OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 4, Issue. 2, Sept 2019. pp. 109-118
doi:10.21108/indojc.2019.4.2.329

Empirical Comparison of Time Series Data Mining Algorithms for Electrical Power prediction

Folorunso S. O.[1], Taiwo, A. I.[2], Olatayo, T. O.[3]
Department of Mathematical Sciences, Olabisi Onabanjo University, Ago Iwoye, Ogun State, Nigeria
sakinat.folorunso@oouagoiwoye.edu.ng[*1]

Abstract

Time series is a sequence of observed data that is usually ordered in time. Time series data mining is the innovative application of the principles and techniques of data mining in the analysis of time series. This research is aimed to apply data mining techniques to forecast time series data. Nigeria electric power consumption data applied from 2001 to 2017. Experiments are conducted with four data mining techniques: Random Regression Forest (RRF), Linear Regression (LR), Support Vector Regression (SVR) and Artificial Neural Network (ANN) which were evaluated based on their forecasting errors generated: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and prediction accuracy on Waikato Environment for Knowledge Analysis (WEKA) platform. The combination of parameters that yields the best results in terms of predefined performance criteria was chosen as optimal for each regressor. A comparative analysis of the regressors' performance was conducted. All the tested regressors have demonstrated the best prediction quality in short periods of time. SVR demonstrated the best results in terms of both error values and time expenses.

Keywords: Time Series Forecast, Time Series Data Mining, Classification Algorithms, Regression analysis, Power consumption

Abstrak

Time series abstrak adalah urutan data yang diamati yang biasanya dipesan dalam waktu. Data time series mining adalah aplikasi inovatif dari prinsip-prinsip dan teknik data mining dalam analisis time series. Penelitian ini bertujuan untuk menerapkan teknik data mining untuk meramalkan data time series. Data Nigeria konsumsi daya listrik diterapkan dari tahun 2001 ke 2017. Percobaan dilakukan dengan empat teknik data mining: Random Regresi Hutan (RRF), Linear Regression (LR), Support Vector Regression (SVR) dan Artificial Neural Network (ANN) yang dievaluasi berdasarkan pada kesalahan peramalan mereka dihasilkan: Berarti Absolute kesalahan (MAE), root mean Square kesalahan (RMSE), Berarti Absolute Persentase kesalahan (MAPE) dan akurasi prediksi di Waikato Lingkungan untuk Analisis Pengetahuan (Weka) platform. Kombinasi parameter yang menghasilkan hasil terbaik dalam hal kriteria kinerja yang telah ditetapkan terpilih sebagai yang optimal untuk setiap regressor. Sebuah analisis komparatif kinerja regressors' dilakukan. Semua regressors yang diuji telah menunjukkan kualitas prediksi terbaik di jangka waktu yang singkat. SVR menunjukkan hasil terbaik baik dari segi nilai kesalahan dan biaya waktu.

Kata kunci: Time Series Prakiraan, Time Series Data Mining, Klasifikasi Algoritma, analisis regresi, Konsumsi daya

I. Introduction

Electricity is one of the basic amenities for the mankind. A range of daily activities such as operating domestic and industrial equipment, lighting, heating, air-conditioning, cooking, washing and many other tasks depends largely on it. Forecasting of electricity consumption is necessary to manage the power system effectively and the demand rate. Power company would require the power consumption forecast to plan their future activities properly, such as building adequate power plants and improve their transmission and distribution networks to

Received on August 2019. Accepted on Sept 2019

meet the necessary demand. Also, forecasting is required to perform daily operations such as unit commitment, energy transfer scheduling and load dispatch of a utility company [1]. Therefore, accurate prediction of electricity consumption is crucial for both, performing daily operations and making future power plans for a power supplying company. Electricity power consumption is a time series data and time series data can be defined as a sequence of time-ordered data $\{TS_t,\ t = 1, \cdots, N\}$, where t represents time, N is the number of observations made during the time period and TS_t is the value measured at time instant t [2].
In order to obtain prediction for time series data using time series models involves the prediction of future values based on the previous values and the current value of the time series. The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred to as short-term prediction. But when multi-step ahead predictions are needed, it is called a long-term prediction problem. Unlike the short-term time series prediction, the long-term prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult [3].

