Ilkogretim Online - Elementary Education Online, Year; Vol 20 (Issue 4): pp. 1171-1178
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.04.131

Prediction of Energy Consumption in a Smart Home Using Deepened K-Means Clustering ARIMA Model

J. Jasmine Christina Magdalene, Asst.Professor, CA,Bishop Heber College,(Affiliated to Bharathidasan University), Tiruchirappalli, India, jjasminebhc@gmail.com
B.S.E.Zoraida, Asst.Professor, CSE, Bharathidasan University, Tiruchirappalli, India, b.s.e.zoraida@gmail.com

Abstract- In this technological era everything is made to act smart. So is a home. A smart home is the place where all the devices are designed to act smartly and it can be programmed in such a way that the maximum benefit is extracted out of it. This helps the mankind in multiple ways and one of the majoradvantages of smart devices in a home is the management of electric energy. When energy management is done in an efficient way, it helps in reducing the price to be paid and scarcity of energy can be avoided at the time of outages. To efficiently manage the energy, forecasting plays a vital role. Based on the forecast, energy production can be planned ahead and energy consumption by various smart devices can also be handled in a better way. In this paper a new model Deepened K-Means Clustering ARIMA (DKMCA) is proposed to predict the amount of energy consumed in a smart home from a smart grid during different seasons in a year. This proposed model removes the ambiguity in the K-Means clustering algorithm and from the clusters obtained, the forecasting model is built using the Model-Based forecasting method ARIMA. The data from a single smart home from the Pecan Project, Texas, USA is taken for this work. The average amount of energy consumed by each smart device in a month from the smart grid for seven years is taken and prediction is done on the average amount of energy that will be consumed from the smart grid during various seasons in a year. The performance of the proposed model (DKMCA) is compared with the ARIMA model. From the result obtained it is found that the RMSE, MAPE, AIC,AICC and MAE of the proposed model is less compared to the ARIMA model. The loglikelihood of the proposed model is also high compared to the ARIMA model. Hence the accuracy of the proposed model is better compared to the standard Model-Based method ARIMA.

Keywords- Smart Home, Energy Management, K-Means Clustering, ARIMA,RMSE, MAPE, Loglikelihood.

I. INTRODUCTION

In today's technological world, the word smart is drawing everybody towards it and everybody is dreaming to have a smart home where there is less tension as the appliances are designed to act smartly. Whatever the changes may come the need for electric energy has not lost its value. These smart devices can function well with the help of electric energy. In a smart home when everything is smart, the electric grid that supplies energy has also been designed to be smart. The amount of energy used by various smart appliances can be known which is not possible with the traditional grid. This smart grid gives the exact amount of energy consumed in a smart home. Managing this energy has become an area of interest as it helps the utilities as well as the consumer. Hence prediction of energy consumption becomes a crucial part in energy management. This prediction of energy consumption will greatly help the utilities to know about the energy demand by each smart home during various months of a year and it also helps the consumers to schedule the devices so that pricing can be lowered and uninterrupted power supply can be obtained. For prediction of energy consumption, the previous data will be of great use. The data that is used for forecasting energy consumption is of time series. A time series data is a data that is sequential and it is recorded at various time intervals. The time series data has four components viz, trend, cyclical, seasonality and randomness [1].

Research in prediction of time series data has gained more interest in recent years and there are a variety of statistical methods to do this. Some of the statistical methods are regression techniques, decomposition models, Markov principles, ARIMA etc., All these models have different level of accuracy. To know the accuracy of a model the minimum error rate given by the forecasted value is taken as a measure. In recent year model-based ARIMA has gained popularity as a stochastic model for time series [2]. The randomness in time series data has paved way for data mining techniques like indexing, classification and clustering [3]. Clustering of data is done to identify the similarity of the data points and put them in a same group and the similarities with the other data points in other groups can be reduced. By doing so the unforeseen patterns in the time series can be identified [4]. In this work, the most famous K-Means clustering algorithm is modified to remove the ambiguity of the data points belongingness in a cluster. After finding the clusters the model-based ARIMA is used to forecast the average amount of energy consumption in the forth coming seasons. This research work is organized as follows: Section II deals with the study of literature followed by preliminaries in Section III. Methodology is dealt in Section IV. The result of this proposed work is given is Section V and this work is concluded in Section VI with the future enhancement.

