2020 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 12-14, 2020

Water Demand Forecasting Using Machine Learning and Time Series Algorithms

Tarek Ibrahim
College of Computer and Information Technology,
Arab Academy for Science & Technology and Maritime Transport,
Cairo, Egypt
Tarek.Omar@yahoo.com,

Yasser Omar
College of Computer and Information Technology,
Arab Academy for Science & Technology and Maritime Transport,
Cairo, Egypt
Dr_Yaser_Omar@yahoo.com,

Fahima A.Maghraby
College of Computer and Information Technology,
Arab Academy for Science & Technology and Maritime Transport,
Cairo, Egypt
Fahima@aast.edu

***Abstract*— nowadays, most of the water distribution networks are still managing their operation using the Instantaneous demand. This means that the machinery's use is determined by the immediate need for water. The network's water reservoirs are packed pumps that start to work when the level of water exceeds a given minimum threshold and stops when it reaches the peak level. Establish a water management strategy focused on predicting future demand is reducing the cost of capture, storage, processing, and distribution. In this paper, we present a comparative study for water demand forecasting using support vector linear regression and AutoRegressive Integrated Moving Average (ARIMA). The study has been carried out on the state of Kuwait daily water consumption. The result shows that ARIMA has MAPE (1.8) and RMSE (9.4) while support vector linear regression has MAPE (0.52) and RMSE (2.59) which indicates the deviation of the forecasted water demand versus the actual water consumption.**

Keywords—water demand, forecasting, machine learning, time series

I. INTRODUCTION

Water is the basic source of life and an important natural economic resource. Water covers almost 70% of earth's surface, and it has been taken for granted that it will always be there for us, however, water shortage already affects multiple areas across different continents, According to the UNESCO recent study it's expected that by 2025, 1.8 billion people living in multiple areas will face severe water shortage, and about 33% of the world population may be under water stress conditions [1].

The sustainability of economy & society development is to a large extent depending on rationalizing the utilization of water resources, for the last couple of decades desalination has become a vital alternative for water supply, It opens the door to tackle unconventional water resources that has great potentiality to provide sustainable water supply. Desalination offers just about 1% of the world's drinking water, but this amount is rising year-on-year [2].

State of Kuwait has a total area of 17,818 km2, Kuwait has a population of 4,62 million (2018), roughly 98 percent of Kuwait Metropolitan Area, 810 km2 or 4.5 percent of the total area, Kuwait is one of few countries in the world without rivers or natural lakes, Kuwait was entirely dependent on distillation plants for its freshwater supplies. For about 30 years multi-stage flash distillation plants have been used successfully in Kuwait. The highest global water consumption per capita was recorded in Kuwait at 500 liters per person per day [3].

As it requires significant energy consumption, desalinating seawater is more expensive than other natural resources as in groundwater or rivers, on the other hand water recycling and water conservation costs $1.09 to $2.49 per thousand gallons [4], the water demand Forecasting reducing capture, and treatment, storage, and distribution costs. Water demand forecasts allowed the Water Distribution Network to minimize energy consumption by 3.1% meanwhile reducing energy costs by 5.2% [7].

Different machine learning methods have been widely used lately in the implementation of effective short-term water demand forecasting such as neural networks, support vector machines, k-nearest neighbors and random forests which helps the operators of water distribution systems make decisions about pumping schedules, storage, treatment, and water distribution.

Upcoming sections of this paper will be structured as follows:

Section II presents the related works, which illustrates the previous water demand forecasting studies using machine learning and time series techniques, Section III introduces the algorithms that are used in this paper. Section IV illustrates the proposed forecasting model structure. Section V illustrates the dataset used in this research. Section VI discusses the evaluation criteria methods. Section VII presents the results discussion. Section VIII the conclusion, Section IX introduces future works. Finally, Section X presents the acknowledgment.

II. RELATED WORK

Many research tackled the topic of water demand forecasting as follows:

Shabani et al. [5] proposed a new strategy to short term water demand forecasting based on a two-stage learning method that merges Gene Expression Programming (GEP) with time-series clustering. The strategy was tested in the town of Milan, Italy's true-life water demand information. Multi-scale modeling was conducted based on the rearrangement of hourly water demand patterns in lead intervals of 3, 6, 12 and 24 hours. The study results showed that when coupled GEP with unsupervised learning algorithms in comprehensive spherical k-means this will propagate more accurate results.