Even though, some time series models like Autoregressive moving average (ARMA), Autoregressive integrated moving average (ARIMA) and Seasonal Autoregressive integrated moving average (SARIMA) are capable of capturing trend, cyclic and seasonal patterns in time series data, but may fail to capture the effect of other independent factors which are non-seasonal and non-cyclic in nature. In hence, this may lead to less accurate prediction and this cause an incorrect decision and conclusion [4].

Time Series Data Mining is a recent approach that can be used to analyse time series data. It can reveal hidden patterns that usually characterized and affect prediction of time series events. Mining time series data had been one of the challenges of data mining [5]. Some of other challenges faced by time series data mining are the large volume data, their unique structure, very high dimensionality, high feature correlation and large amount of noise [6]. The key insight that allows meaningful time series data mining is that virtually all time series data mining algorithms avoid operating on the original raw data instead, they consider some higher-level representation or abstraction of the data.

This approach has been used by several researchers over the years and these include the works of [7] where they applied Support Vector Machine (SVM) and Back Propagation-Artificial Neural Network (BP-ANN) data mining models to five different benchmark time series datasets with 10-fold cross validation. Their result indicated that SVM outperform BP_ANN based on the values of Root Mean Squared Error (RSME), and Mean Absolute Error (MAE) respectively. [8] proposed SVM to solve the system level electricity load prediction problem. Based on the performance of SVM, they concluded that its outperformed Reduced Error Prunning Tree (REPTree), Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) and Artificial Neural Network (ANN) with three layers for predicting the total electricity load with a satisfactory accuracy of Coefficient of Variation of the Root Mean Square Error (CV_RMSE) of 15.2% and Normalized Mean Bias Error (N_MBE) of 7.7%. [9]applied Auto Regressive Integrated Moving Average (ARIMA), Linear Regression and Segment based forecasting and Artificial Feed-Forward Network with multiple hidden layer to predict sales and analysed the impact of promotions on sales activities. From their results, Artificial Neural Network outperformed the other models as it had the minimum error. [7] applied eleven forecasting models Persistence (P), Scaled Persistence (SP), Auto Regressive Moving Average (ARMA), Multi-Layer Perceptron (MLP), Regression Trees (RT), Boosted Regression Trees (RT-Boosted), Bagged Regression Trees (RT-Bagged), Pruned Regression Trees (RT-Prunned), Random Forest (RF), Gaussian Processes (GP) and Support Vector Regression (SVR)) to forecast the time series of horizontal global solar irradiation measurements (GHI) in Ajaccio. The performance of these models was evaluated based on Normalized Root Mean Squared Error (NRSME), Mean Absolute Error (MAE) and skill score related to the smart persistence.

This research involves supervised learning technique where we will infer a function from the labelled training examples (training data). Each of these training examples is a pair consisting of an input object (typically a vector) and an output value (scalar value). Depending on the type of the output values, the problem of inferring falls into two categories and these are regression and classification. Here, the focus is on regression rather than classification. By decomposing a time series, this becomes a standard regression problem, where the current predicted value depends on a block of past values. Therefore, this study aimed at forecasting time series data (electricity power consumption) using data mining regression models. Four different data mining models that will be considered are (Random Forest Regression (RFR) [10], Linear Regression (LR), Support Vector Regression (SVR) [11] and Multi-Layer Perceptron Regression (MLR). These models will be used to extract

patterns and rules from the time series data rather than just finding their statistical properties, identify their components, forecast a 1-step ahead (1 year) of electricity consumption and comparison of their forecast accuracy.

II. MATERIALS AND METHODS

The stages in this study is depicted by figure 1. The dataset will be preprocessed into a supervised learning format which is easy to be used by the data mining algorithms. This is made possible based on information on some of the features of the time series data. The preprocessing can have big impact on the subsequent data mining forecasting performance [12].

