science Achievement Scholarship of Thailand for providing financial support.

Lastly, I would like to acknowledge my family for their unwavering support and encouragement throughout this journey.

I am grateful to all of these individuals for their invaluable contributions to this work and for playing a critical role in its successful completion.

Thanrada Chaikajonwat

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Inception and Importance

Currently, large quantities of data are collected daily. Data mining is the process of discovering interesting patterns and knowledge from data. It can be applied to a variety of applications such as the retail and telecommunications industries, science and engineering, intrusion detection and prevention, recommender systems, and financial data analysis. Data clustering has been extensively studied in the field of data mining. Additionally, data mining is at the heart of business intelligence (Han, Kamber, and Pei, 2012).

In Thailand, convenience stores are prevalent (Wang, 2017), especially in tourist and highly populated areas. The Eastern region of Thailand is a popular tourist destination, known for its proximity to the Gulf of Thailand. Some popular destinations in this region include Pattaya, Koh Lan, Koh Samet, and Koh Kut. As a result, it is common for convenience store franchises to open branches in these areas. Logistics management is crucial for these franchises to remain competitive in the long term. One logistical decision that must be made is determining the location of distribution centers (DCs) to distribute products to franchised convenience stores in Eastern Thailand (Netherlands embassy in Bangkok, 2017 and Bank of Thailand, 2021). This study aims to investigate the demand-based location problem in data mining using a case study of DCs in Eastern Thailand.

In addition to studying locational decisions on the upstream side, it is also interesting to examine the impact of special events on customer demand data on the downstream side. In early 2020, the COVID-19 pandemic affected many countries around the world, including Thailand (World Health Organization, 2020). The Thai government implemented a lockdown strategy to control the spread of the virus, which had a significant impact on the economy. Many people lost their jobs or had reduced salaries, leading to a decrease in unnecessary consumption (The World Bank, 2021). Car purchases were among the items on the unnecessary buying list, as cars are depreciated assets. Additionally, the nationwide lockdown significantly impacted

businesses, including the automotive industry. Therefore, this study also investigates the forecasting of car sales data during the COVID-19 pandemic.

In the following section, the objectives of the study will be outlined.

1.2 Objectives of the Study

The aim of this study is to apply data mining to business data management for the purposes of setting up a distribution center based on customer demands and Holt's forecasting with events. Therefore, this study aims to address the following research objectives related to the two issues mentioned above.

1) To apply data mining to the location problem and sales forecasting.

2) To apply and modify K-means clustering analysis from data mining to solve the location problem based on weighted demands in a case study of Thailand's convenience store franchises.

3) To compare the efficiency and effectiveness of different distance metrics combined with the modified centroid calculation used in the clustering for the location problem.

4) To apply and modify the typical Holt's forecasting method to better suit Thailand's car sales data containing the event component such as the COVID-19 global pandemic.

5) To compare the accuracies of the typical Holt's method and the modified Holt's methods in sales forecasting with events.

In the subsequent section, the scope of the study will be outlined, beginning with the location problem based on weighted demands, followed by Holt's forecasting with events.

1.3 Scope of the Study

The scope of this study is divided into two parts: the demand-based location problem and Holt's forecasting with events.

For the demand-based location problem, K-means clustering analysis is adapted to find suitable locations. Three different distance metrics (Euclidean, Manhattan, and Chebyshev) and two centroid location calculations (typical average and proposed demand-weighted average) are employed. This part is examined using a case study of locating eight DCs to distribute to 260 convenience stores in Eastern Thailand. All experiments in this part are coded in R-programming.

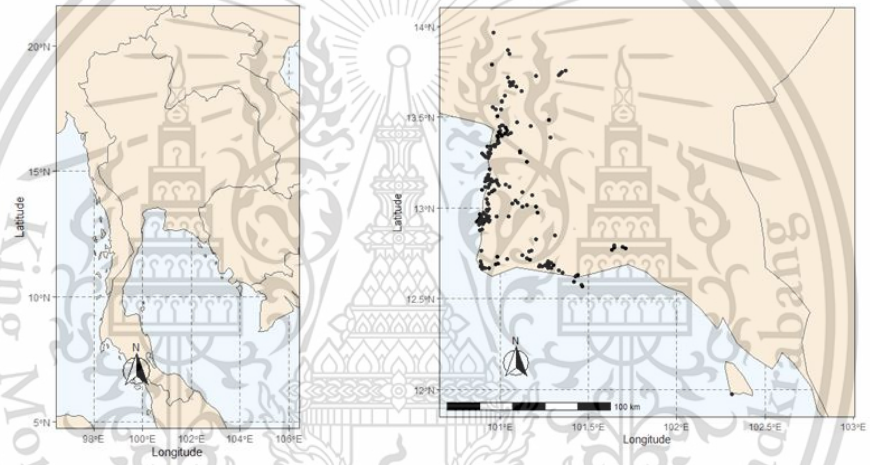

Figure 1.1 Thailand Map and the Convenience Store Locations in Eastern Thailand

For the Holt's forecasting with events problem, the typical Holt's method and modified Holt's methods with seasonality, events, and both seasonality and events are used. This part examines a case study of Thailand's monthly car sales from January 2015 to December 2021 using data from the Office of Industrial Economics, Ministry of Industry, Thailand (The Office of Industrial Economics, 2021). All experiments in this part are conducted on Microsoft Excel 365.

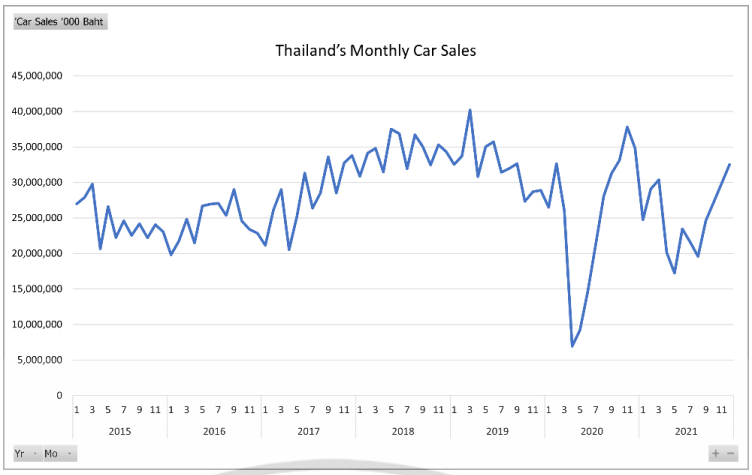

Figure 1.2 Thailand's Monthly Car Sales from January 2015 to December 2021

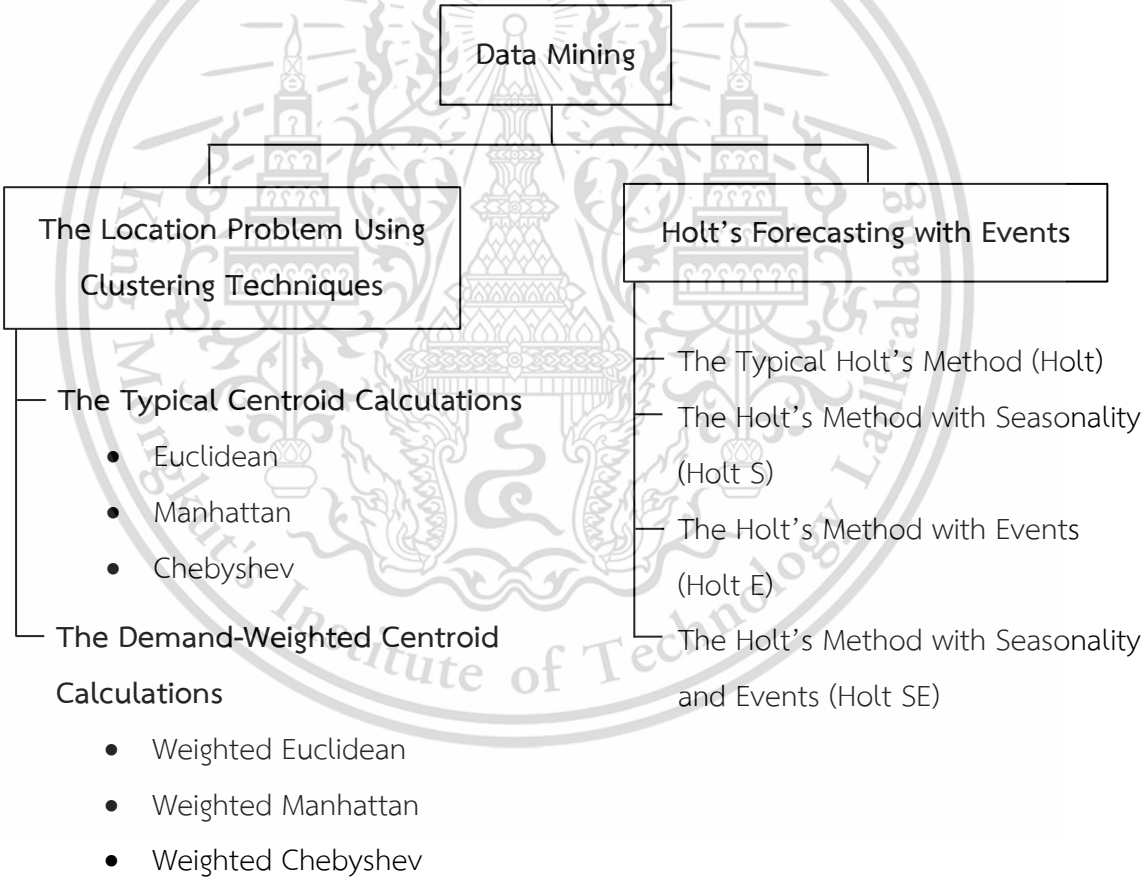

Figure 1.3 Scope of the Study

In the following section, the benefits of the study will be summarized.

1.4 Benefits of the Study

In this section, the benefits of the study upon completion are obtained as follows:

1) Applications of data mining to the location problem based on weighted demands and Holt's forecasting with special events.
2) Suitable locations for DCs with high efficiency and effectiveness values.
3) More appropriate forecasting methods for sales data having events.
4) New techniques that can be applied to other real-world instances or problems.

1.5 Process of the Study

Presented below is the process of this study:

1) Study data mining for the location problem and sales forecasting.
2) Study the K-means clustering algorithm, distance metrics, the average centroid location calculations and the effectiveness and efficiency measurement.
3) Modify the K-means algorithm, distance metric, and centroid calculation method for the demand-based location problem.
4) Use R-programming to determine suitable locations for eight DCs to distribute to 260 convenience stores.
5) Study Thailand's monthly car sales data during the COVID-19 pandemic.
6) Study the typical Holt's method, the moving averages, the centered moving averages and accuracy measurement such as MAPE and SMAPE.
7) Modify the Holt's method to forecast car sales data.
8) Use Microsoft Excel to forecast monthly car sales data with events.
9) Examine the research for potential improvements.
10) Summarize the study and write the thesis.

Table 1.1 The Research Schedule

Activity	The Time Frame (Month of Year)					
	2021		2022			
	8-9	10-12	1-4	5-6	7-10	11-12
Step 1	✓					
Step 2		✓				
Step 3			✓			
Step 4			✓			
Step 5				✓		
Step 6				✓		
Step 7					✓	
Step 8					✓	
Step 9						✓
Step 10						✓

The remainder of this thesis is structured as follows. In Chapter 2, the fundamental background and previous work on the topic of data mining will be described. Chapter 3 will explain the methods used in this research, including the clustering algorithm and forecasting method. Chapter 4 presents and interprets the findings of the study. Lastly, Chapter 5 provides conclusions and suggestions for future research.

Chapter 2
Fundamentals Background and Literature Reviews

In this chapter, the fundamentals background and previous work on data mining for the location problem based on weighted demands and forecasting methods with events are reviewed and summarized.

2.1 Fundamentals Background

In this section, the comprehensive fundamentals background in this study is divided into two distinct subsections. The first subsection thoroughly discusses the fundamentals of the location problem. The second subsection, on the other hand, in-depth examines the fundamental background of Holt's forecasting with events.

2.1.1 Fundamental Background of the Modified K-Means Clustering for Demand-Weighted Locations

In this subsection, the fundamental background of the location problem will be outlined, beginning with logistics management, followed by the facility location problem, distribution center, data clustering, the typical K-means clustering algorithm, the distance metrics and the typical average for centroid location calculations, respectively.

Logistics Management

The science of logistics management has been around since World War II. In the 1930, logistics management originated in the military and typically involved forward logistics, such as the transportation of supplies, food, clothing, medicines, weapons, and other items from storage to front-line operations. After World War II, it took some time for countries to recover, but by the 1960s, the science of logistics management had fully entered the industrial business sector. Logistics is part of the supply chain that manages the flow of goods and services from upstream to downstream in the supply chain to ensure that businesses run efficiently and effectively. (Leenawong, 2022b).

Logistics can be classified into four types as follows:

1. Military logistics, which originated as a form of military transport during World War II.

2. Business logistics, which deals with the transportation of consumer goods and is an important part of modern logistics management.

3. Service logistics, which refers to specialized logistics services in industries such as hotels and tourism.

4. Event logistics involves organizing various types of events, such as motor shows, exhibitions, banquets, and weddings. This requires the input of resources and human resources, which must be devoted to completing the event in a limited time and then withdrawn after the event ends. Additionally, event logistics also includes the logistics of responding to accidents and disasters, such as dispatching aid to areas affected by floods, tsunamis, the COVID-19 pandemic, or other catastrophic events in a timely manner.

Facility Location Problem

The Facility Location Problem (FLP) involves determining the number, size, and location of facilities, as well as allocating services from these facilities to customers both within and outside the organization. The goal of FLP is to minimum transportation costs and delivery times for goods or services. (Farahani and Hekmatfar, 2009)

Some typical examples of facility location problems include determining the location of:

1. A new warehouse relative to production facilities and customers.
2. A new classroom building on a college campus.
3. A fire station, hospital, or library in a city area.
4. A component in an electrical network.

Distribution Center

A distribution center (DC) is a warehouse that serves both as a warehouse and a link between manufacturers and retailers. It is a logistics provider for managing the storage and transportation of goods. Customers can have their needs met timely and accurately through a DC. Most DCs are outsourced or are operated by third-party

logistics service providers (3PLs), who receive goods from manufacturers to store in their warehouses. In the quantity control management of distribution and transportation technology, the owner or manufacturer of the goods is responsible for transportation. This helps to reduce the cost of transporting goods from the manufacturer to the retailer or customer, as the manufacturer only needs to transport goods to one DC. From there, the DC distributes the goods to retailers. Retailers do not need to store large amounts of inventory, which reduces their inventory cost and overall cost. Currently, many retailers are competing based on price and service speed. Many retail stores can offer competitive prices to consumers. (Aemod, 2018)

Data Clustering

Data clustering is a popular technique used to classify data. As data sizes have increased, manual classification has become difficult and costly, making automatic classification an essential part of data mining. In data clustering, similar samples are placed in the same group, known as a cluster, while dissimilar samples are placed in different clusters. Clustering has many applications in data mining, including pattern classification, pattern recognition, network analysis, information retrieval, image segmentation, and more. There are various clustering methods, including partitioning methods that use a distance-based metric to cluster points based on their similarity. These algorithms produce single-level partitions and non-overlapping, spherical-shaped clusters. K-means and K-medoids are popular partitioning algorithms, and K-means will be discussed in the next chapter. (Aggarwal and Reddy, 2013)

The Typical K-Means Clustering Algorithm

K-means clustering is the most commonly used clustering algorithm and one of the most efficient partitioning clustering algorithms. The general steps of the K-means clustering algorithm are explained step by step as follows: (Han, Kamber, and Pei, 2012)

Step 1: Choose the number K of representative clusters to be recognized in the dataset.

Step 2: Randomly select K representative points as the initial centers or “centroids” of the K clusters.

Step 3: Calculate the distance between each point and all the centroids from step 2, then determine the closest centroid and assign that point to the cluster.

Step 4: Once all the clusters are formed, update the centroids of all clusters.

Step 5: Repeat steps 3 to 4 until all the points in each cluster do not change. The algorithm stops and the last set of centroids are used as the chosen locations.

The Distance Metrics

The following notations are used in this subsection:

K refers to the number of clusters/centroids/DCs.

N refers to the total number of convenience stores.

$X_i = (x_i,y_i)$ 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.

1. Euclidean Distance

Euclidean distance is computed by taking the square root of the sum of the squares of the differences between the coordinates of a pair of objects. For a fixed $j = 1,2,\ldots,N$, (Singh et al., 2013)

$$Dist_{Eucl}(S_j,X_i) = \sqrt{\left(r_j - x_i\right)^2 + \left(s_j - y_i\right)^2}\ ;\ i = 1,2,\ldots,K. \quad (2.1)$$

2. Manhattan Distance

Manhattan distance is computed by summing the absolute differences between the coordinates of a pair of objects. For a fixed $j = 1,2,\ldots,N$, (Singh et al., 2013)

$$Dist_{Manh}(S_j,X_i) = |r_j-x_i|+|s_j-y_i|;\ i = 1,2,\ldots,K. \quad (2.2)$$

3. Chebyshev Distance

Chebyshev distance is the maximum value distance that is computed as the absolute magnitude of the differences between the coordinates of a pair of objects. For a fixed $j = 1,2,\ldots,N$, (Singh et al., 2013)

$$Dist_{Cheb}(S_j,X_i) = \max(|r_j\text{-}x_i|,|s_j\text{-}y_i|);\ i = 1,2,\ldots,K. \quad (2.3)$$

The Typical Average for Centroid Location Calculations

The calculation for the new location of each centroid can be performed as follows.

$$X_i = \left(\frac{\sum_{j=1}^{n_i} r_j^i}{n_i}, \frac{\sum_{j=1}^{n_i} s_j^i}{n_i} \right);\ i = 1,2,\ldots,K \quad (2.4)$$

where $S^i_j = (r^i_j,s^i_j)$ refers to the location of store j that is assigned to cluster i, and n_i refers to the number of the convenience stores in cluster i.

In the next subsection, the fundamental background of Holt's forecasting with events will be explained.

2.1.2 Fundamental Background of Holt's Forecasting with Events

In this subsection, time-series data and the typical Holt's method will be presented.

Time-Series

One of the first and most important problems in supply chain and logistics management is business forecasting, which can be achieved with the use of data analysis tools like Excel. The goal of business forecasting is to use a variety of tools and techniques to accurately forecast sales information and apply the results to implementation planning. (Leenawong, 2022a)

Sales data used for forecasting is time-series data, which refers to quantitative data recorded over time. Examples include monthly domestic sales data, quarterly export data, annual budget information, etc.

Time series are often used in forecasting and may include some or all of the following components: trend, seasonal, cyclical and irregular. (Wilson, Keating, and John Galt Solutions, Inc., 2009)

In time series, the trend component represents the long-term change in the level of the data. If the series moves upward, it indicates a positive trend, as shown in Figure 2.1. Conversely, if the series moves downward, it indicates a negative trend, as demonstrated in Figure 2.2. However, if there is neither a positive nor negative trend, the data are considered stationary, which means the series is essentially flat in the long term, as depicted in Figure 2.3.

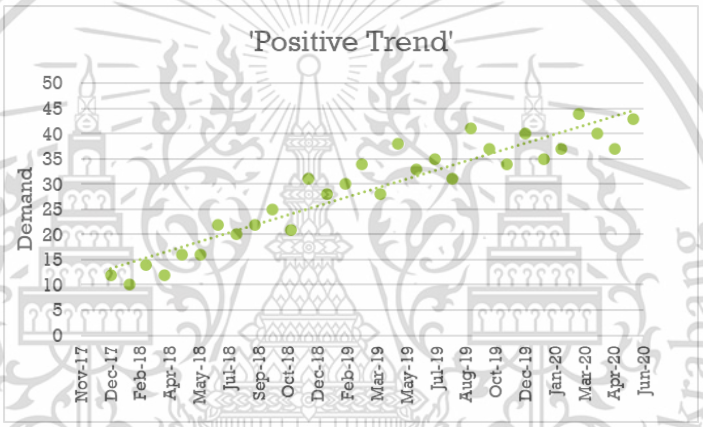

Figure 2.1 Positive Trend

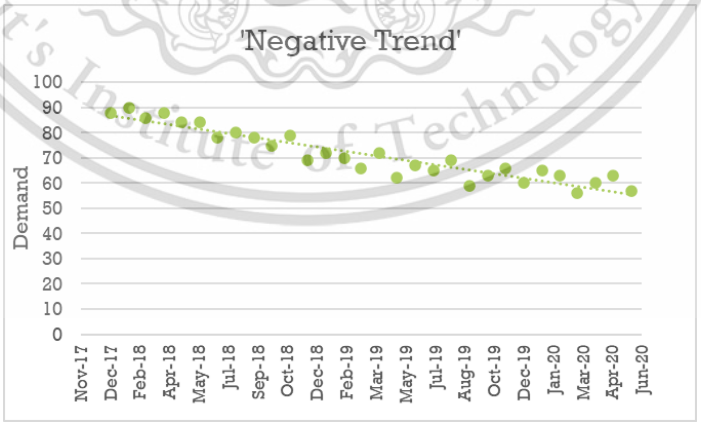

Figure 2.2 Negative Trend

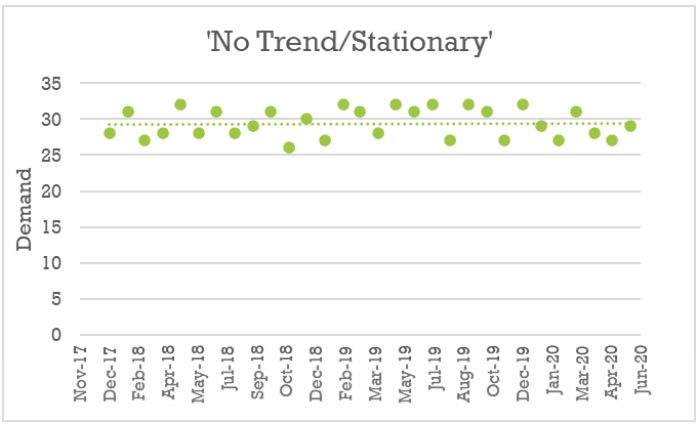

Figure 2.3 No Trend/Stationary

The seasonality component refers to a recurring pattern of systematic information fluctuation, persisting in duration and often manifesting consistently across subsequent time periods, as illustrated in Figure 2.4.