II. REVIEW OF LITERATURE

Prediction in energy management in a smart city and smart homes have become a sought-after area in the field of research. Many researches are being done and many authors have come out with different ideas in this research area. Pasapitch Chujai et.al have used ARMA and ARIMA models to forecast the energy consumption during various periods of a year. They have concluded that ARMA is better for one-day forecasting and ARIMA proves to be better for monthly and quarterly forecasting [5].Cristina Nichiforov et.al have presented two approaches namely ARIMA and Non-linear Auto Regressive neural network (NAR) model for forecasting energy consumption and concludes that ARIMA is better than NAR [6]. Junwei Miao has used the ARIMA model to forecast energy consumption in China and concludes that it has shown better accuracy [7]. Suat Ozturk et.al, have used ARIMA for forecasting the energy consumption of coal, natural gas and oil in Turkey[8]. Sen et.al have applied ARIMA model in their case study on an Indian Pig iron organization to forecast energy consumption.[9].Warut Pannakkong et al., have given a novel hybrid model combining ARIMA, ANN and K-Means for forecasting time series data and it is compared with the ARIMA and ANN[10]. Grzegorz Dudek has used K-Means clustering technique to forecast the next day electric load curve [11]. The works that have been carried out in this area shows that ARIMA and K-Means can be used for time series forecasting. The principles in these models are also considered for forecasting the amount of energy consumption in a smart home.

III. PRELIMINARIES

A. K-Means Clustering

The K-Means clustering algorithm is a well-known clustering algorithm in datamining which is used to form clusters based on similarity. This is one of the simplest, unsupervised, non-hierarchical learning algorithms and it was proposed by MacQueen in the year 1967[12]. This algorithm is a partitioning clustering algorithm which is used to classify the data points into different clusters based on the value of k. Once the initial clustersare formed, the centroid is calculated for each cluster. Centroid is the simple average of all the data points in that particular cluster. The distance of each data point to each of the centroid is calculated using Euclidian distance and the data point is moved to the cluster which has the smallest distance. Once the data point is moved, the centroid is calculated again and the same process is carried out till there is no shuffling of data points. There are some drawbacks in the k-means algorithm. The initial value k is defined by the user and there is no criteria to identify this [13]. If the value is not chosen properly the result will not be an accurate one. The elbow graph is used to overcome this to some extent. The other drawback is for every iteration there is no guarantee that the same data points will fall in the same cluster which will also have an impact on the result obtained.

B. Auto Regressive Integrated Moving Average (ARIMA)

The ARIMA is a very popular and efficient forecasting model when time series data are considered.Time series data are data that are collected in sequence. The sequence may be yearly, monthly, weekly, hourly etc.,The ARIMA model is a group of models namely the Auto Regressive (AR), Moving Average (MA) and Auto Regressive Moving Average (ARMA). The integration part I in the ARIMA model is the differencing which is used to convert the non-stationary data into a stationary data [4ARIMA KNN]. A stationary data is the data that doesn't show trend. The ARIMA is denoted by ARIMA (p,d,q) where p,d,q are the order of the data. The value of p is obtained from the Partial Auto Correlation Function (PACF) plot and the value of q is obtained from the Auto Correlation Function (ACF) plot. The lag difference is taken as the value for d.

C. Problem Definition

The consumption of energy by the smart devices in a smart home is of great importance when energy management is considered.Energy management has to be done efficiently so that the devices can be scheduled accordingly avoiding the peak hours where charges are high. This also helps in uninterrupted power supply during the time of emergencies. Since energy consumption by various smart appliances varies based on seasonality, there arises a need for forecasting the average amount of energy consumed from a smart gird during different seasons.By knowing the forecasted energy consumption, the utilities and the consumers can have a better idea on how to provide uninterrupted power supply to houses and how the pricing can be made which will be beneficial for both of them in different ways.

