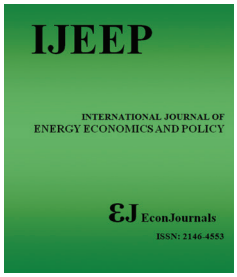
AbuBaker, Maher

Article

Household electricity load forecasting toward demand response program using data mining techniques in a traditional power grid

Provided in Cooperation with:
International Journal of Energy Economics and Policy (IJEEP)

Reference: AbuBaker, Maher (2021). Household electricity load forecasting toward demand response program using data mining techniques in a traditional power grid. In: International Journal of Energy Economics and Policy 11 (4), S. 132 - 148. https://www.econjournals.com/index.php/ijeep/article/download/11192/5900. doi:10.32479/ijeep.11192.



International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2021, 11(4), 132-148.

Household Electricity Load Forecasting Toward Demand Response Program Using Data Mining Techniques in a Traditional Power Grid

Maher AbuBaker*

An-Najah National University, Nablus, Palestine. *Email: abubaker@najah.edu

Received: 13 February 2021 **Accepted:** 28 April 2021 **DOI:** https://doi.org/10.32479/ijeep.11192

ABSTRACT

At present, the continuous increase of household electricity demand is strategic and crucial in electricity demand management. Household electricity consumers can play an important role in this issue. The rationalization of electricity consumption might be achieved by using an efficient Demand Response (DR) program. In this paper a new methodology is suggested using a combination of data mining techniques namely K-means clustering, K-Nearest Neighbors (K-NN) classification and ARIMA for electricity load forecasting using consumers' electricity prepaid bills data set of an ordinary electricity grid with prepaid electricity meters. As a result of applying this methodology, various DR programs are recommended as an attempt to assist the management of electricity system to manage the electricity demand issues from demand-side in an efficient and effective manner, which can be put into practice. A case study has been carried out in Tulkarm District, Palestine. The performance of applying the suggested methodology is measured, and the results are considered very well.

Keywords: Demand Response, K-means Clustering, K-Nearest Neighbor Classification, ARIMA Model, Prepaid Electricity Meters
JEL Classifications: Q4, Q41, Q47, Q49

1. INTRODUCTION

1.1. Background

Improvement of the electricity management system is necessary to allow effective and efficient management of electricity distribution in Palestine (West Bank and Gaza Strip). Palestine relies on external sources of electricity supply mainly from Israel. According to the Palestinian Central Bureau of Statistics in 2017 (PCBS, 2017), the quantity of electricity imported and purchased in Palestine nearly 92% of supply comes from the Israeli Electricity Company (IEC). Palestinian territories face significant energy security challenges as a result of the limitations of electricity supply quantities and the complete control of electricity pricing by IEC. The IEC power supply to West Bank begun experiencing power shortages during peak winter and summer months. Actually, rolling blackouts are the only available solution by IEC to rationing the limited power supply (World Bank Group, 2016). Rationalization in household electricity consumption is very important and mandatory. Rationalization does not mean not using or minimizing electrical appliances, but optimizing the use of electricity in the correct, safe and secure ways. Therefore, it contributes to improve the quality of service and participates in meeting the need for significant growth in residents, industrial firms, agricultural farms, and companies. The day by day increase in electricity demand is increasing the importance of energy efficiency through the efficient system operation (Seunghyeon et al., 2017). Many studies tried to solve the problem of increasing the energy efficiency from demand (customer) side management, while others tried to solve it from supplier side management (Palensky and Dietrich, 2011; Wang et al., 2014; Divshali and Choi, 2016; Seunghyeon et al., 2017). In this study the author trying to solve this problem from the demand side because the utility providers in the Palestinian

territories have no control over supply side management. Tulkarm Municipality (TM) is the only utility provider in Tulkarm district. It is taken as a sample for this study. TM relies completely on a conventional ordinary electricity grid using electricity prepaid meters. The complexity of this study that it depends on an offline data set of electricity consumption, unlike other studies, which are depending on online two-ways (data and information) electricity smart gird (Gharavi and Ghafurian, 2011; Fang et al., 2012; Cardenas et al., 2014; Wang et al., 2015; 2016). TM electricity consumers' prepaid bills (ECPB) data is the only available source of electricity consumption data in TM (See Appendix A). Two years ECPB sample data set for the years 2018 and 2019 are used in this study. Smart grids and smart metering infrastructure enable the generation and storing of a massive load data with a temporal measurement of 15 min (Lu et al., 2019). For conventional electricity billing, the hidden value of smart meter readings is detected by using data mining techniques such as data cleaning, preparation, compression, clustering, forecasting, and so on so forth (Wang et al., 2015).

1.2. Study Objectives

The main aim of this study is to propose a methodology of household electricity demand forecasting using the ECPB data set. This methodology proposes a combination of data mining and statistical techniques such as K-means clustering, autoregressive integrated moving average (ARIMA) model, and K-Nearest Neighbors (K-NN) classification algorithm. It is a hybrid model comprising of clustering technique (K-means) and ARIMA. Power load (demand) forecasting in the short-term for months, weeks, or shorter is more accurate than long-term load forecasting (Fan et al., 2019). K-means clustering main objective is to make electricity consumers' segmentation. It is used to produce clustered weekly electricity consumers load data by dividing weekly electricity consumers load data into a collection of similar weekly load data called clusters. It is used due to its mathematical ideas' simplicity, fast convergence and easy implementation (Xiao-Yu et al., 2017). ARIMA, artificial neural network (ANN), and support vector machine (SVM) models are the most popular models for stochastic time series (Kohiro et al., 2004; Pan and Lee, 2012). The clustered weekly electricity consumers load data is used for load forecasting using ARIMA. ARIMA model is used to produce more accurate 2-weeks demand (load) forecasting for each cluster; consequently, for each electricity consumer belongs to a cluster. K-NN is a popular classification algorithm in data mining and statistics. On the one hand, K-NN is simple to implement and has significant classification performance, but on the other hand, it is unsuitable for the K-NN algorithm to assign a fixed K to all test samples. Instead, assign different K values to different test samples and find the best K by using the cross-validation method is a solution (Zhang et al., 2018). K-NN is used to classify the electricity consumers by using their forecasted 2-weeks demand (load) come from ARIMA model. The classification process, which is recognition about loads, determines the consumers who should assign the loads in the same class with similar patterns, while loads in different classes are differing. Based on the classification, differentiated demand response (DR) programs will be designed for different user classes. The DR programs are an attempt to make demand elastic (Mathieu et al., 2013; Wang et al., 2015). DR is an important means for the new-generation energy systems to deal with power generation uncertainty and load demand fluctuation (Jiangsu, 2019). One of the aspects of demand side management (DSM) is DR, which changes the role of electricity consumers from passive to active by changing electricity consumption pattern to reduce peak load (Tahir et al., 2018). The main advantage of DR is to improve the efficiency of the usage of the available electricity resources. We have two DR programs classes, price-based and incentive-based, that can be used to allow electricity consumers to have active participation in distribution network management (Zita et al., 2011).

1.3. Proposed DR Programs

In this paper, a special case, both incentive and price-based DR is recommended to shift the electricity consumption to periods of lower demand on a weekly basis. The recommended DR is a bit different from what is usually accepted about DR in the literature. DR in the literature refers to the shift of electricity consumption to lower demand within a day (hours) because of the advance metering infrastructure (DOE and NETL, 2007; Mathieu et al., 2013; Wang et al., 2014; 2015; Huang et al., 2019). U.S. Department of Energy (DOE) and National Energy Technology Laboratory (NETL) on Jan, 2007 are defined DR as the changes in the usage of electricity from normal consumption pattern due to changes in the price of electricity over time (DOE and NETL, 2007). Electricity consumers dynamically change their consumption behavior in response to time-of-use electricity price signals or real time dispatching commands to reduce peak demand and shift electricity consumption between different time periods (Huang et al., 2019). The price-based DR programs can be categorized into time-of-use price, peak price, real-time price, multi-step price and direct energy market participation. The incentives-based can be categorized into direct load control, interruptible load, demand-side bidding, emergency demand response (Hongtu et al., 2010). Due to the lack of price signal and market mechanism to promote demand response in Tulkarm, demand response might be achieved by the recommended weekly-based DR of this study and supported by an online energy reporting system (OERS).

1.4. Proposed OERS

In this regard, Web and mobile-based OERS are introduced. OERS plays a vital role in improving the effectiveness of the recommended DR programs. OERS enables household electricity consumers to participate in DR programs easily by manual control of the appliances regarding different parameters such as electricity prices and end-user preferences. The success of the price and incentive-based approaches of the DR programs significantly rely on the number of electricity consumers to be involved in DR programs. Therefore, various types of incentives increase their willingness to be enrolled in a DR program and be involved in DR weekly events. Measuring the performance of the recommended DR is not the focus of this study, dedicated further study will be used for this purpose. The following sections are as follows. Section 2 presents the literature review and the state of the art in the field of electricity consumers load forecasting using statistical models, data mining techniques and the main novelties of this study. Section

3 presents the methodology of this study. Section 4 presents the implementation of the study. Section 5 presents the results and discussion of the study. Finally, Section 6 presents the conclusion followed by the references.

2. LITERATURE REVIEW

Because of the importance of accurate electricity load forecasting in all time-horizon for demand-side management and planning, the literature mentioned many studies using various statistical and data mining techniques to deal with this issue (*Dai* and *Wang, 2007; Abdul Razak* et al.*, 2008; Qingle* and *Min, 2010)*. The state-of-the-art, methodologies used in electricity load forecasting for different applications were comprehensively reviewed (*Fan* et al., *2019)*. Hybrid models comprising clustering techniques and statistical models such as ARIMA, SARIMA, simple exponential smoothing, hidden Markov model and artificial neural network (ANN) etc. were used and proved good performance (*Nazarko* et al., *2005; Patil* et al., *2017;* Seunghyeon et al., *2017; Nepal* et al., *2019)*. Table 1 describes some studies dealing with load forecasting and its applications.

Most studies in Table 1 rely on a massive data produced from advanced metering systems. High-frequency data about the load are generated and stored with a temporal measurement of 15 min (Lu et al., 2019). For conventional electricity billing, data mining is used to extract hidden value of smart meter readings (Wang et al., 2015). The electricity consumer behavior in different situations such as social behavior in various weather conditions also can be extracted and detected using data mining techniques. The main novelty of this research in comparison with the previous mentioned studies that a conventional offline ECPB data set is used with limited short-term electricity consumption features (See Appendix A). ECPB is the only source of electricity consumption data in TM. This data set is used for weekly electric load (demand) forecasting using a novel hybrid model of K-means clustering and ARIMA for weekly load (demand) forecasting. The forecasted load is used for designing various DR programs. K-NN is used to classify electricity consumers according to their electricity demand forecasts on weekly basis.

