Indian Journal of Science and Technology, *Vol 10(10), DOI: 10.17485/ijst/2017/v10i10/97008, March 2017*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Sustainability in Oman: Energy Consumption Forecasting using R

Fathimath Zuha Maksood[1*] and Geetha Achuthan[2]

Department of Electrical and Computer Engineering, Caledonian College of Engineering, Al Hail, Muscat,
Sultanate of Oman; fzuha95@gmail.com, geetha_achuthan@caledonian.edu.com

Abstract

Objective: Smart city projects are still in their initial research stages in Oman. This paper aims to prove the effectiveness of smart cities by using Data Mining Techniques (DMT) to predict energy consumption in Oman. **Methods:** Data collected from thirteen residential and eight industrial meters are used for electricity consumption forecast. Detailed data analysis is carried out using K-means clustering and time-series forecasting in R. Energy consumption data is modeled using average, naive, seasonal naive, Seasonal decomposition of Time Series by Loess (STL) +Random Walk with Drift (RWD), Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal component (TBATS) and Autoregressive Integrated Moving Average (ARIMA) models. **Findings:** Even though the dataset isn't characterized by seasons or trends, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) error measures suggest that electricity consumption for residential sector is more accurately forecasted using TBATS model. Energy consumption for small, medium and large scale industries, on the other hand are more accurately predicted by TBATS, Average and STL + RWD models respectively. **Applications:** The obtained results confirm the efficiency in forecasting energy consumption in Oman using time series models in order to initiate smart city implementation.

Keywords: Data Mining, Energy Consumption, Smart City, Clustering, Time-Series Forecasting, R

1. Introduction

In this technologically sound era, smart devices and interfaces often find a way to solve existing challenges. This digitized collection, generally referred to as Information and Communication Technology (ICT), is responsible for the exceeding amount of data generated and stored at present. It is said that all daily transactions create data which are logged into data models. This data can be integrated and mined according to pre-defined variables to acquire interesting relationships related to transportation, waste management, energy efficiency, etc. Analysis of these relationships is expected to produce certain traits used to predict user characteristics.

Smart Cities can be described as cities where human and social capital, education, infrastructure, communication and energy are planned and utilized in the most optimized manner. Population explosion, exponential growth of data, rise in pollution levels, and massive investments in infrastructure requirements can affect the economy, population and standard of living. Hence, Internet of Things (IoT) has developed to offer a variety of tools and sensors to aggregate data from urban areas for mining and analysis. As predictive analysis has gained majority of the importance in smart city implementation, historical data from these machines are used to forecast parameters for an upcoming time period. Energy sector has gained attention for its prediction potential in the current era. [1]employed k-means clustering and time series forecasting to forecast future energy consumption using big data from multiple cities. The utilization of a simulated dataset for the study could render weaknesses as its validity from the source compared to real profiles had not been confirmed[2]. The authors defended this issue stating that, auto-insertion methods would eliminate the risk of erroneous noise rather than in the case of real data; a theory which is opposed in other works[3]. Furthermore, adoption of R and Hadoop could have provided better solutions for data storage and processing than using a large number of tools such as WEKA, Hadoop, DFS, PIG, etc.

** Author for correspondence*

Public Authority for Electricity and Water (PAEW) has estimated an increase in demand for electric power in recent years to about 8 to 10 percent[4]. Even though Oman can host itself as a solar energy hub and arrive at a plausible solution for this energy consumption issue, 2012 observed 97.5% of the nation's electricity being generated by gas while the remaining 2.5% was generated using diesel[5]. Adoption of approaches such as green environment schemes for electricity consumption has only gained minimal results. Therefore, energy consumption rates are on the rise with production being entirely dependent on the nation's precious oil and gas reserves. Recent strategies for smart living have promoted sustainability and efficiency mostly in European cities such as Frankfurt, London, Copenhagen, etc.[6]. Sultanate of Oman is a growing hub of technology with sufficient resources and intellect which can be used to combine different sectors and create a well-developed smart city. Even though some initiatives are being undertaken to promote the concept, research in this area is marginal. A scoping study revealed methods to analyze residential energy consumption in Oman[7], while analysis of Typical Meteorological Year in Seeb was performed[8]. Further articles posed sectors to implement smart initiatives without providing back end technical strategies such as data mining or manual data analysis[9,10]. Therefore, this paper aims at analyzing the energy sector in Oman using data mining in R to trigger sustainability and energy efficiency as the latter can halve the amount of gas consumed in a year[11]. Electricity consumption is forecasted using historical values by fitting multiple time-series models. The best forecasting model is then chosen based on prediction accuracy measures for residential as well as industrial sectors in Oman.

In order to facilitate data mining operations, this work has utilized R (3.2.2v) programming language to act as an interface between the user and data. RStudio (0.99.489v) IDE is used to house the code, variables as well as implement the analyses process. R and RStudio are preferred as they are gaining widespread recognition in the data analysis field globally and are used by 49% of the data analysts[12]. Two major data mining techniques are utilized in this research, namely clustering and forecasting. A centroid-based clustering algorithm called k-means segments the given observations into a predetermined 'k' number of groups, after which, time-series forecasting is used to predict the values of each parameter at a given time frame. Various time-series models can be employed to fit the dataset in order to perform the task. This research commences with simple time-series model to predict whether the nation's energy consumption follows a constant level, and further proceeds to newer and more flexible models. The different models used are average, naive, seasonal naive, STL + RWD, ARIMA, and models.

Electricity consumption values in kWh of thirteen residential and eight industrial sectors were acquired from private Omani electricity companies. The data was recorded at half-hour intervals for a period of five months. The data acquired in pdf format is cleansed, formatted and saved in csv files for easier processing and analyses, and is then imported to R environment for further algorithmic evaluation and model fitting. Clustering and Time series forecasting are performed sequentially using each model and the results are determined using plots. The accuracy of each prediction is evaluated using error rates such as MAE and RMSE. Furthermore, the model fitting accuracy is diagnosed by analyzing residuals using histograms as well autocorrelation functions. The acquired predictive knowledge is expected to be used to control energy usage effectively, explore the possibilities of data analysis in the environmental sector, and predict interesting relationships in resolving energy efficiency problems in the future. Results of this study can be used by electricity regulation authorities as well as the government to confine the wastage of resources. Determining the amount of energy to be generated can minimize the excessive consumption of oil and gas, which in turn will help save millions of Omani rials in budget for the country, leading to smart economy and environment. Additionally, governmental organizations can recognize energy wastage points, while inhabitants can determine their energy consumption and compare it with their monthly bills. All these measures can lead to an organized regulatory procedure for eliminating excessive and unwanted energy consumption.

This article is structured as follows: Section 2 describes the major tools and techniques used for implementation. Section 3 explains the project methodology and planning. Section 4 discusses the implementation of energy consumption forecasting research by explaining the different tasks which were performed such as preprocessing, clustering, forecasting and accuracy measures. A detailed analysis of the results obtained through data mining is presented in Section 6. Section 7 presents the conclusions retrieved from the acquired results with the possible recommendations and future work.

2. Design and Data Flow: Overview

The Gane-Sarson inspired data flow diagram illustrated in Figure 1 represents the flow of data and implementation of the research. As seen, the generalized methodology pertains to data import, cleansing, preprocessing, mining, prediction accuracy, visualization and storage.

Electricity consumption data of thirteen residential flats and eight industries is acquired from private Omani electricity distribution companies. These data are imported and stored in pdf files from the organization's customized monitoring format. The non-required attributes present in the files apart from the consumption values in kWh are eliminated at the beginning of the process. Furthermore, erroneous values which are read during machine's dysfunctional behavior and null values are replaced by the global mean of each meter reading. These processes are executed using Excel and RStudio. Various packages such as ggplot, k means, vegan, etc. required for the research are then imported to the application interface. Data from each meter is processed separately and the aggregated data is clustered into k groups based on their consumption values. These clusters are fitted into six different models which support univariate time series data and forecasted one after the other. The result for each model is visualized and the best fitting models are determined for each cluster. Finally, the prediction accuracy is measured and results are analyzed, which are then stored in the global environment.

