PAPER • **OPEN ACCESS**

A Statistical Data Selection Approach for Short-Term Load Forecasting Using Optimized ANFIS

To cite this article: M Mustapha *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **884** 012075

A Statistical Data Selection Approach for Short-Term Load Forecasting Using Optimized ANFIS

M Mustapha[1,6], M W Mustafa[2], S Salisu[3], I Abubakar[4] and A Y Hotoro[5]
[1]Department of Electrical Engineering Kano University of Science and Technology, Wudil, Kano, NIGERIA
[2]School of Electrical Engineering, Universiti Technologi MALAYSIA
[3]Department of Electrical Engineering Ahmadu Bello University Zaria, Kaduna, NIGERIA
[4]Department of Electrical Enginerring Bayero University, Kano, NIGERIA
[5]Department of Physics Kano University of Science and Technology, Wudil, Kano, NIGERIA

E-mail: mamunu33@gmail.com

Abstract. Volume of the forecasting data and good data analysis are the key factors that influence the accuracy of forecasting algorithm because it depends on data identification and model parameters. This paper focuses on data selection approach for short-term load forecasting. It involves formulating data selection algorithm to identify factors (variables) that influence energy demand at utility level. Correlation Analysis (CA) and Hypothesis Test (HT) are used in the selection, where Wavelet Transform (WT) is applied to bridge the gap between the forecasting variables. This results to three groups of data; data without CA, HT and WT, data with CA, HT but without WT and data with CA, HT and WT. An optimized adaptive neuro-fuzzy inference system (ANFIS) using Cuckoo Search Algorithm (CS) is used to conduct the forecasting. The essence is to reduce the computational difficulty associated with the gradient descent (GD) algorithm in traditional ANFIS. With the three data groups, it is observed that CHW data can give satisfactory results more than the NCNHNW and NCNHW data. Also the numerical results shows that CHW data selection approach can give a MAPE of 0.63 against the bench-mark approach with MAPE of 3.55. This indicates that it is good practice to select the actual data and process it before the forecasting.

1. Introduction

Proper data selection and analysis is the key stage in obtaining good forecasting results. This gives the reason why Power system organizations are gathering the relevant data because of its significant influence in their business activities [1]. Such information received from data analysis give a clue on which method to be used or how to use it. It is also easy to determine when (the time at which) the consumption is low or high in the load profile or how the consumption is related with these variables [2]. Jain and Satish [3] reported that variation in the load consumption corresponds to time of the day, time of the week, time of the month or temperature of the

[6] To whom any correspondence should be addressed.

forecasting area and behavior of the customers towards electricity usage. It can be observed in Figure 2.1 that a day type contributes to the diversity in the weekly load pattern. This is because the load profile shows a significant difference between weekdays and weekends, and the days adjacent to them [4].

Metrological changes are the main factors that influence the load consumption in STLF. Depending on the forecasting location (area), these factors vary from temperature, relative humidity, wind speed, wind direction, cloud cover and light intensity. Changes in the weather condition determine the characteristics of the load curve because people tend to use one appliance or the other due to these changes. Also, decrease in temperature means more usage of room heaters, and increase in temperature means more usage of air conditioners [5] and this affects the load pattern within a short period of time.

It is difficult to determine the exact relationship between the load consumption and these forecasting variables [6] because different variables affect the load in different way. The degree of the effect may be high or low, or even negative [7]. On the other hand, the relationship between the variables (factors) and the load is highly non-linear. Such relationship, therefore, limits the performances of the forecasting algorithms [8]. Methods used in the recent time did not consider the effect of these variables on the load demand. Some factors have no effect on the load, in some cases their effect is insignificant, therefore using these factors may make the model more complex, and thus drive it out of performance. Unlike in the usual practice where seasons, weeks and days are considered as inputs in some research works [5, 9], in this paper, a daily and seasonal forecasting approach is formulated to take care of the seasonal and daily weather variations. When these data are gathered and analyzed thoroughly, a good and accurate load forecasting will surely save the utility economically. It has been reported that decrease in 1% of MAPE will reduce the generation cost by 0.1% to 0.3% [10]. Zhu [11] suggested that these factors should be reduced, thereby making the model simple and easy to use, and subsequently, give room for determining the actual parameters that influence the load consumption.