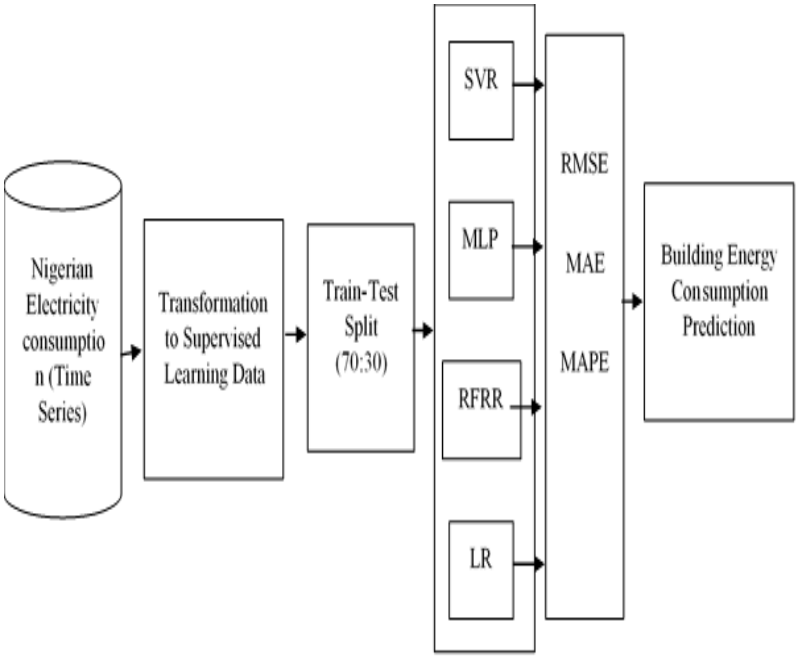


Fig. 1. Block diagram of the stages for energy consumption prediction

A. Data Mining Process

A time series may be formalized as a sequence of scalar random observations $S = s_0, \cdots, s_{i-1}, s_i, s_{i+1}$. The lag of the series is given by the delay used to form the model training and testing data. Then, time series forecasting means predict a future value of the sequence given by (1)

$$\bar{S}_{i+1} = F\left[s_i, s_{i-k}, s_{i-2k}, \cdots, s_{i-(d-1)k}\right] \quad (1)$$

where d is the lag, k is the step lag and F is the data mining model [13]. The data mining models used in this study are (LR, RFR, SVR and MLR)

B. Data input selection

The research formulation is given by (2)

$$E_{(t+h)} = f\left(E_t, E_{t-1}, \cdots, E_{t-n} + \epsilon(t+h)\right) \quad (2)$$

where f is a function of time, h is a step ahead and $\epsilon(t+h)$ is a random white noise.

The electricity consumption data at future time step $(t+h)$, $E(t+h)$ is forecasted based on the observed data E at the times $(t, t-1, \cdots, t-n)$. In other words, the objective f is to calculate the value of n and to obtain $\epsilon(t+h)$ as low as possible in absolute value. The choice of n, that is, the dimension of the input matrix, is made by an auto mutual information method [14]. This auto mutual information is a property of the time series and depends on each dataset. It determinates the degree of statistical dependence of the variables specific to each site.

Another step of pre-process is the hold-one-out (70:30) sampling. The dataset is divided into 70:30 ratio samples and 70% of the sample size is used one time for the training and 30% is used one time for the test. This method allows to have results independent of the set of data used for the training (only one data set being able to have some particularities that disturbs the robustness of the conclusions).

C. *Random Regression Forest (RRF)*

Random Regression forests involve the modified tree learning algorithm that selects at each candidate split in the learning process a random subset of the features. This process is sometimes called feature bagging. This is done to attain the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated [15].

D. *Linear Regression (LR)*

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression [16]. A linear regression is defined as (3)

$$y = a + BX + e \tag{3}$$

where y is the dependent variable, B is the slope, X is the independent variable and e is the random error.

E. *Support Vector Regression (SVR)*

Support Vector Regression (SVR) is the regression process performed by a Support Vector Machine which tries to identify the hyperplane that maximizes the margin between two classes and minimize the total error under tolerance [11]. In order for an efficient SVM to be constructed, a penalty of complexity is also introduced, balancing forecasting accuracy and computational performance. Since in the present study accuracy is far more important than complexity, forecasts were produced using an ϵ-regression SVM which maximizes the borders of the margin under suitable conditions to avoid outlier inclusion, allowing the SVM decide the number of the support vectors needed. The kernel used in training and predicting is the radial basic one, mainly due to its good general performance and the few parameters it requires. Following the suggestions of [17] implementation in Waikato Environment for Knowledge Learning (WEKA) [18], all default values were used. The SVR model used is constructed exploiting the *SVM* function in WEKA data mining tool.