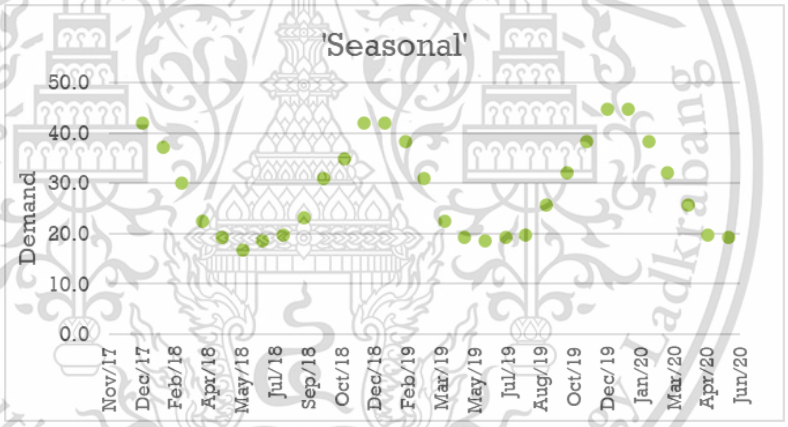

Figure 2.4 Monthly Data with Consistent Seasonality Each Year

The cyclical component is a mercurial long-term fluctuation of data, evocative of the unpredictable seasons. Its timing and trajectory elude forecast, leaving uncertain the duration and magnitude of its crests and troughs. Such variability is depicted in Figure 2.5.

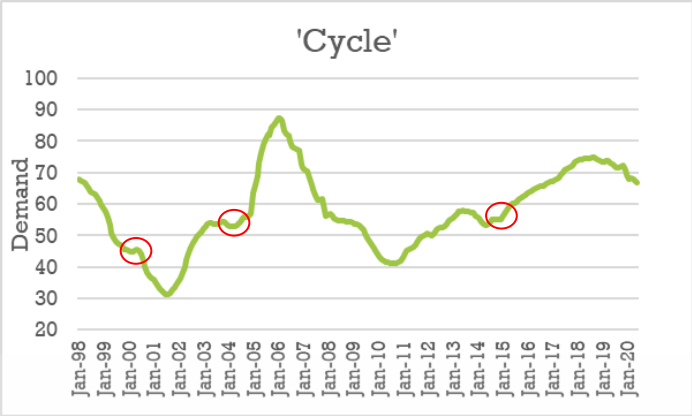

Figure 2.5 Irregular Long-Term Cycle Data

Finally, the irregular component of time series movements in data pertains to short-term fluctuations that resist pattern recognition and defy predictability. Although commonplace, these sporadic and unpredictable spikes in the data graph often deviate from the expected ups and downs. Typically, they are smoothed out or excluded altogether during the forecasting process. As the circle is in Figure 2.5.

The Typical Holt's Method

The common notations used in this typical Holt's method are defined as follows:

A_t represents the actual data for period t,

L_t represents the level estimate for period t,

T_t represents the trend estimate for period t,

α represents the level smoothing constant; $0 \leq \alpha \leq 1$,

and β represents the trend smoothing constant; $0 \leq \beta \leq 1$.

The typical Holt's forecasting method can be executed through the following three steps (Leenawong, 2022b).

Step 1: Computing the Level Estimate.

$$L_t = \alpha A_t + (1 - \alpha)\,(L_{t-1} + T_{t-1}) \tag{2.5}$$

and the initial value for L_t is $L_2 = A_2$.

Step 2: Computing the Trend Estimate.

$$T_t = \beta\,(L_t - L_{t-1}) + (1 - \beta)\,T_{t-1} \tag{2.6}$$

and the initial value for T_t is $T_2 = A_2 - A_1$.

Step 3: Computing the Holt's Estimate.

$$H_{t+m} = L_t + mT_t \tag{2.7}$$

where H_{t+m} refers to Holt's forecasted value for period $t + m$,
and m refers to the future period m^{th}; $m \geq 1$.

In the next section, previous work regarding the demand-based location problem and Holt's forecasting with events will be presented.

2.2 Literature Reviews

In this section, literature reviews of the location problem and Holt's forecasting method will be presented.

2.2.1 The Modified K-Means Clustering for Demand-Weighted Locations

The location problem is first reviewed. In addition, given the emphasis on applying K-means clustering to the location problem, previous work on the K-means clustering algorithm is also reviewed. Both topics are studied and presented as follows.

The location problem, which was first conceptualized by Alfred Weber in 1909 as he sought to minimize the total distance between a single warehouse and several customers, has since been advanced through a number of other applications. For example, Hakimi (1964) aimed to locate switching centers in a communication network and police stations in a highway system.

In addition to the researchers previously mentioned, there are also numerous others who have delved into the location problems, as exemplified by the research described below.

Drezner et al. (2003) study the best location of a central warehouse to determine the number and the locations of local warehouses. An example problem contains six local warehouses. They perform three types of experiments. They solve a small illustrative example problem using Excel, perform a sensitivity analysis on the

parameters of this problem, and solve large randomly generated problems using the generalized Weiszfeld algorithm. The sensitivity of the solutions is investigated by the four models to the parameters of the example problem. Those models are proportional backorder cost rate, given backorder cost rate, the maximin model and the Weber model. The models are demonstrated on an example problem with up to 10,000 demand points. Each model is then solved by Excel Solver in less than half a second. It turns out that the location solutions for the four models are quite different from one another. Hence, the decision maker needed to decide which model is the most suitable one for the situation at hand. In addition, mathematical results show that disregarding inventory costs made the models less exact.

Yang et al. (2007) investigate the logistics location problem under fuzzy environment from another point of view, in which setup cost, turnover cost and demand of each customer are supposed to be fuzzy variables. The aim is to minimize the total relevant cost comprising setup cost, turnover cost and transportation cost, should be minimized. Decision makers need to complete the two tasks as follows: (i) choose the sites of the six distribution centers from the ten potential distribution centers to serve seven customers, (ii) determine the number of products transported from the manufactory to each selected distribution center and also from the selected distribution centers to each customer. Then, they develop the mathematical model to solve the problem. Tabu search algorithm, genetic algorithm and fuzzy simulation algorithm are integrated to seek the approximate best transportation and assignment plan of the distribution centers.

Dantrakul et al. (2014) study facility location problem to minimize the sum of the setup and transportation costs by using greedy, p-median and p-center algorithms. Those two costs are considered a function of the number of opened facilities. All presented methods are demonstrated and examined on the networks representing the road transportation system of six provinces in Northern Thailand. The performances of the presented methods are tested using 100 random data sets. The experimental results show that the developed greedy algorithm is proper for solving the problem when the setup cost is greater than the transportation cost. In contrast, the p-median-based methods are more efficient for the opposite case when the setup cost is lower.

Sharma and Jalal (2017) develop a model to utilize the facility by maximizing the number of customers to result in maximized profit. The proposed approach consists of two parts. In the first part, K-means clustering is used and for each cluster, mixed integer linear programming (MILP) is implemented, in the second part. Numerical examples for clustering and without clustering is considered. The numerical results show that the profit starts to decrease as the number of clusters increased. If the profit kept decreasing, it indicated that the solution process will stop.

Chen (2019) analyzes the location problem of distribution centers in Jiaji Logistics by using the Baumer Walvar model. The aim of this work is to optimize the total distribution center costs, including four cost components, namely, the transportation cost from the factories to distribution centers, and from distribution centers to the customers, the distribution centers' fixed costs, and the distribution centers' change fee. The entire cargo of Jiaji logistics is transported from five factories (Chongqing, Chengdu, Xi'an, Zhengzhou, and Lanzhou) to four customers (Guangzhou, Shanghai, Hangzhou, and Tianjin). The company want to select the optimal five distribution centers out of the predetermined eight distribution centers (Wuhan, Nanchang, Guiyang, Changsha, Shijiazhuang, Beijing and Nanjing). The economies of scale were also taken into account. The results show that the minimum cost is 7,301,620 yuan and the optimal locations of distribution centers are Nanchang, Nanjing, Guiyang, Changsha and Shijiazhuang.

Regarding previous research on K-means clustering algorithms, distance metrics, and performance measurement, the following works are of particular interest.

Singh et al. (2013) study and compare distortion in Minkowski K-means for different values of P. All the experiments are accomplished on dummy data. The results acquired by K-means based on Minkowski distance metrics at P=1 are same as results using Manhattan distance metric because formula for Manhattan distance metric is derived by taking P=1. Similarly, the results at P=2 are same as results using Euclidian distance metric because formula for Euclidean distance metric is derived by taking P=2 in Minkowski distance metric formula. As a conclusion, the experimental

results show that Euclidean distance provides the best performance while Manhattan distance yields the worst.

Sinwar and Kaushik (2014) focus on the study of two popular distance metrics, specifically, Euclidean and Manhattan, on the simple K-means clustering. The datasets used during the overall process of comparative are two real and one synthetic datasets, namely, Iris, Diabetes, and BIRCH respectively. Besides, the numbers of clusters used in this work are 2, 3, 4, 5, 6 and 7. They have used WEKA 3.7.10 as their development tool for clustering of data. The theoretical analysis and experimental results show 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) use 2 methods, K-means and K-medoid methods together with Euclidean, Canberra, and Chebyshev distance algorithm to analyze and compare object clustering from real big data. The sample dataset includes six variables collected from three college classes having 147,679 students at Medan State University. The Davies-Bouldin index is used to compare the performance measurement. In consideration of both the number of clusters produced and the time required, K-means method yields better results than K-medoid method, both in terms of the Euclidean, Canberra, and Chebyshev distance algorithms which the ratio 1:110.7. Nevertheless, The Canberra distance algorithm in both K-means and K-medoid methods obtains very large and undefined for the Davies-Bouldin index. Therefore, the Canberra distance is not suggested for the big data clustering. While the Chebyshev distance in K-means yields better results than that in K-medoid in terms of both accuracy and quality of cloning which is to produce five clusters with a 0.1 second processing time.

In the following subsection, a review of literature on forecasting methods and the incorporation of event components will be presented.

2.2.2 Holt's Forecasting with Events

Lastly, previous research on forecasting methods and the incorporation of event components will be presented below.

Wirotcheewan et al. (2011) study an appropriate forecasting model for the advanced demand of the automotive wheels, parts, and accessories (WPA). They use a linear programming (LP) model to calculate the optimal quantity for export so as to obtain from the maximum profit. The WPA demands for export from Thailand are collected from the top five highest-demand countries, i.e., Japan, China, South Korea, Germany, and Indonesia during the year 1997 to 2008. Time-series forecasting models used in this study are naïve, moving average, single exponential smoothing, and exponential smoothing with trend or Holt's method and artificial neural networks (ANNs). The forecasting accuracy measurement is the mean absolute percentage error (MAPE). The results show that the ANNs model outperforms the other models for Japan, Germany, and Indonesia, while exponential smoothing with trend is best suited for China. However, all five models yield MAPEs greater than 50% for South Korea. After that, the forecasted results are used in finding the optimal quantities for export to those five countries using LP models.

Rattanametawee et al. (2016) propose a method to dealing with multiple linear regression that integrates the seasonality and the effects of special events for subcompact car sales in Thailand. The monthly input data are collected from January 2005 to September 2015, totaling to 129 data. The two special events considered are the 2011 Thailand flood and the government's tax-incentive first-car buyer program in 2011–2012. For the methodology, they use three different multiple linear regressions, specifically, the regular regression (Model 1), regression containing seasonality (Model 2), and the proposed regression containing seasonality and special events (Model 3). As a result, the model 1 that does not take into account the seasonal and the special event effects have a low adjusted R2 of just 42.84% and a high MAPE of 29.82%. The model 2 that integrates the seasonal effect has a higher adjusted R2 of 66.65% and a lower MAPE of 22.04%. The last regression (Model 3) that includes both seasons and special events is shown to be the most talented model with the highest adjusted R2 of 85.89% and the lowest MAPE of 15.80%. From the foregoing, it can be concluded

that the proposed regression model including seasonality and special events (Model 3) outperforms the other two models and achieves the highest adjusted coefficient of determination or R-square and also the highest accuracy in terms of MAPE.

Booranawong and Booranawong (2018) use double exponential smoothing or Holt's method, multiplicative Holt–Winters method and additive Holt–Winters method with the optimal initial values and smoothing constants to forecast lime, Thai chili and lemongrass prices in Thailand from October 2016 to December 2016. This research collects the input price data at the Simummuang market from January 2011 to September 2016. The accuracy measurement for comparisons is MAPE. The results show that Holt's method attains the smallest MAPEs for forecasting Thai chili and lemongrass prices, while multiplicative Holt–Winters method and additive Holt–Winters method produce smaller MAPEs than that of Holt's method for forecasting lime prices possessing the seasonal component.

Muchayan (2019) studies forecasting methods to predict the net asset value (NAV) price movements of the Cipta Ovo Equitas mutual fund. The methods are two different double exponential smoothing methods, namely, Brown's and Holt's. The dataset of NAV price is collected from the Ciptadana Asset Management, PT in Indonesia over the period January 2019 to January 2020. The measurement of effectiveness is MAPE. The results show that Holt's method yields a smaller MAPE than Brown's method.

Sharif and Hasan (2019) apply the Holt's method with different parameters to develop a stock indicator that helps investor predict the next day's share value. The daily stock closing prices or opening price of different companies are gathered from the Dhaka Stock Exchange (DSE) in year 2016. The experimental results notices that the investor can use the Holt's method for short term prediction. In addition, the different smoothing constants have an impact on the prediction values and the suitable values of smoothing constants for this dataset are $\alpha = 0.5$ and $\beta = 0.1$.

Suppalakpanya et al. (2019) apply several exponential smoothing methods for forecasting monthly crude palm oil productions in Thailand. The input data is collected from the database of the Department of Internal Trade, Ministry of Commerce, Thailand during January 2018 to March 2018. This work compares five different

forecasting methods for various ranges of the input data. The first three methods are the conventional method, namely, double exponential smoothing, the multiplicative Holts–Winters, and the additive Holt–Winters methods. In addition, the proposed modified methods are the improved additive Holts–Winters and the extended additive Holts–Winters methods. For the input ranges, they implement all the five methods on four different ranges, i.e., 3-year data (2015–2017), 6-year data (2012–2017), 9-year data (2009– 2017), and 12-year data (2006–2017). The MAPE is used as an accuracy measure. The experimental results show that the additive Holt–Winters and extended additive Holts–Winters methods yield the two lowest MAPEs of 6.94 and 7.05, respectively, when the 12-year data are applied.

Rattanametawee and Leenawong (2020) propose a new time-series decomposition to include the effects of special events, both positive and negative impact, on the dataset. In addition, the dataset has the conventional trend, seasonal, and cyclical components. A case study of subcompact monthly car sales data in Thailand from 2011 to 2018 is investigated as for that time period comprises the 2011 nationwide big flood reflecting, the negative impact, and the nation's tax-incentive first-car buyer scheme reflecting, the positive impact, on the dataset. An accuracy measure of the proposed forecasting method is MAPE. The experimental results illustrate that the proposed forecasting model is very promising results with a low MAPE of just 8.17%.

Based on the literature review, the researcher is interested in studying and experimenting with data mining for the location problem and sales forecasting using the K-means clustering algorithm and Holt's method, respectively. The methodology for these experiments will be explained in the next chapter.

Chapter 3
Research Methodology

This chapter presents two methods to data mining. The first involves an examination of the modified K-means clustering technique for identifying demand-weighted locations in a case study of a convenience store franchise in Thailand. The second approach involves analyzing the use of events for Holt's forecasting in a case study of Thailand's monthly car sales data during the COVID-19 pandemic.

3.1 The Modified K-Means Clustering for Demand-Weighted Locations

This section examines the modified K-means clustering approach for identifying demand-weighted locations, which is divided into two subsections: the typical and proposed K-means clustering algorithms, and the effectiveness and efficiency measurement. These topics will be described in detail in the following paragraphs.

3.1.1 The Typical and Proposed K-Means Clustering Algorithm

In the process of locating DCs for a convenience store franchise, points representing the stores and centroids representing the locations of the DCs serving the stores in the same clusters are being clustered. It is natural to also consider the varying demands at the served stores. In this specific case, these demands are being used as weights when calculating the updated centroids after the clusters are formed at each iteration.

The modified K-means clustering algorithm, which takes into account the varying demands of the stores, is applied and explained in the context of the application. Three distance metrics (Euclidean, Manhattan, and Chebyshev) are also experimented within the algorithm. By combining these metrics with both typical and demand-weighted centroid calculations, six different combinations are tested to determine the best algorithm. The notations used in this thesis are defined as follows.

K refers to the number of clusters/centroids/DCs. Here, $K = 8$.

N refers to the total number of convenience stores. Here, $N = 260$.

n_i refers to the number of the convenience stores in cluster i, $i = 1,2,\ldots,8$.

$X_i = (x_i,y_i)$ refers to the location of centroid i representing DC i, where x_i and y_i are the latitude and longitude of centroid i, $i = 1,2,\ldots,8$, 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, $j = 1,2,\ldots,260$, respectively.

$S^i_j = (r^i_j,s^i_j)$ refers to the location of store j that is assigned to cluster i, $i = 1,2,\ldots,8$.

w_j refers to the demand at convenience store j.

Next, the K-means algorithm using each of the three distance metrics (Euclidean, Manhattan, and Chebyshev), as well as the modified demand-weighted K-means algorithm using each of these metrics, is proceeded with in detail.

Step 1: Randomly select eight initial centroids X_i; where $i = 1, 2, \ldots, 8$, representing eight initial DCs.

Step 2: Calculate the distance between each centroid X_i and the fixed convenience store S_j; where $j = 1, 2, \ldots, 260$, using the formula below.

$$Dist_{Eucl}(S_j,X_i) = \sqrt{\left(r_j - x_i\right)^2 + \left(s_j - y_i\right)^2}\ ;\ i = 1,2,\ldots,8, \tag{3.1}$$

$$Dist_{Manh}(S_j,X_i) = |r_j\text{-}x_i| + |s_j\text{-}y_i|;\ i = 1,2, \ldots, 8, \tag{3.2}$$

$$Dist_{Cheb}(S_j,X_i) = \max(|r_j\text{-}x_i|,|s_j\text{-}y_i|);\ i = 1,2, \ldots, 8. \tag{3.3}$$

Then, select the centroid i that is closest to store j. Assign this store S_j to cluster X_i accordingly. Now, S_j becomes S^i_j, and is grouped in cluster i, which is served by centroid i or DC i. Repeat this process for all remaining stores.

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

$$X_i = \left(\frac{\sum_{j=1}^{n_i} r_j^i}{n_i}, \frac{\sum_{j=1}^{n_i} s_j^i}{n_i} \right) \quad \text{for } i = 1,2,\ldots,8. \tag{3.4}$$

In contrast, the proposed demand-weighted average takes into account the effect of the varying demands of each store when computing the new location of each centroid i, as follows.

$$X_i = \left(\frac{\sum_{j=1}^{n_i} w_j r_j^i}{\sum_{j=1}^{n_i} w_j}, \frac{\sum_{j=1}^{n_i} w_j s_j^i}{\sum_{j=1}^{n_i} w_j} \right) \quad \text{for } i = 1,2,\ldots,8. \tag{3.5}$$

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: To measure the effectiveness, calculate the total distribution cost from the DCs to their served stores and compute the Davies–Bouldin index (DBI). For efficiency measurement, determine the number of iterations to reach the final clusters. Details of these measures are given in the next subsection.

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 steps of the algorithms have been stated, the effectiveness and efficiency of the six clustering methods will be measured and compared. These issues will be explained in more detail in the following subsection.

3.1.2 The Effectiveness and Efficiency Measurement

In this subsection, the measures of effectiveness and efficiency are presented. To assess effectiveness, the distribution cost and Davies-Bouldin index (DBI) are used. To assess efficiency, the number of iterations required to reach the final clusters is considered.

Distribution Cost

The focus of this study is on minimizing the overall distribution cost when locating the DCs. This cost typically depends on the transportation rate, shipment weight, and traveling distance. For the purposes of this study, it is assumed that the transportation rate is $1 per kilometer per unit of shipment weight. In order to calculate the transportation distance between the store and its relevant DC, the Euclidean metric is used. Therefore, the following equation represents the distribution cost from DC X_i to store S_j.

$$Distribution\ cost = \$1 \times l_{ij} \times D_{Euclidean}(S_j, X_i) \tag{3.6}$$

where l_{ij} refers to the shipment load is the demand at each store S_j served by DC X_i.

Davies–Bouldin Index (DBI)

The Davies-Bouldin Index (DBI), developed by David L. Davies and Donald W. Bouldin in 1979, is a widely used metric for evaluating the performance of clustering algorithms. It is an internal evaluation method that assesses the quality of the clustering based on variables and features inherent to the dataset. The following outlines the process for calculating the DBI (Davies and Bouldin, 1979):

Step 1: Calculate the average distance between all stores in the cluster and DC.

Let A_i represents the average distance between all stores and DC in the cluster i.

Then A_i is computed by the following formula.

$$A_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left\|S_j^i - X_i\right\| = \frac{1}{n_i}\sum_{j=1}^{n_i}\sqrt{\left(r_j^i - x_i\right)^2 + \left(s_j^i - y_i\right)^2}\ ;\ i = 1,2,\ldots,K \tag{3.7}$$

where $Xi = (x_i, y_i)$ 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,

K refers to the number of clusters/centroids/DCs,

$S_j = (r_j, s_j)$ 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,

N refers to the total number of convenience stores,

n_i refers to the number of the convenience stores in cluster i.