978-1-7281-5263-9/20/$31.00 ©2020 IEEE

Lopez et al. [6] presented a multi-model predictor for water demand forecasting called Qualitative Multi-Model Predictor Plus (QMMP+). The quantitative element was predicted and the pattern mode was assessed using a Nearest Neighbor (NN) classifier and a calendar. Every period was executed concurrently with the NN classifier and the Calendar, and a probabilistic method was used to select the most suitable model for forecasting. In comparison with other methods such as Radial Basis Function Artificial Neural Networks, Autoregressive Integrated Moving Average and Double Seasonal Holt-Winters, the suggested model QMMP+ provides the highest outcomes when applied to the Barcelona Water Distribution Network. QMMP+ has shown that water usage patterns unique modeling therapy increases predictive precision.

Candelieri [7] introduced a fully data-driven and machine-learning-based strategy to characterize first and then forecast short-term hourly water demand with app depending in two different data sources, first one is urban water demand obtained from Supervisory Control and Data Acquisition (SCADA) and the second is individual water usage from Automatic meter reading (AMR). A real case was carried out on the Milan water distribution network data. Clustering was provided by clustering data in a different time scales then apply various SVM regression models over these clusters of data, the obtained results measured by Mean Absolute Percentage Error (MAPE) to determine the best and worst forecasting models.

Pacchin et al. [8] presented a model for forecasting water requirements in Castelfranco-Emilia city in Italy over a 24-hour time window using a two factors whose value is displayed at each forecast phase. The first factors reflect the percentage between the 24-hour average water supply following the moment the prediction is produced and the 24-hour average water demand. The second ratio reflects the connection between average water demand in a generic hour dropping over the 24-h forecast period and average water demand over that period, The results shows that the forecasting accuracy is generally high, with RMSE values ranging from 4 to 6 L/s and corresponding MAE percentages ranging from 5 to 7 percent.

Gagliardi et al. [9] proposed a technique for forecasting short-term water demand in the Harrogate and Dales area of Yorkshire in the United Kingdom based on the use of the Markov chain. Two models were created and displayed based on homogeneous (HMC) and non-homogeneous Markov chains (NHMC). The artificial neural network (ANN) and naïve techniques were applied to three real-life case studies to predict the corresponding water requirements from 1 to 24 hours ahead. In results, ANN and HMC models deliver better prediction accuracy compared to the naïve and NHMC models.

José and Boccelli [10] used k-nearest neighbor (KNN) and classical seasonal autoregressive model (SAR) techniques to forecast the time sequence of short-term water demand in Hillsborough County in Florida, USA. Comparing the real consumption and predicted demand for the SAR model showed that in spite of the model linearity, end results show that forecasted demand is almost like the actual consumption. On the other hand, the KNN model was inaccurate and resulted in a relatively large number of prediction errors.

Anele et al. [11] submitted an overview of short term water demand (STWD) forecasting techniques for water demand in South-Eastern Spain. This study shows that invariant time series (UTS) models such as (ARMA) and time series regression (TSR) models such as (ARMAX) can be combined with other techniques in a hybrid model such as ARMA and Feedforward back propagation neural network (FFBP-NN) may be counted as one of the accurate models for STWD forecasting.

Tiw et al. [12] made a Comparison of the daily urban water demand forecast using limited data extreme learning machine in combination with wavelet analysis (W) wavelet extreme learning machine (ELMW) or bootstrap (B) bootstrap-based extreme learning machine (ELMB) methods to the similar traditional artificial neural network-based models (i.e., ANN, ANNB, ANNW). ELMW model has been found to perform much better than ANN, ANNB, ANNW, and ELM models.

Al-Zahrani and AbMonasar [13] predict daily water demand in the future for the town of Al Khobar, Saudi Arabia by using time series models and Artificial Neural Networks (ANNs) depend on the daily water consumption and climate information. The result shows that using of the ANNs General Regression Neural Network (GRNN) model method with time series models is more efficient in water demand forecasting.

The above researches presented sound findings about water demand forecasting, moreover, the researchers followed multiple forecasting algorithms to get an accurate prediction of water demand in different cities and areas.

This study is covering an entire county taking into consideration the impact of population growth and applying two different forecasting techniques.