F. *Artificial Neural Network (ANN-MLP)*

For this study, MLP with one hidden layer using WEKA's Optimization class by minimizing the given loss function plus a quadratic penalty with the BFGS method [18]. The activated loss function used here was Sigmoid and the Squared Error Note that all attributes are standardized, including the target. There are several parameters. The ridge parameter is used to determine the penalty on the size of the weights. The number of hidden units can also be specified. Note that large numbers produce long training times. Finally, it is possible to use conjugate gradient descent rather than BFGS updates, which may be faster for cases with many parameters. To improve speed, an approximate version of the logistic function is used as the default activation function for the hidden layer, but other activation functions can be specified. In the output layer, the sigmoid function is used for classification. If the approximate sigmoid is specified for the hidden layers, it is also used for the output layer. For regression, the identity function is used activation function in the output layer. Also, if delta values in the backpropagation step are within the user-specified tolerance, the gradient is not updated for that particular instance, which saves some additional time. Parallel calculation of loss function and gradient is

possible when multiple CPU cores are present. Data is split into batches and processed in separate threads in this case. Note that this only improves runtime for larger datasets. Nominal attributes are processed using the unsupervised Nominal to Binary filter and missing values are replaced globally using Replace Missing Values.

G. Evaluation Metrics

Metrics are used to measure the performance of models upon evaluation on the dataset. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used in this research. Here, the actual consumption value at time t is denoted as P_t, N as the number of instances and the forecast value at the same instance as F_t. Based on this notation, error measurements were calculated in the following forms (4), (5), (6):

$$RMSE = \frac{1}{N}\sqrt{\sum_{t=0}^{N}(P_t - F_t)^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{t=0}^{N}|P_t - F_t| \tag{5}$$

$$MAPE = \frac{100}{N}\sum_{t=0}^{N}\left|\frac{P_t - F_t}{P_t}\right| \tag{6}$$

III. RESULTS AND DISCUSSION

The data used in this study is the yearly electricity consumption rate of Nigeria ranging from year 2000 – 2017 [19]. This dataset consists yearly electricity consumption measured for 17 successive years. This entry consists of total electricity generated annually plus imports and minus exports, expressed in kilowatt-hours. The discrepancy between the amount of electricity generated and/or imported and the amount consumed and/or exported is accounted for as loss in transmission and distribution. The Original dataset consist only of year and the value of the consumed electricity energy depicted by Table 1. The transformed supervised learning dataset is represented by Table 2. The target variable is the Electricity power consumption of the next year (E_{t+h})

TABLE 1: THE DESCRIPTION OF THE DATASET

Feature Variable	Notation	Type
Year	Y	Date
Electricity - consumption (billion kWh)		Numeric

TABLE 2: THE SUPERVISED LEARNING DATASET

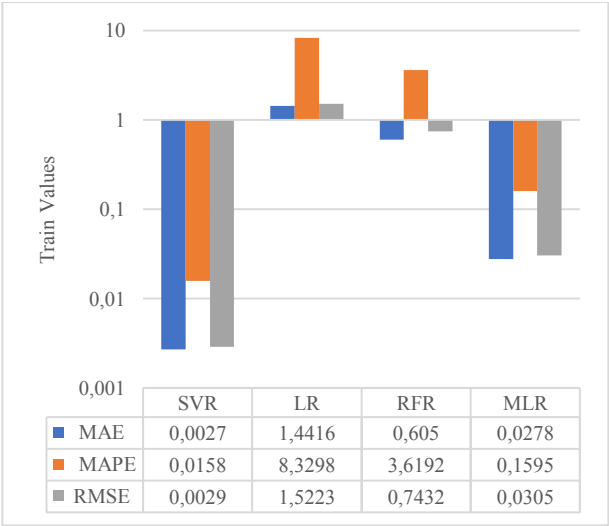
Feature Variable	Notation	Type
Year	Y	Date
Electricity - consumption (billion kWh)		Numeric
Lag_Electricity - consumption (billion kWh)-2	E_t Lag2	Numeric
Lag_Electricity - consumption (billion kWh)-3	E_t Lag3	Numeric
Lag_Electricity - consumption (billion kWh)-4	E_t Lag4	Numeric
Lag_Electricity - consumption (billion kWh)-5	E_t Lag5	Numeric
Lag_Electricity - consumption (billion kWh)-6	E_t Lag6	Numeric
Lag_Electricity - consumption (billion kWh)-7	E_t Lag7	Numeric
Year^2		Numeric