Step 2: Calculate the distance between DCs.

Let $M_{h,i}$ represents the distance between DCs X_h and X_i, computed by the following formula.

$$M_{h,i} = \|X_h - X_i\| = \sqrt{(x_h - x_i)^2 + (y_h - y_i)^2}, \quad (3.8)$$

where $X_i = (x_i, y_i)$ refers to the location of centroid i representing DC i;

x_i and y_i are the latitude and longitude of centroid i, $i = 1,2,\ldots,K$, respectively,

$X_h = (x_h, y_h)$ refers to the location of centroid h representing DC h;

x_h and y_h are the latitude and longitude of centroid h, $h = 1,2,\ldots,K$, respectively.

Step 3: For each pair of DCs X_h and X_i, calculate

$$R_{h,i} = \frac{A_h + A_i}{M_{h,i}}, \quad (3.9)$$

where A_i refers to the average distance between all stores and DC in the cluster i,

A_h refers to the average distance between all stores and DC in the cluster h.

Then, identify

$$D_i = \max_{h \neq i} R_{h,i}. \quad (3.10)$$

Step 4: Finally, calculate DBI using the following formula.

$$DBI = \frac{1}{K} \sum_{i=1}^{K} D_i \quad (3.11)$$

The Expected Number of Iterations to the Final Clusters

To measure efficiency, the number of iterations required to reach the final clusters for each instance is counted. These final clusters are defined as those in which the stores served by the DCs remain unchanged from the previous iteration. The expected number of iterations is then calculated by averaging these values across all instances.

3.2 Holt's Forecasting with Events

This section presents the methodology of sales forecasting methods, which includes the standard Holt's method as well as three modified versions: the Holt's method with seasonality, the Holt's method with events, and the Holt's method with both seasonality and events. Additionally, the accuracy measurements for these methods are described.

3.2.1 The Typical Holt's Method

The common notations to be used in this typical Holt's method are defined as follows.

A_t represents the actual data for period t,

α represents the level smoothing constant; $0 \leq \alpha \leq 1$,

and β represents the trend smoothing constant; $0 \leq \beta \leq 1$.

The typical Holt's forecasting method can be executed by the following three steps.

Step 1: Computing the Level Estimate.

The initial value for the level estimate (L_t) is $L_2 = A_2$ and the level estimates are updated using the following formula:

$$L_t = \alpha A_t + (1 - \alpha)\,(L_{t-1} + T_{t-1}) \tag{3.12}$$

where L_t refers to the level estimate for period t,

T_t refers to the trend estimate for period t.

Step 2: Computing the Trend Estimate.

The initial value for the trend estimate (T_t) is $T_2 = A_2 - A_1$ and the trend estimates are updated using the following formula:

$$T_t = \beta\,(L_t - L_{t-1}) + (1 - \beta)\,T_{t-1}. \tag{3.13}$$

Step 3: Computing the Holt's Estimate.

$$H_{t+m} = L_t + mT_t \tag{3.14}$$

where H_{t+m} refers to Holt's forecasted value for period $t + m$,

and m refers to the future period m^{th}; $m \geq 1$.

In the next subsection, the forecasting steps of the Holt's method with seasonality will be described.

3.2.2 The Holt's Method with Seasonality

In this subsection, the method has been modified from the typical Holt's method to take into consideration the seasonality in the data. Some steps have been changed and seven additional steps have been inserted to account for the seasonal component, bringing the total number of steps to ten. Additionally, the notations used in some steps of the typical Holt's method have been slightly adjusted to reflect the seasonality in the data. In conclusion, the Holt's method with seasonality can be executed by the following.

Step 1: Finding the Moving Averages (MAs).

Let $A_{y,m}$ represents the actual data of year y, month m, where y = 1, 2, ..., 7 refer to the years 2015, 2012, ..., 2021 and m = 1, 2, ..., 12 refer to the months January, February, ..., and December, respectively, and $\overline{A}_{y,m}$ represents the moving average of year y, month m, when the number of periods to average is 12.

Therefore, the values start at $\overline{A}_{1,6}$ go up until $\overline{A}_{7,6}$. More explicitly,

$$\overline{A}_{1,6} = \frac{A_{1,1} + A_{1,2} + \ldots + A_{1,12}}{12},$$

$$\overline{A}_{1,7} = \frac{A_{1,2} + A_{1,3} + \ldots + A_{1,12} + A_{2,1}}{12},$$

$$\overline{A}_{1,8} = \frac{A_{1,3} + A_{1,4} + \ldots + A_{1,12} + A_{2,1} + A_{2,2}}{12},$$

$$\vdots$$

$$\overline{A}_{7,6} = \frac{A_{7,1} + A_{7,2} + \ldots + A_{7,11} + A_{7,12}}{12}.$$

Step 2: Finding the Centered Moving Averages (CMAs).

Let $\overline{C}_{y,m}$ represent the centered moving average (CMAs) in year y, month m, when the number of periods to average is 2.

Therefore, the values start at $\overline{C}_{1,7}$ go up until $\overline{C}_{7,6}$ and can be computed as follows.

$$\overline{C}_{1,7} = \frac{\overline{A}_{1,6} + \overline{A}_{1,7}}{2},$$

$$\overline{C}_{1,8} = \frac{\overline{A}_{1,7} + \overline{A}_{1,8}}{2},$$

$$\overline{C}_{1,9} = \frac{\overline{A}_{1,8} + \overline{A}_{1,9}}{2},$$

$$\vdots$$

$$\overline{C}_{7,6} = \frac{\overline{A}_{7,5} + \overline{A}_{7,6}}{2}.$$

Step 3: Computing the Seasonal Factors.

Let $SF_{y,m}$ represent the seasonal factor in year y, month m, computed by the following formula,

$$SF_{y,m} = \frac{A_{y,m}}{\overline{C}_{y,m}}. \tag{3.15}$$

Step 4: Computing the Unscaled Seasonal Indices.

Let SI_m represent the unscaled seasonal index of month m, that is,

SI_1 = the average of $SF_{2,1}$, $SF_{3,1}$, ..., $SF_{7,1}$,

SI_2 = the average of $SF_{2,2}$, $SF_{3,2}$, ..., $SF_{7,2}$,

SI_3 = the average of $SF_{2,3}$, $SF_{3,3}$, ..., $SF_{7,3}$,

$\vdots$

SI_{11} = the average of $SF_{1,11}$, $SF_{2,11}$, ..., $SF_{6,11}$,

SI_{12} = the average of $SF_{1,12}$, $SF_{2,12}$, ..., $SF_{6,12}$.

Step 5: Computing the Scaled Seasonal Indices.

Let S_t represent the scaled seasonal indices for period t when $t = 1, 2, \ldots, 84$. Since the number of periods in the seasonality cycle is 12,

$$S_t = S_{t+12(1)} = S_{t+12(2)} = \ldots,$$

more explicitly,

$$S_1 = S_{13} = \ldots = S_{61} = S_{73} = \frac{SI_1 \times 12}{\sum_{m=1}^{12} SI_m},$$

$$S_2 = S_{14} = \ldots = S_{62} = S_{74} = \frac{SI_2 \times 12}{\sum_{m=1}^{12} SI_m},$$

$$\vdots$$

$$S_{11} = S_{23} = \ldots = S_{71} = S_{83} = \frac{SI_{11} \times 12}{\sum_{m=1}^{12} SI_m},$$

$$S_{12} = S_{24} = \ldots = S_{72} = S_{84} = \frac{SI_{12} \times 12}{\sum_{m=1}^{12} SI_m}.$$

Step 6: Removing Seasonality from the Data (De-seasonalization).

In this step, to remove seasonality from the data, each actual data point is divided by its corresponding seasonal index. From this step onward, the year y and month m become irrelevant, and $A_{y,m}$ can be represented simply as A_t for $t = 1, 2, \ldots, 84$.

Let D_t represents the de-seasonalized data at period t, computed from the actual data divided by the relative seasonal index, that is,

$$D_t = \frac{A_t}{S_t}. \tag{3.16}$$

Step 7: Computing the Level Estimate.

Since step 6, the data has been removed of seasonality, the level estimates are computed similarly to Step 1 of the typical Holt's method with the replacement of A_t by D_t.

The initial value for the level estimate (L_t) is $L_2 = D_2$ and the level estimates are updated using the following formula:

$$L_t = \alpha D_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \tag{3.17}$$

where L_t refers to the level estimate for period t,

T_t refers to the trend estimate for period t.

Step 8: Computing the Trend Estimate.

Similar to Step 2 of the typical Holt's method with the replacement of A_t by D_t, the initial value for the trend estimate (T_t) is $T_2 = D_2 - D_1$, and the trend estimates are updated using the following formula.

$$T_t = \beta\,(L_t - L_{t-1}) + (1 - \beta)\,T_{t-1}. \tag{3.13}$$

Step 9: Computing the Holt's Estimate.

$$H_{t+m} = L_t + mT_t \tag{3.14}$$

where H_{t+m} refers to Holt's forecasted value for period $t + m$,

and m refers to the future period m^{th}; $m \geq 1$.

Step 10: Computing the Holt's Estimate with Seasonality (Re-seasonalization).

In this final step, the Holts' estimate from step 9 are multiplied by their corresponding seasonal indices. The result, represented by F_{t+m}, is the final forecasted data for period $t + m$, fully re-seasonalized from our Holt's estimation.

$$F_{t+m} = H_{t+m} \times S_{t+m}. \tag{3.18}$$

The following section provides an in-depth analysis of the forecasting steps of the modified Holt's method, that considers event component. This approach provides a more accurate and comprehensive prediction of future trends.

3.2.3 The Holt's Method with Events

The proposed Holt's method, which incorporates the impact of global events such as the COVID-19 pandemic, builds upon the traditional method by modifying certain steps and adding one additional step for a total of four. Notations used in previous steps are also adjusted to account for the effects of the event on the data. As a result, this enhanced method provides a more accurate and comprehensive outlook on future trends.

To accurately account for the effects of global events such as the COVID-19 pandemic, an additional smoothing constant, δ, is introduced in Holt's method, with a value between 0 and 1. Our focus will be on highlighting the modifications made to the traditional forecasting steps and the one additional step added. By following these four steps, the enhanced Holt's method with events can be executed.

Step 1: Computing the Level Estimate.

The initial value for the level estimate (L_t) is $L_2 = A_2$ and the level estimates are updated using the following formula:

$$L_t = \alpha A_t + (1 - \alpha)\,(L_{t-1} + T_{t-1}) \tag{3.12}$$

where L_t refers to the level estimate for period t,

T_t refers to the trend estimate for period t.

Step 2: Computing the Trend Estimate.

The initial value for the trend estimate (T_t) is $T_2 = A_2 - A_1$ and the trend estimates are updated using the following formula:

$$T_t = \beta\,(L_t - L_{t-1}) + (1 - \beta)\,T_{t-1}. \tag{3.13}$$

Step 3: Computing the Event Estimate.

This new step has been introduced in the calculations to account for an additional smoothing constant, δ, for the event component; $0 \leq \delta \leq 1$. This requires a recursive formula with an initial value.

Let $E^k{}_t$ represents the event factor for data with flag k at period t,

where flag $k = 0$ refers to the normal sales period,

$k = 1$ refers to the panic-state, lockdown, COVID-19 superspreading wave-1 period,

$k = 2$ refers to the COVID-19 relief period after any wave, and

$k = 3$ refers to the period in which any later COVID-19 superspreading wave occurs.

For flag $k = 0$, the initial value and updated formula for estimating events is $E^0{}_t = 1$.

For flags $k = 1, 2, 3$, the initial value for estimating events is $E^k{}_1 = 1$. The updated formula for these flags is as follows:

$$E^{k}_{t} = \delta\left(\frac{A_t}{L_t}\right) + (1-\delta)E^{k}_{t-}; t = 2, 3, \ldots \quad (3.19)$$

Here, E^{k}_{t-} refers to the last occurrence of the event factor with the same flag k, prior to period t. In cases where the last occurrence prior to the event factor having the same flag k is not available, $E^{k}_{t-} = E_{t-1}$.

The updated event formula aims to take into account the latest event factor $\left(\frac{A_t}{L_t}\right)$ and the previous event factor with the same flag k (E^{k}_{t-}) by using a smoothing constant (δ) to determine the current event estimate, except when the flag k=0, in which case the event estimate is fixed at 1.

Step 4: Computing the Holt's Estimate with Events.

The typical Holt's method has been modified to take into account the multiplicative effect of the event component, resulting in a new estimate, HE_{t+m}, that represents the Holt's forecasted value with the event component for period $t+m$. This is computed using the following formula:

$$HE_{t+m} = (L_t + mT_t)E^{k}_{t+m}, \quad (3.20)$$

where m refers to the future period m^{th}; $m \geq 1$.

The next subsection details the steps of the modified Holt's method, which takes into account both seasonal and event components for forecasting.

3.2.4 The Holt's Method with Seasonality and Events

In this subsection, the typical Holt's method is modified to take into account both the associated seasonal and event components. The forecasting steps from the previous subsections on the Holt's method with seasonality and the Holt's method with events are combined and modified here. To deal with both seasonality and events simultaneously, a set of notations from the Holt's method with seasonality is altered to include the flag k, as defined in the Holt's method with events above, resulting in the following new set of notations:

$A_{y,m}$ becomes $A^k_{y,m}$, $\overline{A}_{y,m}$ becomes $\overline{A}^k_{y,m}$,

$\overline{C}_{y,m}$ becomes $\overline{C}^k_{y,m}$, $SF_{y,m}$ becomes $SF^k_{y,m}$.

Step 1: Finding the Moving Averages (MAs).

Let $A^k_{y,m}$ represents the actual data of year y, month m and flag k

where y = 1, 2, ..., 7 refer to the years 2015, 2012, ..., 2021,

m = 1, 2, ..., 12 refer to the months January, February, ..., and December,

$k = 0$ refers to the normal sales period,

$k = 1$ refers to the panic-state, lockdown, COVID-19 superspreading wave-1 period,

$k = 2$ refers to the COVID-19 relief period after any wave,

$k = 3$ refers to the period in which any later COVID-19 superspreading wave occurs.

and $\overline{A}^k_{y,m}$ represents the moving average of year y, month m and flag k, when the number of periods to average is 12.

Therefore, the values start at $\overline{A}^k_{1,6}$ go up until $\overline{A}^k_{7,6}$. More explicitly,

$$\overline{A}^k_{1,6} = \frac{A^k_{1,1} + A^k_{1,2} + \ldots + A^k_{1,12}}{12},$$

$$\overline{A}^k_{1,7} = \frac{A^k_{1,2} + A^k_{1,3} + \ldots + A^k_{1,12} + A^k_{2,1}}{12},$$

$$\overline{A}^k_{1,8} = \frac{A^k_{1,3} + A^k_{1,4} + \ldots + A^k_{1,12} + A^k_{2,1} + A^k_{2,2}}{12},$$

$$\vdots$$

$$\overline{A}^k_{7,6} = \frac{A^k_{7,1} + A^k_{7,2} + \ldots + A^k_{7,11} + A^k_{7,12}}{12}.$$

Step 2: Finding the Centered Moving Averages (CMAs).

Let $\overline{C}^k_{y,m}$ represent the centered moving average (CMAs) in year y, month m and flag k, when the number of periods to average is 2.

Therefore, the values start at $\overline{C}^k_{1,6}$ go up until $\overline{C}^k_{7,6}$ and can be computed as follows.

$$\bar{C}^k_{1,7} = \frac{\bar{A}^k_{1,6} + \bar{A}^k_{1,7}}{2},$$

$$\bar{C}^k_{1,8} = \frac{\bar{A}^k_{1,7} + \bar{A}^k_{1,8}}{2},$$

$$\bar{C}^k_{1,9} = \frac{\bar{A}^k_{1,8} + \bar{A}^k_{1,9}}{2},$$

$$\vdots$$

$$\bar{C}_{7,6} = \frac{\bar{A}_{7,5} + \bar{A}_{7,6}}{2}.$$

Step 3: Computing the Seasonal Factors.

Let $SF^k_{y,m}$ represent the seasonal factor in year y, month m and flag k computed by the following formula,

$$SF^k_{y,m} = \frac{A^k_{y,m}}{\bar{C}^k_{y,m}}. \tag{3.21}$$

Step 4: Computing the Unscaled Seasonal Indices of the Normal Sales Periods Only.

Step 4 of Holt's method with seasonality has been modified to exclude the sales data during events other than regular sales periods, i.e., flags $k = 1, 2, 3$. This is because the seasonal component should not be affected by abnormal factors such as events. Then, the event component can be taken care of later.

Similar to Step 4 of Holt's method with seasonality, the unscaled seasonal index (SI_m) for month m is still computed by averaging the seasonal factors. However, this time, only the seasonal factors with flag $k = 0$ are used, that is,

$$SI_m = \text{the average of } SF^0_{y,m} \tag{3.22}$$

Step 5: Computing the (Scaled) Seasonal Indices of the Normal Sales Periods Only.

The formula for S_t remains the same as in Step 5 of the Holt's method with seasonality, however, in this case, it represents the scaled seasonal index for period t with flag $k = 0$ only. This is because S_t is the SI_m scaled to sum up to 12, and the new SI_m formula in Step 4 above is computed only from the seasonal factors with flag $k = 0$.

Step 6: Removing Seasonality from the Data (De-seasonalization).

The formula for de-seasonalized data at period t, D_t, remains the same as in Step 6 of Holt's method with seasonality. Each actual data point is divided by its corresponding seasonal index. However, D_t here is seasonally removed by the seasonality from the normal sales periods only, not from all the sales periods, as D_t in Holt's method with seasonality. From this step onward, the year y and month m and flag k become irrelevant, and $A^k_{y,m}$ can be represented simply as A_t for $t = 1, 2, \ldots, 84$. De-seasonalized data at period t is computed by the following formula,

$$D_t = \frac{A_t}{S_t}. \tag{3.16}$$

Subsequently, steps 7 through 8 calculate the level and trend estimates in exactly the same manner as the Holt's method with seasonality, after replacing A_t with D_t.

Step 7: Computing the Level Estimate.

The initial value for the level estimate (L_t) is $L_2 = D_2$ and the level estimates are updated using the following formula:

$$L_t = \alpha D_t + (1 - \alpha)\,(L_{t-1} + T_{t-1}) \tag{3.17}$$

where L_t refers to the level estimate for period t,

T_t refers to the trend estimate for period t.

Step 8: Computing the Trend Estimate.

The initial value for the trend estimate (T_t) is $T_2 = D_2 - D_1$ and the trend estimates are updated using the following formula:

$$T_t = \beta\,(L_t - L_{t-1}) + (1 - \beta)\,T_{t-1}. \tag{3.13}$$

Step 9: Computing the Event Estimate.

This new step has been introduced in the calculations to account for an additional smoothing constant, δ, for the event component; $0 \leq \delta \leq 1$. This requires a recursive formula with an initial value.

Let E^k_t represents the event factor for data with flag k at period t,
where flag $k = 0$ refers to the normal sales period,

$k = 1$ refers to the panic-state, lockdown, COVID-19 superspreading wave-1 period,

$k = 2$ refers to the COVID-19 relief period after any wave, and

$k = 3$ refers to the period in which any later COVID-19 superspreading wave occurs.

For flag $k = 0$, the initial value and updated formula for estimating events is $E^0_t = 1$.

For flags $k = 1, 2, 3$, the initial value for estimating events is $E^k_1 = 1$. The updated formula for these flags follows the same procedure as step 3 in Holt's method with events. By replacing " A_t " with " D_t", this formula can be calculated as followed:

$$E^k_t = \delta\left(\frac{D_t}{L_t}\right) + (1-\delta)E^k_{t-}\ ;\ t = 2, 3, \ldots \qquad (3.23)$$

Here, E^k_{t-} refers to the last occurrence of the event factor with the same flag k, prior to period t. In cases where the last occurrence prior to the event factor having the same flag k is not available, $E^k_{t-} = E_{t-1}$.

Step 10: Computing the Holt's Estimate with Events.

This step is the same as step 4 in Holt's method with events. It can be calculated as follow:

$$HE_{t+m} = (L_t + mT_t)E^k_{t+m}, \qquad (3.20)$$

where m refers to the future period m^{th}; $m \geq 1$.

Step 11: Computing the Holt's Estimate with Seasonality and Events (Re-seasonalization).

In this final step, the Holts' estimate from step 10 are multiplied by their corresponding seasonal indices. The result, represented by FE_{t+m} , is the final forecasted data for period $t + m$, fully re-seasonalized from our Holt's estimation.

$$FE_{t+m} = HE_{t+m} \times S_{t+m}. \qquad (3.24)$$

3.2.5 The Accuracy Measurement

In this study, the mean absolute percentage error (MAPE) and the symmetric mean absolute percentage error (SMAPE) are used to compare the accuracy of different methods. These forecast accuracy metrics are considered good measures because they estimate the discrepancy as a percentage and can be used to compare forecast accuracy on other datasets. The formulas for these measures are as follows: (Leenawong, 2022b)

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|A_t - F_t|}{|A_t|}\times 100\%\right) \tag{3.25}$$

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|A_t - F_t|}{(|A_t| + |F_t|)/2}\times 100\%\right) \tag{3.26}$$

where A_t refers to the actual data for period t,

F_t refers to the forecasted data for period t,

and n refers to the number of data.

Chapter 4
Results and Discussion

In this section, the results of this research are divided into two parts. The first part is the modified K-means clustering for demand-weighted locations in the case study of Thailand's convenience store franchise. The second part is Holt's forecasting with events in the case study of Thailand's monthly car sales data during the COVID-19 pandemic.