III. Machine Learning and Time Series Algorithms

A. Support Vector Machine Regression

Two types of forecasting algorithms has been used, Support Vector Machine as a regression technique retaining all the primary characteristics that characterize the algorithm. The Support Vector Regression (SVR), The loss-insensitive function ε is used in the training stage for Support Vector Regression (SVR) algorithm [14]. It describes the mapping between forecasted water consumption and actual water consumption. using the training data set as follows :

$$\{(x_i\ y_i)\}_{i=1}^{N}\ (x_i \in I', Yi \in I) \quad (1)$$

The main target for the SVR regression function is to find the mapping function f: $I' \rightarrow I$ and to make $f(x_i) \approx y_i$ which will build a Linear mapping relationship. Linear regression function can be written as:

$$y = f(x) = w^t \phi(x) + b \quad (2)$$

$\phi(x)$ is the nonlinear function, w is the support vector weight, b stands for bias, "w and b" parameters are obtained by minimizing the risk function as follows:

$$\frac{1}{2} w^t w + \eta \sum_{i=1}^{N} | y_{i\_} (w^t \phi(x) + b) | \varepsilon \quad (3)$$

"η and ε" are experiential parameters where η is the control parameter that identified by the user for nonnegative constant and ε is insensitive loss function can be represented as :

$$| y_{i-} (w^t \phi(x) + b) | \varepsilon = \begin{cases} 0 & | y_{i-} (w^t \phi(x) + b) | < \varepsilon \\ y_{i-} (w^t \phi(x) + b) | - \varepsilon & | y_{i-} (w^t \phi(x) + b) | \geq \varepsilon \end{cases} \quad (4)$$

If the value of ε is greater than the predictive then the loss function value will be zero, otherwise, linear punishment will be applied by using soft margin loss function[15] which can be calculated by joining the two positive slack variables ξ_i and ξ_i^* . The minimization of (3) is equal to minimizing the risk function as follows:

$$\frac{1}{2} w^t w + C \left(\sum_{i=1}^{N} (\xi_i - \xi_i^*)\right) \quad (5)$$

Subject to:

$$\begin{cases} y_{i-} (w^t \phi(x) + b) | \leq \varepsilon + \xi_i \\ w^t \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

“C” is the regularization constant parameter that balance between maximizing the margin and minimizing the number of the training set where ξ_i and ξ_i^* refer to upper and lower constraints of the model outputs.

Eventually, regression function can be calculated by utilizing the kernel techniques SV “ support vector” as follow:

$$y = \sum_{i=1}^{N} (a_i - a_i^*) K\langle x_i, x\rangle + b \quad (6)$$

a_i and a_i^* the Lagrange multipliers should be greater than zero, where $0 \leq a_i, a_i^* \leq C$ and k $\langle x_i, x\rangle$is kernel function [16]. The SVR output is two-dimensional, each coordinate dimension will be used as one of SVR outputs and will be trained autonomously.

B. Auto Regressive Integrated Moving

The second forecasting algorithm is AutoRegressive Integrated Moving Average (ARIMA) was proposed by Box and Jenkins[17] which is one of the forecasting algorithms of time series type that used historical data to estimate future data.

ARIMA model is classified as an "ARIMA (p,d,q)"model [18], where:p is the number of autoregressive terms, d is the amount of stationary nonseasonal variations, and q is the amount of lagged prediction mistakes in the prediction equation.

The conceptual of ARIMA is a combination of 3 methods to forecast the outcome as follows:AR–Auto-Regressive using a linear combination of past values of the variable to predict data from equation 1.

$$x_t = c + \sum_{i=1}^{P} \phi_i x_{t-i} \varepsilon_t \quad (7)$$

Where c is constant, ϕ_i is order of autoregressive in terms of i, x_{t-i} is time-series in term ε_t is the error of model I–Integrated data to make stationary data from raw data, for instance, differencing raw data.

MA–Moving Averagelearned data from historical data and predict data(method liked to Auto Regressivebut used error instead) from equation 2.

$$x_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (8)$$

where μ is constant, ε_t is error of model, θ_i is an order of moving average in term of i and ε_{t-i} is error in term of t-i.

IV. PROPOSED MODEL

This section of the paper illustrates a structured methodology used to produce the most optimum results of the study. The roadmap of this methodology shows the structure of the proposed model as shown in figure 1 below.