Usage of big data architectures such as Hadoop, Weka, etc. wasn't required for implementation as the dataset characteristics did not portray volume, velocity, veracity and variety; the integration of these architectures were further subdued by the adoption of R. The dataset considered in this study is divided into two folders based on residential and industrial consumption files. Since the readings are stored as monitored, the residential files consisting of consumption values of five months are split into three. Each of these files contains four parameters, namely, the apparent power, reactive lead, lag and active power. Since the scope this work is confined to active power values, the other attributes are nullified or eliminated during the course of the research. The files are named after the meter name and the month and the year for which the readings are stored in the particular file. As thirteen residential meters are considered in the scope, thirty nine excel files pertaining to thirteen smart meters are stored with each file containing half-hourly data of consumption during the specified months.

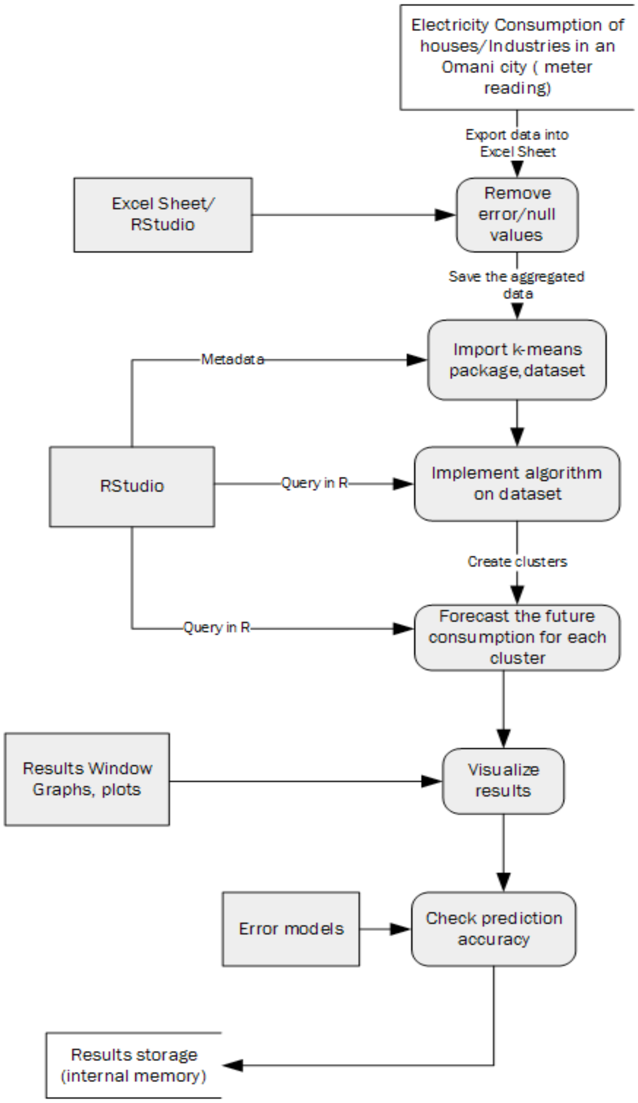
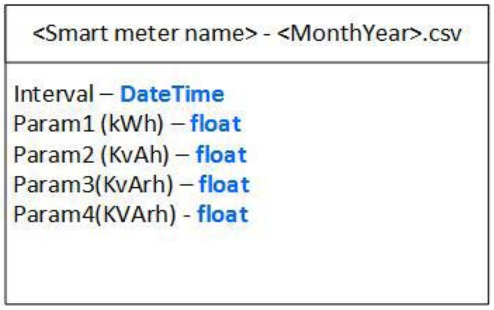

Electricity Consumption of houses/Industries in an Omani city (meter reading)
Export data into Excel Sheet
Excel Sheet/ RStudio
Remove error/null values
Save the aggregated data
Metadata
Import k-means package,dataset
RStudio
Query in R
Implement algorithm on dataset
Create clusters
Query in R
Forecast the future consumption for each cluster
Results Window Graphs, plots
Visualize results
Error models
Check prediction accuracy
Results storage (internal memory)

Figure 1. Gane-Sarson DFD of implementation.

Industrial data, on the other hand is obtained in two formats. Pdf files which store the active and apparent power at half hour intervals for a period of three months is acquired along with excel sheets housing the industrial consumption values at hourly intervals. The availability of two different formats of data poses a requirement for additional processing which is explained further in this paper. The blueprint of data and file formats in the acquired dataset are presented in Figure 2.

The volume of data in each of the files depends upon the months for which the electricity consumption is recorded. Residential files housed 7248 rows of data, while industrial sector had 1849 records per meter. Even though four variables are primarily present in the dataset, the scope of the work is only confined to modeling and predicting active power of each of these sectors.



<Smart meter name> - <MonthYear>.csv

Interval – DateTime
Param1 (kWh) – float
Param2 (KvAh) – float
Param3(KvArh) – float
Param4(KVArh) - float

Figure 2. Blueprint of unprocessed dataset.

3. Data Processing and Implementation

Raw data from sensors are cleansed before data mining algorithms are applied to it. This is due to the existence of erroneous or null values which would sometimes arise during smart meter dysfunctions or power outages. When the dataset is clustered, residential and industrial sectors (clustered separately) are grouped based on similar consumption values. The existence of nulls or errors can lead to groups which entirely represent faulty data and would ultimately reduce the accuracy of prediction. Hence, data converted into a format of comma separated values is imported to the R environment. Using code in R, the existence of faults is determined and these error values are replaced by the global mean of consumption of the given residential flat/industry. This action eliminates the existence of non-numeric values which are not fully supported by clustering techniques. Clustering and time-series forecasting DMTs which are majorly employed in this research are explained further.

3.1 K-means Clustering

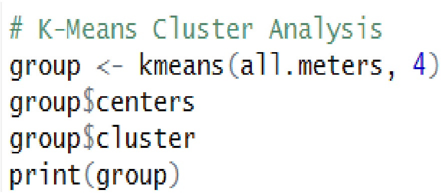
K-means clustering is used to group a given number of observations into k clusters with each data point being closer to the cluster mean it belongs to, i.e. the clustering objective mainly defines an arrangement to minimize the sum of the squares of each point in a given cluster. The cluster center in context is the mean of the Voronoi set[13]. Equation for this technique is given by,

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \qquad (3.1)$$

where, x – observation, k – no. of clusters, S – clusters, μ_i – mean of each cluster

K-means cluster analysis can be implemented using various algorithms; the Hartigan and Wong algorithm (1979) is used in R by default. Other common algorithms include MacQueen (1967), Lloyd (1957) and Forgy (1965). This research involves forecasting energy-related data, possessing around 95k rows of observations. Due to the otherwise inefficient and time consuming process, k-means clustering is used to group identical residential and industrial sectors with respect to electricity consumption. It is also stated that clustering prior to forecasting can improve the efficiency of the process by 20%[1].

The pre-requisite for implementing k-means algorithm is the specification of k so that k number of clusters can be generated by the function. It produces an object of "k means" class specifying the clusters, centers, sum of squares as well as the within sum of squares. Hartigan and Wong algorithm is used as it is one of the simplest unsupervised learning techniques compared to its contemporaries. An implementation of k-means is as shown in Figure 3.



```
# K-Means Cluster Analysis
group <- kmeans(all.meters, 4)
group$centers
group$cluster
print(group)
```

Figure 3. K-means cluster analysis in R.