4.1 The Modified K-Means Clustering for Demand-Weighted Locations in the Case Study of Thailand's Convenience Store Franchise

In this section, the results of the experiments and the accompanying discussion are presented. The six previously mentioned clustering methods, which were derived from a combination of three distance metrics and two calculation methods for centroid locations, were experimented in order to solve the location problem. Specifically, the Euclidean, Manhattan, and Chebyshev distance metrics, along with the demand-weighted versions of these methods (i.e., Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev) were applied to identify the optimal eight DC locations for distributing goods to 260 convenience stores in Eastern Thailand.

In this study, experiments were conducted using 10,000 diverse instances for comparison. Each instance was assigned new initial centroids, which were then utilized in all six methods. Following the completion of these 10,000 instances for each approach, the expectations of effectiveness and efficiency measurement were calculated across all of the instances.

It is worth mentioning that all experiments conducted in this research were executed on an Intel® Core™ i5-1035G4 processor with 8GB of DDR4 memory. The programming language used was R programming, utilizing RStudio version 1.3.1093 as the development environment.

Table 4.1 The Results of the Optimal Solution for the Locations of Eight Centroids or DCs

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)

As seen in Table 4.1, all eight optimal centroids or DC locations were obtained after implementing all six clustering methods. Despite their proximity, these locations are not easily distinguishable. Therefore, it is necessary to measure the effectiveness and efficiency of the six clustering methods for comparison purposes.

4.1.1 The Effectiveness Measurement

The effectiveness of each approach is measured by two criteria: the expected distribution cost and the expected DBI.

The Expected Distribution Cost

The distribution cost of each approach is calculated using equation 3.6. To ensure accuracy, the expectation is averaged over 10,000 instances for each clustering approach. Table 4.2 reports the expected distribution cost of each approach and Figure 4.1 visually presents the results.

Table 4.2 The Expected Distribution Cost

Clustering Approach	The Expected Distribution Cost
Weighted Chebyshev	$1,559.66
Chebyshev	$1,564.61
Weighted Euclidean	$1,569.07
Euclidean	$1,581.26
Manhattan	$6,650.62
Weighted Manhattan	$6,805.71

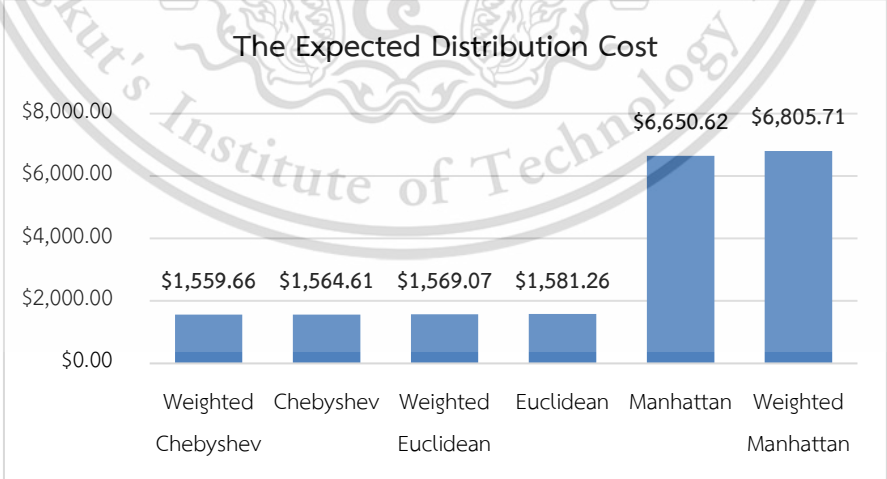

Figure 4.1 The Expected Distribution Cost from Each Method of Clustering, Depicted Across 10,000 Instances

As shown in Table 4.2 and Figure 4.1, the Weighted Chebyshev method yields the lowest expected distribution costs with a value of $1,559.66, followed closely by the Chebyshev method at $1,564.61. In contrast, the Weighted Manhattan and Manhattan methods produce the highest expected distribution costs at $6,805.71 and $6,650.62 respectively. Furthermore, the Weighted Euclidean and Euclidean methods obtain expected distribution costs of $1,569.07 and $1,581.26, respectively. It is clear that the expected distribution costs of the Weighted Manhattan and Manhattan methods are significantly higher than those of the other methods.

The Expected Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) of each approach is calculated using equation 3.11. To ensure accuracy, the expectation is averaged over 10,000 instances for each clustering approach. Table 4.3 reports the expected DBI of each approach and Figure 4.2 visually presents the results.

Table 4.3 The Expected Davies-Bouldin Index (DBI)

Clustering Approach	The Expected DBI
Weighted Chebyshev	0.6779
Chebyshev	0.6793
Weighted Euclidean	0.6891
Euclidean	0.6906
Weighted Manhattan	2.0939
Manhattan	2.1905

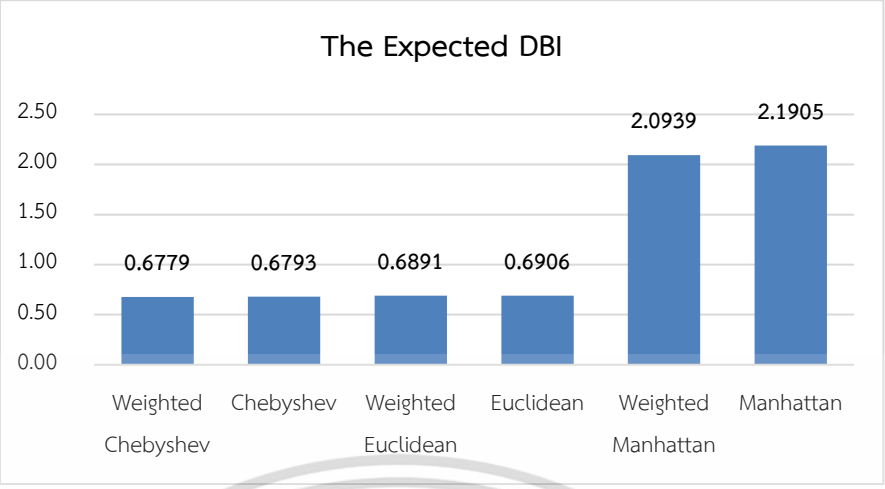

Figure 4.2 The Expected DBI from Each Method of Clustering, Depicted Across 10,000 Instances

As shown in Table 4.3 and Figure 4.2, the Weighted Chebyshev method continues to yield the lowest expected DBI with a value of 0.6779, followed closely by the Chebyshev method at 0.6793. Additionally, the Weighted Euclidean and Euclidean methods generate the expected DBI of 0.6891 and 0.6906, respectively. On the contrary, the Weighted Manhattan and Manhattan methods produce the highest expected DBI at 2.0939 and 2.1905, respectively. Notably, the expected DBI of these methods are notably higher than those of the other methods.

4.1.2 The Efficiency Measurement

The Expected Number of Iterations to the Final Clusters

To measure the efficiency of all six clustering methods, the expected number of iterations to reach the final clusters is determined and compared. These results are obtained by averaging over 10,000 instances for each clustering method. Table 4.4 reports the expected number of iterations of each approach and Figure 4.3 presents this information in a visual format for easy comparison.

Table 4.4 The Expected Number of Iterations to the Final Clusters

Clustering Approach	The Expected Number of Iterations
Euclidean	8.65
Weighted Chebyshev	8.88
Chebyshev	8.97
Weighted Euclidean	9.10
Manhattan	14.84
Weighted Manhattan	16.01

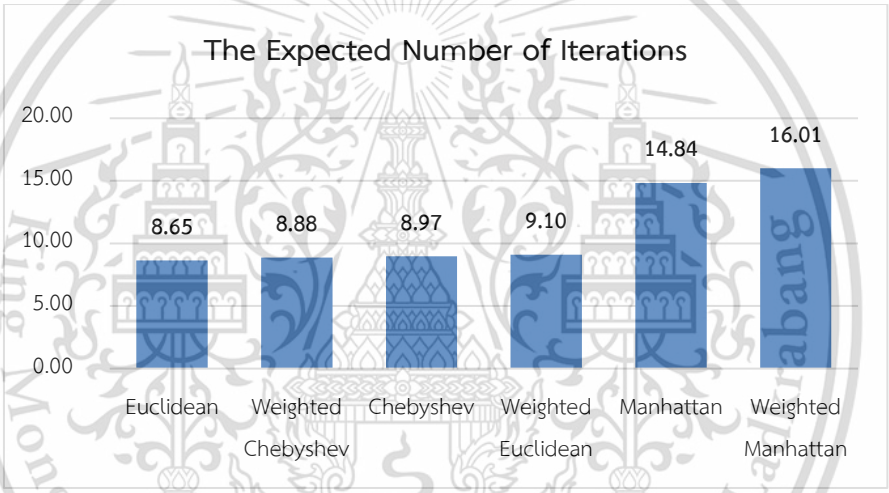

Figure 4.3 The Expected Number of Iterations to the Final Clusters from Each Method of Clustering, Depicted Across 10,000 Instances

As shown in Table 4.4 and Figure 4.3, the Euclidean method obtains the lowest expected number of iterations with a value of 8.65, followed closely by the Weighted Chebyshev, Chebyshev and Weighted Euclidean methods with values of 8.88, 8.97, and 9.10, respectively. On the contrary, the Weighted Manhattan and Manhattan methods produce the highest expected number of iterations at 16.01 and 14.84, respectively. It is obvious that the expected number of iterations of these methods are notably higher than those of the other methods.

As shown in Table 4.1, the optimal locations obtained from the six clustering methods are relatively similar. As such, the effectiveness and efficiency measurements presented in the previous subsections provide a clearer differentiation between the six methods. However, it is still important to consider the combined results across all methods for a more comprehensive analysis, as presented here.

Table 4.5 Summary of the Effectiveness and Efficiency of All Six Different Clustering Methods.

Clustering Method	Effectiveness		Efficiency
	The Expected Distribution Cost	The Expected DBI	The 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

Table 4.5 presents a summary of the effectiveness and efficiency results of all six different clustering methods. It is evident that the Weighted Chebyshev method outperforms the rest in terms of effectiveness, as seen in its low the expected distribution cost and the expected DBI. However, in terms of efficiency, the Euclidean method takes the lead, followed closely by the Weighted Chebyshev method. Therefore, it can be concluded that the Weighted Chebyshev method is the optimal choice for the case study of locating the DCs to serve their convenience stores with varying demands.

4.2 Holt's Forecasting with Events in the Case Study of Thailand's Monthly Car Sales Data During the COVID-19 Pandemic

In this research, Holt's methods are used to forecast monthly car sales data from January 2015 to December 2019. The methods include the typical Holt's Method, Holt's method with seasonality, Holt's method with events, and Holt's method with seasonality and events. The results of all methods are presented below. Additionally, to evaluate the accuracy of these four forecasting methods, this research employs the mean absolute percentage error (MAPE) and the symmetric mean absolute percentage error (SMAPE) as evaluation metrics. Note that all experiments in this research were conducted on an Intel® Core™ i5-1035G4 processor with 8GB of DDR4 memory using Excel Microsoft 365.

The following table presents the results for each forecasting method. The "Car Sales (000 Baht)" column displays the actual car sales data in thousand Thai baht. The other columns for each method offer a comprehensive insight into the different steps involved in the forecasting process.

4.2.1 The Result of the Typical Holt's Method

Table 4.6 presents the results of the typical Holt's method, where the "Level" and "Trend" columns denote the level and trend estimates calculated using equations 3.12 and 3.13, respectively. The final column, "Holt", illustrates the forecasted value obtained by applying equation 3.14, providing a detailed understanding of the method's forecasting performance.

Table 4.6 Forecasting Thailand's Monthly Car Sales Figures Utilizing the Typical Holt's Method

Period	Year	Month	Car Sales (000 Baht)	Level	Trend	Holt
1	2015	1	26,977,962			
2	2015	2	27,902,177	27,902,176.73	924,215.20	
3	2015	3	29,774,249	29,667,544.26	938,473.10	28,826,391.93
4	2015	4	20,635,767	21,758,163.83	788,498.08	30,606,017.36
5	2015	5	26,625,639	26,166,449.66	849,855.08	22,546,661.91
6	2015	6	22,234,255	22,772,592.21	777,922.28	27,016,304.74
7	2015	7	24,605,388	24,486,635.77	793,789.95	23,550,514.49
⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	26,502,130	26,809,163.98	297,175.42	29,229,514.99
62	2020	2	32,657,259	32,032,366.50	380,673.76	27,106,339.41
63	2020	3	26,042,182	26,759,378.97	284,841.70	32,413,040.26
64	2020	4	6,960,650	9,221,548.85	-17,260.45	27,044,220.68
65	2020	5	9,204,255	9,204,258.61	-17,260.95	9,204,288.40
66	2020	6	14,674,115	14,056,405.28	65,277.66	9,186,997.65
67	2020	7	21,454,177	20,628,724.25	175,574.88	14,121,682.93
68	2020	8	28,104,688	27,282,849.83	285,389.17	20,804,299.13
69	2020	9	31,278,826	30,861,108.34	341,204.76	27,568,239.00
70	2020	10	33,105,585	32,891,324.71	369,834.25	31,202,313.09
71	2020	11	37,805,574	37,293,988.41	438,192.49	33,261,158.96
72	2020	12	34,824,993	35,152,268.75	394,461.84	37,732,180.89
73	2021	1	24,738,911	25,955,596.53	231,887.89	35,546,730.59
74	2021	2	29,069,221	28,744,810.73	275,235.69	26,187,484.41
75	2021	3	30,379,730	30,226,664.39	295,688.40	29,020,046.43
76	2021	4	20,168,056	21,333,686.71	139,936.46	30,522,352.79
77	2021	5	17,224,505	17,702,847.77	76,020.15	21,473,623.17
78	2021	6	23,469,074	22,828,501.75	161,613.67	17,778,867.92
79	2021	7	21,615,843	21,770,551.04	140,941.52	22,990,115.43
80	2021	8	19,576,829	19,839,652.40	105,822.91	21,911,492.56
81	2021	9	24,635,676	24,107,678.34	176,374.09	19,945,475.32
82	2021	10	27,218,794	26,888,416.99	220,519.22	24,284,052.43
83	2021	11	29,930,900	29,613,218.34	262,967.91	27,108,936.21
84	2021	12	32,538,148	32,238,478.74	303,009.81	29,876,186.24

The data from Table 4.6 was used to generate a graph below, which compares the actual monthly car sales to the forecasted values produced by the Holt's method.

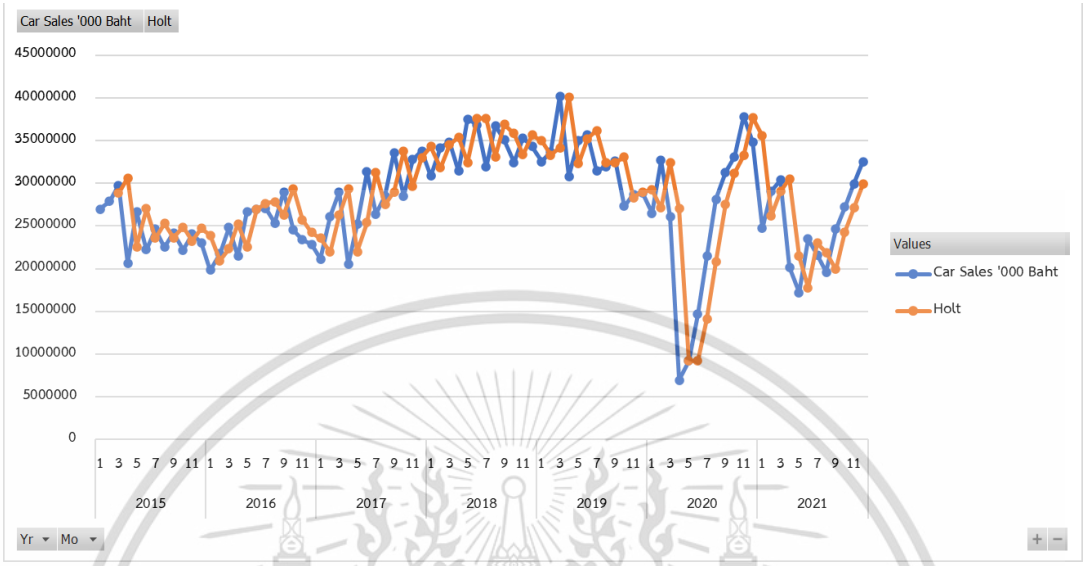

Figure 4.4 Actual monthly car sales and the Holt forecast

4.2.2 The Result of the Holt's Method with Seasonality

The results of the Holt's method with seasonality are presented in Tables 4.7 and 4.8. Table 4.7 deals with the seasonality of the car sales data used in the Holt's method with Seasonality. The first step is to calculate the moving averages, displayed in the "MA" column, starting at Period 6 and ending at Period 78. The second step is to compute the centered moving averages, displayed in the "CMA" column, by taking the average of two consecutive periods, starting at Period 7 and ending at Period 78. Next, the "SF", "SI unscaled", and "SI" columns refer to the seasonal factor, the unscaled seasonal indices, and the scaled seasonal indices, respectively. As shown in Table 4.7, different seasonal factors are obtained for different periods, while the unscaled and scaled seasonal indices for corresponding months are the same value in the same month each year. These seasonal indices are then used to remove the seasonal component in the sales data, resulting in the de-seasonalized car sales data (Deseason).

In Table 4.8, the "Level", "Trend", and "Holt S" columns refer to the level estimate, the trend estimate, and the forecasted value calculated using equations 3.17, 3.13, and 3.14, similar to the typical Holt's method, with the replacement of the actual data by the de-seasonalized data. The final column, "Reseason", illustrates the final forecasted value obtained by applying equation 3.18, after fully re-seasonalizing the Holt's estimation.

Table 4.7 Dealing with Seasonality in the Car Sales Data for the Holt's Method with Seasonality

Period	Year	Month	Car Sales (000 Baht)	MA	CMA	SF	SI Unscaled	SI	DeSeason
1	2015	1	26,977,962				0.924	0.928	29,085,208.77
2	2015	2	27,902,177				1.064	1.069	26,107,588.89
3	2015	3	29,774,249				1.102	1.107	26,899,595.55
4	2015	4	20,635,767				0.762	0.765	26,978,206.09
5	2015	5	26,625,639				0.873	0.877	30,375,252.54
6	2015	6	22,234,255	24,564,721.20			0.988	0.992	22,423,839.45
7	2015	7	24,605,388	23,967,856.58	24,266,288.89	1.014	0.967	0.971	25,331,101.47
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	26,502,130	23,917,024.70	24,333,568.57	1.089	0.924	0.928	28,572,209.82
62	2020	2	32,657,259	23,595,908.11	23,756,466.40	1.375	1.064	1.069	30,556,837.79
63	2020	3	26,042,182	23,481,330.89	23,538,619.50	1.106	1.102	1.107	23,527,853.81
64	2020	4	6,960,650	23,962,827.37	23,722,079.13	0.293	0.762	0.765	9,100,017.24
65	2020	5	9,204,255	24,722,380.43	24,342,603.90	0.378	0.873	0.877	10,500,464.15
66	2020	6	14,674,115	25,217,869.42	24,970,124.93	0.588	0.988	0.992	14,799,236.66
67	2020	7	21,454,177	25,070,934.52	25,144,401.97	0.853	0.967	0.971	22,086,948.20
68	2020	8	28,104,688	24,771,931.36	24,921,432.94	1.128	1.025	1.029	27,318,767.73
69	2020	9	31,278,826	25,133,393.68	24,952,662.52	1.254	1.105	1.110	28,181,235.54
70	2020	10	33,105,585	26,234,010.91	25,683,702.29	1.289	1.002	1.006	32,904,058.03
71	2020	11	37,805,574	26,902,365.10	26,568,188.00	1.423	1.082	1.087	34,787,725.22
72	2020	12	34,824,993	27,635,278.36	27,268,821.73	1.277	1.057	1.061	32,825,032.01
73	2021	1	24,738,911	27,648,750.56	27,642,014.46	0.895	0.924	0.928	26,671,266.02
74	2021	2	29,069,221	26,938,095.62	27,293,423.09	1.065	1.064	1.069	27,199,572.20
75	2021	3	30,379,730	26,384,499.74	26,661,297.68	1.139	1.102	1.107	27,446,618.71
76	2021	4	20,168,056	25,893,933.88	26,139,216.81	0.772	0.762	0.765	26,366,743.34
77	2021	5	17,224,505	25,237,711.02	25,565,822.45	0.674	0.873	0.877	19,650,183.78
78	2021	6	23,469,074	25,047,140.56	25,142,425.79	0.933	0.988	0.992	23,669,187.59
79	2021	7	21,615,843				0.967	0.971	22,253,382.78
80	2021	8	19,576,829				1.025	1.029	19,029,381.69
81	2021	9	24,635,676				1.105	1.110	22,195,966.54
82	2021	10	27,218,794				1.002	1.006	27,053,102.82
83	2021	11	29,930,900				1.082	1.087	27,541,650.71
84	2021	12	32,538,148				1.057	1.061	30,669,517.60

Table 4.8 The Final Forecasts of the Holt's Method with Seasonality

Period	Year	Month	Level	Trend	Holt S	Reseason
1	2015	1				
2	2015	2	26,107,588.89	-2,977,619.88		
3	2015	3	26,899,595.55	-2,503,553.21	23,129,969.01	25,601,777.23
4	2015	4	26,978,206.09	-2,178,821.37	24,396,042.34	18,660,657.03
5	2015	5	30,375,252.54	-1,477,602.54	24,799,384.72	21,738,073.02
6	2015	6	22,423,839.45	-2,291,746.27	28,897,650.00	28,653,332.16
7	2015	7	25,331,101.47	-1,637,921.14	20,132,093.18	19,555,326.39
⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	28,572,209.82	-360,742.92	26,613,567.87	24,685,393.04
62	2020	2	30,556,837.79	-65,790.05	28,211,466.90	30,150,671.46
63	2020	3	23,527,853.81	-941,478.43	30,491,047.75	33,749,505.32
64	2020	4	9,100,017.24	-2,637,517.25	22,586,375.38	17,276,433.56
65	2020	5	10,500,464.15	-2,129,704.55	6,462,499.99	5,664,749.28
66	2020	6	14,799,236.66	-1,321,261.93	8,370,759.61	8,299,988.25
67	2020	7	22,086,948.20	-238,601.00	13,477,974.72	13,091,842.58
68	2020	8	27,318,767.73	449,356.82	21,848,347.21	22,476,891.57
69	2020	9	28,181,235.54	501,309.49	27,768,124.55	30,820,307.26
70	2020	10	32,904,058.03	1,032,205.22	28,682,545.03	28,858,216.29
71	2020	11	34,787,725.22	1,139,284.72	33,936,263.24	36,880,247.31
72	2020	12	32,825,032.01	749,181.26	35,927,009.94	38,115,968.41
73	2021	1	26,671,266.02	-118,930.53	33,574,213.28	31,141,733.97
74	2021	2	27,199,572.20	-37,534.30	26,552,335.50	28,377,494.40
75	2021	3	27,446,618.71	-1,745.54	27,162,037.90	30,064,737.37
76	2021	4	26,366,743.34	-137,330.70	27,444,873.17	20,992,723.27
77	2021	5	19,650,183.78	-964,731.80	26,229,412.64	22,991,573.93
78	2021	6	23,669,187.59	-337,979.27	18,685,451.98	18,527,474.11
79	2021	7	22,253,382.78	-473,526.16	23,331,208.32	22,662,789.69
80	2021	8	19,029,381.69	-819,424.75	21,779,856.62	22,406,430.61
81	2021	9	22,195,966.54	-318,145.84	18,209,956.94	20,211,536.68
82	2021	10	27,053,102.82	332,695.50	21,877,820.70	22,011,815.24
83	2021	11	27,541,650.71	352,295.43	27,385,798.32	29,761,527.01
84	2021	12	30,669,517.60	701,350.15	27,893,946.14	29,593,466.63

The data from Tables 4.7 and 4.8 were used to generate a graph below, which compares the actual monthly car sales to the forecasted values produced by the Holt's method with seasonality.