D.Notations

1. K-Means:

Let En indicate the average amount of energy consumed by various smart appliances in a smart home from the smart grid.

$$C = \frac{x1+x2+\dots+xn}{n}$$

where c is the centroid

x1,x2,…,xn are the data points in the cluster

n is total number of data points in the cluster

Euclidian Method:

$$d = \sqrt{(p-q)^2}$$

where d is the distance of the data point from the centroid

p is the data point

q is the centroid

2.Auto Regressive Integrated Moving Average (ARIMA):

The formula that is used in this algorithm is:

$Y_t = C+\phi_1 y_{t-1} + \phi_2 y_{t-2+} \phi_3 y_{t-3+\dots\dots+} \phi_p y_{t-p+} e_{t\ -}$ AR

$Y_t = (Y_{t\ -} Y_{t-1}) - (Y_{t-1-} Y_{t-2}) - I$

$Y_t = C+ e_t + \Theta_1 e_{t-1} + \Theta_2 e_{t-2} + \dots\dots + \Theta_q e_{t-q}$ - MA

where C is the constant

ϕ is the lag's coefficient

e_t is the error term

p is the autoregressive model

Y_t as predictors

q as moving average order

E. Root Mean Square Error (RMSE)

The Root Mean Square Error is used to find the difference between the actual value and the predicted value. This acts as a measure to find the accuracy of the forecasting model and it also helps in comparing the prediction errors of various models for a given dataset. The lower the RMSE value the better is the model.

F.Mean Average Percentage Error (MAPE)

The MAPE is used to measure the accuracy of a statistical based forecasting model. It is represented in the ratio form and often the accuracy is represented in percentage. The lower value indicates that the error is less and hence the accuracy is better.

G.Loglikelihood

The loglikelihood is used to measure the fitness of a statistical forecasting model for a specific dataset. The loglikelihood talks about the efficiency of the parameter taken for building the prediction model. The higher the value of the loglikelihood the better is the model that is developed for forecasting and the probability of choosing the parameter increases.

IV. METHODOLOGY

In General energy consumption in a smart home varies based on the various appliances used in a home. During summer season air conditioners will be used more whereas in winter season heaters are used mostly. Since each device consumes energy at various levels, the consumption of electric energy from the smart grid during various seasons varies. As there is variation in the consumption of energy from the smart grid, prediction of energy consumption is considered to be very important. The prediction of energy consumption during various seasons will help the utilities to know when the consumption will be high so that the generated from various sources is sufficient to meet the needs of the consumers. Also, it helps the consumers to schedule the various smart appliances, so that pricing can be reduced. Since prediction depends on the past values data-driven forecasting models like Holt Winter's will not prove good, hence there arises the need for forecasting using model-based forecasting techniques. Model-based forecasting techniques are statistical based approaches which helps in increasing the accuracy of the forecasting model. In this work the data is clustered using K-Means clustering algorithm and model-based ARIMA forecasting technique is used to forecast the amount of energy consumed in various seasons.

The dataset to implement this work is taken from a single home with a number 370 from Pecan Project, Austin USA. The energy consumption of the smart home from the smart grid is taken for seven years (2013 – 2019) to forecast the amount of energy consumed by various appliances in the house from the grid for various seasons. The energy consumption from the smart grid is taken as the input variable and it is denoted by the function f(Ec). The energy consumed from the smart grid in each month i.e., January to December for the years 2013 -2019 is considered as the input. The data is clustered using K-Means Clustering and the belongingness of the data point in that particular cluster is identified by iterating and identifying the mode of the data point. Once the data point's cluster is known, then the model-based ARIMA technique is used to forecast the amount energy consumed during various seasons. In this paper, five years data is taken as the training data and two years data is taken as the test data. Based on this the forecasting is done for the various seasons in the upcoming year. The efficiency of this Deepened K-Means Clustering Arima (DKMCA) is evaluated by comparing the RMSE, MAPE and the loglikelihood with the standard model-based ARIMA model. This work is carried out using R tool.