3. METHODOLOGY

The main objective of this methodology is to forecast weekly household electricity demand (load) by using a hybrid clustering approach namely K-means clustering and time series ARIMA model to assist TM in managing the electricity critical-peak demand on a weekly basis. Figure 1 is depicted the workflow of this methodology. It comprises the following steps:

- Step 1: Electricity consumers' prepaid bills (ECPB) data set collection and preparation phase
- Step 2: Data preprocessing phase. Preprocessing data mining techniques are applied to the data set. Electricity consumers' weekly load (ECWL) data set is created as a result of the implementation of an aggregation algorithm that is seen in Algorithm 1 (Appendix A)
- Step 3: Features reduction phase. Features reduction is applied to the ECWL data set by using principal component analysis (PCA)
- Step 4: Clustering phase. K-means clustering is applied to the ECWL data set to classify electricity consumers based on the weekly distribution of 2-year electricity load. Elbow method and silhouette analysis method are used to specify number of clusters K. The two methods are used for verification purpose
- Step 5: Forecasting of the next 2-weeks consumers' electricity load using the ARIMA model. The clustered electricity consumers' weekly load data is the input of the time series ARIMA model
- Step 6: Classification of electricity consumers according to their electricity demand forecasts using K-Nearest Neighbors (K-NN)
- Step 7: According to the classification process for each electricity consumer, the changes in consumer behavior in electricity consumption such as passive consumption, changes in the consumer segment (moving from one class to another) will be determined

Accordingly, the OERS will be activated using the different price and incentive-based DR programs that are designed for this issue.

- Step 8: Step 2 through step 7 will be repeated on weekly basis.

This methodology starts with data preparation and preprocessing. Data standardization (normalization) is a central step in data

Figure 1: Methodology workflow

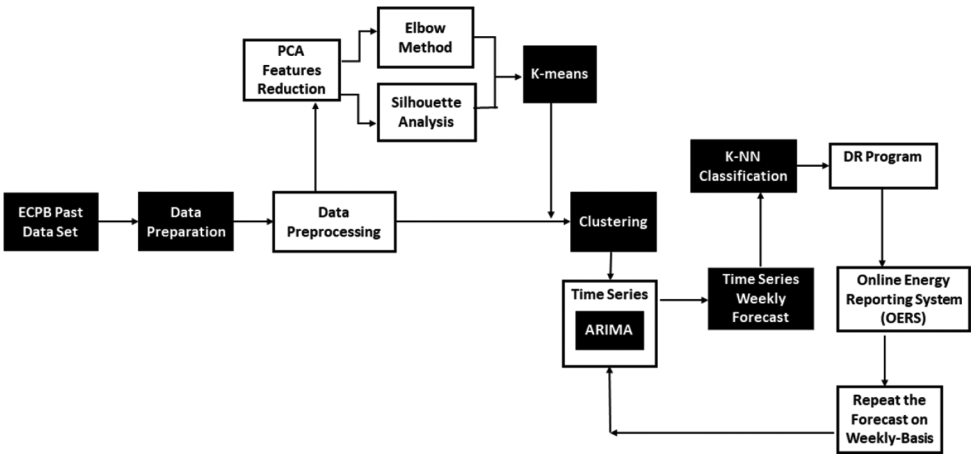


Table 1: Related studies of electricity load forecasting and its applications

Ref.	Load forecasting method	Clustering algorithm	Classification algorithm	Description
Seunghyeon et al., 2017	ARIMA	K-means	Bayesian classification	The performance of the proposed model was also compared with the Neural Network based forecasting. The proposed model shows better performance than the Neural Network
Wang et al., 2016	Fast Search and Find of Density Peaks (CFSFDP)	CFSFDP		In this paper, instead of focusing on the shape of the load curves, a novel clustering approach was used focusing on clustering of electricity consumption behavior dynamics, where "dynamics" refer to transitions and relations between consumption behaviors, or rather consumption levels, in adjacent periods. potential applications of the proposed method to demand response targeting, abnormal consumption behavior detecting and load forecasting were analysed and discussed.
Wang et al., 2015	Review of load profiling methods	Direct clustering k-means, Fuzzy k-means, Hieratical clustering and Self-organizing map (SOM) Indirect Clustering Dimension reduction based: PCA, Sammon Map and Deep Learning Time Series based: DFT, DWT, SAX, and HMM	-	A state-of-the-art, comprehensive review of data mining techniques from the perspectives of different technical approaches used in electricity load profiling.
Lu et al., 2019	Hidden Markov model	Davies–Bouldin index-based adaptive k-means algorithm	-	A Davies–Bouldin index-based adaptive k-means algorithm is proposed to cluster electricity consumers into several groups. Then, a hidden Markov model was used to extract the representative dynamic weekly load features for each cluster using the probabilistic transitions of different load levels of each cluster. The short-term load forecasting methods were evaluated by an invented feasible tool based on dynamic characteristics of load patterns, which realizes the pre-check for the forecasting results without future real measurements in the forecasting horizon
(Fan et al., 2019)	Weighted K-NN, Back-propagation neural network and ARMA models	-	W-K-NN	A novel short-term load forecasting model was proposed using weighted K-NN algorithm. It showed higher satisfied accuracy. Forecasting errors were compared with back-propagation neural network and ARMA models. The comparison illustrated a reflection of variation trend and good fitting ability of the proposed model
(BinMajid et al., 2008)	SARIMA	-	-	half hourly load data for 6 weeks had been plotted according to day-type to forecast the load demand for a day ahead. MAPEs obtained were ranging from 1.07% to 3.26%.
Patil et al., 2017	Electricity price forecasting : ARIMA and Simple Exponential Smoothing	K-means	K-NN	K-means and k-NN were used. The price data was classified by day of the week using k-means; then, the data was classified according to a month of the year. Using the classified data, short-term electric price forecasting using the ARIMA was performed. The MAPE for all the models was within an acceptable range
Nepal et al., 2019	hybrid model comprising a clustering technique and ARIMA	K-means	-	The combination of clustering and the ARIMA model has proved to increase the performance of forecasting rather than that using the ARIMA model alone
Nazarko et al., 2005	ARIMA	Fuzzy clustering approach	-	This work illustrates possibilities of ARIMA modelling with clustering approach to electrical load forecasting. The study aimed to demonstrate the proposed method efficiency. The results showed that it is possible to combine the fuzzy clustering and ARIMA models for load profile clustering. It is revealed that these models give the similar results as fuzzy coefficient approach. But in practise, estimation of ARIMA models demands in many cases statistical experience and sophisticated tools. Therefore, the first of examined method seems to be more advantageous

(*Contd...*)

Table 1: (*Continued*)

Ref.	Load forecasting method	Clustering algorithm	Classification algorithm	Description
Lee et al., 2018	Simple moving average (SMA), Weighted moving average (WMA), Simple exponential smoothing (SES), Holt linear trend (HL), Holt-Winters (HW) and Centered moving average (CMA)	-	-	UTHM (Public university in Malaysia) electricity consumption was forecasted. HW gives the smallest MAE and MAPE, while CMA produces the lowest MSE and RMSE. As a result, HW might forecast better in this problem
Li et al., 2018	ARIMA	Data-driven Linear Clustering (DLC) method	-	A (DLC) method is proposed to solve the long-term system load forecasting problem caused by load fluctuation. Firstly, data was preprocessed by the proposed linear clustering method, then optimal ARIMA models were constructed for the sum series of each obtained cluster to forecast their respective future load. Finally, the load forecasting result is obtained by summing up all the ARIMA forecasts. The errors were analysed both theoretically and practically. The result of analysis proved that the proposed DLC method can reduce random forecasting errors while guaranteeing modelling accuracy

preprocessing. It refers to convert the data attributes from one dynamic range into a specific range in order to enhance the accuracy of the clustering algorithm (BinMohamad and Usman, 2013). Many standardization techniques are used in the literature such as max-min, Z-score, Bob-Cox, natural logarithm, etc. In this study natural logarithm is used for standardizing data set features. In order to visualize the weekly loads of all consumers in 2D visualization, PCA is applied which in turns reduce the dimensionality of large data sets with minimum information loss (Jolliffe and Cadima, 2016). It allows us to compare electricity consumers' weekly loads at a glance (AbuBaker, 2019). PCA is implemented to find the dimensions in the data that maximize the variance of features included in the data set. The ratio of the explained variance is reported and the PCA component or dimension which is a composition of the data set original features is considered as a new feature of the space.

One of the important techniques in data mining is clustering or cluster analysis (Qinpei and Pasi, 2013). It used to find data segmentation and pattern information by dividing the data into groups or clusters such that each group has similar characteristics. Similarity of a group means that the more similar data points (distance) are located in the same group or cluster (Taylor, 2010; Badase et al., 2015). K-means is an unsupervised learning problem based on the category of centroid-based clustering. A data point at the center of a cluster is called a centroid. Clusters are represented by a central vector in centroid-based clustering. K-means clustering is an unsupervised iterative algorithm in which the concept of similarity is computed as a function of distance i.e., how close the distance of a data point is to the centroid of the cluster. The objective function of K-means clustering is minimizing the sum of squared distances by partitioning a data set $X=\{x_1, x_2,\ldots, x_n\}$ of n objects into a set of k clusters (Trupti and Prashant, 2013). The objective function is presented as in Formula 1.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} X_i^{(j)} - C_j^{\,2} \quad (1)$$

Where $X_i^{(j)} - C_j^{\,2}$ is the squared distance between a data point $X_i^{(j)}$ and the centroid C_j, which is an indicator of the distance of the n data points from their respective centroids (AbuBaker, 2019). The optimal number of clusters (k) is arguable (Weron, 2006). The literature has been mentioned several methods to find the optimal number of clusters such as rule of thumb, elbow, information criterion approach, an Information theoretic approach, choosing k using the silhouette, and cross-validation (Trupti and Prashant, 2013). The main idea behind K-means clustering segmentation method is to identify clusters such that the total within-cluster variation or sum of square (WCSS) are minimized. The idea behind elbow method is that a line chart plot showing WCSS in the y-axis of each value of k, if the line chart plot is like the elbow in the arm then the point corresponding to the elbow in the x-axis might be chosen as the optimal number of clusters (AbuBaker, 2019). The idea behind silhouette analysis is to analyze the separation distance among clusters; it is a plot of a measure from -1 to 1 to determine how close every point in a cluster to the points of the neighboring cluster. This analysis allows us visually determine the optimal number of clusters by trying different values of k then choosing the best k (AbuBaker, 2019).