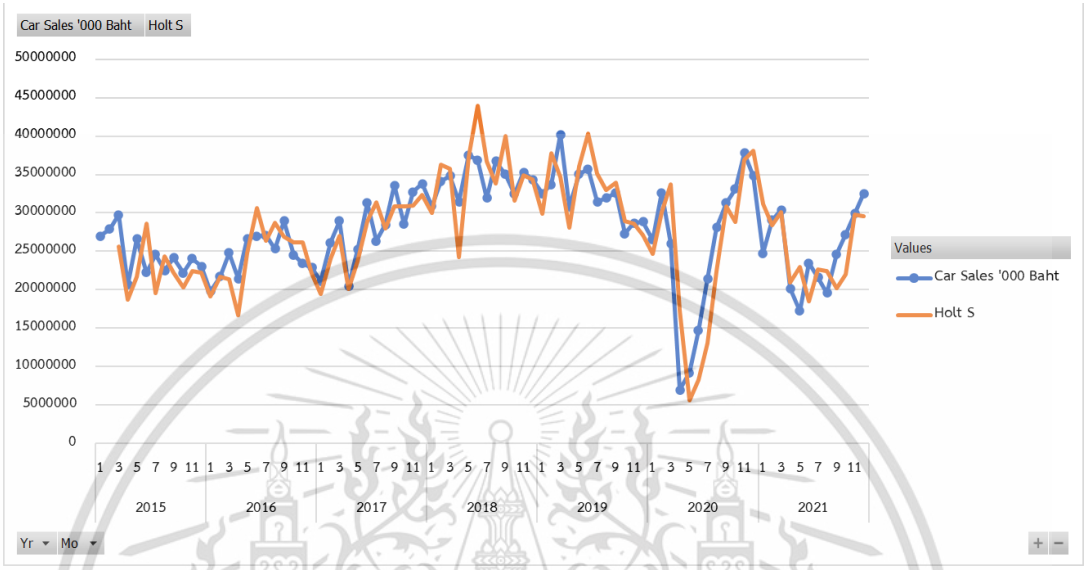

Figure 4.5 Actual monthly car sales and the Holt S forecast

4.2.3 The Result of the Holt's Method with Events

The results of Holt's method with events are presented in Table 4.9, showcasing a detailed understanding of the method's forecasting performance. Similar to the typical Holt's method, the "Level" and "Trend" columns represent the level and trend estimates calculated using equations 3.12 and 3.13 respectively. The "Event" column illustrates the event estimate, calculated using equation 3.19 and based on the last occurrence of the event factor with the same flag k, as represented in the "E_{t-}" column. Finally, the "Holt E" column displays the final forecasted value obtained by applying equation 3.20.

Table 4.9 Forecasting Thailand's Monthly Car Sales Figures Utilizing the Holt's Method with Events

Period	Year	Month	Flag	Car Sales (000 Baht)	Level	Trend	E_{t-}	Event	Holt E
1	2015	1	0	26,977,962				1.000	
2	2015	2	0	27,902,177	27,902,176.73	924,215.20	1.000	1.000	
3	2015	3	0	29,774,249	28,980,200.10	973,964.27	1.000	1.000	28,826,391.93
4	2015	4	0	20,635,767	28,442,073.65	484,880.34	1.000	1.000	29,954,164.37
5	2015	5	0	26,625,639	28,553,520.93	364,093.87	1.000	1.000	28,926,953.99
6	2015	6	0	22,234,255	27,833,110.01	13,312.10	1.000	1.000	28,917,614.80
7	2015	7	0	24,605,388	27,320,501.39	-156,796.33	1.000	1.000	27,846,422.11
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	0	26,502,130	28,276,318.79	-976,945.20	1.000	1.000	28,619,981.29
62	2020	2	0	32,657,259	28,168,794.42	-695,732.08	1.000	1.000	27,299,373.58
63	2020	3	0	26,042,182	27,240,874.31	-770,833.02	1.000	1.000	27,473,062.34
64	2020	4	1	6,960,650	23,304,263.97	-1,794,799.89	1.000	0.299	7,906,221.83
65	2020	5	1	9,204,255	19,512,705.13	-2,440,649.17	0.299	0.472	10,146,137.46
66	2020	6	1	14,674,115	16,682,943.54	-2,566,507.12	0.472	0.880	15,016,373.72
67	2020	7	1	21,454,177	15,307,127.12	-2,181,379.64	0.880	1.402	19,785,327.23
68	2020	8	2	28,104,688	15,556,371.19	-1,395,197.32	1.402	1.807	23,713,437.57
69	2020	9	2	31,278,826	16,938,845.04	-496,762.92	1.807	1.847	26,149,651.47
70	2020	10	2	33,105,585	19,146,058.71	377,835.06	1.847	1.729	28,430,119.76
71	2020	11	2	37,805,574	22,490,451.08	1,337,364.45	1.729	1.681	32,818,906.45
72	2020	12	2	34,824,993	25,612,321.00	1,914,560.60	1.681	1.360	32,398,606.83
73	2021	1	3	24,738,911	27,074,479.27	1,768,231.61	1.360	0.914	25,152,286.95
74	2021	2	2	29,069,221	28,879,466.54	1,780,120.19	1.360	1.007	29,032,223.81
75	2021	3	2	30,379,730	30,614,174.57	1,765,431.68	1.007	0.992	30,424,794.58
76	2021	4	3	20,168,056	30,398,045.34	1,124,498.19	0.914	0.663	21,482,753.70
77	2021	5	3	17,224,505	29,202,409.40	374,053.59	0.663	0.590	18,592,993.68
78	2021	6	3	23,469,074	28,585,420.67	53,502.14	0.590	0.821	24,282,735.40
79	2021	7	3	21,615,843	27,499,291.84	-315,110.13	0.821	0.786	22,511,650.86
80	2021	8	3	19,576,829	25,949,741.08	-714,388.47	0.786	0.754	20,508,107.00
81	2021	9	3	24,635,676	25,138,043.38	-745,863.02	0.754	0.980	24,731,040.17
82	2021	10	2	27,218,794	24,850,853.31	-597,505.80	0.992	1.095	26,716,416.17
83	2021	11	2	29,930,900	25,174,640.50	-299,514.70	1.095	1.189	28,835,546.25
84	2021	12	2	32,538,148	26,118,599.82	102,685.48	1.189	1.246	30,989,047.21

The data from Table 4.9 was used to generate a graph below, which compares the actual monthly car sales to the forecasted values produced by the Holt's method with events.

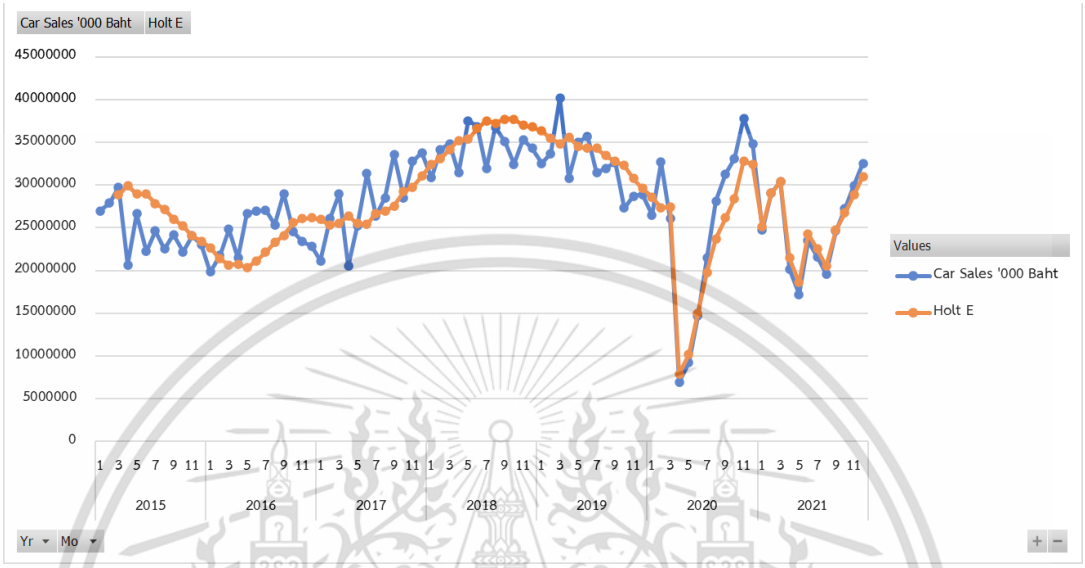

Figure 4.6 Actual monthly car sales and the Holt E forecast

4.2.4 The Result of the Holt's Method with Seasonality and Events

The results of the Holt's method with events are presented in Tables 4.10 and 4.11. Similar to the table results of the Holt's method with seasonality, the "MA", "CMA", and "SF" columns in Table 4.10 refer to the moving averages, the centered moving averages, and the seasonal factor, respectively. However, the "SI unscaled" column refers to the unscaled seasonal indices calculated using equation 3.22. This is calculated by averaging the seasonal factors only with flag k=0. The "SI" column refers to the scaled seasonal indices of the normal sales periods only because the "SI unscaled" column is computed only from the seasonal factors with flag k=0. The last column "DeSeason" is similar to the table result of the Holt's method with seasonality calculated using equation 3.16.

In Table 4.11, the "Level" and "Trend" columns are exactly the same as those of the Holt's method with seasonality after the replacement of the actual data with the de-seasonalized data. The "Event", "E_{t-}", and "Holt SE" columns are similar to the Holt's method with events, with the replacement of the actual data by the de-seasonalized data. Finally, the "Reseason" column is calculated using equation 3.24.

It is similar to the "Reseason" column in the Holt's method with seasonality, re-seasonalizing the Holt's estimation.

Table 4.10 Dealing with Seasonality in the Car Sales Data for the Holt's Method with Seasonality and Events

Period	Year	Month	Flag	Car Sales (000 Baht)	MA	CMA	SF	SI Unscaled	SI	DeSeason
1	2015	1	0	26,977,962				0.9296	0.9178	29,395,602.24
2	2015	2	0	27,902,177				1.0643	1.0508	26,554,084.21
3	2015	3	0	29,774,249				1.0950	1.0811	27,541,580.65
4	2015	4	0	20,635,767				0.8765	0.8653	23,847,031.31
5	2015	5	0	26,625,639				1.0466	1.0333	25,768,470.42
6	2015	6	0	22,234,255	24,564,721.20			1.1011	1.0870	20,453,895.19
7	2015	7	0	24,605,388	23,967,856.58	24,266,288.89	1.0140	0.9903	0.9777	25,167,107.50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	0	26,502,130	23,917,024.70	24,333,568.57	1.0891	0.9296	0.9178	28,877,128.64
62	2020	2	0	32,657,259	23,595,908.11	23,756,466.40	1.3747	1.0643	1.0508	31,079,424.73
63	2020	3	0	26,042,182	23,481,330.89	23,538,619.50	1.1064	1.0950	1.0811	24,089,369.01
64	2020	4	1	6,960,650	23,962,827.37	23,722,079.13	0.2934	0.8765	0.8653	8,043,840.84
65	2020	5	1	9,204,255	24,722,380.43	24,342,603.90	0.3781	1.0466	1.0333	8,907,939.10
66	2020	6	1	14,674,115	25,217,869.42	24,970,124.93	0.5877	1.1011	1.0870	13,499,117.14
67	2020	7	1	21,454,177	25,070,934.52	25,144,401.97	0.8532	0.9903	0.9777	21,943,956.93
68	2020	8	2	28,104,688	24,771,931.36	24,921,432.94	1.1277	1.0040	0.9912	28,353,320.27
69	2020	9	2	31,278,826	25,133,393.68	24,952,662.52	1.2535	1.0758	1.0621	29,448,765.08
70	2020	10	2	33,105,585	26,234,010.91	25,683,702.29	1.2890	0.9447	0.9327	35,495,335.99
71	2020	11	2	37,805,574	26,902,365.10	26,568,188.00	1.4230	1.0143	1.0014	37,754,524.32
72	2020	12	2	34,824,993	27,635,278.36	27,268,821.73	1.2771	1.0126	0.9997	34,835,988.33
73	2021	1	3	24,738,911	27,648,750.56	27,642,014.46	0.8950	0.9296	0.9178	26,955,898.22
74	2021	2	2	29,069,221	26,938,095.62	27,293,423.09	1.0651	1.0643	1.0508	27,664,742.76
75	2021	3	2	30,379,730	26,384,499.74	26,661,297.68	1.1395	1.0950	1.0811	28,101,659.06
76	2021	4	3	20,168,056	25,893,933.88	26,139,216.81	0.7716	0.8765	0.8653	23,306,536.84
77	2021	5	3	17,224,505	25,237,711.02	25,565,822.45	0.6737	1.0466	1.0333	16,669,990.77
78	2021	6	3	23,469,074	25,047,140.56	25,142,425.79	0.9334	1.1011	1.0870	21,589,838.94
79	2021	7	3	21,615,843				0.9903	0.9777	22,109,314.01
80	2021	8	3	19,576,829				1.0040	0.9912	19,750,017.97
81	2021	9	3	24,635,676				1.0758	1.0621	23,194,291.94
82	2021	10	2	27,218,794				0.9447	0.9327	29,183,603.23
83	2021	11	2	29,930,900				1.0143	1.0014	29,890,483.35
84	2021	12	2	32,538,148				1.0126	0.9997	32,548,420.87

Table 4.11 The Final Forecasts of the Holt's Method with Seasonality and Events

Period	Year	Month	Level	Trend	E_{t-}	Event	Holt SE	Reseason
1	2015	1				1.000		
2	2015	2	26,554,084.21	-2,841,518.04	1.000	1.000		
3	2015	3	24,585,859.63	-2,418,254.32	1.000	1.000	23,712,566.17	25,634,833.91
4	2015	4	22,550,636.44	-2,232,608.62	1.000	1.000	22,167,605.31	19,182,494.35
5	2015	5	21,561,124.69	-1,630,110.35	1.000	1.000	20,318,027.83	20,993,891.35
6	2015	6	20,050,269.16	-1,572,310.49	1.000	1.000	19,931,014.34	21,665,861.30
7	2015	7	20,003,570.54	-832,884.20	1.000	1.000	18,477,958.67	18,065,537.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	27,968,469.87	-928,188.07	1.000	1.000	27,699,998.74	25,421,812.91
62	2020	2	27,961,499.82	-481,696.51	1.000	1.000	27,040,281.80	28,413,057.51
63	2020	3	26,706,537.99	-856,479.06	1.000	1.000	27,479,803.31	29,707,463.49
64	2020	4	21,788,947.66	-2,824,799.13	1.000	0.369	9,543,084.10	8,258,003.25
65	2020	5	16,670,602.24	-3,936,424.22	0.369	0.534	10,133,495.95	10,470,578.87
66	2020	6	12,908,639.70	-3,851,866.96	0.534	1.046	13,316,675.08	14,475,792.87
67	2020	7	11,995,986.94	-2,427,302.62	1.046	1.829	16,567,326.39	16,197,550.25
68	2020	8	13,852,945.94	-350,827.11	1.829	2.047	19,584,568.67	19,412,830.15
69	2020	9	17,139,112.64	1,411,933.73	2.047	1.718	23,199,609.81	24,641,323.94
70	2020	10	22,415,575.33	3,284,975.22	1.718	1.584	29,375,807.40	27,398,058.15
71	2020	11	28,449,732.24	4,617,435.51	1.584	1.327	34,106,193.08	34,152,309.63
72	2020	12	33,470,587.34	4,812,962.99	1.327	1.041	34,416,111.62	34,405,249.19
73	2021	1	35,700,022.74	3,560,791.15	1.041	0.755	28,906,633.86	26,529,208.34
74	2021	2	36,616,067.26	2,278,947.95	1.041	0.756	29,662,943.02	31,168,865.47
75	2021	3	36,433,345.92	1,085,837.80	0.756	0.771	30,000,386.43	32,432,378.60
76	2021	4	34,277,667.74	-485,244.72	0.755	0.680	25,510,552.36	22,075,276.95
77	2021	5	29,887,264.59	-2,377,978.31	0.680	0.558	18,848,141.09	19,475,109.96
78	2021	6	26,159,222.26	-3,032,320.91	0.558	0.825	22,704,079.43	24,680,301.13
79	2021	7	22,894,817.52	-3,144,806.20	0.825	0.966	22,333,435.23	21,834,961.84
80	2021	8	19,750,012.84	-3,144,805.46	0.966	1.000	19,750,016.46	19,576,827.12
81	2021	9	18,107,997.33	-2,416,440.39	1.000	1.281	21,269,388.36	22,591,151.02
82	2021	10	18,768,723.64	-925,013.89	0.771	1.555	24,398,897.91	22,756,222.99
83	2021	11	20,591,249.28	406,650.48	1.555	1.452	25,902,124.83	25,937,148.28
84	2021	12	23,632,257.67	1,683,458.53	1.452	1.377	28,920,151.78	28,911,023.99

The data from Tables 4.10 and 4.11 were used to generate a graph below, which compares the actual monthly car sales to the forecasted values produced by the Holt's method with seasonality and events.

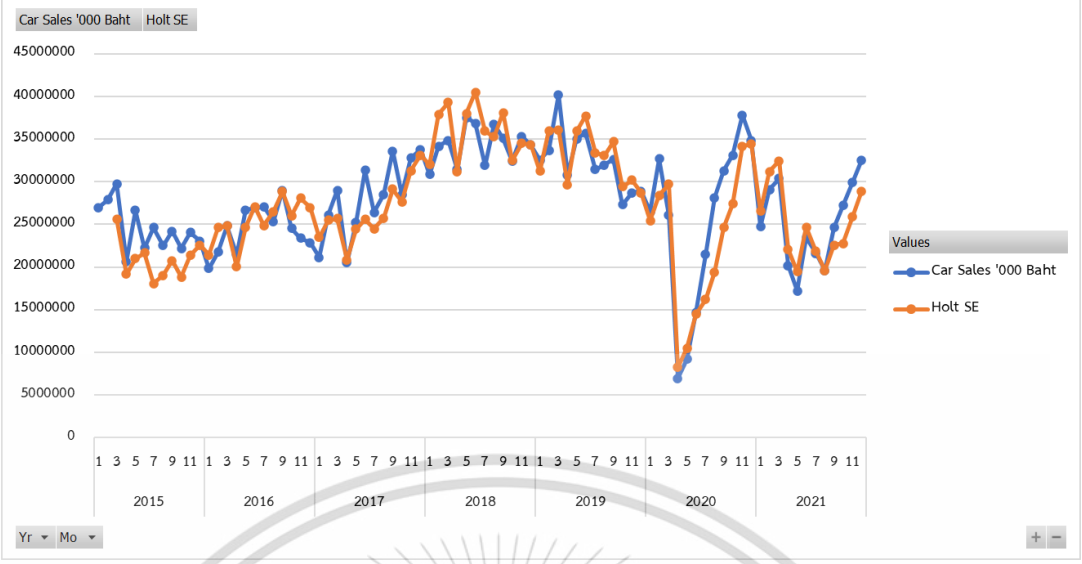

Figure 4.7 Actual monthly car sales and the Holt SE forecast

Table 4.12 summarizes and compares the Mean Absolute Percentage Error (MAPE) and the Symmetric Mean Absolute Percentage Error (SMAPE) for the typical Holt's method, Holt's method with seasonality, Holt's method with events, and Holt's method with seasonality and events. The methods are denoted by Holt, Holt S, Holt E, and Holt SE respectively.

Table 4.12 Accuracy comparison

Method	MAPE	SMAPE
Holt	16.27%	13.91%
Holt S	12.37%	11.99%
Holt E	9.47%	9.33%
Holt SE	**8.64%**	**8.90%**

As shown in Table 4.12, MAPE and SMAPE obtained from all four methods are demonstrated. The typical Holt's method, denoted by Holt, achieves the worst forecast on Thailand's monthly car sales data containing the COVID-19 pandemic period, in terms of both accuracy measures, possessing the highest MAPE and SMAPE at 16.27% and 13.91%, respectively. An improvement from Holt can be observed when addressing seasonality in Holt S, resulting in a lower MAPE and SMAPE at 12.37% and

11.99%. Even higher improvement can be obtained through Holt E, indicating that, in the pandemic case, the event effects are stronger on these car sales data than the seasonal effects. Holt E obtains MAPE and SMAPE at 9.47% and 9.33%, respectively. Additionally, the method combining both seasonal and event effects, Holt SE, achieves the lowest MAPE and SMAPE at 8.64% and 8.90%, respectively, indicating that this proposed modified Holt's method best fits Thailand's car sales data during the COVID-19 pandemic.