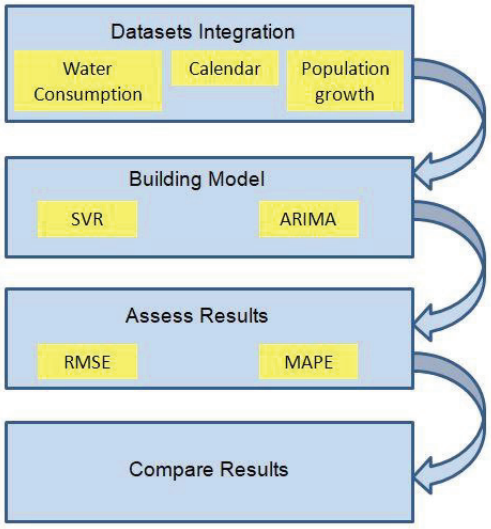

Fig. 1. Proposed Model

First step in the model is data integration throw gathering daily water consumption data provided by Kuwaiti ministry of water and electricity along with Kuwait's daily population and calendar data which indicate whether the day is a working day or a day off in one integrated dataset.

The second step is building two models using machine learning technique (Support vector regression) and time series analysis (ARIMA) followed by two experiments that run over the same data, one using support vector regression and the other using the ARIMA to forecast daily water demand.

The third step is assessing the results using two of the most popular evaluation methods RMSE and MAPE to determine the deviation of the forecasted water demand versus the actual water consumption.

The fourth and final step is comparing the two approaches results and illustrates which technique shows better results depending on the previous evaluation methods results.

V. WATER DEMAND DATASET

Kuwait is one of the few countries in the world without natural lakes or rivers. Kuwait has been fully dependent on seawater Distillation Plants as freshwater suppliers.

The daily Water consumption data of the state of Kuwait provided by Kuwaiti ministry for water and electricity between Jan 2016 to Jun 2018 has been used as experimental data in implementing the two algorithms that have been discussed in the previous section.

The water consumption data illustrated in Table 1 consists of four columns: date of consumption, the consumption rate, working day and population.

TABLE I. SAMPLE OF WATER CONSUMPTION DATA

Date	Consumption	Work Day	Population
1-Jan-17	379.197	FALSE	4411124
2-Jan-17	390.229	TRUE	4411369.473
3-Jan-17	384.303	TRUE	4411614.945
4-Jan-17	391.985	TRUE	4411860.418
5-Jan-17	385.404	TRUE	4412105.89
6-Jan-17	371.508	FALSE	4412351.363
7-Jan-17	382.38	FALSE	4412596.835

The date of consumption which is the date that the water sensors capture the consumption rate on it, consumption rate which is the consumed water amount in a specific day, the working day which has the value of true or false to filter working days from the off days. Finally, the population which represents the summation of the annual population at the beginning of each year plus daily increasing illustrated in Table 2.

TABLE II. POPULATION OF KUWAIT FOR THE PAST 3 YEARS

Year	Population	Yearly Increasing	Daily Increasing
2016	4411124	172118	471.55
2017	4500476	89352	244.80
2018	4621638	121162	331.95

VI. EVALUATION CRITERIA

To assess the performance of the different models we used the Mean Absolute Percentage Error (MAPE) and Root mean squared error (RMSE) where:

MAPE is a predictive accuracy measure of a forecasting method in statistics, for example in trend estimation.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \quad (9)$$

RMSE is a quadratic scoring rule that also measures the average magnitude of the error.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \quad (10)$$

where $\hat{y}_i$ is the predicted water demand, y_i is the real water demand, y is the mean water demand, and n is the number of observations.

VII. RESULTS DISCUSSION

Two approaches have been applied, in the first approach, water demand was forecasted using a time series technique. The second approach water demand was forecasted using a machine learning technique, Both approaches have been applied using python code over visual studio code.

The best model criteria were identified based on RMSE and the MAPE to determine the best daily forecasting model. Therefore, the best model should have the minimum RMSE and MAPE which gives an estimation of prediction error with values close to 0 being the best possible outcome.

After many experiments, ARIMA(5,1,1) was found to be the best time series model since it has the minimum deviation, ARIMA shows MAPE (1.804) and RMSE (9.418) deviation of the forecasted water demand versus the actual water consumption.

Support vector linear regression shows MAPE (0.526) and RMSE (2.590) deviation of the forecasted water demand versus the actual water consumption.