Number of clusters for residential data is determined by the plot shown in Figure 4.The knee of the graph for the former is at four indicating that it is optimal to group the dataset into four clusters for implementation. In the case of industrial consumption, three clusters are considered optimal. K-means clustering is applied to these datasets and are stored into a variable of k means class which contains information about each of the clusters, with their sizes and means. This illustrates the existence of multiple residences or industries which have the same trend of energy consumption and can be forecasted as a group rather than individually utilizing memory space when implemented in real time. Moreover, any deflection in consumption levels during the course of a long-term period can be analyzed effectively.

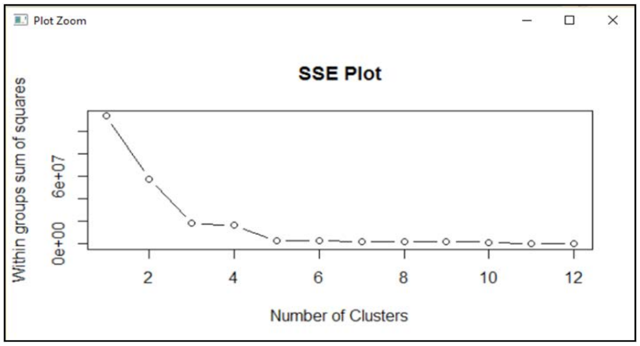


Figure 4. SSE plot for residential data.

3.2 Time Series Forecasting

Data aligned to a time frame is referred to as time series data. They can be distributed at a continuous interval or as a sequence of data points with common intervals. Analysis of time series data mainly involves forecasting, prediction and graphical plotting. Weather forecasting, earthquake prediction, econometrics and intelligent transport forecasting are some of the major applications of time series analysis. This work aims at forecasting energy consumption of residential/industrial areas for a given period of time using historical data. Since values are fed ahead of prediction, it is characterized as a supervised learning method. 80% of the data is used as training set in order to fit the model, while the remaining 20% is considered as the test set. After time-series forecasting is implemented, prediction accuracy is determined by comparing the results with the test set. Therefore, time series forecasting involves two steps, viz. fitting the time series data into the required model and applying the forecast function. Prediction accuracies are also often determined for forecasting in order to evaluate authenticity. In this research, data is fitted into various models[14,15] in order to determine the best possible method for forecasting energy consumption and are discussed in 4.2.1–4.2.6.

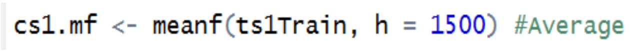
3.2.1 Average Method

Average method is one of the simplest models in time series prediction. Even though it can be easily determined by the observation values, it illustrates efficiency for short term prediction. All forecast values in this model are determined by the mean of the training data set, i.e. if a forecast of 2 weeks is generated from the previous 8, the value of forecast will be equal to the mean of eight weeks of data. Therefore, the forecast is determined by,

$$\hat{y}_{t+h|t} = \bar{y} = \frac{(y_1 + y_2 + .. + y_t)}{t} \tag{3.2}$$

$where t - no. of observations, \quad y_{1..t} - observations, \quad \hat{y}_{t+h|t} - forecast, \quad \bar{y} - mean of y_{1..t}$

The average method is employed to forecast energy consumption and evaluate its prediction accuracy compared to complicated predictive models. Equation (3.2) shows how the mean is calculated as the sum of the observations per the number of observations in the dataset. R uses a function called meanf () to convert a time series dataset into an i.i.d model and forecast it as shown in Figure 5. The i.i.d model is determined by the sample mean and normal i.i.d error. The function returns forecast and prediction intervals which are often plotted for analysis.



```
cs1.mf <- meanf(ts1Train, h = 1500) #Average
```

Figure 5. Average forecasting method in R.

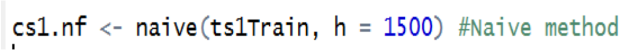
3.2.2 Naive Method

Naive method is a simple forecasting method appropriate only for time series data. In this method, all the forecasts are set to the last value under observation, i.e. if a training set of 100 observations are set for prediction, the forecast for the prediction horizon h will be set to the value of the 100th observation. This model was formulated due to its extensive advantages in economic and market time series analysis. The forecast by naive model is given by Equation (3.3).

$$\hat{y}_{t+h|t} = y_t \tag{3.3}$$

$where, t - no. of observations, \hat{y}_{t+h|t} - forecast, y_t - last observation$

Naive method denoted by naive () (or rwf() alternatively) in R.Random walk pertaining to ARIMA (0, 1, 0) model is illustrated in Figure 6. As in the case of average forecasting model, this model is used to forecast energy consumption and its accuracy is compared to complicated predictive methods.



```
cs1.nf <- naive(ts1Train, h = 1500) #Naive method
```

Figure 6. Naive or Random Walk model in R.

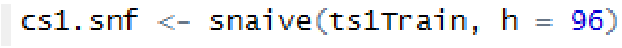
3.2.3 Seasonal Naive Model

Datasets can possess trend, cyclic and seasonal characteristics. Seasonal data are distributed across various seasons in a year and each of those seasons possesses similar observations. It can be based on the weather, monthly data, academic or financial year, etc. In order to easily identify and forecast these data sets, seasonal naive models are used. As this model is a variation of the naive method (4.2.2), it sets forecasts to be equal to the final observation of the comparable season of the year, i.e., for a dataset of two years, the observations of each month in the second year will be forecasted to the last observation of the same month in the previous year. Since the values do not undergo change, the forecast for time t + h where h is the forecast horizon is given by Equation (3.4).

$$y_{t+h-km},\ where\ k = \left\lfloor \frac{h-1}{m} \right\rfloor + 1 \qquad (3.4)$$

$$m = seasonal\ period$$

The residential energy consumption dataset is not large enough to house seasonal qualities and hence, this method is employed in order to compare results with naive and average methods. In R, the seasonal naive method was applied for a prediction horizon equal to the frequency of time period at which the observations are recorded as shown in Figure 7.

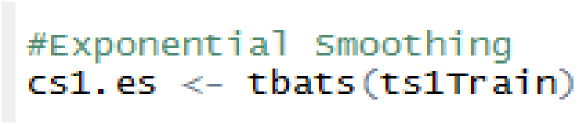


```
cs1.snf <- snaive(ts1Train, h = 96)
```

Figure 7. Seasonal Naive method.

3.2.4 Exponential Smoothing Model – TBATS

The characteristics of data are not always similar and some of the data mining algorithms are mainly based on seasonal data. Decomposition of time series into trend, cyclic and seasonal characters is sometimes impossible for small sets of data and renders unimpressive results. Therefore, in 2011, an exponential smoothing model based on trigonometric seasonality was introduced. This approach allows modeling datasets with varying characteristics, such as linear and non-linear time series, single, multiple, high period, and non-integer seasonality as well as dual calendar affects. These advantages make it a complex framework housing a large number of features which are made easy for statisticians. This method is called TBATS, an acronym for all the techniques used to create the model, which include Trigonometric seasonality, Box-Cox transformations, ARMA errors, and Trend and Seasonal components. Various complex equations are used to determine the seasonality of the dataset. The tbats function in R is as shown in Figure 8.



```
#Exponential Smoothing
cs1.es <- tbats(ts1Train)
```

Figure 8. TBATS modeling.

In this research, the dataset is spread over 3-5 months and does not provide seasonal or trendy behavior. Therefore, TBATS model, being an adapting model was expected to provide better and impressive prediction accuracy compared to the other models. Hence, it is employed for decomposing, modeling and then forecasting energy consumption data.