A. Input Parameter:

The energy consumption by the smart appliances in a smart home has an impact on the average amount of consumption of energy in a month and this is considered as the input parameter. Hence the following is taken as the input.

(i). Energy consumption from a smart grid

1)Energy Consumption from a Smart Grid:

The energy sources in a smart home are electrical grid, windmill, solar etc., but most of the energy is taken from the smart electrical grid. The consumption of energy by various appliances in a smart home varies. Some appliances like heaters and air conditioners are widely used during certain seasons. Even the power consumption by lights will be higher in winter when it will become dark early. Hence the energy consumed by the appliances in a smart home will vary during various seasons.

B.Data Preprocessing

To build the forecasting model the data has to be preprocessed. For this work the timeseries datasetfrom a single smart home in Pecan Project, Texas, USA is taken. The average amount of energy consumed in kw in an hour is taken for a period of seven years from 2013 to 2019. This hourly data is converted into a monthly data and some of the missing values are filled with the previous value. By doing so the data is preprocessed and all the null values are removed. Table 1 shows the average amount of energy consumed from the grid by various appliances for a period of seven years. The energy consumption is represented in kw.

TABLE I
AVERAGE AMOUNT OF ENERGY CONSUMPTION FROM A SMART GRID IN KW

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2013	0.55	0.3	0.32	0.15	0.72	1.47	1.49	2.46	1.81	0.99	0.72	0.47
2014	0.36	0.47	0.29	0.47	1.23	1.52	0.64	0.94	0.96	0.64	0.5	0.8
2015	0.55	0.46	0.57	0.54	0.85	1.08	1.38	1.36	0.84	0.81	0.76	0.77
2016	0.33	0.34	0.44	0.55	0.82	1	0.76	1.12	1.16	0.63	0.56	0.77
2017	0.66	0.53	0.36	0.41	0.8	0.85	1.89	0.58	0.97	0.46	0.6	0.71
2018	0.55	0.8	0.31	0.35	0.79	1.09	1.48	1.37	1.24	0.75	0.61	0.56
2019	0.4	0.53	0.29	0.35	0.72	0.83	1.03	1.85	1.8	0.88	0.63	0.61

C. Construct and Decompose the Timeseries Data

The data that is preprocessed is a timeseries data as it is obtained sequentially related to time. This data is then changed into a yearly timeseries data. Twelve is taken as the frequency for converting the data into a yearly data. The data that is obtained is decomposed to find the trend, seasonality, cyclical and randomness of the data. This decomposition helps in selecting the best model which can be used for forecasting using this data set. Fig 1. Shows the trend, cyclical, seasonality and randomness of the dataset.