Figure-2.A and Figure- 2.B shows the water consumption for the last 6 months comparing the actual and predicted consumption of the last two weeks using the two approaches respectively.

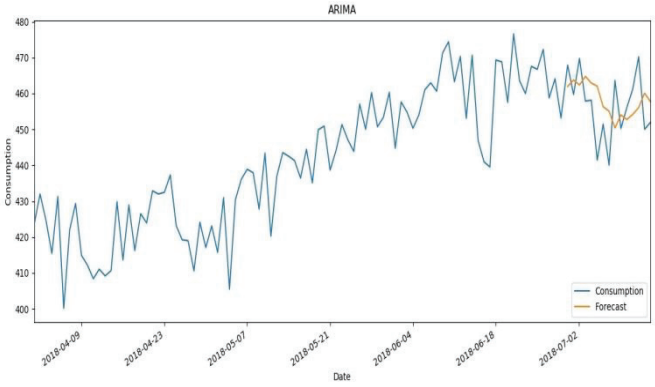

Fig. 2. A ARIMA actual consumptionVS Forcasted

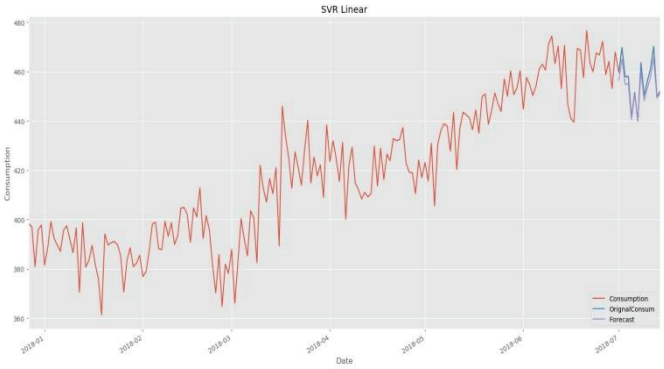

Fig. 2 B SVR actual consumptionVS Forcasted

Figure-3.A and Figure-3.B shows the water consumption comparing the real and predicted consumption of the last two weeks using the two approaches respectively.

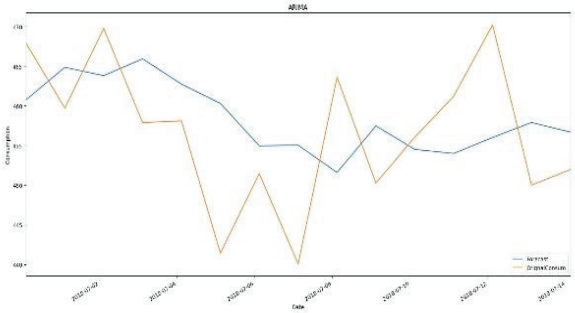

Fig. 3. A ARIMA last 2 weeks Forecasting

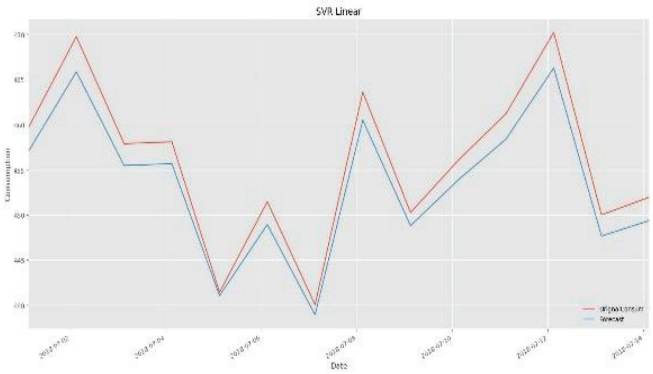

Fig. 3. B SVR last 2 weeks Forecasting

VIII. CONCLUSIONS

In this paper, we present a comparative study between two predicting approaches, Machine learning and time series to maintain accurate water demand forecasting aiming to reduce costs for capture, treatment, storage and distribution of water.

The study has been carried out on the state of Kuwait daily water consumption, the dataset included Kuwait daily water consumption population growth factor and defining the day whether it is a working day or day off.

The final results based on evaluation methods MAPE and RMSE shows the upcoming accuracy numbers using ARIMA [MAPE (1.804) & RMSE (9.418)] which is too high compared to Support Vector Regression [MAPE (0.526) & RMSE (2.590)] which has the ability to deal with other features such as population growth and day status more efficiently.