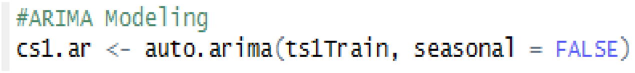
3.2.5 Autoregressive Integrated Moving Average Model

Exponential smoothing is one of the most widely used approaches in data analysis, followed by ARIMA models. Traditional regression approaches require a dependent and an independent variable for regressive forecasts. ARIMA solves the bottleneck by providing a method to fit the model to a time series dataset. In this method, differencing (smoothing out data differences) is combined with auto regression and moving averages to develop an adapting model. ARIMA model is given by,

$$y'_t = c + \emptyset_1 y'_{t-1} + .. + \emptyset_p y'_{t-p} + \theta_1 e_{t-1} + .. + \theta_q e_{t-q} + e_t \qquad (3.5)$$

$$where, y'_t - differencedseries, laggedvaluesanderrorsontherightside$$

ARIMA model implementation often requires three important arguments commonly names p, q and d representing order of the autoregressive part, order of the moving average part, and the degree of first differencing involved respectively. This research uses an automated ARIMA function which thoroughly traverses through different combinations of p, d, and q and models the time series based on the best fitting model. Automatic ARIMA implementation in R is as shown in Figure 9.

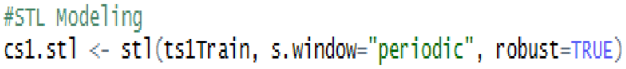


```
#ARIMA Modeling
cs1.ar <- auto.arima(ts1Train, seasonal = FALSE)
```

Figure 9. ARIMA modeling in R.

3.2.6 STL + Random Walk with Drift

Random walk with Drift is a variation of the Random walk model, otherwise referred to as the naive method of time series analysis. In the Random walk model, all the forecasts were set to the value of the last observation; while in Random walk with a drift model, a random number is added or subtracted from the previous observation and set to the forecast value. This research employs random walk with drift model to forecast a decomposed time series data set based on STL. This is a robust and versatile method for decomposition into trend, cyclic and seasonal differences with Loess to estimate nonlinear relationships. It can be implemented for data of ts, msts or numeric and univariate forms. This model can be of multiplicative and additive types. Modeling STL in R is as shown in Figure 10. S.window and t.window are arguments of concern while creating the model as its periodicity and robustness should be determined. In this research, energy consumption data is forecasted using this model after which the forecasting accuracy is determined by various other approaches.



```
#STL Modeling
cs1.stl <- stl(ts1Train, s.window="periodic", robust=TRUE)
```

Figure 10. STL Modeling in R.

3.3 Prediction Accuracy

Forecasting is said to be viable only if the forecast values pertain to the original values at the given time frame. The difference or residual between both these values should be as minimized as possible. Hence, every prediction algorithm implementation should be followed by a process to find the prediction accuracy. In R, the accuracy () method with forecast and test set as the arguments will provide a series of error results using various formulae. The most common forecast accuracy methodologies involve

- Mean Absolute Error (MAE) given by $mean\,|error_i|$
- Root Mean Squared Error (RMSE) calculated by $\sqrt{mean(error_i^2)}$

The implementation of this work progressed through three stages, viz. clustering, time series forecasting and determining the best model using prediction accuracy and residual evaluation. Time series forecasting involved modeling the dataset to fit well-known time-series models and using the forecast function to predict for a given time frame. A variety of R packages were used through the course of implementation. Time series packages such as xts, plotting packages such ggplot, vegan, permute and lattice were used for enhancing visualization.

4. Results and Discussion

The results achieved by the implementation of this work cover a wide scope and is characterized by graphical visualization and statistics. Prediction of energy in residential and industrial areas paves way to creating a sustainable environment in Oman. Hence, this section aims to analyze the results of these domains and discuss its features.

4.1 Electricity Consumption Forecasting in Residential Areas

Dataset possessing about 95000 rows which was collected from thirteen smart meters was used to predict electricity consumption in Omani residential areas. Due to high density, the data was clustered for optimizing resource usage and minimizing execution time. Several models were used to model and forecast the consumption in each cluster. This section deals with the electricity consumption prediction of each of the four clusters formed while mining residential data.

4.1.1 Forecast Visualization

Forecast visualization refers to visualizing the forecast graphs against the test set. This method gives a raw idea about the correctness of prediction of each model for all the clusters. The forecasts and original data of residential energy consumption using TBATS model is depicted in Figure 11. It is evident from the figure that the model has fitted all four clusters with very minimal errors, and hence produced near perfect results throughout. The green graph indicates the predicted values while black represents the original data. The alignment of predicted values to the test set explains the adaptability of the model to fit any season, trend and rapid changes. The usage of trigonometric functions to exponentially smooth out the dataset has allowed this model to visually create better predictions than the STL + RWD model. When clusters are compared using the prediction from this model, it can be noted that cluster IV is the least aligned prediction, while cluster III acquired excellent forecast values. Even though the model is flexible, it can be seen that its performance declines during random outbursts in data, but still renders better results than other time-series models (from visualizing other graphs).

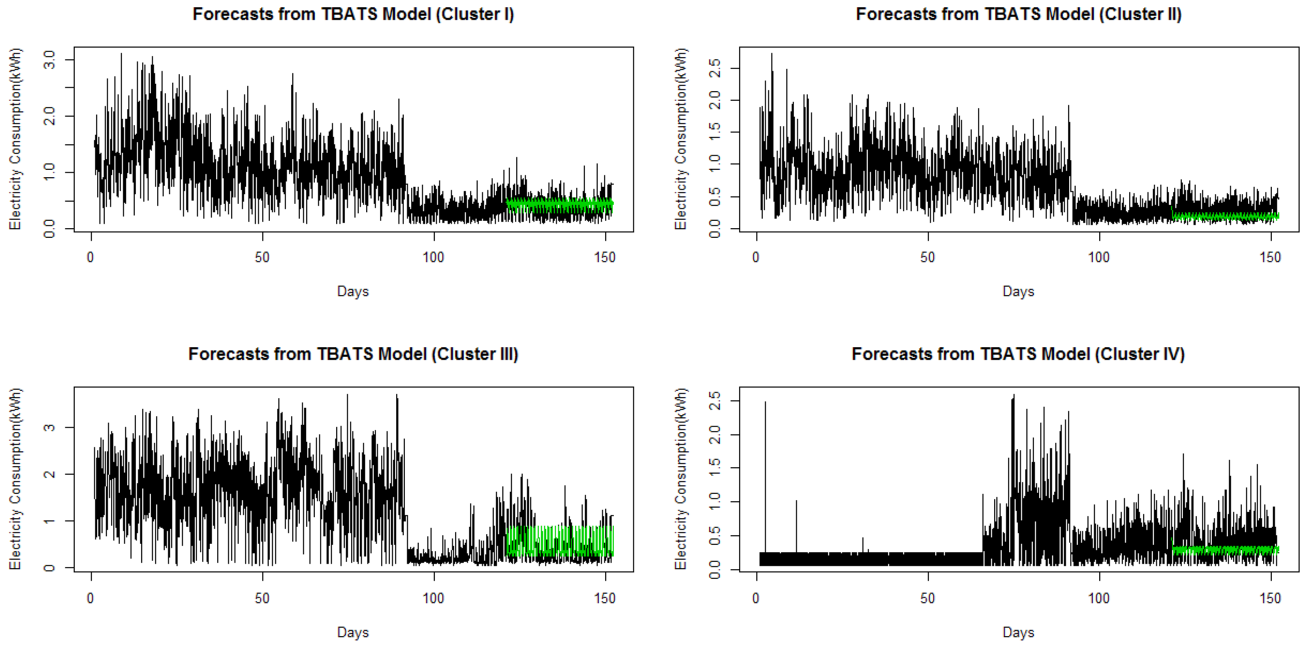


Figure 11. Forecasts from TBATS model c.

4.1.2 Residual Diagnostics

Residuals are defined as the difference between the observed value and the forecast for the particular observation. In other words, it determines how close the forecast is to the original test set. As expected, a residual of zero indicates that the model precisely fits and predicts the given dataset. In this section, residual of TBATS model fitted to residential data is diagnosed to analyze its effectiveness in predicting various cluster consumptions.

Residual plot shown in Figure 12 from TBATS model does not indicate concurrency to zero, but the uniform distribution displayed by all clusters conforms to a healthy prediction characteristic by the model. All clusters have spikey features referring to the difference from original data, but the spikes adhere to a uniform distribution as each cluster accommodates residuals with a maximum difference of 1.