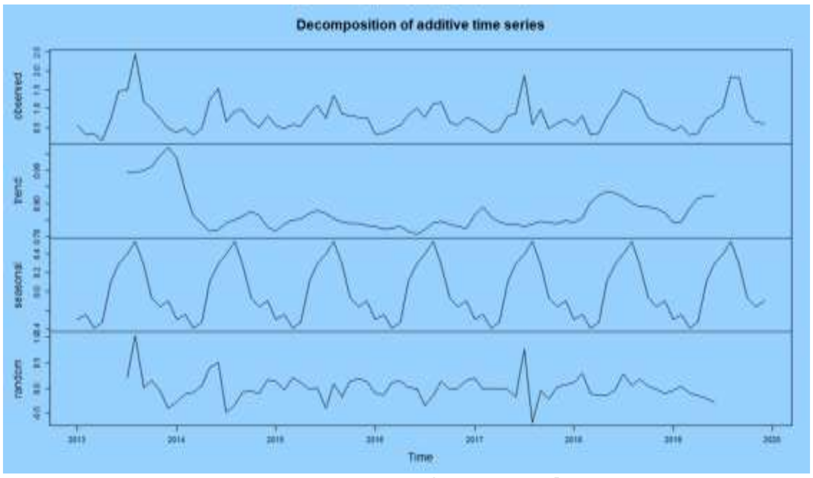

Fig 1: Decomposition of time series data

D.Construct the Model

To construct the proposed model, the initial step is to go for K-Means clustering. The number of clusters is decided using the elbow graph that is shown in Fig 2. Based on the elbow graph the number of clusters is initialized as 4. Once the number of clusters is initialized, the centroid for each cluster is calculated using the Euclidian method. Then the distance from the data point in a cluster to all the centroids in the clusters are calculated and the datapoint is shifted to the cluster with the least distance. Likewise, all the data points in all the clusters are shuffled until there is no shuffling. One of the major disadvantages of K-Means is, each time this algorithm is executed there is no guarantee that the clusters will have the same data points. To avoid this and to find the exact belongingness of a data point to a particular cluster, it is iterated many numbers of times and the mode is taken. By this we can remove the ambiguity on to which cluster a particular data point belongs to. Once the clusters are formed the model-based ARIMA technique is used to forecast the energy consumed in the forthcoming season. The p and q values in the ARIMA is identified using the Partial Auto Correlation Function (PACF) and Auto Correlation Function (ACF) plots respectively. Fig 3 shows the PACF plot and Fig 4 shows the ACF plot. The difference is taken as 1. By using these values, the average energy consumption from a smart grid for the forth coming seasons is forecasted. Fig 5 shows the graphical representation of the forecasted value.