Auto regression integrated moving average (ARIMA) model is one of the time series analysis techniques that can reflect trends. The main purposes of ARIMA model, like any time series data model, are for searching and prediction (Seunghyeon et al., 2017). In this paper, it is used for prediction purposes. Box and Jenkins (1979) (Weron, 2006) introduced a general model that uses autoregressive model in addition to the moving average parts, and it includes the differencing in the formulation, forming an autoregressive integrated moving average (ARIMA) or Box–

Jenkins model (Weron, 2006). The first part of the model is Auto Regression (AR) model, that is a time series model assumes that data have an internal autocorrelation, trend or seasonal variation i.e., internal structure. This structure is detected or explored by forecasting methods. If the electricity load is assumed to be a linear combination of past loads, then future load values can be forecasted by using the AR model. The order of the model is how many lagged past values are included in the model and denoted as AR(p) for example AR(1) is the simplest first-order AR model (Weron, 2006). The second part of the model is moving average (MA), which is a simple time series method for smoothing previous load history. The idea behind moving averaging is that electricity load (demand) observations that are close to one another are also likely to be similar in value (Samsul and Saiful, 2013). MA with order q denoted as MA(q) is the number of moving average orders in the model (Patil et al., 2017). ARIMA model has three types of parameters. The first parameter is the autoregressive parameters Ø1,…, Øp. The second parameter is the number of differencing passes at lag 1 (d). The third one is the moving average parameters (θ1,…, θq). Box and Jenkins ARIMA(p,d,q) notation is formulated as in Formula 2:

$$(B)\ L_t=\theta(B)\varepsilon_t \quad (2)$$

where L_t is the electricity load at time t, and (B) are functions of the backshift operator and ε_t is the error term (Patil et al., 2017).

The main idea of K-NN is to find out the closest K training samples (K is the number of training samples) to a target object in order to assign the dominant category of the target object as the dominant category of the closest k training samples (Fan et al., 2019). The K-NN approach depends mainly on three key elements; (1) labeled objects; (2) stored records; (3) metric to measure the similarity such as the distance between objects (Patil et al., 2017). Despite of K-NN algorithm is non-parametric, lazy algorithm, simple, understandable and is widely used machine learning algorithm, it has a problem in selecting number of neighbors (K). The literature dealt with this problem and has shown that no optimal number of neighbors suitable for all kind of data sets. For instance, many methods for choosing the number of neighbors (K) are used in (Zhang et al., 2018). In this study a mix of square root and cross validation methods is used by testing the classification accuracy-score for different K values from 2 to the square root of the number of training samples, afterward select K which has the maximum classification accuracy-score.

The change of electricity consumers demand with the change in the price of one kWh over time is known as demand response (DR). Generally, DR programs are categorized into two main categories: (a) Time-based such as time-of-use, real time pricing, and critical peak pricing program, and (b) incentive-based such as interruptible/curtailable service, direct load control, emergency demand response program, capacity market program, demand bidding/buy back, and ancillary service markets (Parvania and Fotuhi-Firuzabad, 2010; Aazami et al., 2013). DR programs can significantly improve power system reliability. Therefore, reliability aspects in DR programs should be included and evaluated in terms of their effects on power system reliability (Kamruzzarnan and Benidris, 2018). The main advantages of DR is to enhance the efficiency of the usage of the available electricity resources. One of the aspects of demand side management (DSM) is DR, which changes the role of electricity consumers from passive to active by changing electricity consumption pattern to reduce peak load (Tahir et al., 2018). As mentioned in the introduction part of this study. A special case, both incentive and price-based DR is recommended to shift the electricity consumption to periods of lower demand on a weekly basis. The recommended DR is a bit different from what is usually accepted about DR in the literature. For this purposes the OERS is introduced. OERS enables household electricity consumers to participate in DR programs easily by manually controlling the appliances regarding different parameters such as electricity prices and end-user preferences. The success of the price and incentive-based approaches of the DR programs significantly rely on the number of electricity consumers to be involved in DR programs. Therefore, various types of incentives increase their willingness to be enrolled in a DR program and be involved in DR weekly events. Because of measuring the performance of the proposed system is not the focus of this study, dedicated further study will be used for this purpose.

4. IMPLEMENTATION

Electricity distribution management system in Tulkarm district is taken as our case study. The proposed methodology is an attempt to sensitize and motivate electricity consumers to change their bad behaviors in electricity consumption.

4.1. Data Preparation

ECPB data set of TM is used as a main source of data for this analysis. TM has about 19,000 electricity consumers using prepaid electricity meters. There are 27 different types of electricity consumers' tariffs such as household, commercial, governmental, agricultural and industrial tariffs. This study is used only the household electricity consumers. There are 13,755 household electricity consumers. A billing transaction processing system captures consumers' prepayment transaction data. This demand side generated data is come from the consumers who are charging their electricity prepaid smart cards in the consumer services centers (vending stations). Each transaction presents a bill that is recorded in a database by using a client-side billing transaction processing system installed at each different vending station. The collected electricity prepaid bills data from vending stations are consolidated and stored in a central database. Electricity prepaid bills data is converted into CSV file forming the ECPB data set. Consumer number, bill number, bill date and time, bill quantity in kWh, unit price and the total amount of money paid in the bill are the only available features (attributes) in the ECPB data set (See Appendix A). A sample of the data set for 19 months is taken between June-2018 and December-2019 for household electricity consumers' prepaid bills, there are exactly 424,753 bills.

4.2. Data Preprocessing

Data integration, cleaning, reduction and transformation are applied on the data sets. Data preprocessing is a central step for using the data mining techniques. As a result of data

transformation, three new attributes (year, month and week number) are added as a new feature, which are derived from the bill date attribute. These attributes are used to determine the weekly load of each consumer. A new electricity consumers' weekly load data set (ECWL) is created for the period between June-2018 and December-2019 by applying the electricity consumers' weekly load calculation algorithm (Appendix A). The general idea of weekly load calculation's algorithm is illustrated in the pseudo code as seen in Algorithm 1.

This algorithm based on the assumption that the consumer smart card is charged by the consumer when the electricity is consumed. The analysis of ECWL data set for the mentioned period shows that the average household electricity consumers' weekly load varies from week to week due to different electricity consumption behavior see Figure 2.

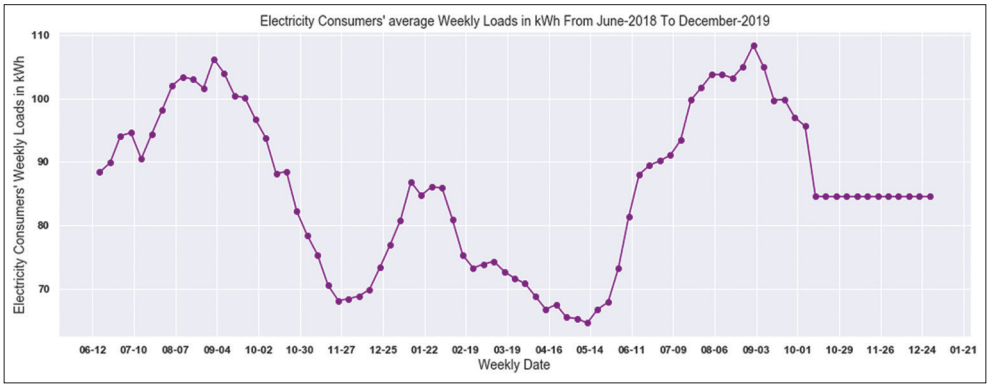
Figure 2 shows the household electricity consumers' loads start increasing in summer from June-2018 reaching the peak in September-2019, this is due to the high temperature of summer in Tulkarm district and the heavy use of air conditioning. Then the electricity loads start decreasing in autumn from October-2018 and November-2018, then return increasing in winter in December-2018 and January-2019 due to the use of heaters and then start decreasing in spring from February-2019 to April-2019 and return increasing in summer 2019. This is similar to the climate of the Mediterranean type, which has long, hot, and dry summers between May and August, and short, cool, and rainy winters between November and March. Figure 3 shows the monthly average electricity consumers' load from the mid of June to December 2018. The maximum average electricity monthly load is 507.33 kWh on September 2018.

The minimum average electricity monthly load is 292.38 kWh on November 2018. The average electricity monthly load on June 2018 represents electricity monthly load starting from the mid of June. Figure 4 shows the monthly average electricity consumers' load in 2019. The maximum average electricity monthly load is 509.88 kWh on September 2019. The minimum average electricity monthly load is 264.41 kWh on May 2019.

Algorithm 1: **Consumers' weekly load calculation pseudo code**

```
Step 1. Read ECPB data set
Step 2. Derive, Year, Month and Week features from BillDate feature
Step 3. Add the derived features to ECPB data set as new features
Step 4. Sort ECPB data set according to (ConsumerID, Year, Month, Week)
Step 5. Repeat
        Read the i^th consumer's bills as one block ; Read the first consumer's bill
        IF there are more consumer bills Then
                WHILE there are more consumer bills
                        PreviousWeek = CurrentWeek ; PreviousYear = CurrentYear;
                        PreviousQuantity = CurrentQuantity ; Read new consumer bill;
                        Gap = CurrentWeek–PreviousWeek
                        IF Gap = 0 Then
                                Assign CurrentQuantity to the consumer's weekly load for the CurrentWeek in the CurrentYear
                        Else IF Gap = 1 Then
                                        Assign PreviousQuantity to the consumer's weekly load for the CurrentWeek in the PreviousYear
                                Else
                                        CurrentLoad = PreviousQuantity/Gap
                                        LowerWeek = PreviousWeek + 1
                                        UpperWeek = CurrentWeek
                                        For Week between LowerWeek and UpperWeek
                                                Assign CurrentLoad to the consumer's weekly load for the Week in the PreviousYear
                                                of that Week
        Else Assign CurrentQuantity to the consumer's weekly load for the CurrentWeek in the CurrentYear
        UNTIL no more consumers in sorted ECPB data set
```

Figure 2: Electricity consumers' weekly load



Figures 5 and 6 show the average electricity consumptions in kWh for each week covering the period of the mid of June 2018 to December 2019.

The maximum average electricity weekly load in 2018 is 106.2 kWh in week 2 September, 2018 and the maximum average weekly load in 2019 is 108.36 kWh in week 1 September, 2019.The minimum average electricity weekly load in 2018 is 68.11 kWh in week 25 November, 2018 and the minimum average weekly load in 2019 is 64.57 kWh in week 12 March, 2019. The average electricity weekly load distribution is clearly seen in Figures 5 and 6.

The analysis of electricity consumption of electricity weekly and monthly load of household electricity consumers gives TM a general view of the electricity demand of Tulkarm district.