Illustration of ACF of TBATS in Figure 13 displays that some correlation exists between the observations, its values are lesser and comparatively negligible in Clusters I, II and III. Cluster IV represents more correlation indicating that TBATS might not fit the cluster as effectively as it does to the other three clusters.

The residual histograms for TBATS model presented in Figure 14 render surprising results. It was observed in the previous two residual diagnostic methods that cluster IV is seemingly outcast while the other three clusters would fit the TBATS model in a fairly good manner. The histogram conversely shows that cluster III has a longer right tail eliminating the possibility of a perfect normal distribution. Hence, from residual histograms, it can be concluded that TBATS model will not fit cluster III as anticipated. The prediction accuracy will further determine the efficiency of this implementation.

4.1.3 Prediction Accuracy

As stated in Section 4.3, traditional methods were used to evaluate the prediction accuracy of all the time-series models. MAE and RMSE were used to determine if the values of 20% of the dataset aligns closely with the forecast made. They are also used to compare the prediction accuracies of each model and finally determine the most effective forecasting model if any, for each cluster.

All the error values for easy analysis of the results are summarized pictorially in Figure 15. Initial visualization indicates that all clusters have managed to create an error with a maximum of 0.79 kWh by cluster 4 while forecasting the average model. Further analysis indicates that, for cluster I, TBATS model renders the minimum MAE as well as RMSE with 0.15 kWh and 0.19 kWh respectively. These values are negligible and hence indicate a good prediction result. Cluster II is considerably the

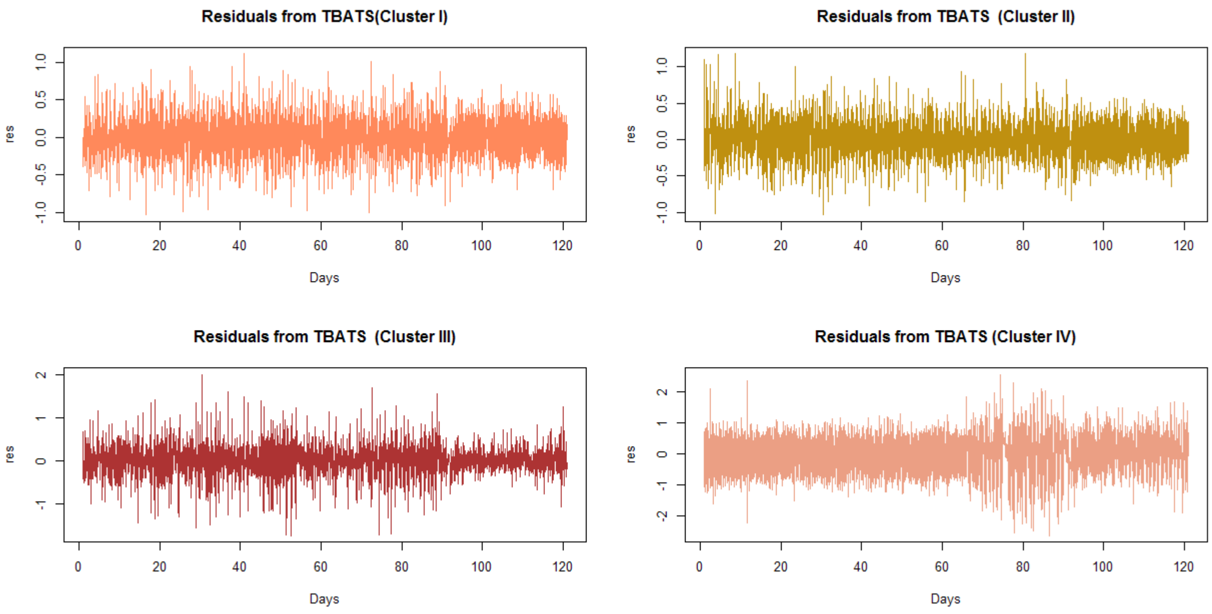


Figure 12. Residuals from TBATS model.

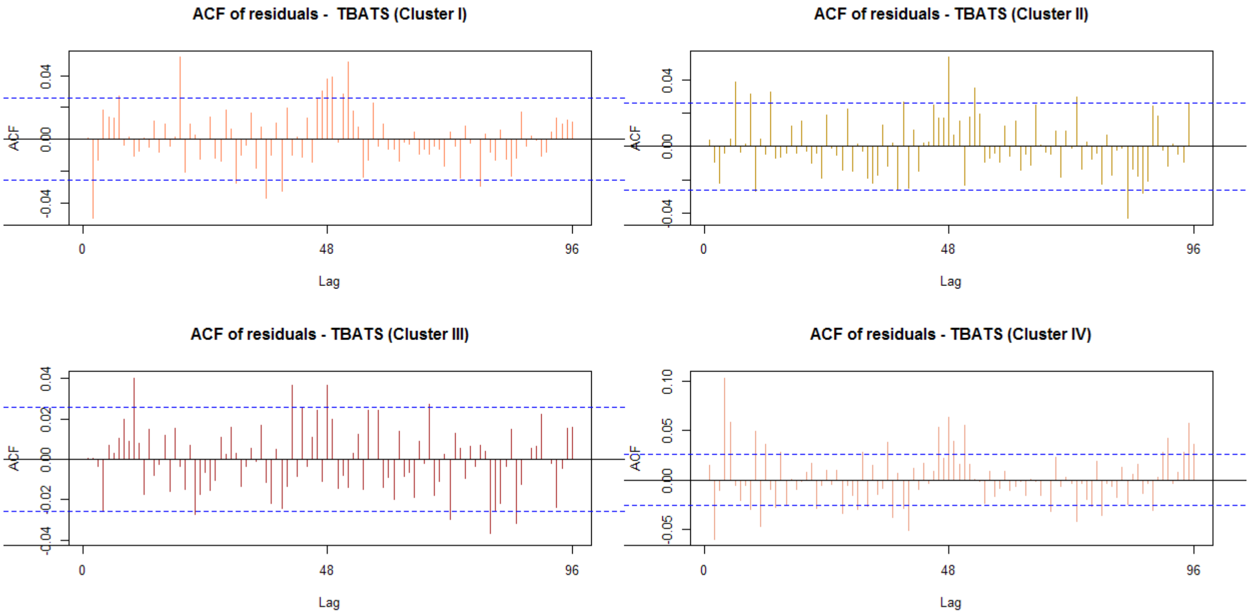


Figure 13. ACF of residuals – TBATS model.

best cluster modeled in the residential sector as all the models very minimal error with the least being TBATS model. An MAE and RMSE of 0.12 kWh and 0.16 kWh illustrate an impressive error rate, indicating the models effectiveness to fit the time series data. Cluster III, on the other hand failed to impressively fit any of the models used in implementation; but as compared to other models, TBATS secured better results at prediction with an MAE and RMSE of 0.3 and 0.33 kWh respectively. The final cluster, viz. cluster IV possesses the greatest variation in errors among the four clusters. Even though ARIMA and TBATS are seen impressive in modeling the cluster data, minimal errors were secured by TBATS with 0.17 and 0.24 kWh for MAE and RMSE respectively.