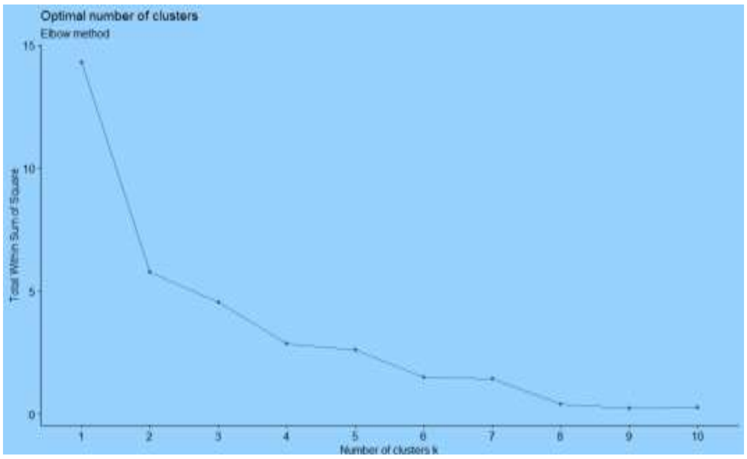

Fig 2. Elbow graph for the dataset

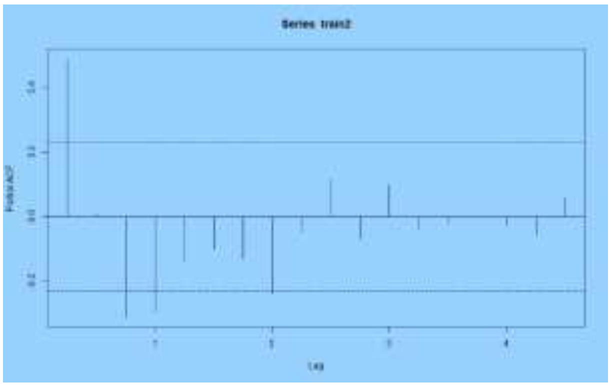

Fig 3. Partial Auto Correlation Function Plot

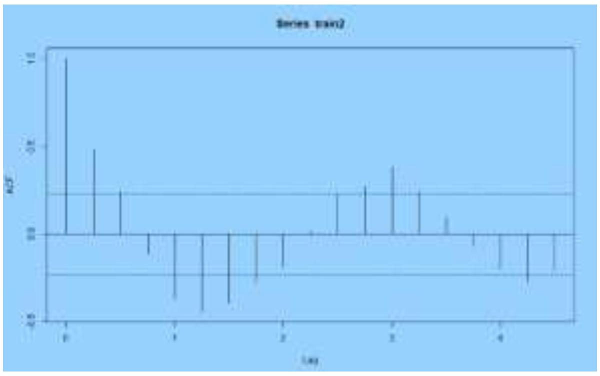

Fig 4. Auto Correlation Function Plot

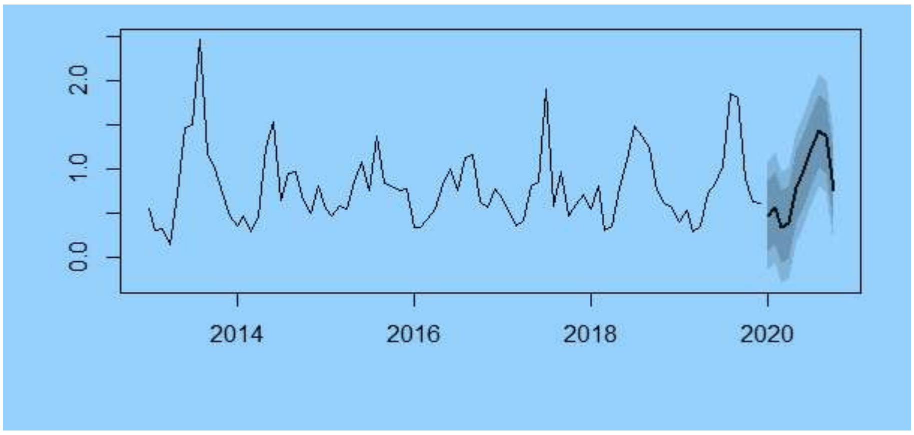

Fig 5. Forecasting using DKMCA

C.Pictorial Representation

Fig 6 represents the conceptual diagram of this work. The energy usage from smart grid for the forthcoming year is predicted using the proposed model.

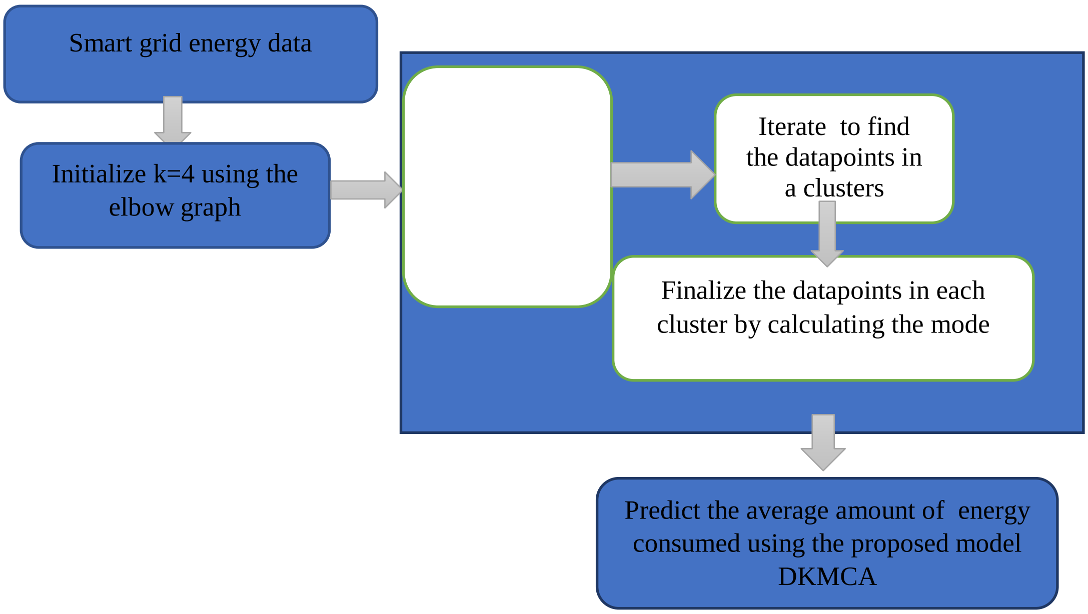
Smart grid energy data
Initialize k=4 using the elbow graph
Iterate to find the datapoints in a clusters
Finalize the datapoints in each cluster by calculating the mode
Predict the average amount of energy consumed using the proposed model DKMCA