We can conclude from the above discussion that when considering adapting a technique for water demand forecasting the machine learning approach (SVM) proved providing higher accuracy and efficiency in comparison to the time serious approach (ARIMA) model.

IX. FUTURE WORK

Currently, a big project running in Kuwait to install smart meters in all Kuwait houses, this will give more accurate and diverse data, on the other hand using other machine learning and time serious technique may lead to better results than the result obtained from this study.

ACKNOWLEDGMENT

We would like to acknowledge the respected Kuwaiti ministry for water and electricity for support and for providing us with the water consumption data of state of Kuwait.

REFERENCES

[1] Unesco.org, 'World Water Assessment Programme [UNESCO WWAP] ', 2019. [online]. Avaliable: http://www.unesco.org/new/en/naturalsciences/environment/water/wwap/wwdr/wwdr4-2012/. [accessed in 5-12-2019].
[2] Iwa-network.org, 'Desalination – Past, Present and Future' , 2019. [online]. Avaliable:
[3] http://www.iwa-network.org/desalination-past-present-future/. [accessed in 5-12-2019].
[4] Beatona.net, 'Technologies Used for Waste water Treatment In Kuwait', 2019. [online]. Avaliable: http://www.beatona.net/CMS/index.php?option=com_content&view=article&id=1658&lang=en&Itemid=84. [accessed in 5-12-2019]
[5] advisian.com, 'The Cost of Desalination' 2019. [online]. Avaliable:
[6] https://www.advisian.com/en-gb/global-perspectives/the-cost-of-desalination. [accessed 5-12-2019].
[7] S. Shabani, A. Candelieri, F. Archetti and G. a. Naser 'Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts Detection ' MDPI, 2018
[8] R. Lopez Farias, V. Puig, H. R. Rangel and J. J. Flores 'Multi-Model Prediction for Demand Forecast in Water Distribution Networks ' MDPI, 2018
[9] A. Candelieri - Department of Computer Science, University of Milano-Bicocca ' Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection' MDPI, 2017
[10] E.Pacchin, S. Alvisi and M. Franchini 'A Short-Term Water Demand Forecasting Model Using a Moving Window on Previously Observed data' MDPI, 2017
[11] F. Gagliardi , S. Alvisi, Z. Kapelan and M. Franchini 'A Probabilistic Short-TermWater Demand Forecasting Model Based on the Markov Chain' MDPI, 2017
[12] P. José, A. Oliveira1 and D. L. Boccelli, A.M.ASCE 'k-Nearest Neighbor for Short Term Water Demand Forecasting' ASCE, 2017
[13] A. O. Anele, Y.Hamam, A. M. Abu-Mahfouz and Ezio Todini 'Overview, Comparative Assessment and Recommendations of Forecasting Models for Short-Term Water Demand Prediction' MDPI, 2017
[14] M. Tiw, J. Adamowski, K. Adamowski 'Water demand forecasting using extreme learning machines', ITP, 2016
[15] M. A. Al-Zahrani & A. AbMonasar 'Urban Residential Water Demand Prediction Based on Artificial Neural Networks and Time Series Models' Springer, 2015
[16] A. S. Abdou, M. A. Aziem, A Aboshosha. 'An efficient indoor
[17] localization system based on Affinity Propagation and Support Vector Regression', Sixth International Conference on Digital Information Processing and Communications
[18] (ICDIPC), 2016
[19] Z. Wu, C. Li, 1. Ng, and K. Leung, 'Location Estimation via
[20] Support Vector Regression,' IEEE Trans. Mob. Comput., vol. 6, no.3, pp. 311-321, Mar. 2007
[21] K. Shi, Z. Ma, R. Zhang, W. Hu, and H. Chen, 'Support Vector
[22] Regression Based Indoor Location in IEEE 802.11 Environments,'
[23] Mob. Inf. Syst., vol. Volume 2015, no. Article ID 295652, p. 14
[24] G. Box and G. Jenkins, Time Series Analysis: Forecasting and Control, 5th ed. San Francisco: Holden-Day: John Wiley & Sons, Inc., 2015.
[25] J. Fattah, L. Ezzine1, Z. Aman, H. El Moussami, and A.Lachhab ' Forecasting of demand using ARIMA model' International Journal of Engineering Business Management Volume 2018