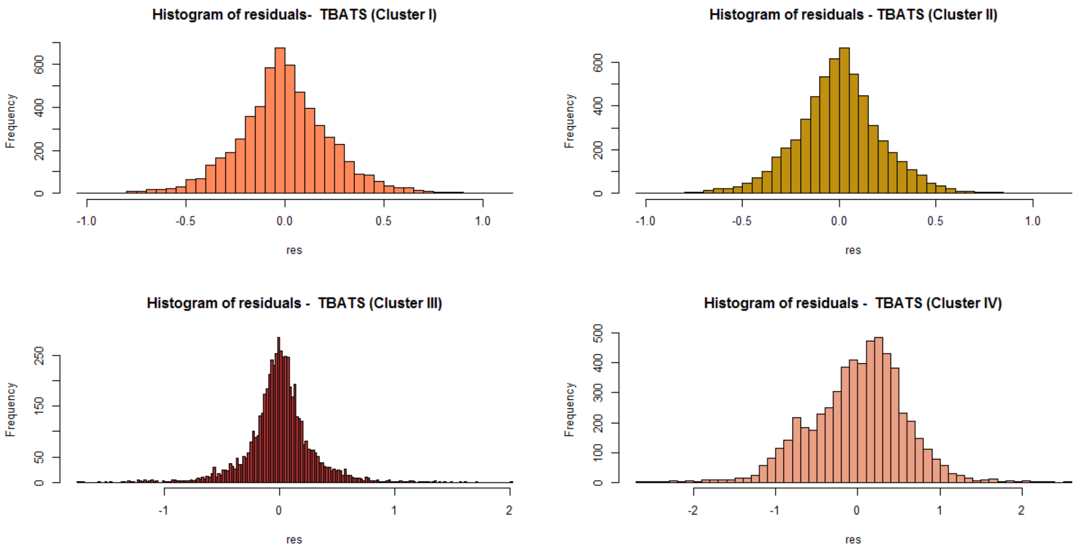


Figure 14. Histogram of residuals - TBATS model.

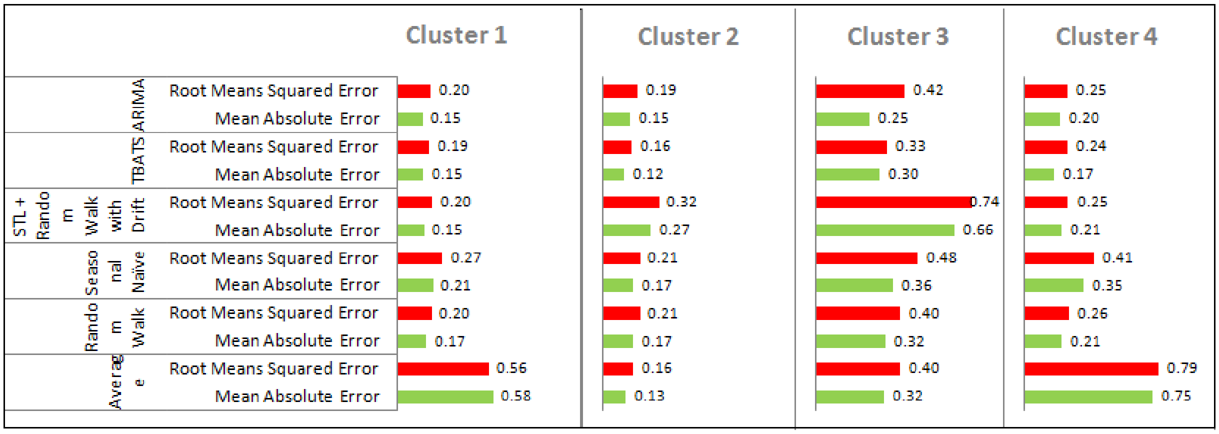


Figure 15. Prediction accuracy of electricity consumption in residential areas.

Therefore visualization, residual diagnostics and prediction accuracy together determined that there exist data models which can accurately forecast energy consumption of the residential sector in Sultanate of Oman. In this research, all four clusters displayed minimal errors when modeled using TBATS time-series model displaying its potential to be adopted as the forecast model for energy consumption in the residential sector in Oman.

4.2 Electricity Consumption Forecasting in the Industrial Sector

Cities are characterized by residential as well as industrial areas. Industries consume more energy than houses, at a rate of MWh. In order to conclude success in mining energy consumption data, it is necessary to determine whether these models can accurately forecast energy in industries as well. This section traverses through the visualization, residual diagnostics and prediction accuracy for industrial energy consumption forecasts.

4.2.1 Forecast Visualization

In this sub-section, the best model for a cluster is determined by visualizing the forecasts of all the models of a particular cluster together. The energy consumption in industrial sector was divided into three clusters and the forecast of Cluster III for various models is as shown in Figure 16. In the first graph, mean (blue), naive (red) and seasonal naive (green) methods is observed to have successfully predicted the energy consumption as all the

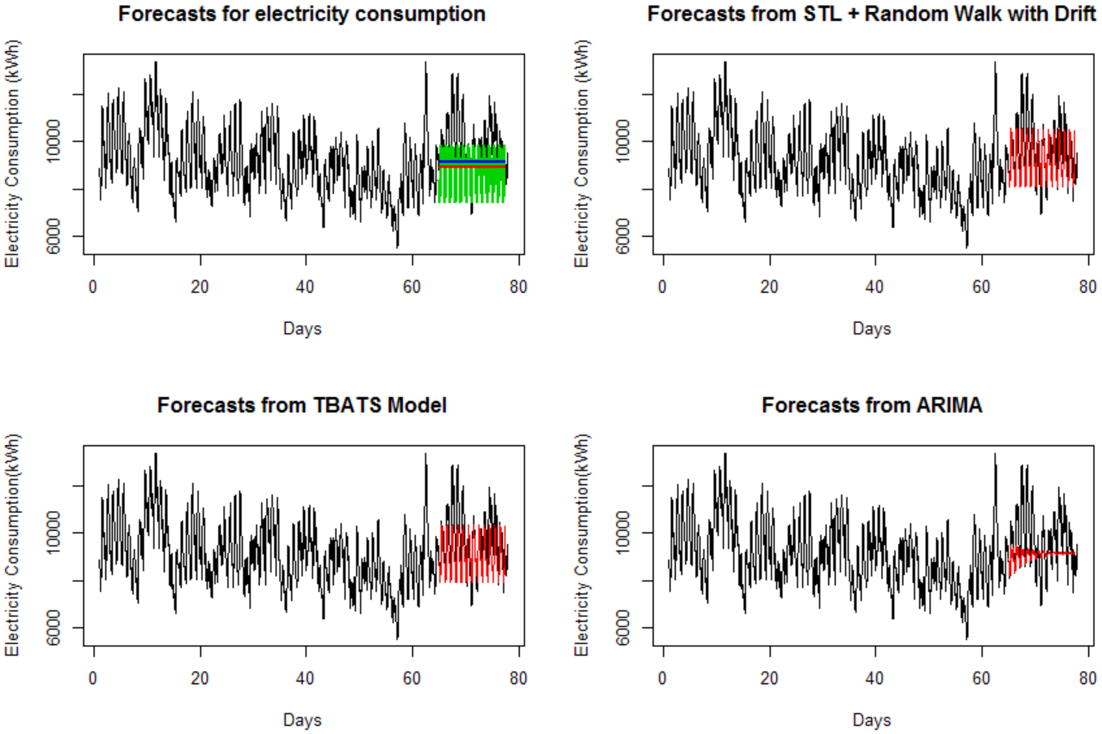


Figure 16. Forecasts from various models (Industrial - Cluster III).

forecasts are aligned to the original dataset. Moreover, seasonal naive method renders almost perfect prediction for the given forecast horizon. The seasonality of the given cluster data might be the reason as to why clear results are acquired. STL + RWD provides almost the same forecast as that of TBATS model indicating the existence of difficulty over determining the effective one for the given cluster. Even though ARIMA renders a varying prediction, it does not visibly align with most of the original dataset and hence, does not seemingly supply more advantages as compared to TBATS and STL + RWD.

4.2.2 Residual Diagnostics

Cluster III had higher data density than the other two clusters in the industrial sector. The residuals for this cluster possess uniform features when calculated from each of the four models taken into consideration as graphically depicted in Figure 17. The seasonal naive model is the most scattered among the group, while the rest follow a uniform fluctuating pattern from the center value, viz. 0.

Seasonal naive method display heavy correlation characteristic; while STL + RWD, TBATS and ARIMA signify minimal correlated features but exceed the limitations at a higher rate than expected in Figure 18.