Figure 3: Average electricity consumers' load monthly distribution in 2018

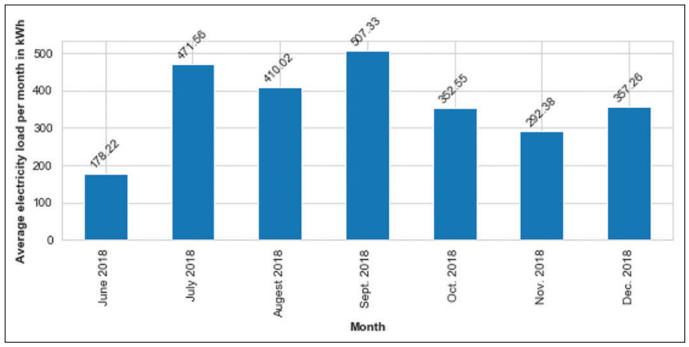
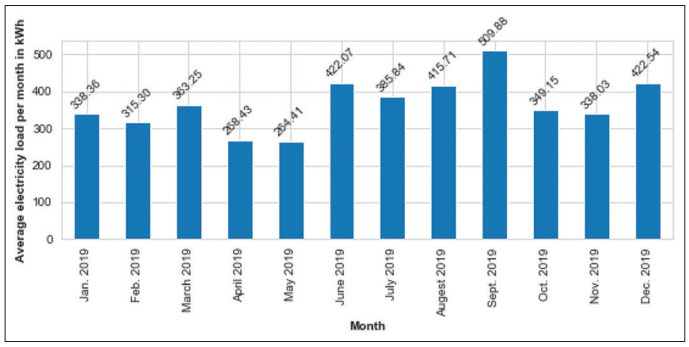


Figure 4: Average electricity consumers' load monthly distribution in 2019



4.3. Features Reduction Using PCA

The ECWL data set contains 83 features for each electricity consumer (Appendix A). Each feature represents the weekly load in kWh of each week's number in that month of the year, for instance W861 attribute represents the weekly load of the week number 1 in June, 2018 and W9711 represents the weekly load of week number 11 in July, 2019. In order to visualize the weekly loads of all consumers in 2D visualization, PCA is applied which in turns reduce the dimensionality of large data sets with minimum information loss (Jolliffe and Cadima, 2016). It allows us to compare electricity consumers' weekly loads at a glance (AbuBaker, 2019). It is implemented to find the dimensions in the data that maximize the variance of features included in the data set. The ratio of the explained variance is reported and the PCA component or dimension which is a composition of the data set original features is considered as a new feature of the space. Figure 7 shows the PC plane with original feature projections.

As seen in Figure 7, the weekly loads span the period of the data set which are represented in the X-Y coordinate system. The dimension reduction is achieved by identifying the principal directions, called principal components, in which the data varies. PCA assumes that the directions with the largest variances are the most important. In Figure 7 the PC1 axis is the first principal direction along which the samples show the largest variation. The PC2 axis is the second most important direction and it is orthogonal to the PC1 axis. It is important to note that PCA method is useful when the variables within the data set are highly correlated. Correlation indicates that there is redundancy in the data which are seen with different shades. Due to this redundancy, PCA can be used to reduce the original features into a smaller number of new features (PCs) explaining most of the variance in the original features. The names of the original features are printed on top of each other because of that they are displayed in black and all of them have the same rank and importance. So, any feature could be represented by the first dimension and also any feature also could be represented by the second dimension.

4.4. Clustering Using K-means Clustering Algorithm

In this study elbow method is used to determine the optimal number of clusters and is verified by using silhouette analysis. Clustering is unsupervised data mining technique; it is a machine learning method for targeting a specific type of electricity consumers with similar characteristics which is effective for consumers'

Figure 5: Average electricity consumers' load weekly distribution in 2018

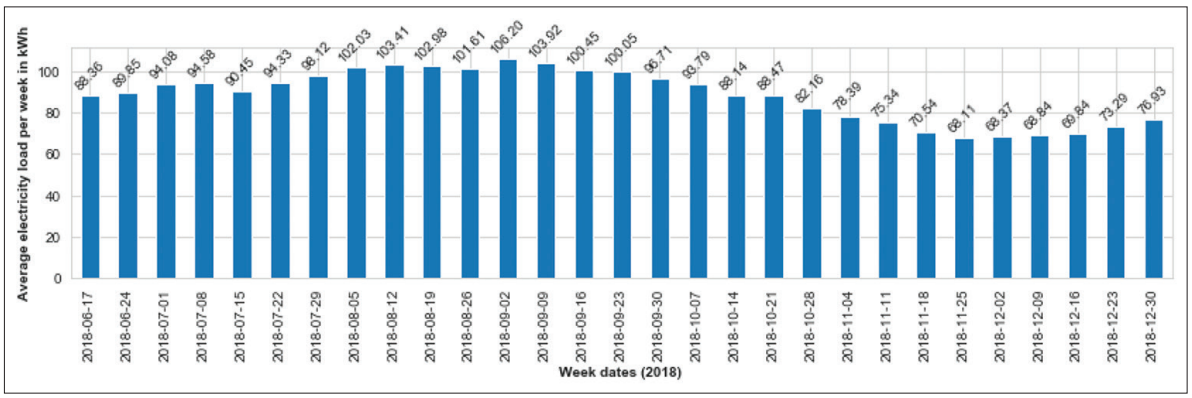


Figure 6: Average electricity consumers' load weekly distribution in 2019

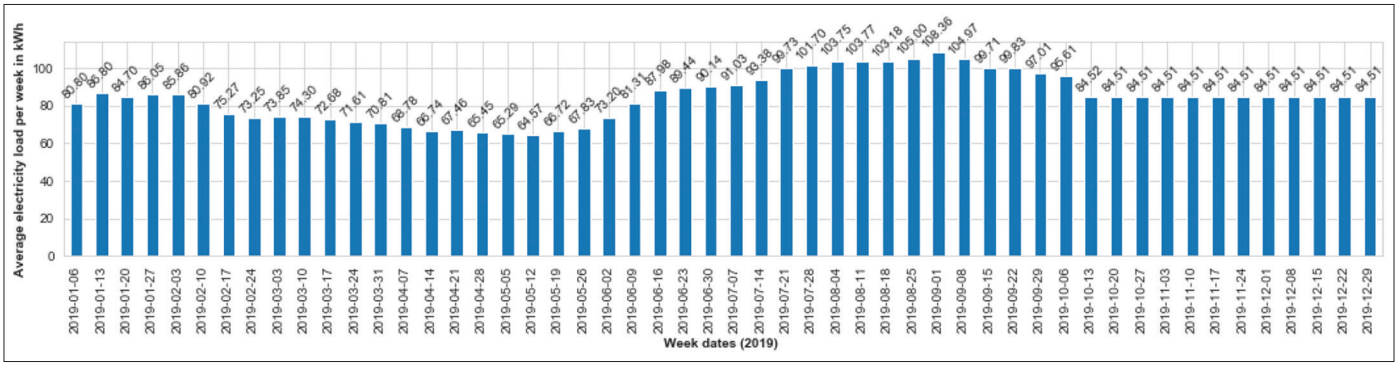


Figure 7: PCA plane with original features projections

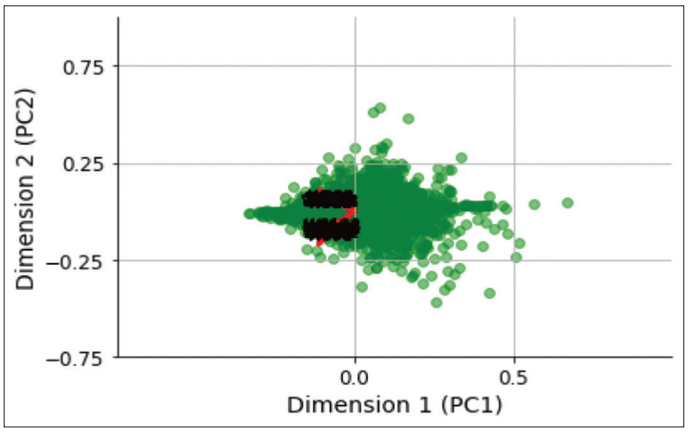
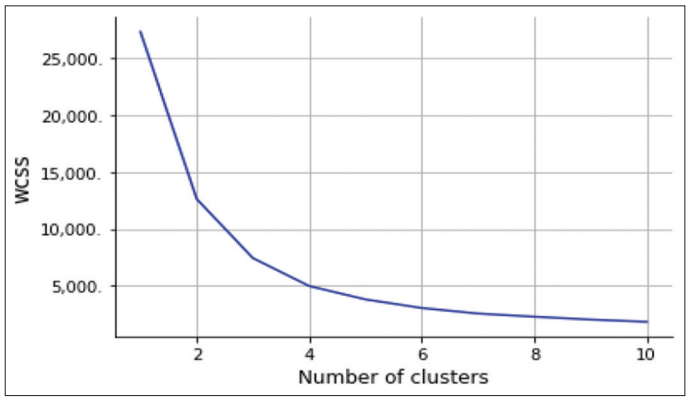


Figure 8: Detrmining k using Elbow method



segmentation. K-means clustering is applied in this study on ECWL data set to classify electricity consumers according to their weekly load. Electricity consumers segment refers to a sub division of consumers who share similar electricity consumption behavior and each segment is called cluster. Many advantages are gained from consumers' segmentation. It helps TM management to well manage electricity demand and load forecasting for each segment. In this paper Python Sklearn library (Buitinck et al., 2013) is applied for applying K-means clustering algorithm for electricity consumers segmentation. In order to apply K-means clustering a set of tests on ECWL features should be implemented. Firstly, K-means clustering assumes that all features should be normally distributed (Spherical shape). Secondly, number of observations should be the same for all clusters (Cluster size). Thirdly, no or little correlation between the features (AbuBaker, 2019). All these tests are applied on the data set after applying the natural logarithm standardization on the data to satisfy the required assumptions. Initially elbow method is used and then it is verified using silhouette analysis method. As a result of applying these methods, the optimal number of clusters is 3. K-means clustering algorithm with (k=3) is applied for electricity consumers' segmentation. The main idea behind K-means clustering segmentation method is to identify clusters such that the total within-cluster variation or sum of square are minimized (WCSS). Figure 8 shows the WCSS of each value of k. if the line chart plot is like the elbow in the arm then the point corresponding to the elbow in the x-axis might be chosen as the optimal number of clusters.

The objective is to choose a small value of k with a low WCSS. As seen in the Figure 8, three clusters (k = 3) are the optimal number of clusters. The idea behind silhouette analysis is to analyze the separation distance among clusters; it is a plot of a measure from –1 to 1 to determine how close every point in a cluster to the points of the neighboring cluster. This analysis allows us visually determine the optimal number of clusters by trying different values of k then choosing the best k. Figure 9 shows silhouette analysis by applying different values of k (n-clusters) starting from k = 2 to 6. The optimal number of clusters can be determine easily using this method by grouping together graphs when k = 3, 5, and 6, this gives the graph with k = 3, and also by grouping together graphs when k = 2 and 4 gives also the graph with k = 3. Therefore, the optimal number of k = 3. Therefore, elbow and silhouette methods agree to be the optimal number of clusters are three cluster.

As a result of applying the K-means clustering algorithm with k=3, the electricity consumers are segmented into three groups (clusters) according to their weekly loads. The cluster type is added to each consumer as a new feature in the ECWL data set. This new labeled data set will be used to solve a classification problem using one of the data mining classification models. The solution of classification problem using classification model is one of the approaches used to measure the accuracy of the implementation of the K-means clustering algorithm on ECWL data set.