Fig 6. Conceptual Diagram

D.Proposed Deepened K-Means Clustering ARIMA (DKMCA) algorithm to predict the average amount of energy consumption in a smart grid based on seasonality

```
Step 1: Read the Variable Ec
Step 2: Initialise k = 4 (with the help of Elbow Graph)
Step 3: Create the cluster based on the k value
Step 4: Compute the centroid of each cluster
$c = x1+x2+...+xn$
$n$
Step 5: Find the distance between each data point to all the other clusters
$d = \sqrt{(p-q)^2}$

Step 6: Move the data point to the nearest cluster
        Goto step 4 and repeat steps until there is no shuffling takes place
Step 7: iterate from step 3 for 1000 times
Step 8: Find the mode of each data point in a cluster and finalise the cluster
Step 9: Build the model based on the cluster using ARIMA
         ARIMA(p,q,d)
Step 10: Predict the average amount of energy consumed using the proposed model
```

Step 12: Stop

V. RESULTS AND DISCUSSIONS

ARIMA is a model-based forecasting model as it considers the previous data for forecasting the future. K-Means is a clustering technique where similar data points are combined and formed as a cluster. In this work a new model Deep K-Means Clustering ARIMA(DKMCA) is proposed to forecast the energy consumption in a smart home from a smart grid. The dataset is taken from a smart home in Pecan Project, Austin, Texas. The average amount of energy consumed in a smart home is considered and the data is taken for seven years from 2013 to 2019.The methodology that is discussed above is followed to forecast the average amount of energy consumption in the forthcoming seasons. The data is divided into training data and test data and based on that the various errors are calculated. The accuracy of the proposed model DKMCA is compared with that of the standard ARIMA model. The RMSE, MAPE and AIC values are taken as measures to understand the performance of the proposed model. The lower the value the more is the accuracy of the model.Table II shows the values of RMSE, MAPE, AIC, loglikelihood for the dataset using ARIMA and the proposed DKMCA model. From the table it is clear that the values of the proposed model is much lesser than ARIMA. The value of MAPEusing ARIMA is 37% whereas for DKMCA it is 20% .The RMSE value is also less for DKMCA proving that the performance of DKMCA is better than ARIMA. The graphical representation of these values is shown in Fig 7. From these representations it is clearly understood that DKMCA proves better for this dataset compared to the model-based ARIMA.

TABLE II

COMPARISON OF THE PERFORMANCE OF ARIMA AND DKMCA

Parameters	ARIMA	DKMCA
AIC	49.08	-9.55
AICC	50.66	-9.33
BIC	61.64	-8.56
Loglikelihood	-18.54	5.78
ME	-0.0093559	-0.0022694
RMSE	0.3127846	0.165485
MAE	0.230121	0.097939
MPE	-16.29267	-0.6547174
MAPE	37.18423	20.63582
ACF1	0.1625264	0.027872