The histograms of the residuals in Figure 19 illustrate that only seasonal naive method has residuals which are normally distributed. The other models display an elongated right tail indicating the potential to be improved.

Residual diagnostics for all three clusters rendered confusing results, and hence prediction accuracy will be used to finally determine the effectiveness of each model for forecasts.

4.2.3 Prediction Accuracy

Prediction accuracy for Industrial energy consumption data is depicted in Figure 20. MAE and RMSE values are grouped based on the clusters and models used for forecasting. Each cluster varies based on their consumption values which result in the variation in MAE and RMSE errors among the three clusters. For cluster I, TBATS model creates the minimal error of 36.4 and 47.02 kWh of energy. This is only considered as a minimal deviation when compared to the industrial consumption values. In cluster 2, the average method renders minimal errors with MAE and RMSE as 202.13 and 542.51 kWh respectively. Average method is a simple forecasting method and is not

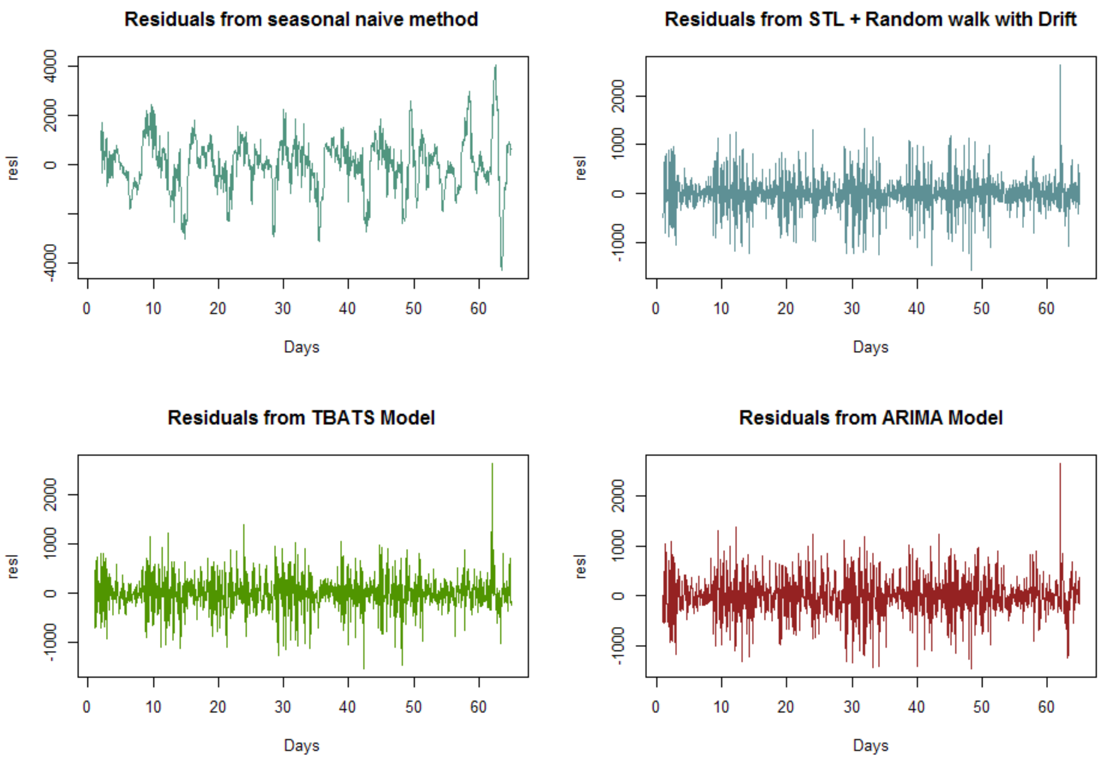


Figure 17. Residuals of various methods (Industrial - Cluster III).

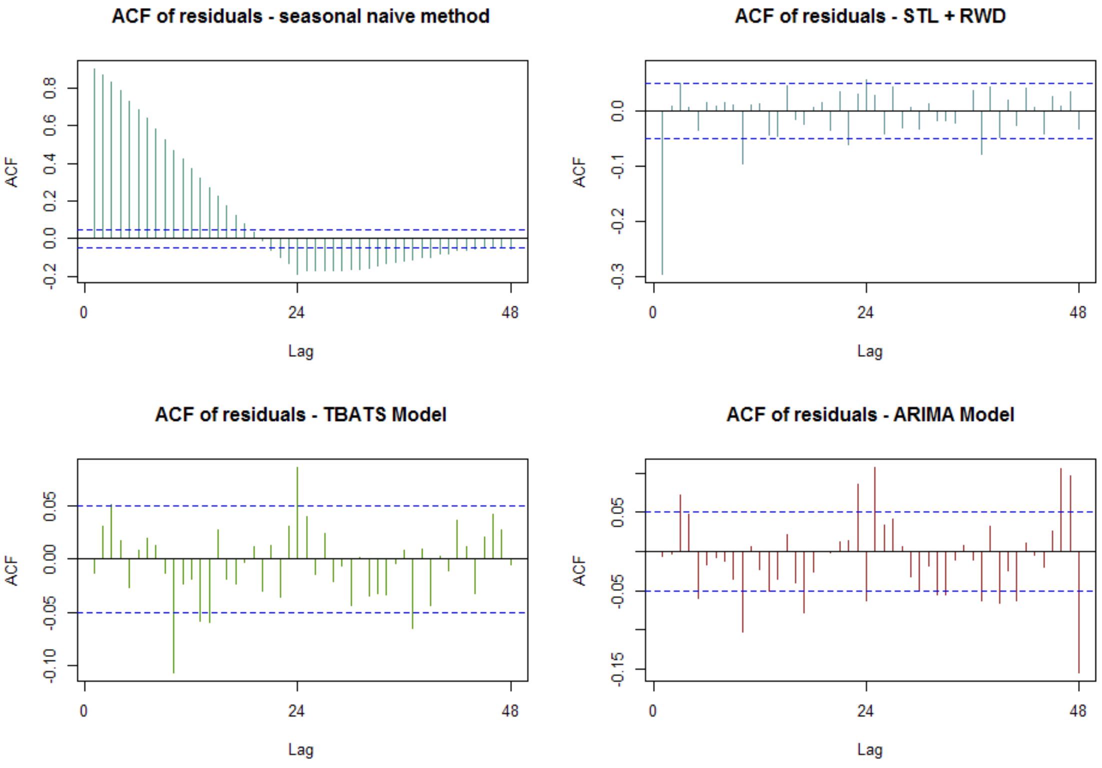


Figure 18. ACF of residuals (Industrial - Cluster III).