Figure 10 illustrates the result of the clustering process. Three electricity consumers' segments are identified and labeled thus:

1. Cluster 0: represents the first electricity consumers' segment. It represents the consumers who have the modest weekly electricity consumption

Figure 9: Determing k using Silhouette analysis method

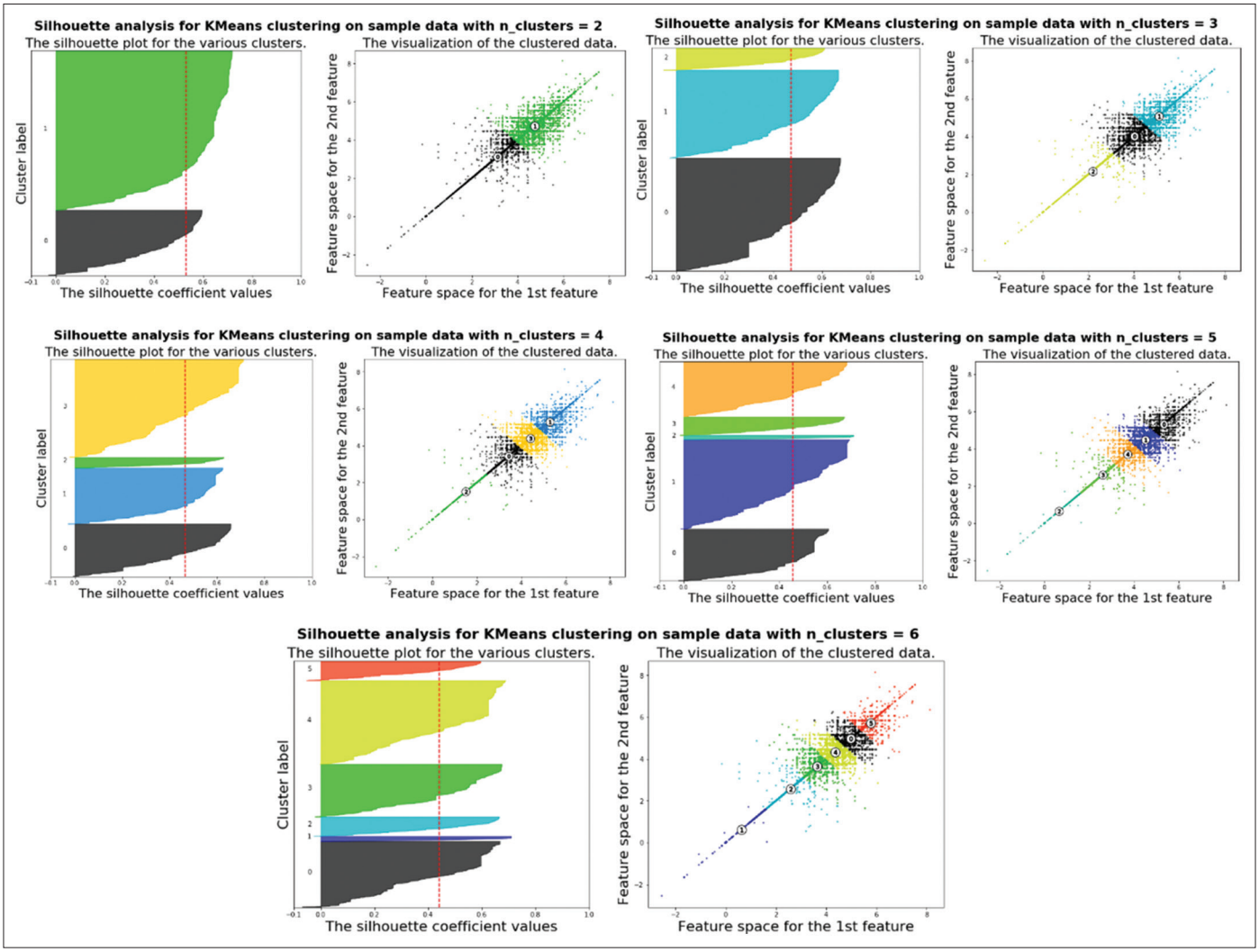
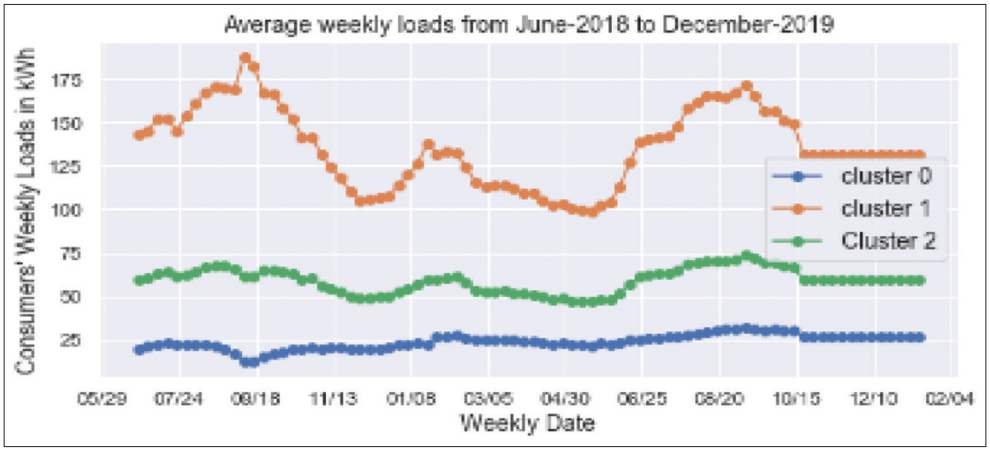


Figure 10: Electricity consumers' segmentation according to weekly electricity load



2. Cluster 1: represents the second electricity consumers' segment. It represents the consumers who have the sizable weekly electricity consumption
3. Cluster 2: represents the third electricity consumers' segment. It represents the consumers who have the moderate weekly electricity consumption.

Figure 11 shows that the average electricity weekly load of the first consumers segment (cluster 0) is between 11.89 kWh and 32.11 kWh. That means any consumer has a weekly electricity load between 11.89 and 32.11 kWh is classified as a cluster 0 consumers. Figure 11a shows the average weekly electricity load of the cluster 0 consumers for the period from the mid of June

2018 to December 2018. Figure 11b shows the average weekly load of the cluster 0 consumers in 2019.

Figure 12 shows that the average electricity weekly load of the second consumers' segment (cluster 1) is between 98.33 kWh and 187.69 kWh. That means any consumer has a weekly electricity load between 98.33 and 187.69 kWh is classified as a cluster 1 consumers. Figure 12a shows the average weekly electricity load of the cluster 1 consumers for the period from the mid of June 2018 to December 2018. Figure 12b shows the average weekly load of the cluster 1 consumers in 2019.

Figure 13 shows that the average electricity weekly load of the third consumers' segment (cluster 2) is between 46.53 kWh and 73.69 kWh. That means any consumer has a weekly electricity load between 46.53 and 73.69 kWh is classified as a cluster 2 consumers. Figure 13a shows the average weekly electricity load of the cluster 2 consumers for the period from the mid of June 2018 to December

Figure 11: Average electricity load of the first consumers' segment (Cluster 0). (a) Cluster 0 average weekly electricity load in 2018. (b) Cluster 0 average weekly electricity load in 2019

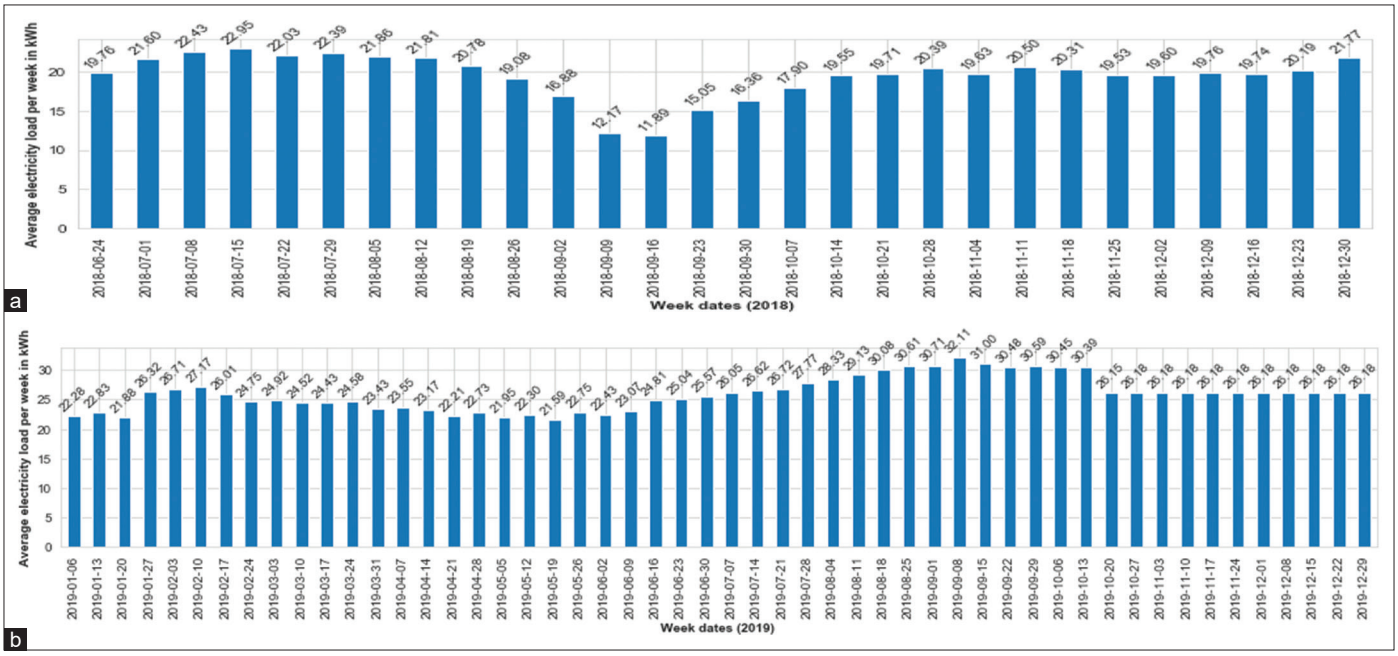
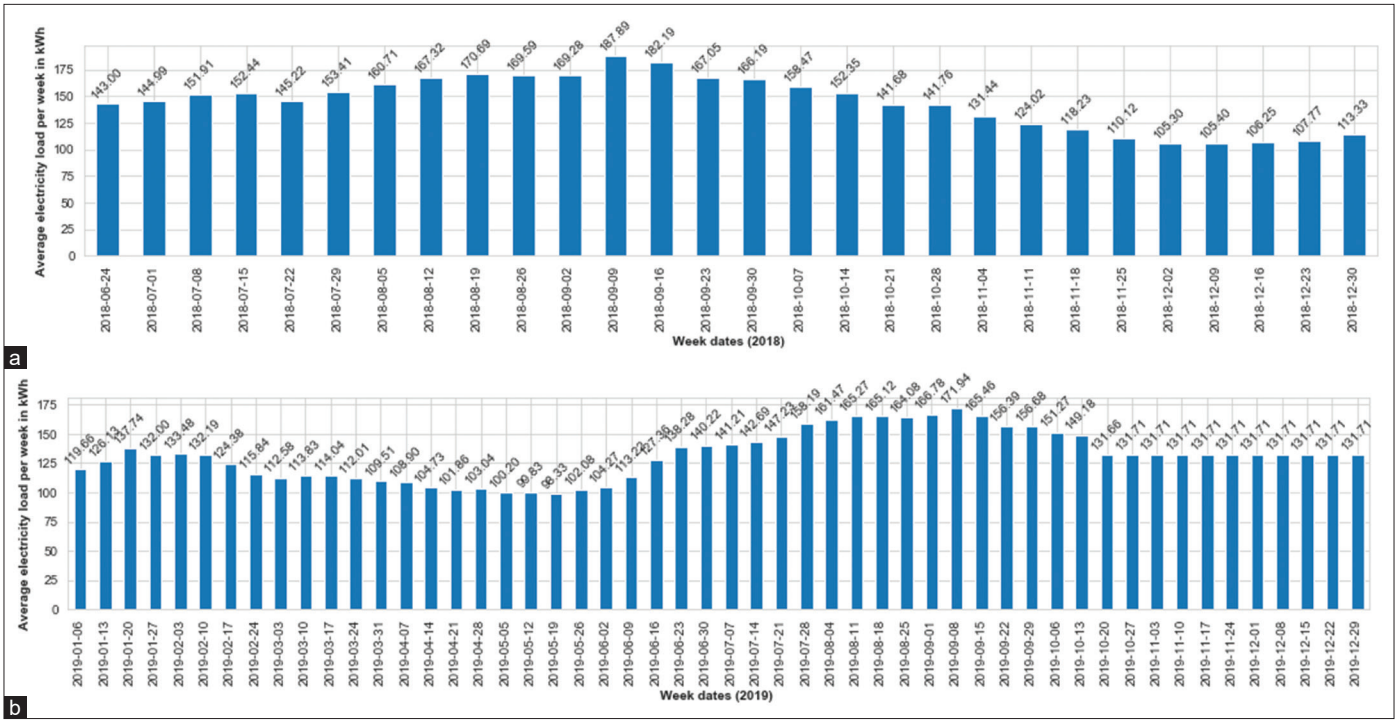