Year^3		Numeric
Year*Lag_Electricity - consumption (billion kWh)-2	Y*Lag2	Numeric
Year*Lag_Electricity - consumption (billion kWh)-3	Y*Lag3	Numeric
Year*Lag_Electricity - consumption (billion kWh)-4	Y*Lag4	Numeric
Year*Lag_Electricity - consumption (billion kWh)-5	Y*Lag5	Numeric
Year*Lag_Electricity - consumption (billion kWh)-6	Y*Lag6	Numeric
Year*Lag_Electricity - consumption (billion kWh)-7	Y*Lag7	Numeric

For modeling of the data mining regression models in this research, an open source machine learning workbench Waikato Environment for Knowledge Learning (WEKA) [18] was used. Weka's time series forecast framework takes a machine learning/data mining approach to modeling time series by transforming the data into a supervised learning data form that standard data mining model can process. It does this by removing the temporal ordering of individual input instances by encoding the time dependency via additional input fields. These fields are sometimes referred to as "lagged" variables. Various other fields are also computed automatically to allow the algorithms to model trends and seasonality. After the data has been transformed to supervised learning, then data mining regression models can be applied to learn and forecast the data.

This is a 1-step ahead forecast so only one unit specified to predict the next year only (2018). The time stamp is year and the periodicity of the data is yearly with a confidence level of 95%. The time series data contains just two variables which are year and the values of the consumed electricity power in billions of kwh. The consumption values were transformed to a supervised learning data with 15 variables consisting of the year, 6-Lagged variables, 2-powers of time and 6- product of time and lagged variables. The dataset was divided into Train-Test data to the ratio 70% :30%. The leading instances with unknown values were removed.

The prediction performance of the four data mining regression models with WEKA default values on the training and testing data are shown in fig. 2 to 3. The 1- step ahead prediction for the year 2018 for each regression model from the actual and predicted values is given in fig. 4 to 7. It can be seen from the four different models built for the forecast that Linear Regression model performed poorest as it had the highest metric values based on MAE, MAPE and RMSE with 1.4416, 8.3298 and 1.5223 values respectively.