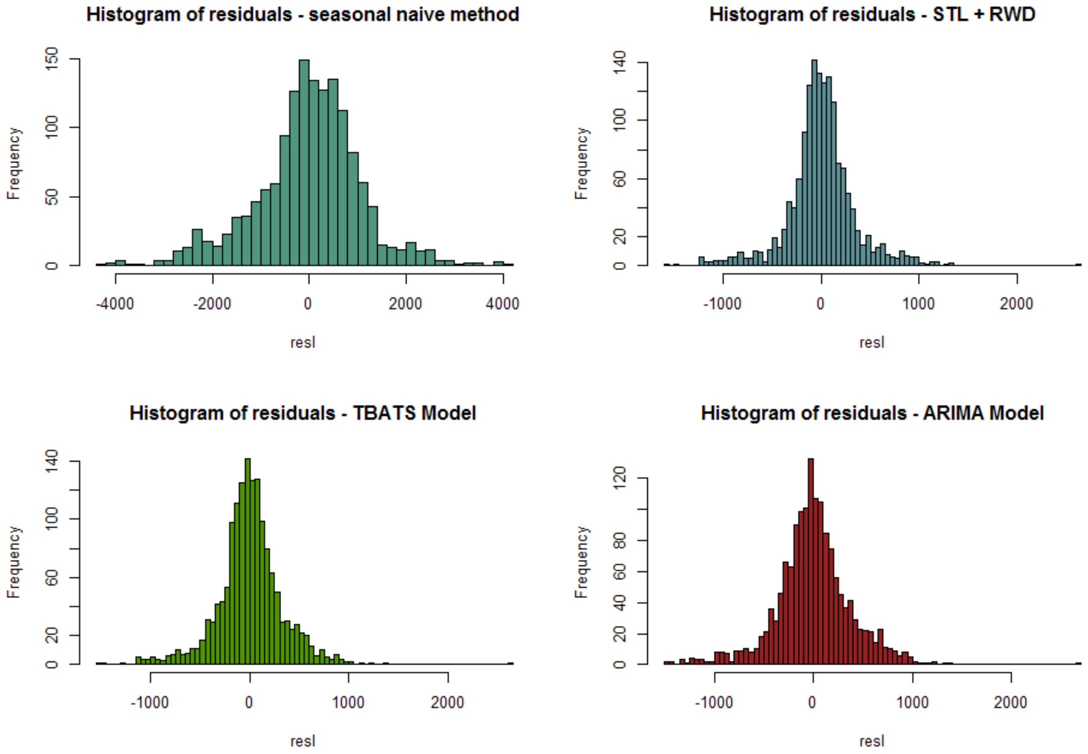


Figure 19. Histogram of residuals (Industrial - Cluster III).

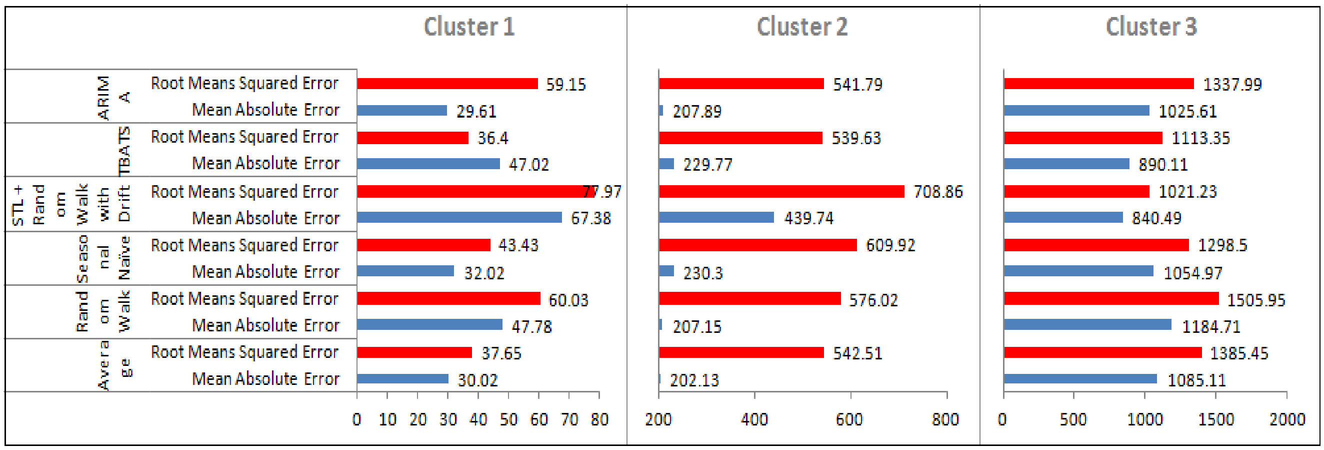


Figure 20. Prediction accuracy for Industrial Data.

commonly used for electricity prediction mechanism, but this model is the best suited for cluster II for the given forecasting horizon. For very large consumption values in cluster III, the least errors were provided by STL + RWD model. Hence, the three clusters which are characterized by varying consumption levels are best fitted by TBATS, Average and STL + RWD models.

This section dealt with forecasting energy consumption in residential and industrial areas and analyzing the results. The effectiveness of using DMTs to predict the consumption of energy in cities and create a sustainable environment is evaluated and rendered positive results. Despite the lack of big data, it was found that TBATS, STL + RWD and ARIMA could fairly predict the consumption of at least 89% of the clusters formed in residential and industrial sectors. All cluster consumptions and predictions are compared and contrasted in this chapter and the best model for electricity forecast is determined.

5. Conclusions

This research presented the effectiveness of smart city implementation in Oman by examining the energy consumption sector through data mining. K - Means clustering technique was used to group residential and industrial energy consumption data into four and three clusters respectively. Energy usages of these clusters were then forecasted using 80% of the total data. The time-series models used for this purpose include Average, Naive, Seasonal Naive, decomposition using STL + Random Walk with Drift, TBATS, and ARIMA models. Algorithms for clustering, modeling and predicting energy consumption were facilitated efficiently. Visualization determined that the trend of energy consumption in residential areas stay uniform during summer and decreases towards the end of the year (winters), while in industries, the energy consumption stayed uniform during the period for which the data was collected. The results indicate that electricity consumed in residential apartments can be fairly predicted by TBATS model. This prediction would allow inhabitants to regulate their consumption and related authorities to generate only the predicted amount of electricity. It would release the pressure on oil and gas reserves and help minimize the budget for the nation. In the industrial sector, consumption in small, medium and large scale industries were predicted by TBATS, Average, and STL + RWD models respectively indicating a trend in adopting forecasting models for different types of industries.

Implementation of this research covered only a small amount of data. With the acquisition of energy consumption values for a period of over a year, Big Data analytics can be used with an architecture integrating Hadoop based applications with R. Larger dataset would also promote the existence of trends and seasonality rendering more accurate results. Forecasts can be further improved if more than one attribute and relationships between multiple attributes are used for prediction. For example, upgraded forecasts can be achieved if data about the temperature at the particular time frame is available along with the energy consumed.

6. Acknowledgments

The authors wish to acknowledge Caledonian College of Engineering for the opportunity to conduct this research and Muscat Electricity Distribution Center for access to an accurate dataset.

7. References

1. Costa C, Santos MY. Improving cities sustainability through the use of data mining in a context of big city data. Proceedings of the 2015 International Conference of Data Mining and Knowledge Engineering, London. 2015,pp. 320-25. PMid:26056886
2. Guzey O. Data Mining in Constrained Random Verification. PhD Dissertation. Santa Barbara: University of California, Department of Electrical and Computer Engineering: Santa Barbara, 2008.
3. Figueiredo V, Rodrigues F, Vale Z, Gouveia JB. An Electric Energy Consumer Characterization Framework based on Data Mining Techniques. IEEE Transactions on Power Systems. 2005 May, 20(2), pp. 596 - 602. Crossref
4. Sultanate of Oman Renewables Readiness Assessment. Crossref. Date accessed: 16/04/2016.
5. Oman Energy Situation. Crossref. Date accessed: 21/03/2016.
6. These are the most sustainable cities in the world. Crossref. Dateaccessed: 8/06/2016.
7. Residential Energy Use In Oman:A Scoping Study. Crossref. Date accessed:12/12/2015.
8. Zurigat YH. Analysis of Typical Meteorological Year for Seeb, Muscat, Oman. International Journal of Low Carbon Technologies. 2007 Apr; 2(4):323-38. Crossref
9. Smart Data & Well-being. 2015. Available from: Crossref
10. What a smart home can do. 2016. Available from: Crossref
11. Energy efficiency can halve gas consumption in Oman. 2016. Available from: Crossref
12. Four main languages for Analytics. Data Mining, Data Science. 2016. Available from: http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html
13. R Documentation. 2016. Available from: Crossref
14. Forecasting: Principles and practice. 2016. Available from: Crossref
15. Sajana T, Rani CMS, Narayana KV. A survey on clustering techniques for big data mining. Indian Journal of Science and Technology. 2016 Jan; 9(3):1-4. Crossref