Figure 12: Average electricity load of the second consumers' segment (Cluster 1). (a) Cluster 1 average weekly electricity load in 2018. (b) Cluster 1 average weekly electricity load in 2019



2018. Figure 13b shows the average weekly load of the cluster 2 consumers in 2019. Figure 14 shows the average weekly load in different periods of time for all consumers' segments. Figure 14a shows the consumers' behavior in weekly electricity consumption during on-month period. Figure 14b shows the consumers' behavior in weekly electricity consumption during 2-month period. Figure 14c shows the average weekly load during 1-month period by numbers.

4.5. Forecasting Consumers' Electricity Load Using ARIMA Model

The forecasting of electricity consumers' load (demand) is carried out using ARIMA model. The clustered ECWL data set (Appendix A) is used for forecasting. It is span the period between the mid of June-2018 and December-2019. Python Scikit-learn library is used for building ARIMA model. Different ARIMA models are built to forecast electricity load. The performance of ARIMA models is evaluated by dividing the data set from mid of June-2018 to December-2019 into train data set from the mid of June-2018 to August-2019 which is 75% of the data set and test data set from September-2019 to December-2019 which is 25% of data set. The forecasted results are compared against the actual electricity weekly load test data set by calculating the mean absolute percent error (MAPE). The MAPE results for the forecasts are as follows:

1. Forecasting the average weekly load of all cluster types of household electricity consumers. The MAPE is 15.9%
2. Forecasting the weekly electricity loads for each cluster type cluster-0, cluster-1 and cluster-2, the MAPE are 19.7%, 9% and 13.4% respectively

Figure 13: Average electricity load of the third consumers' segment (Cluster 2). (a) Cluster 2 average weekly electricity load in 2018. (b) Cluster 2 average weekly electricity load in 2019

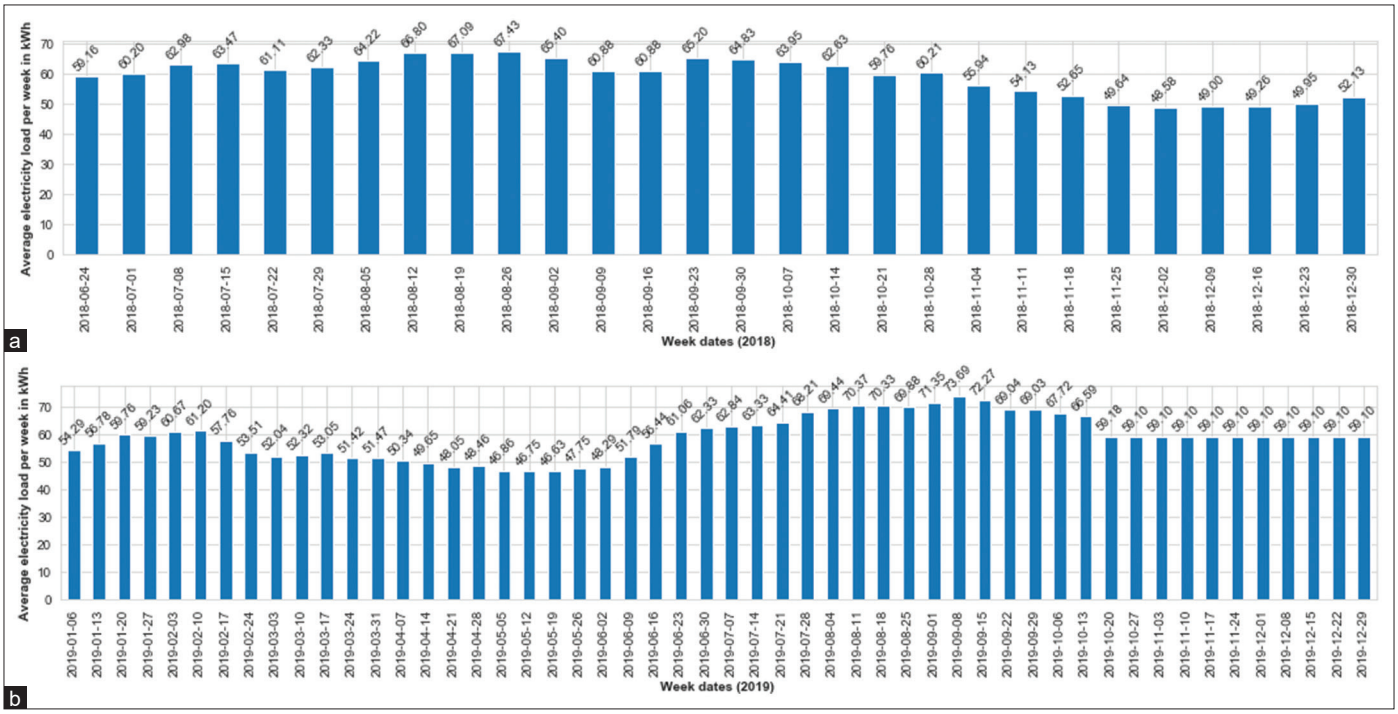
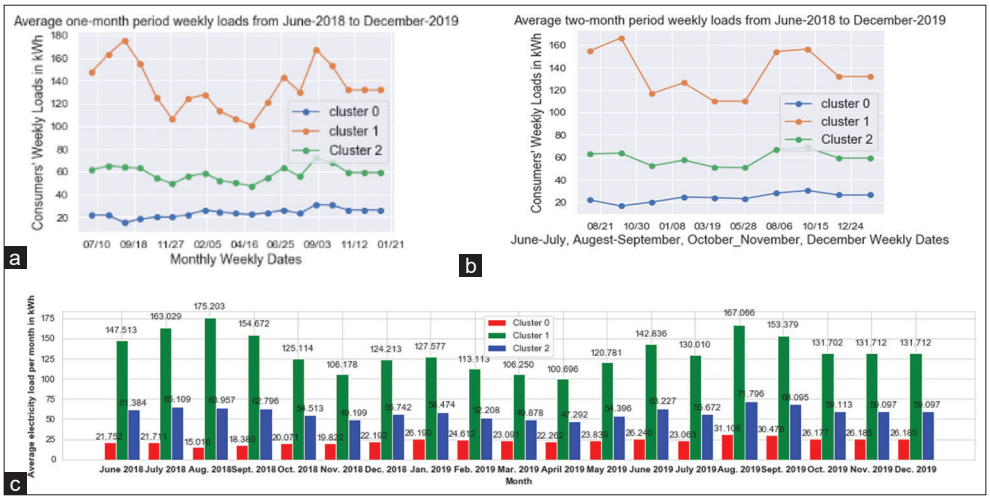


Figure 14: Different periods average weekly load of all consumers' segments. (a) Average one-monthly period weekly load. (b) Average 2-month period weekly load. (c) Average 1-month period weekly load



The performance of ARIMA models forecast are considered good.

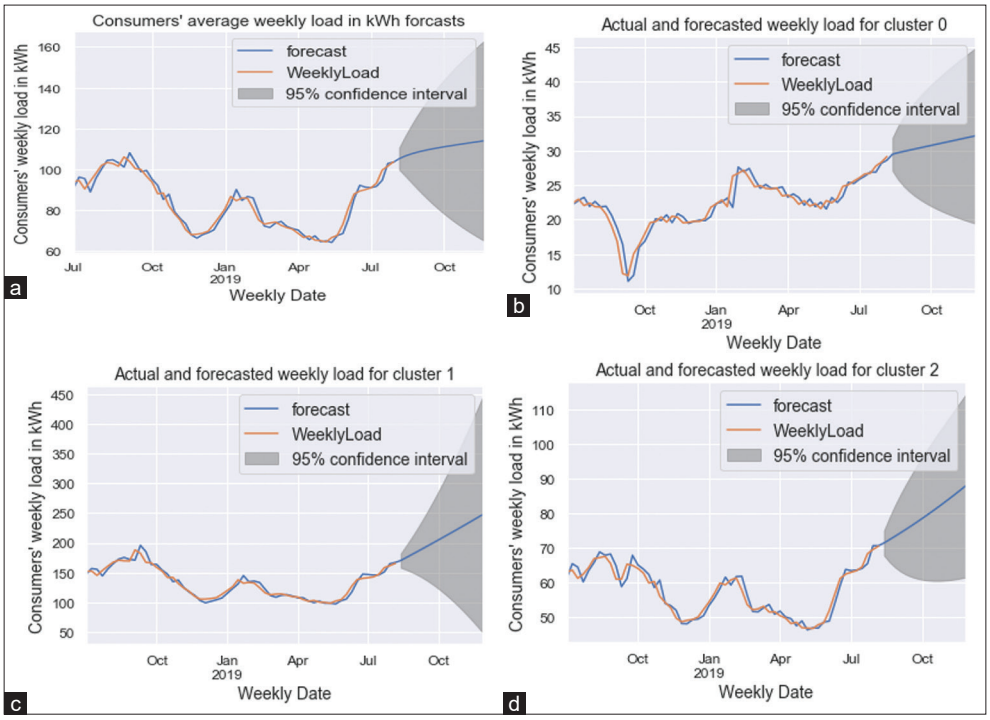
Figure 15 shows the result of applying ARIMA model for forecasting the average weekly load of electricity consumers' segments (cluster 0, 1, and 2). Figure 15a shows the actual average weekly electricity load and the forecasted average electricity load of all electricity consumers' segments covering the period from the mid of June 2018 to December 2019. It appears in the figure starting from July 2018. The forecasting of the average weekly load of all consumers' segments appears in the gray shaded area. This area is the prediction interval called the confidence interval. The gray cone around the predicted values gives the reader a spatial feeling for the range of possible values of the observation may take in the future. The predicated values might reach approximately an average weekly electricity load from 60 to 160 kWh. Figure 15b shows the actual average weekly electricity load and the forecasted average electricity load of the first electricity consumers' segment (cluster 0) covering the period from the mid of June 2018 to December 2019. It appears in the figure starting from July 2018. The forecasting of the average weekly load of the first consumers' segment appears in the gray shaded area. The predicated values might reach approximately an average weekly electricity load from 15 to 44 kWh.