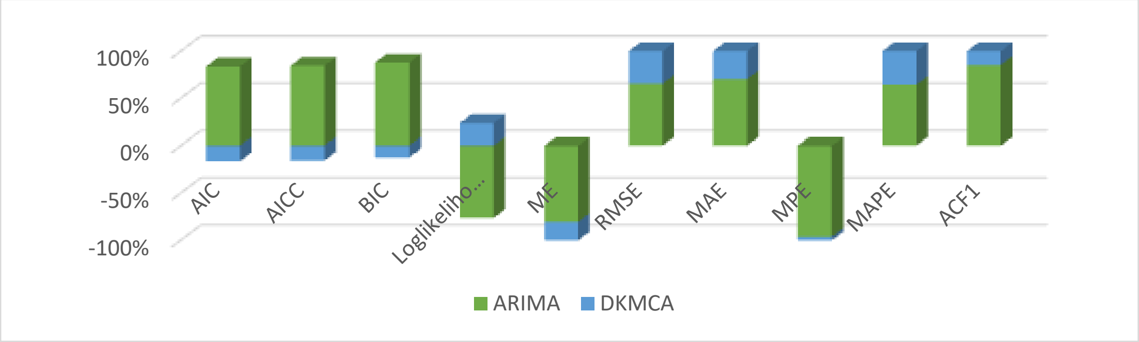

Fig 7. Graphical representation on the comparison of ARIMA and DKMCA

VI. CONCLUSION

In this work a new algorithm Deepened K-Means Clustering ARIMA is proposed to forecast the average amount of energy consumption from a smart grid in a smart home in the forthcoming seasons. The result of this work is discussed in the previous section and based on this it can be concluded that the proposed model can be used for predicting time series data where there is seasonality. The performance has improved and it can be noted based on values that are obtained for RMSE, MAPE, AIC, loglikelihood. The accuracy of this work is identified by comparing the error values of the proposed model with that of the standard ARIMA model. In this model the forecasting is done for energy drawn from a smart grid. In the future energy consumption from various renewable resources can also be considered and combined with that of the smart grid energy consumption to forecast the amount of energy consumed during various seasons. This will greatly help the consumers as well the utilities to schedule and use the energy so that the energy requirements are satisfied without uninterrupted energy supply.

References

[1] Bruce L. Bowerman, Richard T. O' Connell, & Anne B. Koehler, "Forecasting, time series, and regression: an applied approach," 4th ed. The United States of America: Thomson Brooks, 2005.

[2] Kohiro JM, Otienio RO, Wafula C." Seasonal time series forecasting: a comparative study of ARIMA and ANN models". Af J Sci Technol. 2004;5(2):41–49.

[3] F. Petitjean, A. Ketterlin, P. Gançarski, A global averaging method for dynamic time warping, with applications to clustering, Pattern Recog. 44 (3) (2011) 678–693.

[4] T.W. Liao," Clustering of time series data! a survey", Pattern Recog. 38 (11) (2005) 1857–1874.

[5] Pasapitch Chujai, Nittaya Kerdprasop, and Kittisak Kerdprasop," Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models", Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, Hong Kong, March 13 - 15, 2013.

[6] Cristina Nichiforov, Iulia Stamatescu, Ioana Fagarasan, Grigore Stamatescu," Energy Consumption Forecasting Using ARIMA and Neural Network Models,
978-1-5386-2059-5/17/$31.00 c 2017 IEEE.

[7]Junwei Miao "The Energy Consumption Forecasting in China Based on ARIMA Model", International Conference on Materials Engineering and Information Technology Applications (MEITA 2015),Published by Atlantis Press,2015.

[8] Suat Ozturk,Feride Ozturk,"Forecasting Energy Consumption of Turkey by Arima Model", DOI: 10.18488/journal.2.2018.82.52.60, 2018.

[9] Sen, Parag & Roy, Mousumi & Pal, Parimal, "Application of ARIMA for forecasting energy consumption and GHG emission: A case study of an Indian pig iron manufacturing organization," Energy, Elsevier, vol. 116(P1), pages 1031-1038,2016.

[10] Warut Pannakkong, Van-Hai Pham, Van-Nam Huynh," A Novel Hybridization of ARIMA, ANNand *K*-Means for Time Series Forecasting", International Journal of Knowledge and Systems Science,Volume 8, Issue 4, October-December 2017.

[11] Grzegorz Dudek**,"** Next day electric load curve forecasting using k-means clustering**",**Rynek Energii 92(1):143-149**,** February 2011.

[12] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, University of California Press, 1967.

[13] Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), ISBN:978-1-4799-1024-3/13/$31.00,IEEE,2013.