	SVR	LR	RFR	MLR
MAE	0,0027	1,4416	0,605	0,0278
MAPE	0,0158	8,3298	3,6192	0,1595
RMSE	0,0029	1,5223	0,7432	0,0305

Fig. 2. Comparison of regression models on train data

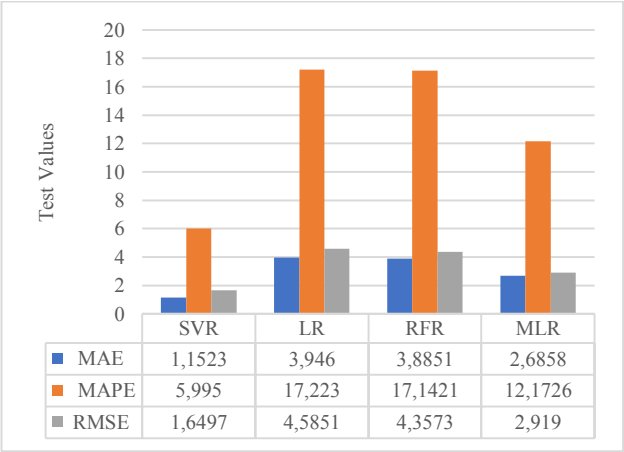


Fig. 3. Comparison of the regression models on test data

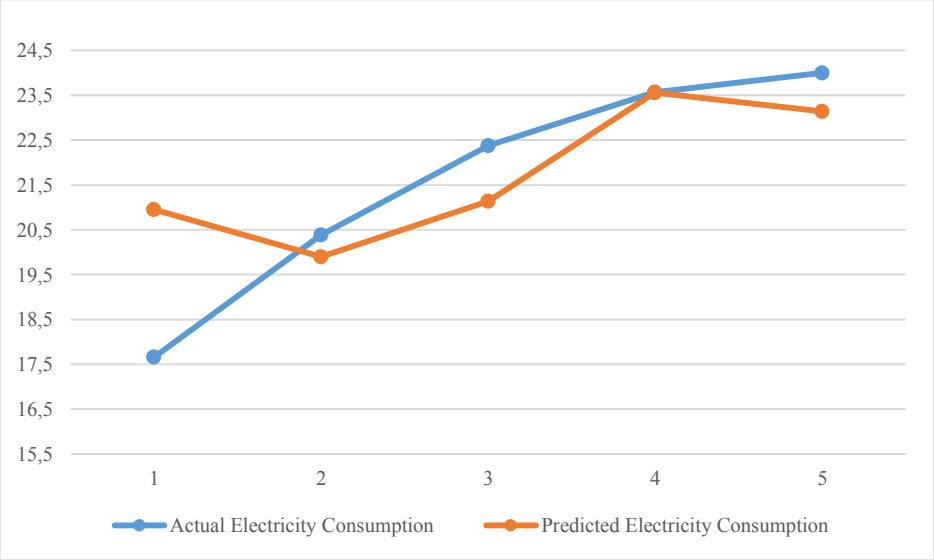
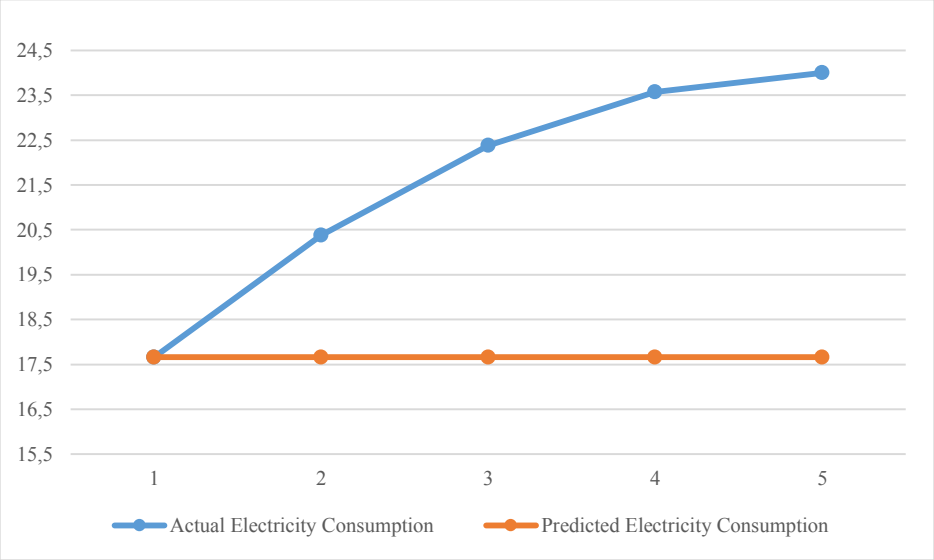


Fig. 4. 1-Step ahead forecast of Support Vector Regression (SVR)

Based on fig. 3 to 7, the support vector regression (SVR) model performed better than other regression models obtained. Though all the four regression models are good models for forecasting Nigerian electricity consumption, the SVR is obviously better suited when the electricity consumption behavior is highly stochastic. Thus, to apply the 'right' forecast model for the 'right' electricity consumption behavior is important. The data mining techniques forecasts were evaluated based on Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) in order to determine a better forecast accuracy measure.