Figure 15c shows the actual average weekly electricity load and the forecasted average electricity load of the second electricity consumers' segment (cluster 1) covering the period from the mid of June 2018 to December 2019. It appears in the figure starting from July 2018. The forecasting of the average weekly load of the first consumers' segment appears in the gray shaded area. The predicated values might reach approximately an average weekly electricity load from 50 to 450 kWh. Figure 15d shows the actual average weekly electricity load and the forecasted average electricity load of the third electricity consumers' segment (cluster 2) covering the period from the mid of June 2018 to December 2019. It appears in the figure starting from July 2018. The predicated values might reach approximately an average weekly electricity load from 60 to 110 kWh.

4.6. Classification Using K-NN

Multiple classes K-NN classifier is applied on the past and forecasted electricity consumers' weekly loads (demand). The purpose of using K-NN classifier is to solve a classification problem. It is used to predict the electricity consumer segment in which consumer belongs. The load forecasts come from the ARIMA model and K-NN implementation will help in predicting the consumer passive electricity consumption or/and the predicted move of a consumer from one cluster to another. Accordingly, the price and incentive-based DR program suitable for the consumer electricity consumption behavior is assigned. The model splits ECWL data set into train set and test set. The main challenge in implementing the K-NN model to classify the type of electricity consumers is to determine the number of nearest neighbors K. In this study K is determined by using the square root of the training set length, then cross-validation is applied for different values of K starting from 2 to the square root of training set. The model splits the data set which is 13,755 electricity consumers into train set and test set. The test set is 20% (2,751 consumers) of the data set and the remaining 80% (11,004 consumers) is the training set. The

Figure 15: Average electricity consumers' weekly load forecasting. (a) Consumers' segements forecast starting from the mid of June 2018 to December 2019. (b) The first electricity consumers segment (Cluster 0) forecast starting from the mid of June 2018 to December 2019. (c) The second electricity consumers segment (Cluster 1) forecast starting from the mid of June 2018 to December 2019. (d) The third electricity consumers segment (cluster 2) forecast starting from the mid of June 2018 to December 2019



square root of the training set length is 104, then cross-validation is applied for different values of K from 2 to 104. Table 2 shows the results of K-NN classification of the data set. The best classification rate (accuracy score) is 86.91% with K equals to the odd number 3. Different accuracy scores are shown in Table 2.

The mean classification rate is 86% which is considered good accuracy. Table 2 shows that the total true positive (TTPall) is equal to 167+969+1255 = 2391, this means that the total number of times over the samples were correctly classified or predicted is 2391 times. The total number of false positive of cluster 2 (TFPi=2) is equal to 71+135 = 206, This means that we have 206 times non-class 2 classified or predicted as class 2. The total number of false negative of cluster 1 (TFNi=1) is equal to 0+135 = 135 times, this means that all class 1 instances that are not classified or predicted as class 1 is 135 times. The total number of true negative of cluster 1 (TTNi=1) is equal to 167+71+17+1255 = 1510 times, this means that all non-class 1 instances that are not classified or predicted as class 1 are 1510 times and so on for the other results. The total number of cases is 2751 case. Therefore, the overall accuracy of K-NN classifier with k=3 is computed as the TTPall over total number of cases (2391/2751 = 86.91%), obviously, the 1-overall accuracy is the overall classification error, which is 13%. So, the best classification rate (accuracy score) is 86.91% with K equals to the odd number 3, the classification error is 13%.

4.7. Forecasting Using ARIMA and K-NN for a Specific Consumer

According to the proposed methodology of this study. An average weekly electricity 2-weeks load (demand) forecasts for each household electricity consumer will be applied on weekly basis. An experiment is carried out for an arbitrary selected household consumer to forecast 2-weeks electricity consumer's load (demand) using ARIMA model for the period between the mid of June-2018 and December-2019. The data of the selected consumer is divided into train data from the mid of June-2018 to August-2019 which is 75% of the consumer's data set and test data from September-2019 to December-2019 which is 25% of the consumer's data set, the forecasted results are compared against the actual electricity weekly load test data set by calculating the mean absolute percent error (MAPE). The performance of the ARIMA model forecast is 18.4% which is considered good performance.

Table 2: Results of K-NN classification

Confusion matrix of K-NN classifier (K=3)				Accuracy Score with different K values	
				K-values	Accuracy score (%)
Cluster	0	1	2	K=3	86.91
0	167	0	71	K=5	86.33
1	0	969	135	K=7	86.80
2	17	137	1,255	K=9	86.87

Figure 16 shows the consumer weekly load within the mentioned period. Figure 16a shows the actual electricity weekly load for the mentioned period. Figure 16b shows the actual and forecasted weekly load for the consumer after applying ARIMA model for the train data and test data. The selected electricity consumer is a cluster 2 consumer. The average actual weekly load for the selected consumer is approximately between (46 and 73 kWh). As seen in Figure 16b the confidence interval shows that the selected consumer might move to cluster 1 with average electricity weekly load between (98 and 187 kWh).

Afterward, K-NN classification is applied for the past and forecasted weekly load of the electricity consumer to classify the consumer to which cluster belongs. The result of the prediction shows that the selected electricity consumer stays in the same cluster (cluster 2), but if the result of the prediction shows that the selected electricity consumer is a cluster 1 consumer (move from cluster 2 to cluster 1), immediately the OERS is activated. An SMS message will be sent to inform the electricity consumer to open the energy system's web site or mobile application dedicated to the electricity consumer's account. The OERS shows all information related to the selected electricity consumer about electricity consumption behavior. The actual and forecasted load and the corresponding prices and incentives DR programs might be used or offered will be shown both visually and in numbers to warn the consumer to change manually his/her electricity consumption behavior.

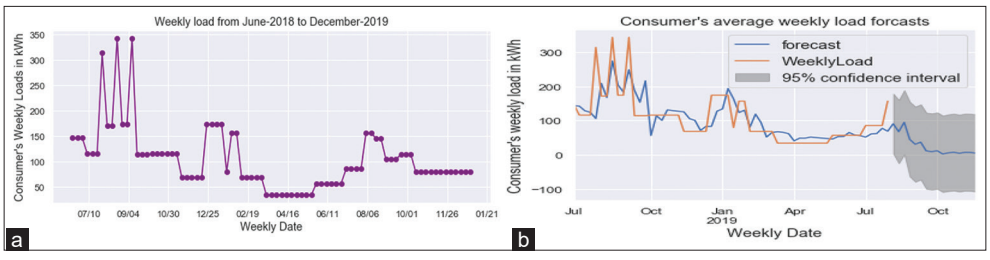
4.8. Experiment's Data Sets and Codes

The experiment's data sets and codes are uploaded to github.com. They are shared to the public and can be downloaded and used for free. Anaconda 3 Python distribution with its library such as Numpy, Scipy, Matplotlib, Scikit-learn and others with Jupyter Notebook (Buitinck et al., 2013; Ryan, 2018) are used as the programming environment. Appendix A contains a link to all data sets and codes related to the experiment. The Jupyter code files and the data sets used in the experiment are illustrated in the readme text file and can be found in the mentioned github url address.

5. RESULTS AND DISCUSSION

As a result of applying the proposed methodology of household electricity demand forecasting using a hybrid model comprising

Figure 16: Average actual and forecasted weekly load for an arbitrary selected consumer. (a) The actual electricity weekly load (b) The actual and forecasted weekly load



clustering technique (K-means) and ARIMA, three consumers' segments (cluster 0, 1, and 2) with different electricity weekly loads are obtained. This meaningful consumers' segmentation will help the managers of TM in the following:

1. The ability to design different price and incentive DR programs that will be suitable for each consumer-segment
2. The ability to manage the forces of demand by predicting the electricity consumption associated with each consumer-segment
3. The ability to forecast the growth in demand each consumer segment.

The forecasting of electricity consumers' load (demand) is carried out using ARIMA model for ECWL clustered data set. The performance of the ARIMA model is measured using MAPE. The data set is divided into 75% train data set and 25% test data set, the forecasted results are compared against the actual electricity weekly load test data set by calculating MAPE which is 15.79%, it is considered good performance. The forecasting of electricity consumers' load (demand) is carried out using ARIMA model for each cluster (cluster-0, cluster-1, cluster-2). The forecasted results are compared against the actual electricity weekly load by calculating MAPE and results are shown Table 3. The performance of the ARIMA model forecasts are considered good.

Multiple classes K-NN classifier is applied on the actual and forecasted electricity consumers' weekly loads (demand). It is used to predict the electricity consumer segment in which consumer belongs. The load forecasts come from the ARIMA model and K-NN implementation is used. This will help in predicting the move of a consumer from one cluster to another according to their weekly loads. The main challenge in implementing the K-NN model to classify the type of electricity consumers is to determine the number of nearest neighbors K. the model splits the data set into train set and test data set. Cross-validation is applied for different values of K from 2 to the square root of the training data set length. The best classification rate (accuracy score) is 86.91% with K equals to the odd number 3. The accuracy score alone is not enough, because it can be misguiding. Confusion matrix calculation is used to measure the performance of the K-NN classification used in this study because we have more than 2 clusters or classes. OERS is introduced in this paper in order to improve the effectiveness of the recommended DR programs. OERS enables household electricity consumers to participate in DR programs easily by manually controlling the appliances regarding different parameters such as electricity prices and end-user preferences. The success of the price and incentive-based approaches of the DR programs significantly rely on the number of electricity consumers to be involved in DR programs. Therefore, various types of incentives increase their willingness to be enrolled in a DR program and be involved in DR weekly events. Because of measuring the performance of the proposed DR is not the focus of this study, dedicated further study will be used for this purpose.