Chapter 5
Conclusions and suggestions

5.1 Conclusions

The aim of this research is to use data mining techniques to efficiently manage and analyze business data. On one hand, optimal distribution centers are to be located for the case when customer demands are also contemplated. Additionally, this research modifies the Holt's forecasting method to forecast sales data where special events occur.

5.1.1 The Modified K-Means Clustering for Demand-Weighted Locations in the Case Study of Thailand's Convenience Store Franchise

To determine the best locations for the distribution centers, a case study of a convenience store franchise in Thailand is conducted. The K-means clustering algorithm is adapted and the final iteration's centroids are used to take into account the unique characteristics of the problem, including varying demands at each store and different shipment sizes. To improve the algorithm's effectiveness, a modification is proposed that adjusts the centroid calculation by weighting it according to the stores' different demands. Additionally, three new distance metrics - Weighted Euclidean, Weighted Manhattan, and Weighted Chebyshev - are introduced to complement the typical distance metrics of Euclidean, Manhattan, and Chebyshev. By experimenting with these six clustering methods on a case study of locating eight DCs to service 260 convenience stores in Eastern Thailand, the resulting locations are not significantly different, but the efficiency and effectiveness of these methods are significant. The clustering approach that best fits this problem is the proposed demand-weighted Chebyshev as shown in Figure 5.1.

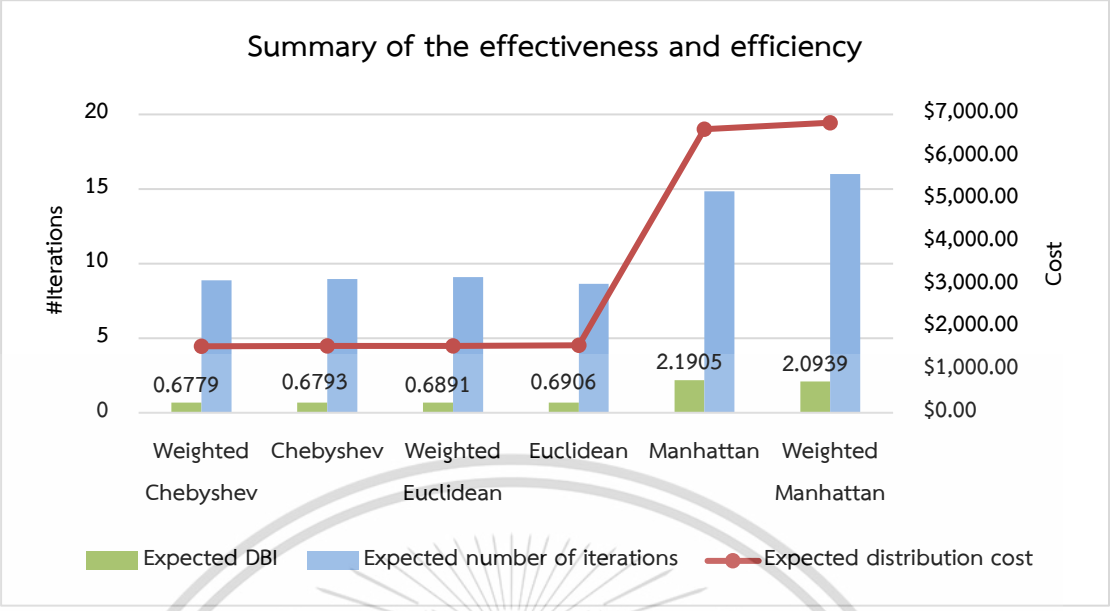

Figure 5.1 Graphical Results from the Six Clustering Approaches Sorted Ascendingly by their Distribution Costs

5.1.2 Holt's Forecasting with Events in the Case Study of Thailand's Monthly Car Sales Data During the COVID-19 Pandemic

The second objective of this research is to modify the typical Holt's forecasting method to better handle time series data that contains an event component, such as the COVID-19 global pandemic. In addition to the traditional Holt's method, three modified methods that incorporate the seasonal and/or event components are proposed: the Holt's method with seasonality, the Holt's method with events, and the Holt's method with seasonality and events. The methods with events incorporate another smoothing constant for the event estimate, denoted by δ and $0 \leq \delta \leq 1$. As for the methods with seasonality, the actual sales data are first de-seasonalized prior to being processed in the next steps. All four of these methods are applied to Thailand's monthly car sales data from January 2015 to December 2021, which also include the COVID-19 pandemic period. In terms of forecasting accuracy, the experimental results show that the Holt's method with seasonality and events best fits the Thailand's car sales data, with the lowest MAPE and SMAPE among all the four methods. It is worth noting that even the Holt's method with events alone can yield the second-best accuracy, which is much better than that of the typical Holt's and the Holt's with seasonality methods as shown in Figure 5.2.

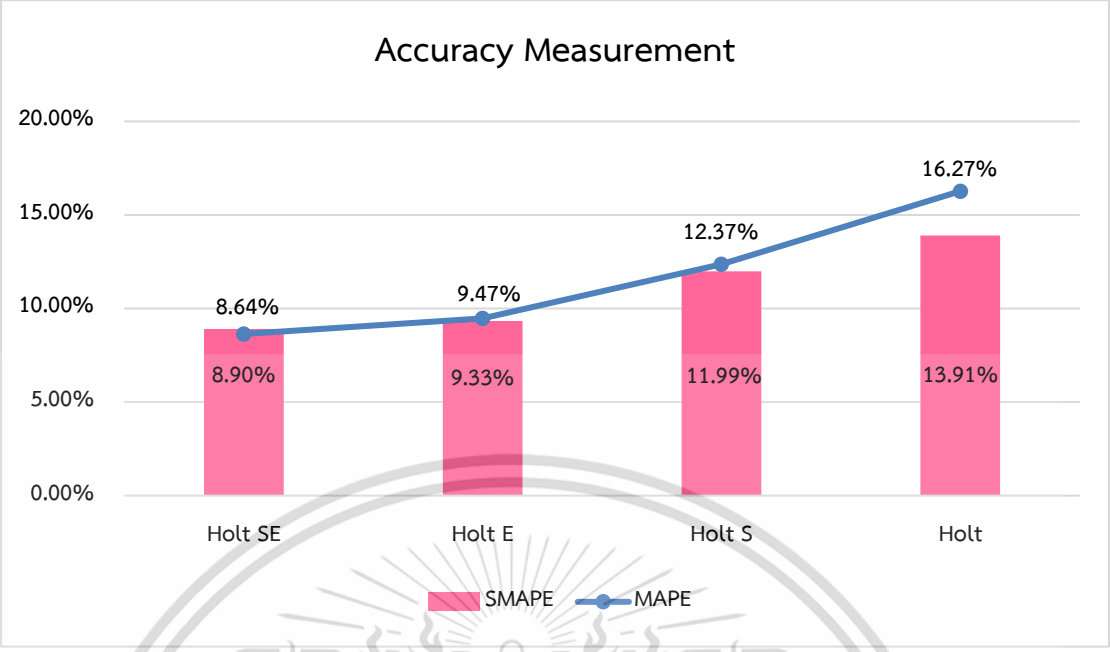

Figure 5.2 Forecasting Accuracy Bar Charts of All Holt's-based Methods Sorted Ascendingly by MAPEs

5.2 Suggestions

5.2.1 The Modified K-Means Clustering for Demand-Weighted Locations in the Case Study of Thailand's Convenience Store Franchise

1. Other than the three-distance metrics employed here, the distances between the convenient stores and their respective distribution centers may be figured from another distance metric or the real world based primarily on existing land routes.

2. 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 in order to obtain higher algorithm efficiency.

5.2.2 Holt's Forecasting with Events in the Case Study of Thailand's Monthly Car Sales Data During the COVID-19 Pandemic

1. The modified Holt's methods proposed in this research can certainly be applied to time series forecasting for other data largely impacted by abnormal events other than the COVID-19 pandemic. As evidenced here, the incorporation of another constant for the event estimate into Holt's method significantly improves the

forecasting accuracy; thus, the forecasted results can be used further in all kinds of planning for the business to stay competitive and beyond.

2. It would be interesting to compare the proposed methods with other time-series models. However, to ensure a fair comparison, a subtle way to incorporate the event component into the model of choice is, most likely, mandatory, and this is currently an ongoing project.

References

Aggarwal, C.C. and Reddy, C.K. 2013. **DATA CLUSTERING Algorithms and Applications**. New York : CRC Press.

Bank of Thailand. 2021. Revitalising Thailand's tourism sector. [Online]. Available : https://shorturl.asia/HVJ9b

Booranawong, T. and Booranawong, A. 2018. "Double exponential smoothing and Holt-Winters methods with optimal initial values and weighting factors for forecasting lime, Thai chili and lemongrass prices in Thailand." *Engineering and Applied Science Research*. 45(1) : 32-38.

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.

Dantrakul, S. Likasiri, C. and Pongvuthithum, R. 2014. "Applied p-median and p-center algorithms for facility location problems." *Expert Systems with Applications*. 41(8) : 3596-3604.

Davies, D.L. and Bouldin, D.W. 1979. "A Cluster Separation Measure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1(2) : 224-227.

Drezner, Z. Scott, C. and Song, J.S. 2003. "The central warehouse location problem revisited." *IMA Journal of Management Mathematics*. 14(4) : 321-336.

Farahani, R.Z. and Hekmatfar, M., editor. 2009. **Facility location: concepts, models, algorithms and case studies.** Heidelberg : Physica.

Gultom, S. Sriadhi, S. Martiano, M. and Simarmata, J. 2018. "Comparison analysis of K-Means and K-Medoid with Euclidean Distance Algorithm, Distance, and Chebyshev Distance for Big Data Clustering." *Proceedings of the Materials Science and Engineering*. 420 : 12092-12098.

Hakimi, S.L. 1964. "Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph." *Operations Research*. 12(3) : 450-459.

Han, J. Kamber, M. and Pei, J. 2012. **Data Mining: Concepts and Techniques**. 3rd ed. San Francisco : Morgan Kaufmann Publishers.

Leenawong, C. 2022a. **Data Analytics with Excel for Logistics & Supply Chain Management**. Bangkok : CU press. (in Thai)

Leenawong, C. 2022b. **Logistics Intelligence and Forecasting with Excel 365**. Bangkok : KMITL. (in Thai)

Muchayan, A. 2019. "Comparison of Holt and Brown's Double Exponential Smoothing Methods in The Forecast of Moving Price for Mutual Funds." *Journal of Applied Science, Engineering, Technology, and Education*. 1(2) : 183–192.

Netherlands Embassy in Bangkok. 2017. **Tourism industry in Thailand.** [Online]. Available : https://bit.ly/3zHNFar

Rattanametawee, W. and Leenawong, C. 2020. "Event Index Computation for Forecasting Case Study: Car Sales in Thailand." *Thai Journal of Mathematics*. 18(4) : 2079–2091.

Rattanametawee, W. Leenawong, C. and Netisopakul, P. 2016. "The Effects of Special Events on Regression for Subcompact Car Sales in Thailand." *Jurnal Teknologi (Sciences & Engineering)*. 78(11) : 161–165.

Sharif, O. and Hasan, M.Z. 2019. "Forecasting the Stock Price by using Holt's Method." *Indonesian Journal of Contemporary Management Research*. 1(1) : 15-24.

Sharma, A. and Jalal, A.S. 2017. "Clustering based hybrid approach for facility location problem." *Management Science Letters*. 7(12) : 577–584.

Singh, A. Yadav, A. and Rana, A. 2013. "K-means with Three different Distance Metrics." *International Journal of Computer Applications*. 67(10) : 13-17.

Sinwar, D. and 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.

Suppalakpanya, K. Nikhom, R. Booranawong, T. and Booranawong, A. 2019. "Study of Several Exponential Smoothing Methods for Forecasting Crude Palm Oil Productions in Thailand." *Current Applied Science and Technology*. 19(2) : 123-139.

The Office of Industrial Economics. 2021. **Industrial Statistics (e-Statistic) : TSIC.** [Online]. Available : https://indexes.oie.go.th/industrialStatistics1.aspx

The World Bank. 2021. **Monitoring the Impact of COVID-19 in Thailand.** [Online]. Available : https://bit.ly/3P4aH12

Wang, M. 2017. "The research of strategy for the 7-Eleven convenience store in Thailand." Master of Business Administration, Thesis of Siam University.

Weber, A. 1909. **Theory of the Location of Industries**. CSISS Classics.

Wilson, J. H. Keating, B. and John Galt Solutions, Inc. 2009. **Business forecasting with forecastX**. 6th ed. New York : McGraw-Hill.

Wirotcheewan, P. Kengpol, A. Ishii, K. and Shimada, Y. 2011. "Modelling and Forecasting for Automotive Parts Demand of Foreign Markets on Thailand." *Asian International Journal of Science and Technology in Production and Manufacturing Engineering*. 4(1) : 1-13.

World Health Organization. 2020. **Novel Coronavirus – Thailand (ex-China).** [Online]. Available : https://bit.ly/3Qavsta

Yang, L. Ji, X. Gao, Z. and Li, K.P. 2007. "Logistics distribution centers location problem and algorithm under fuzzy environment." *Journal of Computational and Applied Mathematics*. 208(2) : 303-315.

Appendices

Appendix A

data

MDPI

Article

Event Forecasting for Thailand's Car Sales during the COVID-19 Pandemic

Chartchai Leenawong * and Thanrada Chaikajonwat

School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand; 63605011@kmitl.ac.th

* Correspondence: chartchai.le@kmitl.ac.th

Abstract: The COVID-19 pandemic that started in 2020 has affected Thailand's automotive industry, among many others. During the several stages of the pandemic period, car sales figures fluctuate, and hence are difficult to fit and forecast. Due to the trend present in the sales data, the Holt's forecasting method appears a reasonable choice. However, the pandemic, or in a more general term, the "event", requires a subtle method to handle this extra event component. This research proposes a forecasting method based on Holt's method to better suit the time-series data affected by large-scale events. In addition, when combined with seasonality adjustment, three modified Holt's-based methods are proposed and implemented on Thailand's monthly car sales covering the pandemic period. Different flags are carefully assigned to each of the sales data to represent different stages of the pandemic. The results show that Holt's method with seasonality and events yields the lowest MAPE of 8.64%, followed by 9.47% of Holt's method with events. Compared to the typical Holt's MAPE of 16.27%, the proposed methods are proved strongly effective for time-series data containing the event component.

Keywords: Thailand's car sales; Holt's method; event component; COVID-19; seasonality

Citation: Leenawong, C.; Chaikajonwat, T. Event Forecasting for Thailand's Car Sales during the COVID-19 Pandemic. *Data* **2022**, *7*, 86. https://doi.org/10.3390/data7070086

Academic Editor: Francisco Guijarro

Received: 30 May 2022
Accepted: 22 June 2022
Published: 25 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The automotive industry has been one of the most blossoming industries in Thailand for decades in both production and sales dimensions, owing to several factors. For the production dimension, Thailand is a crucial automobile part manufacturer and at the same time, an assembly hub for the Southeast Asia region [1], even though Thailand does not own any internationally renowned car brands.

Regarding the sales dimension, Bangkok, the capital city, is always among the top worst traffic cities in the world [2,3]. The majority of middle-class Thai citizens are in need of vehicles to commute between home in the suburban areas and workplace in the city on account of insufficient and inconvenient public transportation systems. Besides pick-up trucks widely used for commercials, the best-selling vehicle types for ordinary Thai citizens are compact and subcompact sedans, with engines up to 1800 c.c. [4].

Beginning in early 2020, Thailand was among the first countries attacked by the coronavirus disease or COVID-19 [5]. As was the case for every other place in the world, Thailand's economy rapidly decreased as the virus developed into a global pandemic. Millions of people were at risk of losing their jobs, earning less salaries, and gaining reduced or no bonuses [6,7]. Consequently, every household tried to save their incomes to make ends meet, especially for the rainy days, and thus decreased the unnecessary consumption. A new car purchase was obviously on the unnecessary buying list, attributable to a certain characteristic of the cars being depreciated assets. To make the matter even worse, all the businesses, automotives included, were heavily affected by the nationwide lockdowns initiated by the government to prevent the disease spreading.

In Figure 1, Thailand's monthly car sales from January 2015 to December 2021 are depicted. During the first COVID-19 superspreading wave across the nation, approximately for the first to the mid-third quarters of 2020, the sales curve declined drastically, especially from March to April, before gradually entering a relief period. However, the nation and the entire world had not experienced the end of the pandemic yet. Superspreading waves 2 and 3 struck in the following months, making car sales unsteady all over again. From an optimistic point of view, as the new car sales figures continue to progress into 2022, the situation will eventually return to normal in the near future.

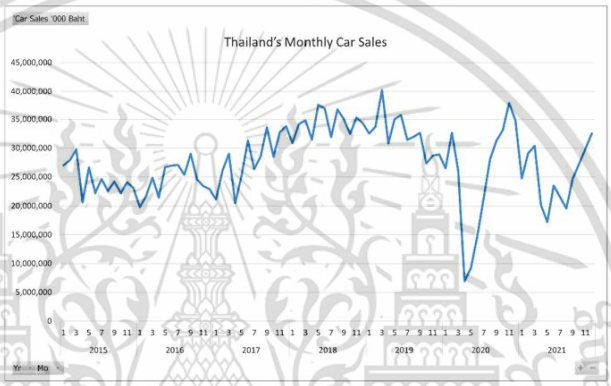

Figure 1. Thailand's monthly car sales from January 2015 to December 2021.

In Figure 2, the monthly car sales data are trimmed to show only from the normal period, in other words, no COVID-19 pandemic effects, i.e., from January 2015 to December 2019. Evidently, the added trendline exhibits an upward trend, in which the time-series Holt's or double exponential smoothing method are, thus, appropriate methods to start with [8]. However, the full car sales data in Figure 1 are severely disturbed by the pandemic. Consequently, the typical Holt's method may not perform as well as intended. Thus, in this research, modifications of the Holt's method are proposed to incorporate the effects of the pandemic, or the more general term, "the event component".

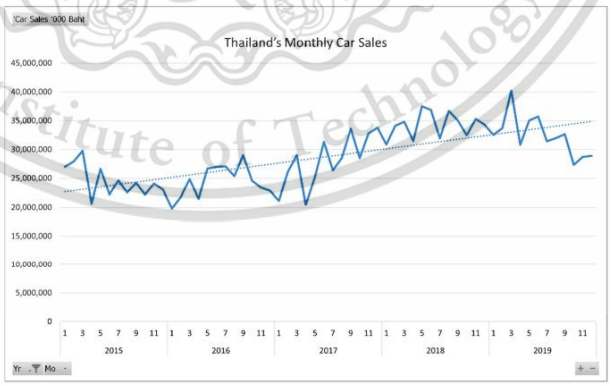

Figure 2. Thailand's monthly car sales from January 2015 to December 2019 with a trendline.

Furthermore, to gain even higher accuracy forecasts, the seasonal component in the car sales data is also integrated into the proposed modifications. In total, the typical Holt's method, along with the three modified methods, namely, Holt's method with seasonality, Holt's method with events, and Holt's method with seasonality and events, will be used for Thailand's monthly car sales from January 2015 to December 2021. Then, their accuracies, using the mean absolute percentage error (MAPE) and the symmetric mean absolute percentage error (SMAPE), will be computed and compared.

2. Literature Review

Previous work regarding forecasting methods and the event component is surveyed and presented as follows.

Wirotcheewan et al. [9] found a suitable forecasting model for the advanced demand of the automotive wheels, parts, and accessories (WPA), and used a linear programming (LP) model to calculate the optimal quantity for export so as to gain from the maximum profit. The WPA demands for export from Thailand are collected from the top five highest-demand countries, i.e., Japan, China, South Korea, Germany, and Indonesia from 1997 to 2008. Time series forecasting models used in the study are naïve, moving average, single exponential smoothing, and exponential smoothing with trend or Holt's method and artificial neural networks (ANNs). The forecasting accuracy measurement is MAPE. The results show that the ANNs model outperforms the other models for Japan, Germany, and Indonesia, while exponential smoothing with trend is best suited for China. Nonetheless, all five models produce MAPEs greater than 50% for South Korea. After that, the forecasted results are used in finding the optimal quantities for export to those five countries using LP models.

Rattanametawee et al. [10] study the effects of special events on regression for subcompact car sales in Thailand. The monthly input data are collected from January 2005 to September 2015, summing to 129 data. The two special events considered are the 2011 Thailand flood and the government's tax-incentive first-car buyer program in 2011–2012. For the methodology, they use three different multiple linear regressions, specifically, the regular regression, regression containing seasonality, and the proposed regression containing seasonality and special events. The results show that the last regression outperforms the other two models and achieves the highest adjusted coefficient of determination or R-square and also the highest accuracy in terms of MAPE.

Booranawong and Booranawong [11] use double exponential smoothing (DES) or Holt's methods, multiplicative Holt–Winters (MHW) and additive Holt–Winters (AHW) with optimal initial values and smoothing constants to forecast lime, Thai chili and lemongrass prices in Thailand from October 2016 to December 2016. This research collects the input price data at the Simummuang market from January 2011 to September 2016. The accuracy measure for comparisons is MAPE. The results show that DES attains the smallest MAPEs for forecasting Thai chili and lemongrass prices, while MHW and AHW generate smaller MAPEs than that of DES for forecasting lime prices possessing the seasonal component.