Fig, 5. 1-step ahead forecast of Linear Regression (LR)

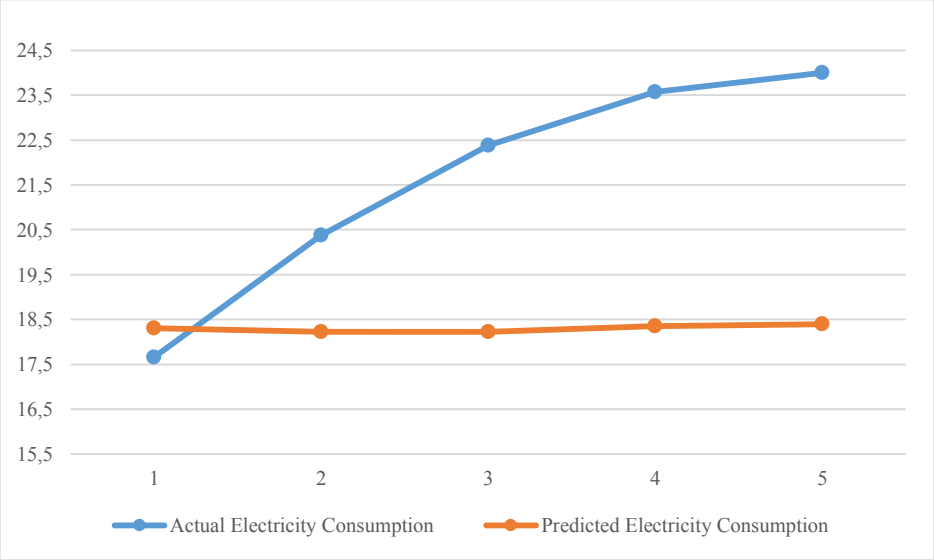


Fig 6. 1-step ahead forecast of Random Forest Regression (RFR)

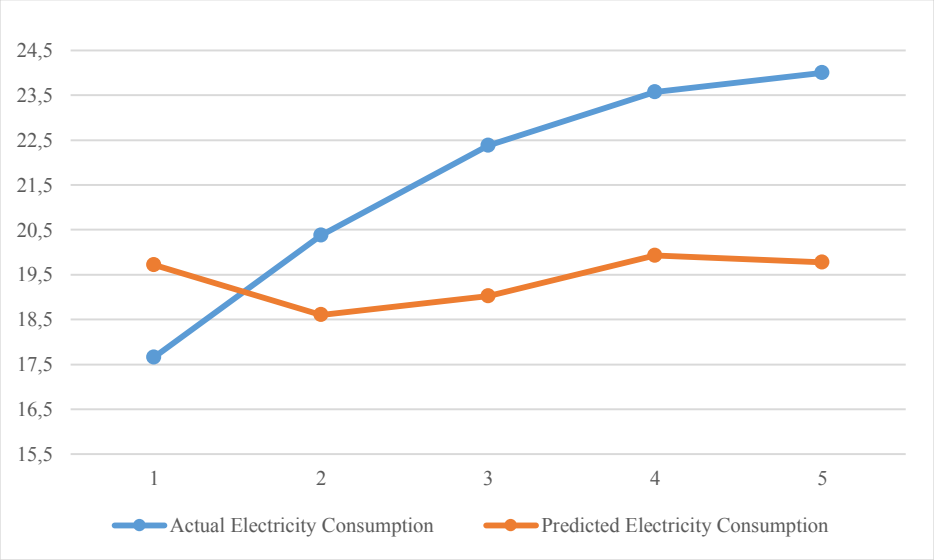


Fig. 7. 1-step ahead Forecast of Multilayer Perceptron Regression (MLR)

IV. CONCLUSION

In this research, four different data mining regression model (SVR, LR, RFR and MLR) were used to forecast a 1-step ahead prediction for electricity power consumption in Nigeria. This is a time series data that was transformed into a supervised learning data for better prediction. The time series data contains just two variables, that is, year and the values of the consumed electricity power in billions of kwh. It was transformed to a supervised learning data with 15 variables consisting of the year, 6-Lagged variables, 2-powers of time and 6-product of time and lagged variables. The dataset was divided into Train-Test data to the ratio 70% :30%. Both training and testing results showed that all proposed models were able to predict the electricity power consumption with good accuracy. By comparing all the four data mining regression models, it is clear that SVR had better prediction accuracy with lowest values, in terms of MAE, MAPE and RMSE with 0.0034, 0.0192 and 0.0034 values respectively for train data. For test data which is 30% of the data, the values for MAE, MAPE and RMSE were 1.1517, 5.9923 and 1.6492 respectively as presented in Fig. 3. Thus, SVR is suggested for predicting the short-term electricity consumption for Nigeria.