Table 3: Forecasting performance

Cluster types	MAPE value (%)
Cluster-0	19.7
Cluster-1	9
Cluster-2	13.4

6. CONCLUSION

In this paper, we conclude that the improvement of electricity management system from demand side can be achieved efficiently and effectively by using a combination of a novel data mining and statistical techniques such as K-means clustering, ARIMA model, and K-NN classification algorithm despite of having a conventional ordinary electricity grid using electricity prepaid meters and has only an offline electricity consumers' prepaid bills data about electricity consumption. This paper proposed a methodology consists of the following:

- Average electricity consumers weekly load preparation from prepaid bills data
- Electricity consumers' segmentation according to the average electricity weekly load using K-means clustering. This will enable electricity providers to design different price and incentive DR programs that will be used for each consumer-segment. They can manage the forces of demand by predicting the electricity consumption associated with each consumer-segment, and also can forecast the growth in demand each consumer segment
- Forecasting average electricity demand by using ARIMA. This will help electricity providers in determining the forecasted critical-peak demand and also the various price and incentive DR programs that encourage electricity consumers to apply
- Classification of electricity consumers based on electricity load forecasts using K-NN. This will help to target the electricity consumers of passive electricity consumption.

REFERENCES

Aazami, R., Mohammadbeigi, N., Mirzaei, H., Mansouri, A., Mohamadian, E. (2013), Power system reliability analysis with emergency demand response program. International Journal of Smart Electrical Engineering, 2(4), 231-236.

AbuBaker, M. (2019), Data mining applications in understanding electricity consumers' behavior: A case study of Tulkarm district, Palestine. Energies, 12(22), 4287.

Abdul Razak, A., Majid, M., Rahman, H., Hassan, M. (2008), Short Term Load Forecasting Using Data Mining Technique. IEEE 2nd International Power and Energy Conference. p139-142.

Badase, P., Deshbhratar, G., Bhagat, A. (2015), Classification and Analysis of Clustering Algorithms for Large Datasets. Coimbatore, India: Proceedings International Conference on Innovations in Information, Embedded and Communication Systems.

BinMohamad, I., Usman, D. (2013), Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G. (2013), API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. Prague, Czech Republic: European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases.

Cardenas, J., Gemoets, L., Rosas, J., Sarfi, R. (2014), A literature survey on smart grid distribution: An analytical approach. Journal of Cleaner Production, 65, 202-216.

Dai, W., Wang, P. (2007), Application of Pattern Recognition and Artificial Neural Network to Load Forecasting in Electric Power System. Haikou: 3rd International Conference on Natural Computation. p381-385.

Divshali, P., Choi, B. (2016), Electrical market management considering power system constraints in smart distribution grids. Energies, 9(6), 405.

DOE and NETL. (2007), U.S. Department of Energy and the National Energy Technology Laboratory. Systems View of the Modern Grid. Available from: https://www.smartgrid.gov/document/systems_view_modern_grid_netl_modern_grid_initiative.

Fan, G., Guo, Y., Zheng, J., Hong, W. (2019), Application of the weighted k-nearest neighbor algorithm for short-term load forecasting. Energies, 12(5), 916.

Fang, X., Misra, S., Xue, G., Yang, D. (2012), Smart grid-the new and improved power grid: A survey. IEEE Communications Surveys and Tutorials, 14(4), 944-980.

Gharavi, H., Ghafurian, R. (2011), Smart Grid: The Electric Energy System of the Future. New York: Institute of Electrical and Electronics Engineers. p917-921.

Hongtu, Z., Zhizhong, Z., Erkeng, Y. (2010), Study on demand response markets and programs in electricity markets. Power System Technology, 34(5), 146-153.

Huang, W., Zhang, N., Kang, C., Li, M., Huo, M. (2019), From demand response to integrated demand response: Review and prospect of research and application. Protection and Control of Modern Power Systems, 4, 12.

Jiangsu Demand Side Management Platform. (2019), Built Second-level Peak-shaving Power Plant: The First Real-Time Automatic Demand Response for Industrial Electric Equipment in the World[Z/OL]. Jiangsu: Jiangsu Demand Side Management Platform.

Jolliffe, I., Cadima, J. (2016), Principal component analysis: A review and recent developments. Philosophical Transactions of The Royal Society A: Mathematical, Physical, and Engineering Sciences, 374(2065), 20150202.

Kamruzzarnan, M., Benidris, M. (2018), Demand Response based Power System Reliability Enhancement. Boise, USA: IEEE Conference on Probabilistic Methods Applied to Power Systems.

Kohiro, J., Otieno, R., Wafula, C. (2004), Seasonal time series forecasting: A comparative study of ARIMA and ANN models. African Journal of Science and Technology, 5(2), 41-49.

Lee, Y., Gaik, T., Yee, C. (2018), Forecasting electricity consumption using time series model. International Journal of Engineering and Technology, 7(4), 218-223.

Li, Y. Han, D., Yan, Z. (2018), Long-term system load forecasting based on data-driven linear clustering method. Journal of Modern Power Systems and Clean Energy, 6, 306-316.

Lu, S., Lin, G., Liu, H., Ye, C., Que, H., Ding, Y. (2019), A weekly load data mining approach based on hidden markov model. IEEE Access, 7, 34609-34619.

Mathieu, J., Haring, T., Ledyard, J., Anderson, G. (2013), Residential demand response program design: Engineering and economic perspectives. Stockholm: 10th International Conference on the European Energy Market. p1-8.

Nazarko, J., Jurczuk, A., Zalewski, W. (2005), ARIMA Models in Load Modelling with Clustering Approach. Russia: IEEE Russia Power Technology.

Nepal, B., Yamaha, M., Yokoe, A., Yamaji, T. (2019), Electricity load forecasting using clustering and ARIMA model for energy management in buildings. Japan Architectural Review, 3(1), 62-76.

Palensky, P., Dietrich, D. (2011), Demand side management: Demand response, intelligent energy systems, and smart loads. IEEE Transactions on Industrial Informatics, 7(3), 381-388.

Pan, X., Lee, B. (2012), A Comparison of Support Vector Machines and Artificial Neural Networks for Mid□term Load Forecasting. Athens, Greece: IEEE International Conference on Industrial Technology. p95-101.

Parvania, M., Fotuhi-Firuzabad, M. (2010), Demand response scheduling by stochastic scuc. IEEE Transactions on Smart Grid, 1(1), 89-98.

Patil, M., Deshmukh, S., Agrawal, R. (2017), Electric Power Price Forecasting using Data Mining Techniques. New Delhi: Proceedings International Conference on Data Management, Analytics and Innovation.

PCBS. (2017), Report of the Palestinian Central Bureau of Statistics. Ramallah, Palestine: Palestinian Central Bureau of Statistics. Available from: http://www.pcbs.gov.ps/site/886/default.aspx.

Qingle, P., Min, Z. (2010), Very Short-Term Load Forecasting Based on Neural Network and Rough Set. Hunan, China: International Conference in Intellegent Computation Technology and Automation. p1132-1135.

Qinpei, Z., Pasi, F. (2013), Centroid ratio for a pairwise random swap clustering algorithm. IEEE Transactions on Knowledge and Data Engineering, 26(5), 1090-1101.

Ryan, S. (2018), Python and R in Big Data and Data Science. DZone Guide to Big Data: Stream Processing, Statistics, and Scalability. Available from: https://www.dzone.com/articles/dzone-research-5.

Samsul, A., Saiful, A. (2013), Electricity load forecasting in UTP using moving averages and exponential smoothing techniques. Applied Mathematical Sciences, 7(80), 4003-4014.

Seunghyeon, P., Sekyung, H., Yeongik, S. (2017), Demand Power Forecasting with Data Mining Method in Smart Grid. Asia: Proceedings IEEE Innovation Smart Grid Technologies - Asia.

Tahir, M., Haoyong, C., Ibn Idris, I., Larik, N., Adnan, S. (2018), Demand response programs significance, challenges and worldwide scope in maintaining power system stability. International Journal of Advanced Computer Science and Applications, 9(6), 121-132.

Taylor, W. (2010), Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research, 204(1), 139-152.

Trupti, K., Prashant, M. (2013), Review on determining of cluster in K-means clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90-95.

Wang, S., Xue, X., Yan, C. (2014), Building power demand response methods toward smart grid. HVAC and R Research, 20(6), 665-687.

Wang, Y., Chen, Q., Kang, C., Xia, Q. (2016), Clustering of electricity consumption behavior dynamics toward big data applications. IEEE Transaction on Smart Grid, 7(5), 2437-2447.

Wang, Y., Chen, Q., Kang, C., Zhang, M, Wang, K., Zhao, Y. (2015), Load profiling and its application to demand response: A review. Tsinghua Science and Technology. International Journal on Information Science, 20(2), 117-129.

Weron, R. (2006), Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. England/Chichester, West Sussex: John Wiley and Sons Ltd.

World Bank Group. (2016), Securing Energy for Development in the West Bank and Gaza. West Bank, Palestine: World Bank Group Report.

Xiao-Yu, L., Li-Ying, Y., Hang, L., Xue-Fei, T. (2017), The parallel implementation and application of an improved K-means algorithm. Journal of University of Electronic Science and Technology of China, 46(1), 61-68.

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R. (2018), Efficient kNN classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems, 29(5), 1774-1785.

Zita, V., Morais, H., Ramos, S., Soares, J., Faria, P. (2011), Using Data Mining Techniques to Support Demand Response Programs Definition in Smart Grids. Detroit, MI: IEEE Power and Energy Society.

APPENDIX

Appendix A

Electricity Consumers' Prepaid Bills		
Attributes	**Description**	**Format**
ConsumerID	The bill number of the prepaid smart card charging transaction	Numeric
BillNo	The bill date of the prepaid smart card charging transaction	Date
BillDate	Consumer or user unique ID	String
ConsumerID	The quantity charged in kWh	Numeric
Quantity	The price of 1 kWh	Numeric
Price	The area number where consumer is located (from 1 to 48)	Numeric
AreaNo	Consumers' Tariff (from 1 to 27) Household is 3	Numeric
TariffNo	The load in kWh in week 4 month 12 year 2019	Numeric
Electricity Consumers' Weekly Load		
Attributes	**Description**	**Format**
ConsumerID	Consumer or user unique ID	String
W861	The load in kWh in week 1 month 6 year 2018	Numeric
W862	The load in kWh in week 2 month 6 year 2018	Numeric
-	-	Numeric
-	-	Numeric
-	-	Numeric
W9124	The load in kWh in week 4 month 12 year 2019	Numeric
Clustered Electricity Consumers' Weekly Load		
Attributes	**Description**	**Format**
ConsumerID	Consumer or user unique ID	String
Cluster	0, 1 or 2 means Cluster number 0, 1, 2	Numeric
W861	The load in kWh in week 1 month 6 year 2018	Numeric
W862	The load in kWh in week 2 month 6 year 2018	Numeric
-	-	Numeric
-	-	Numeric
-	-	Numeric
W9124	The load in kWh in week 4 month 12 year 2019	Numeric

Experiment's data sets and Jupyter codes link: https://github.com/1175maher/Electricity_Demand_Forecasting