Muchayan [12] uses two different double exponential smoothing methods, namely, Brown's and Holt's, to predict the net asset value (NAV) price movements of the Cipta Ovo Equitas mutual fund from the Ciptadana Asset Management, PT in Indonesia. The NAV price data are collected over the period January 2019 to January 2020. The measurement of effectiveness is MAPE. The results show that Holt's method yields a smaller MAPE than Brown's method.

Sharif and Hasan [13] develop a stock indicator that helps buyers predict the next day's share value by applying the Holt's method. The stock closing prices of different companies are gathered from the Dhaka Stock Exchange (DSE) in 2016. This study finds that the Holt's method is appropriate for short-term prediction. In addition, this study shows that different smoothing constants have an impact on the prediction values and the proper values of smoothing constants for this dataset are $\alpha = 0.5$ and $\beta = 0.1$.

Suppalakpanya et al. [14] study several exponential smoothing methods for forecasting crude palm oil productions in Thailand from January 2018 to March 2018. The monthly crude palm oil productions data are collected from the database of the Department of Internal Trade, Ministry of Commerce, Thailand. This research compares five different forecasting methods for various ranges of the input data. The first three methods are double exponential smoothing (DES), the multiplicative Holts–Winters (MHW), and the additive Holt–Winters (AHW) methods. Additionally, their proposed modified methods are the improved additive Holts–Winters (IAHW) and the extended additive Holts–Winters (EAHW) methods. For the input ranges, they implement all the five methods on four different ranges, i.e., 3-year data (2015–2017), 6-year data (2012–2017), 9-year data (2009–2017), and 12-year data (2006–2017). The accuracy measurement is MAPE. The results show that the AHW and EAHW methods yield the two lowest MAPEs of 6.94 and 7.05, respectively, when the 12-year data are applied.

Rattanametawee and Leenawong [15] propose a new time-series decomposition to incorporate the effects of special events that impact the dataset. They use a case study of subcompact monthly car sales data in Thailand from 2011 to 2018. This dataset has the conventional trend, seasonal, and cyclical components. The special events investigated in the study are the tax-incentive program for first car buyers having a positive impact on the data, and the national big flood having, in contrast, a negative impact. MAPE is used as an accuracy measure of the proposed forecasting method, resulting in a relatively low value at 8.17%.

With regard to the organization of this paper, in the next section, the typical Holt's method along with the modified methods are explained in detail and their formulas are shown explicitly, including the accuracy measurement. In Section 4, the experimental results and their discussion are provided. Then, in the final Section 5, this research is concluded.

3. Materials and Methods

In this section, all forecasting methods are explicitly described to be implemented on Thailand's monthly car sales data from January 2015 to December 2021. The data are collected from the Office of Industrial Economics, Ministry of Industry, Thailand [16]. Note that all the experiments are conducted on Microsoft Excel 365.

3.1. The Typical Holt's Method

First of all, common notations to be used in this typical Holt's method and its subsequent modified methods are defined as follows.

A_t represents the actual data for period t,
α represents the smoothing constant for the level estimate; $0 \le \alpha \le 1$,
and β represents the smoothing constant for the trend estimate; $0 \le \beta \le 1$.

The typical Holt's forecasting method can be executed by the following three steps [17].

Step 1: Computing the Level Estimate.

$$L_t = \alpha A_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad (1)$$

where L_t refers to the level estimate for period t, T_t refers to the trend estimate for period t, and the initial value for L_t is $L_2 = A_2$, and for T_t is $T_2 = A_2 - A_1$.

Step 2: Computing the Trend Estimate.

$$T_t = \beta (L_t - L_{t-1}) + (1 - \beta) T_{t-1} \quad (2)$$

Step 3: Computing the Holt's Estimate.

$$H_{t+m} = L_t + mT_t \quad (3)$$

where H_{t+m} refers to Holt's forecasted value for period t + m, and m refers to the future period m^{th}; $m \geq 1$.

In the next section, the forecasting steps of the Holt's method with seasonality will be described.

3.2. The Holt's Method with Seasonality

The forecasting steps of the typical Holt's method are modified to take into consideration the seasonality in the data. Some steps are changed and two additional steps are inserted to account for the seasonal component and it, thus, comes to a total of five steps. Note that certain notations used in some steps of the typical Holt's method are slightly altered to reflect the seasonality in the data as well. In conclusion, the Holt's method with seasonality can be executed by the following five main steps.

Step 1 (Additional): Dealing With Seasonality.

In this new step, for the original data with a seasonal component, it is advised that the seasonal component is first removed or de-seasonalized before performing Holt's method. Then, after obtaining the forecasted values from Holt's method, these Holt's forecasts must be incorporated back or re-seasonalized in the very last stop to reach the final forecasts. In doing so, seasonal indices are required and they are determined by the following sub-steps, 1.1–1.5. In the last sub-step 1.6, the seasonal indices are used to remove the seasonal component from the original data.

Step 1.1: Finding the Moving Averages (MAs).

$A_{y,m}$ represents the actual data of year y, month m, where y = 1, 2, ..., 7 refer to the years 2015, 2012, …, 2021 and m = 1, 2, ..., 12 refer to the months January, February, ..., and December, respectively, and $\overline{A}_{y,m}$ represents the moving average of year y, month m, when the number of periods to average is 12.

Therefore, the values start at $\overline{A}_{1,6}$ up until $\overline{A}_{7,6}$. More explicitly [18],

$$\overline{A}_{1,6} = \left(A_{1,1} + A_{1,2} + \ldots + A_{1,12}\right)/12,$$
$$\overline{A}_{1,7} = \left(A_{1,2} + A_{1,3} + \ldots + A_{1,12} + A_{2,1}\right)/12,$$
$$\overline{A}_{1,8} = \left(A_{1,3} + A_{1,4} + \ldots + A_{1,12} + A_{2,1} + A_{2,2}\right)/12,$$
$$\vdots$$
$$\overline{A}_{7,6} = \left(A_{7,1} + A_{7,2} + \ldots + A_{7,11} + A_{7,12}\right)/12.$$

Step 1.2: Finding the Centered Moving Averages (CMAs).

$\overline{C}_{y,m}$ represents the centered moving average of year y, month m, when the number of periods to average is 2.

Therefore, the values start at $\overline{C}_{1,7}$ up until $\overline{C}_{7,6}$. More explicitly [19],

$$\overline{C}_{1,7} = \left(\overline{A}_{1,6} + \overline{A}_{1,7}\right)/2,$$
$$\overline{C}_{1,8} = \left(\overline{A}_{1,7} + \overline{A}_{1,8}\right)/2,$$
$$\overline{C}_{1,9} = \left(\overline{A}_{1,8} + \overline{A}_{1,9}\right)/2,$$
$$\vdots$$
$$\overline{C}_{7,6} = \left(\overline{A}_{7,5} + \overline{A}_{7,6}\right)/2.$$

Step 1.3: Computing the Seasonal Factors.

$SF_{y,m}$ represents the seasonal factor of year y, month m, computed by the following formula:

$$SF_{y,m} = A_{y,m} / \overline{C}_{y,m} \quad (4)$$

Step 1.4: Computing the Unscaled Seasonal Indices.

SI_m represents the unscaled seasonal index of month m, that is,

SI_1 = the average of $SF_{2,1}$, $SF_{3,1}$, …, $SF_{7,1}$,
SI_2 = the average of $SF_{2,2}$, $SF_{3,2}$, …, $SF_{7,2}$,
SI_3 = the average of $SF_{2,3}$, $SF_{3,3}$, …, $SF_{7,3}$,
⋮
SI_{11} = the average of $SF_{1,11}$, $SF_{2,11}$, …, $SF_{6,11}$,
SI_{12} = the average of $SF_{1,12}$, $SF_{2,12}$, …, $SF_{6,12}$.

Step 1.5: Computing the (Scaled) Seasonal Indices.

S_t represents the scaled seasonal index for period t when t = 1, 2,…, 84. Since the number of periods in the seasonality cycle is 12,

$$S_t = S_{t+12(1)} = S_{t+12(2)} = \ldots,$$

more explicitly,

$$S_1 = S_{13} = \ldots = S_{61} = S_{73} = SI_1 \times 12 / \sum_{m=1}^{12} SI_m,$$

$$S_2 = S_{14} = \ldots = S_{62} = S_{74} = SI_2 \times 12 / \sum_{m=1}^{12} SI_m,$$

⋮

$$S_{11} = S_{23} = \ldots = S_{71} = S_{83} = SI_{11} \times 12 / \sum_{m=1}^{12} SI_m,$$

$$S_{12} = S_{24} = \ldots = S_{72} = S_{84} = SI_{12} \times 12 / \sum_{m=1}^{12} SI_m.$$

Step 1.6: Removing Seasonality From the Data (De-seasonalization).

In this step, to remove seasonality from the data, each of the actual data are divided by each corresponding seasonal index. From this step onward, because the year y and month m are irrelevant, $A_{y,m}$ can become just the actual data at period t, when t = 1, 2,…, 84 (notationally, A_t, for simplicity).

D_t represents the de-seasonalized data at period t, computed from the actual data divided by the relative seasonal index, that is,

$$D_t = A_t / S_t. \quad (5)$$

Steps 2 to 4: Computing Level, Trend, and Holt's Estimates

Similar to the typical Holt's steps 1–3, the computations for the level, trend, and Holt's estimates can be determined with the replacement of A_t by D_t.

Step 5 (Additional): Computing the Holt's Estimate With Seasonality (Re-seasonalization).

This additional and last step multiplies each Holts' estimate from step 4 with the corresponding seasonal index. F_{t+m} represents the final forecasted data for period t + m, re-seasonalized from the Holt's estimation,

$$F_{t+m} = H_{t+m} \times S_{t+m}. \quad (6)$$

In the next section, the forecasting steps of the modified Holt's method that takes into account the event component will be described.

3.3. The Holt's Method with Events

For the proposed Holt's method that incorporates the event component, or more precisely, in our case, the global COVID-19 pandemic, the forecasting steps of the typical Holt's method are modified. Some steps are changed and one additional step is inserted in order to account for the event component and it, thus, comes to a total of four steps.

Note that certain notations used in some steps of the typical Holt's method are slightly altered to reflect the event component in the data as well.

In addition to the two smoothing constants previously defined in Holt's method, one more smoothing constant for the event estimate is then defined as δ and $0 \le \delta \le 1$. Only the changes in the forecasting steps of the typical Holt's method, plus one additional step, are addressed in detail as follows. Thus, the Holt's method with events can be executed by the following four steps.

Steps 1 to 2: Computing Level and Trend Estimates.

These steps are exactly the same as steps 1 and 2 in the typical Holt's method.

Step 3 (Additional): Computing the Event Estimate.

This new step is inserted in order to add another smoothing constant, δ, for the event component. Thus, a recursive formula with an initial value is needed.

The formula for event estimates is as follows:

$$\begin{aligned} E^k{}_t &= 1 \ ; k = 0, \\ E^k{}_t &= \delta\,(A_t / L_t) + (1 - \delta)\,E^k{}_{t-} \ ; k = 1, 2, 3, \end{aligned} \tag{7}$$

where $E^k{}_t$ refers to the event factor for data with flag k at period t, flag k = 0 represents the normal sales period, k = 1 represents the panic-state, lockdown, COVID-19 superspreading wave-1 period, k = 2 represents the COVID-19 relief period after any wave, and k = 3 represents the period in which any later COVID-19 superspreading wave occurs, and $E^k{}_{t-}$ refers to the last occurrence prior to period t of the event factor having the same flag k.

The idea behind the above event formula update is that, except for flag k = 0 having the event estimate fixed at 1, the smoothing constant for the event component, δ, identifies where the current event estimate with flag k should be, between the latest event factor, A_t/L_t, and the previous event factor with the same flag k, $E^k{}_{t-}$.

As for the initial values of $E^k{}_t$ when k = 1, 2, 3, two mutually exclusive cases are as follows:

Case 1: when t > 1, the initial value is $E^k{}_1 = 1$.

Case 2: when $t \neq 1$, the initial value is $E^k{}_t = E_{t-1}$, where E_{t-1} refers to the event factor immediately preceding the current period t.

Step 4: Computing the Holt's Estimate With Events.

Modified from the counterpart step 3 of the typical Holt's method to include the multiplicative effect of the event component, the new HE_{t+m} representing the Holt's forecasted value with the event component for period t + m is then computed by the following formula:

$$HE_{t+m} = (L_t + mT_t)E^k{}_{t+m}, \tag{8}$$

where m = the future period m^{th}; $m \ge 1$.

In the next section, the forecasting steps of the modified Holt's method that takes into account both seasonal and event components will be described.

3.4. The Holt's Method with Seasonality and Events

In this section, the typical Holt's method is modified to encapsulate the associated seasonal and event components. Forecasting steps of the Holt's methods with seasonality and with events from the above sections are combined here with some modification. To deal with seasonality and events simultaneously, a set of notations from Holt's method with seasonality to include flag k, as defined in Holt's method with events above, is altered to the following new set of notations:

$$A_{y,m} \text{ becomes } A^k{}_{y,m}, \quad \bar{A}_{y,m} \text{ becomes } \bar{A}^k_{y,m},$$

$\bar{C}_{y,m}$ becomes $\bar{C}^{k}_{y,m}$, $SF_{y,m}$ becomes $SF^{k}_{y,m}$.

Step 1: Dealing With Seasonality.

In this new main step 1, step 1 of Holt's method with seasonality is modified in detail as follows.

Steps 1.1 to 1.3: Computing MAs, CMAs, and the Seasonal Factors.

From Holt's method with seasonality, steps 1.1 to 1.3 remain exactly the same after the change in $A_{y,m}$, $\bar{A}_{y,m}$, $\bar{C}_{y,m}$, and $SF^{k}_{y,m}$ to $A^{k}_{y,m}$, $\bar{A}^{k}_{y,m}$, $\bar{C}^{k}_{y,m}$, and $SF^{k}_{y,m}$, respectively.

Step 1.4: Computing the Unscaled Seasonal Indices of the Normal Sales Periods Only.

From step 1.4 of Holt's method with seasonality, it is modified to exclude the sales data during the events other than the normal period, i.e., flags k = 1, 2, 3, because the seasonal component should not be affected by abnormal factors, such as the events. Afterwards, the event component can be taken care of in the later step.

Similar to step 1.4 of Holt's method with seasonality, SI_m representing the unscaled seasonal index of month m is still computed from the average of the seasonal factors, but this time, only those with flag k = 0, that is,

$$SI_m = \text{the average of } SF^{0}_{y,m}. \quad (9)$$

Step 1.5: Computing the (Scaled) Seasonal Indices of the Normal Sales Periods Only.

Even though the formula for S_t is still the same as shown in step 1.5 of Holt's method with seasonality, S_t now represents the scaled seasonal index for period t with flag k = 0 only. This is because S_t is practically the SI_m scaled to sum up to 12 and the new SI_m formula in step 1.4 above is computed only from the seasonal factors with flag k = 0.

Step 1.6: De-seasonalization.

In this step, the formula for the de-seasonalized data at period t, D_t, from step 1.6 of Holt's method with seasonality, stays unchanged. However, D_t here is seasonally removed by the seasonality from the normal sales periods only, not from all the sales periods, as D_t in Holt's method with seasonality.

Steps 2 to 3: Computing Level and Trend Estimates.

These main steps 2 and 3 are exactly the same as those of Holt's method with seasonality after the replacement of A_t with D_t.

Steps 4 to 5: Computing the Event Estimate and Holt's Estimate With Events.

For these steps, the counterpart steps 3 and 4 of Holt's with events are adopted without any changes. More precisely, the event component has now been fully incorporated into the Holt's estimate, becoming HE_{t+m}.

Step 6: Computing the Holt's Estimate With Events and Seasonality (Re-seasonalization).

This step is comparable to step 5 of Holt's method with seasonality, except that the Holt's estimate, H_{t+m}, is now the Holt's estimate with events, HE_{t+m} to be used in the re-seasonalization process. The final forecasted value for period t + m incorporating both events and seasonality, FE_{t+m}, can then be obtained by

$$FE_{t+m} = HE_{t+m} \times S_{t+m}. \quad (10)$$

3.5. The Accuracy Measurement

In this research, to compare among the different methods, the forecasted results are measured by MAPE and SMAPE. Their formulae are expressed as follows [16].

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\left|A_t - F_t\right|}{\left|A_t\right|} \times 100\%\right) \quad (11)$$

and

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\left|A_t - F_t\right|}{\left(\left|A_t\right| + \left|F_t\right|\right)/2} \times 100\%\right) \quad (12)$$

4. Results and Discussion

In this section, the four forecasting methods from the previous section are implemented on the case study of Thailand's car sales data, according to the corresponding steps of each method. In addition, the results from all methods are compared and discussed.

4.1. Implementation of the Holt's Method with Seasonality and Events on the Car Sales Data

The typical Holt's method and its three modified ones, i.e., the Holt's method with seasonality, the Holt's method with events, and the Holt's method with seasonality and events, are implemented on Thailand's monthly car sales data from January 2015 to December 2021 using Microsoft Excel 365. Nonetheless, to save space, only the Holt's method with seasonality and events is demonstrated here, following the steps from Section 3.4.

In Table 1, period, Yr, and Mo refer to the t^{th} data period, the year, and month, respectively. k refers to the assigned data flag taking the values of 0, 1, 2, and 3, corresponding to different periods of the pandemic. The flag assignment for these car sales data considers the unpredictable pandemic waves and the government prevention policies that reflect the number of infected cases. Car sales '000 baht refers to the actual car sales data in thousand Thai baht. Next, computation steps start with the main step 1 (dealing with seasonality) from Section 3.4. It consists of six sub-steps for computing the moving averages (MA and CMA), finding the seasonal indices (SF, SI unscaled, SI), and also removing the seasonality from the car sales data.

After the first three sub-steps from Section 3.4, different seasonal factors (SF) are obtained for different periods. The next two sub-steps are then performed to obtain the unscaled (SI unscaled) and the scaled seasonal indices (SI) for corresponding the months, equivalent to seasons. As a reminder, these SI unscaled and SI are averages of SFs over the normal sales periods only, i.e., flag k = 0.

Afterwards, these seasonal indices (SI) are used in the final sub-step for removing the seasonal component in the sales data, resulting in the de-seasonalized car sales data (deseason).

In Table 2, deseason or the seasonally adjusted car sales are subsequently used in place of the actual car sales for computation in the remaining steps, that is, computing level (L), trend (T), and event (E) estimates in steps 2 to 4. However, to help simplify the calculation of the event estimate in step 4, Et- referring to E^{k}_{t-} is separately obtained beforehand. The Holt's estimate with events (HE) is then figured in step 5. Lastly, the final forecast (FE) is obtained by re-seasonalizing HE with the corresponding seasonal index SI.

Table 1. Dealing with seasonality in the car sales data for the Holt's method with seasonality and events.

Period	Yr	Mo	k	Car Sales ('000 Baht)	MA	CMA	SF	SI Unscaled	SI	Deseason
1	2015	1	0	26,977,961.53				0.9296	0.9178	29,395,602.24
2	2015	2	0	27,902,176.73				1.0643	1.0508	26,554,084.21
3	2015	3	0	29,774,248.84				1.0950	1.0811	27,541,580.65
4	2015	4	0	20,635,767.24				0.8765	0.8653	23,847,031.31
5	2015	5	0	26,625,638.70				1.0466	1.0333	25,768,470.42
6	2015	6	0	22,234,255.04	24,564,721.20			1.1011	1.0870	20,453,895.19
7	2015	7	0	24,605,387.66	23,967,856.58	24,266,288.89	1.0140	0.9903	0.9777	25,167,107.50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	0	26,502,129.77	23,917,024.70	24,333,568.57	1.0891	0.9296	0.9178	28,877,128.64
62	2020	2	0	32,657,258.86	23,595,908.11	23,756,466.40	1.3747	1.0643	1.0508	31,079,424.73
63	2020	3	0	26,042,182.41	23,481,330.89	23,538,619.50	1.1064	1.0950	1.0811	24,089,369.01
64	2020	4	1	6,960,649.53	23,962,827.37	23,722,079.13	0.2934	0.8765	0.8653	8,043,840.84
65	2020	5	1	9,204,254.83	24,722,380.43	24,342,603.90	0.3781	1.0466	1.0333	8,907,939.10
66	2020	6	1	14,674,115.15	25,217,869.42	24,970,124.93	0.5877	1.1011	1.0870	13,499,117.14
67	2020	7	1	21,454,176.54	25,070,934.52	25,144,401.97	0.8532	0.9903	0.9777	21,943,956.93
68	2020	8	2	28,104,687.94	24,771,931.36	24,921,432.94	1.1277	1.0040	0.9912	28,353,320.27
69	2020	9	2	31,278,826.07	25,133,393.68	24,952,662.52	1.2535	1.0758	1.0621	29,448,765.08
70	2020	10	2	33,105,584.68	26,234,010.91	25,683,702.29	1.2890	0.9447	0.9327	35,495,335.99
71	2020	11	2	37,805,573.95	26,902,365.10	26,568,188.00	1.4230	1.0143	1.0014	37,754,524.32
72	2020	12	2	34,824,993.37	27,635,278.36	27,268,821.73	1.2771	1.0126	0.9997	34,835,988.33
73	2021	1	3	24,738,910.91	27,648,750.56	27,642,014.46	0.8950	0.9296	0.9178	26,955,898.22
74	2021	2	2	29,069,220.98	26,938,095.62	27,293,423.09	1.0651	1.0643	1.0508	27,664,742.76
75	2021	3	2	30,379,730.20	26,384,499.74	26,661,297.68	1.1395	1.0950	1.0811	28,101,659.06
76	2021	4	3	20,168,056.27	25,893,933.88	26,139,216.81	0.7716	0.8765	0.8653	23,306,536.84
77	2021	5	3	17,224,505.15	25,237,711.02	25,565,822.45	0.6737	1.0466	1.0333	16,669,990.77
78	2021	6	3	23,469,074.27	25,047,140.56	25,142,425.79	0.9334	1.1011	1.0870	21,589,838.94
79	2021	7	3	21,615,842.91				0.9903	0.9777	22,109,314.01
80	2021	8	3	19,576,828.62				1.0040	0.9912	19,750,017.97
81	2021	9	3	24,635,675.60				1.0758	1.0621	23,194,291.94
82	2021	10	2	27,218,794.28				0.9447	0.9327	29,183,603.23
83	2021	11	2	29,930,899.65				1.0143	1.0014	29,890,483.35
84	2021	12	2	32,538,147.92				1.0126	0.9997	32,548,420.87

Table 2. The final forecasts of the Holt's method with seasonality and events.