Future research work to be considered is building data mining and machine learning models to predict long-term and econometric data.

REFERENCES

[1] M. Hayati and Y. Shirvany, "Artificial Neural Network approach for short term load forecasting for illam region," *World Academy of Science, Engineering and Technology,* vol. 28, pp. 280-284, 2007.

[2] A. Anguera, J. M. Barreiro, J. A. Lara and D. Lizcano, "Applying data mining techniques to medical time series: An empirical case study in electroencephalography and stabilometry," *Computational and Structural Biotechnology Journal,* vol. 14, pp. 185 - 199, 2016.

[3] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji and A. Lendasse, "Methodology for long-term prediction of time series," *Neurocomputing 70 (2007) 2861–2869,* vol. 70, p. 2861–2869, May 2007.

[4] A. I. Taiwo, "Spectral and Fourier Parameter Estimation of Periodic Autocorrelated Time Series Data," Ago- Iwoye, Nigeria, 2017.

[5] Q. Yang and X. Wu, "10 Challenging Problems in Data mining," *International Journal of Information Technology and Decision Making,* vol. 5, no. 4, pp. 597 - 604, 2005.

[6] K. Chakrabarti, E. Keogh, M. Pazzani and S. Mehrotra, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Database," *ACM Transactions on Database Systems,* vol. 27, no. 2, pp. 188-228, June 2002.

[7] A. Fouilloy, C. Voyant, G. Notton, M. L. Nivet and L. J. Duchaud, "Machine Learning Methods for Solar Irradiation Forecasting: A Comparison in a Mediterranean Site," 2017.

[8] Y. Fu, Z. Li, H. Zhang and P. Xu, "Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices," in *Procedia Engineering*, 2015.

[9] S. S. Kathait, A. Kaur, S. Tiwari and A. Varshney, "Integrating Neural Networks with Time Series Forecasting:," *International Journal of Engineering Research And Management (IJERM),* vol. 04, no. 1, pp. 98-100, January 2017.

[10] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[11] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat Comput,* vol. 14, pp. 199-222, 2004.

[12] N. K. Ahmed, A. F. Atiya, N. E. Gayar and H. El-Shishiny, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting," *Econometric Reviews,* 2010.

[13] A. T. Sergio and T. B. Ludermir, "Deep Learning for Wind Speed Forecasting in Northeastern Region of Brazil," in *Brazilian Conference on Intelligent Systems (BRACIS)*, Anderson Tenorio Sergio, Teresa B. Ludermir, Anderson Tenorio Sergio, and Teresa B. Ludermir, " ," 2015 B, 2015.

[14] A. H. Jiang, X. C. Huang, Z. H. Zhang, J. Li, Z. Y. Zhang and H. X. Hua, "Mutual Information Algorithms," *Mech Syst Signal Process,* vol. 24, pp. 2947-2960, 2010.

[15] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 832 - 844, August 1998.

[16] D. A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, 2009.

[17] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya and K. R. K. Murthy, "Improvements to the SMO Algorithm for SVM Regression," *IEEE Transactions on Neural Networks,* 1999.

[18] R. R. Bouckaert, E. Frank, M. A. Hall, B. Holmes, P. Reutemann and I. A. Witten, "WEKA — Experiences with a Java Open-Source Project. Journal of Machine Learning Research," vol. 11, pp. 2533-2541, 2010.

[19] "https://www.indexmundi.com/g/g.aspx?c=ni&v=81," 2018. [Online]. Available: https://www.indexmundi.com/g/g.aspx?c=ni&v=81.. [Accessed 2018].

[20] R. Samsudin, A. Shabri and P. Saad, "A Comparisonof Time Series Forecasting using Support Vector Machine and Artificial Neural Network Model," *Journal of Applied Sciences,* vol. 10, no. 11, pp. 950-958, 2010.