Period	Yr	Mo	*k*	Deseason	L	T	Et-	E	HE	FE
1	2015	1	0	29,395,602.24				1.0000		
2	2015	2	0	26,554,084.21	26,554,084.21	−2,841,518.04	1.0000	1.0000		
3	2015	3	0	27,541,580.65	24,585,859.63	−2,418,254.32	1.0000	1.0000	23,712,566.17	25,634,833.91
4	2015	4	0	23,847,031.31	22,550,636.44	−2,232,608.62	1.0000	1.0000	22,167,605.31	19,182,494.35
5	2015	5	0	25,768,470.42	21,561,124.69	−1,630,110.35	1.0000	1.0000	20,318,027.83	20,993,891.35
6	2015	6	0	20,453,895.19	20,050,269.16	−1,572,310.49	1.0000	1.0000	19,931,014.34	21,665,861.30
7	2015	7	0	25,167,107.50	20,003,570.54	−832,884.20	1.0000	1.0000	18,477,958.67	18,065,537.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	2020	1	0	28,877,128.64	27,968,469.87	−928,188.07	1.0000	1.0000	27,699,998.74	25,421,812.91
62	2020	2	0	31,079,424.73	27,961,499.82	−481,696.51	1.0000	1.0000	27,040,281.80	28,413,057.51
63	2020	3	0	24,089,369.01	26,706,537.99	−856,479.06	1.0000	1.0000	27,479,803.31	29,707,463.49
64	2020	4	1	8,043,840.84	21,788,947.66	−2,824,799.13	1.0000	0.3692	9,543,084.10	8,258,003.25
65	2020	5	1	8,907,939.10	16,670,602.24	−3,936,424.22	0.3692	0.5344	10,133,495.95	10,470,578.87
66	2020	6	1	13,499,117.14	12,908,639.70	−3,851,866.96	0.5344	1.0457	13,316,675.08	14,475,792.87
67	2020	7	1	21,943,956.93	11,995,986.94	−2,427,302.62	1.0457	1.8293	16,567,326.39	16,197,550.25

68	2020	8	2	28,353,320.27	13,852,945.94	−350,827.11	1.8293	2.0467	19,584,568.67	19,412,830.15
69	2020	9	2	29,448,765.08	17,139,112.64	1,411,933.73	2.0467	1.7182	23,199,609.81	24,641,323.94
70	2020	10	2	35,495,335.99	22,415,575.33	3,284,975.22	1.7182	1.5835	29,375,807.40	27,398,058.15
71	2020	11	2	37,754,524.32	28,449,732.24	4,617,435.51	1.5835	1.3271	34,106,193.08	34,152,309.63
72	2020	12	2	34,835,988.33	33,470,587.34	4,812,962.99	1.3271	1.0408	34,416,111.62	34,405,249.19
73	2021	1	3	26,955,898.22	35,700,022.74	3,560,791.15	1.0408	0.7551	28,906,633.86	26,529,208.34
74	2021	2	2	27,664,742.76	36,616,067.26	2,278,947.95	1.0408	0.7555	29,662,943.02	31,168,865.47
75	2021	3	2	28,101,659.06	36,433,345.92	1,085,837.80	0.7555	0.7713	30,000,386.43	32,432,378.60
76	2021	4	3	23,306,536.84	34,277,667.74	−485,244.72	0.7551	0.6799	25,510,552.36	22,075,276.95
77	2021	5	3	16,669,990.77	29,887,264.59	−2,377,978.31	0.6799	0.5578	18,848,141.09	19,475,109.96
78	2021	6	3	21,589,838.94	26,159,222.26	−3,032,320.91	0.5578	0.8253	22,704,079.43	24,680,301.13
79	2021	7	3	22,109,314.01	22,894,817.52	−3,144,806.20	0.8253	0.9657	22,333,435.23	21,834,961.84
80	2021	8	3	19,750,017.97	19,750,012.84	−3,144,805.46	0.9657	1.0000	19,750,016.46	19,576,827.12
81	2021	9	3	23,194,291.94	18,107,997.33	−2,416,440.39	1.0000	1.2809	21,269,388.36	22,591,151.02
82	2021	10	2	29,183,603.23	18,768,723.64	−925,013.89	0.7713	1.5549	24,398,897.91	22,756,222.99
83	2021	11	2	29,890,483.35	20,591,249.28	406,650.48	1.5549	1.4516	25,902,124.83	25,937,148.28
84	2021	12	2	32,548,420.87	23,632,257.67	1,683,458.53	1.4516	1.3773	28,920,151.78	28,911,023.99

4.2. Numerical Results

Since the number of periods, or in this case months, is 12, the first MA, as shown in Table 1, starts at Period 6 and ends at Period 78. In addition, since the number of periods, 12, is even, the MA should be centered by taking another average over the two consecutive periods. The first CMA then starts at Period 7 and ends at Period 78.

Consequently, SF, being car sales divided by CMA, starts at Period 7 and ends at Period 78 as well. One can observe that SFs are different throughout all the time periods. However, for each month with flag $k = 0$, when SF is averaged, the resulting SI unscaled, as reported in Table 1, is identical for the successive years. For example, as shown in Table 1, SI unscaled for Mo 1, Yr 2015 is 0.9296, equivalent to those of Mo 1, Yr 2020 and Mo 1, Yr 2021. All 12 SI unscaled values are summarized in Table 3.

Even though each SI unscaled seems valid as an indicator of how much the car sales are in each month, the seasonality requires that the summation of all seasonal indices must equal the number of seasons, or 12 in our case. The SI unscaled summed to 12.1547 is, therefore, scaled to SI to sum up to 12, as shown in Table 3. Evidently by SI, the car sales were the highest in June with SI = 1.0870, most decreased in April with SI = 0.8653, and varied in other months. It is, thus, suggested that the seasonal effects are eliminated from the car sales data before proceeding further with the forecasting.

Table 3. Summary of each month's SI unscaled and SI.

Mo	SI Unscaled	SI
1	0.9296	0.9178
2	1.0643	1.0508
3	1.0950	1.0811
4	0.8765	0.8653
5	1.0466	1.0333
6	1.1011	1.0870
7	0.9903	0.9777
8	1.0040	0.9912
9	1.0758	1.0621
10	0.9447	0.9327
11	1.0143	1.0014
12	1.0126	0.9997
Sum	12.1547	12.0000

The deseason column in Table 1 shows the de-seasonalized car sales data calculated from the actual car sales divided by SI. So, deseason will be treated as the new car sales with no seasonality left in the data. Then, the Holt's method with events can be applied to this deseason. As a result, the HE column in Table 2 reports the car sales forecasts that already include the event effects, but not the seasonal effects yet. After multiplying each HE by each corresponding SI to incorporate back the seasonality, the last FE column in Table 2 reveals the desired final forecasts.

4.3. Forecasting Accuracy Comparison and Discussion

Similar to FE in Table 2, the final forecasts from all other methods can be obtained. Table 4 summarizes and compares MAPEs and SMAPEs of the typical Holt's method, denoted by Holt, with the three modified Holt's method, i.e., Holt's with seasonality, Holt's with events, and Holt's with seasonality and events, denoted by Holt S, Holt E, and Holt SE, respectively. Both MAPEs and SMAPEs of all four methods are also visualized by bar chart in Figure 3.

According to Table 4 and Figure 3, among all four methods, Holt performs the worst forecast on Thailand car sales data covering the COVID-19 pandemic period, in terms of both accuracy measures, possessing the highest MAPE and SMAPE at 16.27% and 13.91%, respectively. An improvement from Holt can be observed with the seasonality addressed in Holt S with lower MAPE and SMAPE at 12.37% and 11.99%. An even higher improvement from Holt can be obtained through Holt E, implying that, in the pandemic case, the event effects are stronger on these car sales data than the seasonal effects. Furthermore, the method combining both seasonal and event effects, Holt SE, achieving the lowest MAPE and SMAPE at 8.64% and 8.90%, respectively, indicates that this proposed modified Holt's method best fits Thailand car sales data during the COVID-19 pandemic.

Table 4. Accuracy comparison.

Method	MAPE	SMAPE
Holt	16.27%	13.91%
Holt S	12.37%	11.99%
Holt E	9.47%	9.33%
Holt SE	8.64%	8.90%

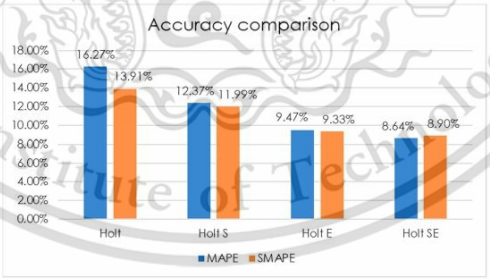

Figure 3. Bar chart of the forecasting accuracy.

5. Conclusions

This study aimed to modify the typical Holt's forecasting method to better suit time series data containing the event component, defined here as unusual impacts for a certain time period, largely caused by irregular incidents or events, such as the COVID-19 global pandemic.

As previously mentioned in the literature review section, there were previous attempts on capturing the effects of such irregular incidents into some forecasting models. More precisely, one attempt works with the regression method [10], while another works with the time-series decomposition (TSD) method [15]. The events of their interest are the2011 Thailand flood, along with the government's tax-incentive program. Notably, their methods are completely different from what is proposed here because, unlike Holt's method, both regression and TSD formulas are not recursive, nor need any smoothing constants.

Regardless, besides the typical Holt's method, three modified methods based on Holt's method including the seasonal and/or event components are proposed and named accordingly as the Holt's method with seasonality, the Holt's method with events, and the Holt's method with seasonality and events. The methods with events incorporate another smoothing constant for the event estimate, denoted by δ and $0 \leq \delta \leq 1$. As for the methods with seasonality, the actual sales data are first de-seasonalized prior to being dealt with in the next steps.

All these four methods are implemented on Thailand's monthly car sales data from January 2015 to December 2021, which also include the COVID-19 pandemic period. The input data obtained from the Office of Industrial Economics, Ministry of Industry, Thailand are flagged with different numbers to refer to various stages of the time period affected by the event of interest, i.e., the pandemic.

In terms of forecasting accuracy, the experimental results show that the Holt's method with seasonality and events best fits the Thailand's car sales data, possessing the lowest MAPE and SMAPE, among all the four methods. It must be noted that even the Holt's method with events alone can also yield second best accuracy, much better than those of the typical Holt's and the Holt's with seasonality methods.

Beyond that, the modified Holt's methods proposed in this research can certainly be applied to time series forecasting for other data largely impacted by abnormal events other than the COVID-19 pandemic. As evidenced here, the incorporation of another constant for the event estimate into Holt's method significantly improves the forecasting accuracy; thus, the forecasted results can be used further in all kinds of planning for the business to stay competitive and beyond.

Last but not least, it would be interesting to compare the proposed methods with other time-series models. However, to ensure a fair comparison, a subtle way to incorporate the event component into the model of choice is, most likely, mandatory, and this is currently an ongoing project.

Author Contributions: Conceptualization, C.L. and T.C.; methodology, C.L. and T.C.; software, C.L. and T.C.; validation, C.L. and T.C.; formal analysis, C.L.; investigation, C.L. and T.C.; resources, C.L. and T.C.; data curation, C.L. and T.C.; writing—original draft preparation, C.L. and T.C.; writing—review and editing, C.L. and T.C.; visualization, C.L. and T.C.; supervision, C.L.; project administration, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available at the following link: https://bit.ly/3QEHls6 accessed on 23 May 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ASEAN Briefing. Available online: https://www.aseanbriefing.com/news/thailands-automotive-industry-opportunities-incentives/ (accessed on 29 April 2022).

2. CNN BUSINESS. Available online: https://money.cnn.com/2017/02/20/autos/traffic-rush-hour-cities/index.html (accessed on 29 April 2022).
3. Mashable. Available online: https://mashable.com/article/bangkok-traffic-jams (accessed on 29 April 2022).
4. Focus2move. Available online: https://www.focus2move.com/thailand-best-selling-car/#:~:text=Thailand%27s%20best%2Dselling%20car%20ranking,units%20sold%20(%2D4.8%25) (accessed on 29 April 2022).
5. World Health Organization. Available online: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON234 (accessed on 28 April 2022).
6. World Health Organization. Available online: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems (accessed on 28 April 2022).
7. The World Bank. Available online: https://www.worldbank.org/en/country/thailand/publication/monitoring-the-impact-of-covid-19-in-thailand#:~:text=Income%3A,income%20groups%20experiencing%20income%20%20declines (accessed on 28 April 2022).
8. Leenawong, C. *Logistics Intelligence and Forecasting with Excel 365*; KMITL: Bangkok, Thailand, 2022; pp. 89–90.
9. Wirotcheewan, P.; Kengpol, A.; Ishii, K.; Shimada, Y. Modelling and Forecasting for Automotive Parts Demand of Foreign Markets on Thailand. *Asian Int. J. Sci. Technol. Prod. Manuf. Eng.* **2011**, *4*, 1–13.
10. Rattanametawee, W.; Leenawong, C.; Netisopakul, P. The Effects of Special Events on Regression for Subcompact Car Sales in Thailand. *J. Teknol. (Sci. Eng.)* **2016**, *78*, 161–165. http://doi.org/10.11113/.v78.9113.
11. Booranawong, T.; Booranawong, A. Double exponential smoothing and Holt-Winters methods with optimal initial values and weighting factors for forecasting lime, Thai chili and lemongrass prices in Thailand. *Eng. Appl. Sci. Res.* **2018**, *45*, 32–38. http://doi.org/10.14456/easr.2018.5.
12. Muchayan, A. Comparison of Holt and Brown's Double Exponential Smoothing Methods in The Forecast of Moving Price for Mutual Funds. *J. Appl. Sci. Eng. Technol. Educ.* **2019**, *1*, 183–192. https://doi.org/10.35877/454RI.asci1167.
13. Sharif, O.; Hasan, M.Z. Forecasting the Stock Price by using Holt's Method. *Indones. J. Contemp. Manag. Res.* **2019**, *1*, 15–24. https://doi.org/10.33455/ijcmr.v1i1.8.
14. Suppalakpanya, K.; Nikhom, R.; Booranawong, T.; Booranawong, A. Study of Several Exponential Smoothing Methods for Forecasting Crude Palm Oil Productions in Thailand. *Curr. Appl. Sci. Technol.* **2019**, *19*, 123–149. http://doi.org/10.14456/easr.2019.6.
15. Rattanametawee, W.; Leenawong, C. Event Index Computation for Forecasting Case Study: Car Sales in Thailand. *Thai J. Math.* **2020**, *18*, 2079–2091.
16. The Office of Industrial Economics, Ministry of Industry, Thailand. Available online: https://indexes.oie.go.th/industrialStatistics1.aspx (accessed on 3 August 2021).
17. Leenawong, C. *Data Analytics with Excel for Logistics & Supply Chain Management*; CU press: Bangkok, Thailand, 2022; p. 47.
18. NIST. Available online: https://www.itl.nist.gov/div898/handbook/ (accessed on 6 February 2022).
19. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018; p. 2.

Appendix B

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.

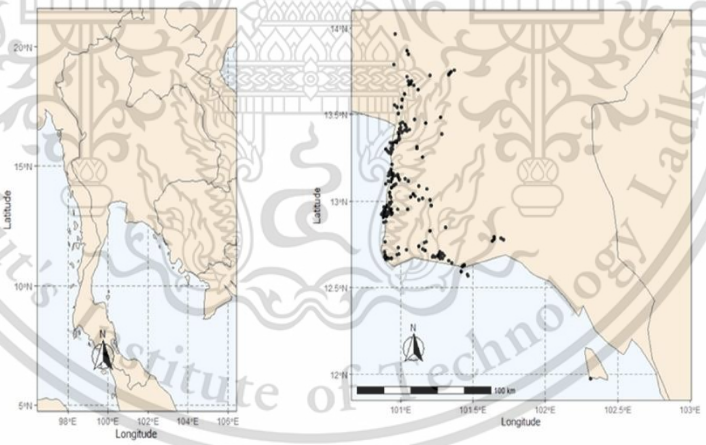

Figure 1. Thailand map and the convenience store locations in Eastern Thailand

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 economics 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.

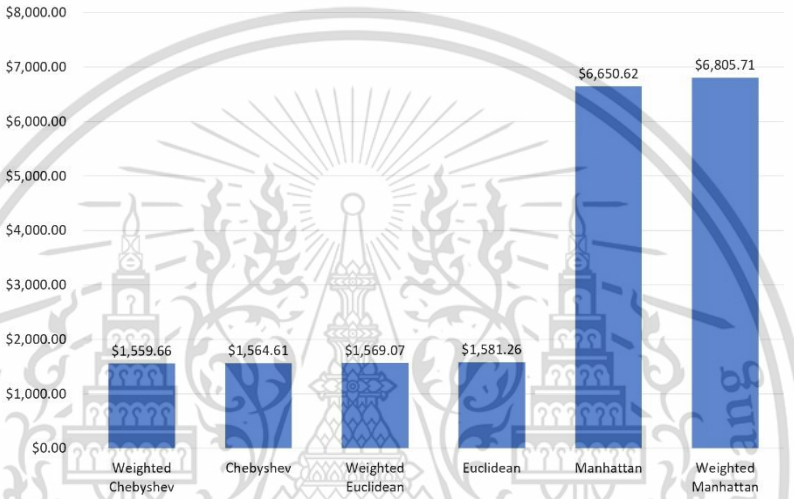

Figure 2. Bar chart of the expected distribution costs over 10,000 instances from each different clustering approach

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.

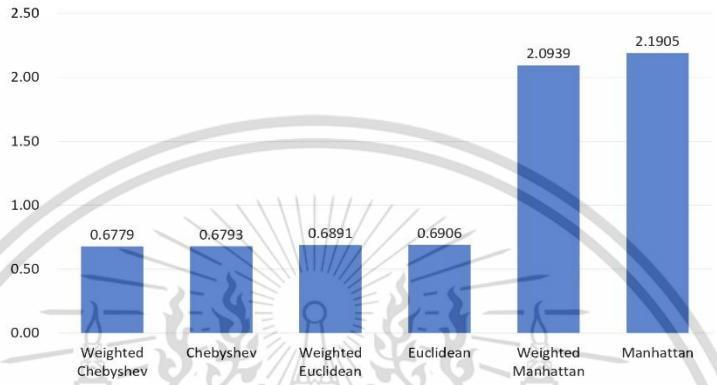

Figure 3. Bar chart of the expected DBI over 10,000 instances from each different clustering approach

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.

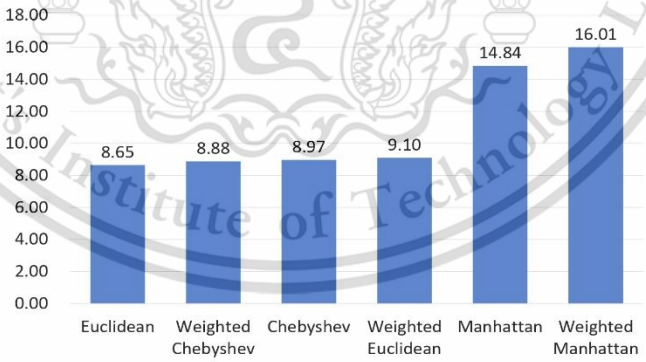

Figure 4. Bar chart of the expected number of iterations to the final clusters over 10,000 instances from each different clustering approach

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

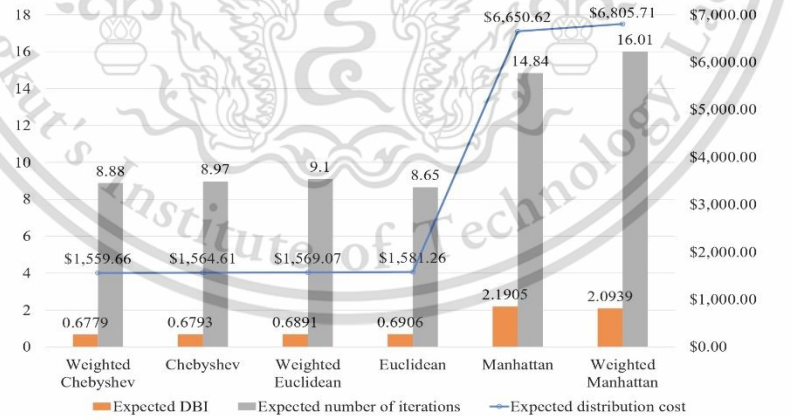

Figure 5. Graphical summary of the effectiveness and efficiency results from all six different clustering approaches

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

Author Biography

Name	Miss Thanrada Chaikajonwat
Date of Birth	6 February 1996
Address	Lam Luk Ka, Pathum Thani
Education	(2017) Bachelor of Science in Mathematics GPA 3.38 Srinakharinwirot University.
	(2019) Master of Science in Statistics GPA 3.69 Kasetsart University.
Scholarship	Science Achievement Scholarship of Thailand (SAST)

Academic Publications

1. Leenawong, C. and Chaikajonwat, T. 2022. "Event Forecasting for Thailand's Car Sales during the COVID-19 Pandemic." *Data*. 7(86) : 1-14.
2. Leenawong, C. and Chaikajonwat, T. 2023. "Modified K-means Clustering for Demand-weighted Locations Case Study: A Thailand's Convenience Store Franchise." *Pertanika Journal of Science & Technology*. 31(2) : 655-670.