2. Correlation analysis

To select the forecasting variables, Correlation analysis (CA) is used to determine how one variable relate with another, and even though there is no causality between the variables, yet their relationship contains some information if the correlation coefficient between them is significant. Bashir and El-Hawary [12] reported that there is strong correlation between the electric energy consumption and the weather variables. Correlation coefficient (R) with the value from -1 to +1 determines the strength of a relationship between two pairs of data [13]. Any value of R close to ±1 shows strong correlation and value close to zero shows weak correlation. Zero value of R indicates no relationship. Equation (1) gives the relationship for computing R [14];

$$R_{xy} = \frac{\operatorname{cov}(x, y)}{\delta_x \delta_y} \tag{1}$$

where $\operatorname{cov}(x, y)$ is the population covariance, δ_x and δ_y are the population individual standard deviations. In this research correlation between the forecasting data is determined according to seasons to make the forecasting easier. Data is divided into seasons, and training data is separated from the testing data. Each group of data is treated separately, meaning that each season is forecasted independently. This is because of the changes in the weather variables of each season. In Nova Scotia winter starts from middle of December to middle of March, spring starts from middle of March to Middle of June, summer starts from middle of June to middle of September, and autumn starts from middle of September to middle of December. This is the reason why data is selected from middle of December in 2010, not January 2011. Equation (1) is used to compute the correlation coefficient R, between each variable and the load consumption in each season.

3. Hypothesis test

The first step is by setting a null hypothesis and alternative hypothesis on the observation, followed by pre-setting an α-value, and finally deduce conclusion based on the computed p-values. In this work 0.0005 is chosen as α-value. When the sample data is large z-test is used, otherwise t-test is applied. For z-test, p-value is computed using equations (2) and (3). The smaller this value is, the more significant the observation [13]. In other words the more this value rises above the pre-set value, determines the significance of the null hypothesis, and otherwise is false and finally rejected. All the computed p-values and correlation coefficient are recorded in Table 1 for all the four seasons.

$$z_i = \frac{alternative\ hypothesis - null\ hypothesis}{standard\ deviation} \tag{2}$$

For two side hypothesis

$$P - value = 1 - P(z_i < Z > z_j) \tag{3}$$

From the correlation and hypothesis test results, all variables with correlation coefficient equal to or greater than ±3 and p-value more than the α-value are rejected, thus not considered for the forecasting. The null hypothesis and alternative hypothesis are stated as follows;

H_0 = The correlation between the load consumption and the forecasting variable is by random chance

H_1 = The correlation between the load consumption and the forecasting variable is NOT by random chance.

H_0 is the null hypothesis and H_1 is the alternative hypothesis.

4. Wavelet transform

Wavelet Transform (WT) is a data processing technique that has wider application in signal processing and data science. It is through decomposition of the data series into a number of approximate and detailed components. The usual process is reconstruction of the signal (data) back for further analysis, after successful data analysis. The decomposition and reconstruction is into levels, based on selection of an appropriate mother wavelet. Equations (4), (5) and (6) present the series of a decomposed signal, $S(t)$, wavelet and inverse wavelet expressions [14].

$$S(t) = A_n(t) + D_n(t) + D_{n-1}(t) + D_{n-2}(t) + ... \tag{4}$$

$$WT_{(a,b)} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t)\psi(\frac{t-b}{a})dt \tag{5}$$

$$s(t) = \frac{1}{c_\psi^2} \int_a \int_b \psi_\psi(a,b) \frac{1}{a^2} \psi(\frac{t-b}{a}) dadb \tag{6}$$

where An(t) is the approximate component and $D_n(t)$, $D_{n-1}(t)$, $D_{n-2}(t)$, etc, are detail components, Ψ(t) is the mother wavelet, a is scale factor, b is the time-shift parameter and Ψ(●) is a scaling function. Wavelet Transform (WT) is applied in different forecasting models with aim of reducing the variability of the forecasting variables. At level two, a DB2 mother wavelet is selected because of its orthogonal features which makes it not lose any information from the decomposed coefficients while reconstruction into original signal [15]. Therefore, both approximate and details components are used in the forecasting and reconstruction also involved the same frequency components.

5. Forecasting data

These data processing and data analysis approaches gave a way for classifying the forecasting data into three. The first class of the data is the raw data selected based on the season and forecasting day, and is called 'NCNHNW'. The second group is a data without CA and HT but with WT referred as 'NCNHW'. The last group is the data with CA, HT and WT referred as 'CHW'. Under this selection there are seven classes of data for seven days in four seasons, thus resulted to 28 data groups for each group. Two different forecasting models are developed to make comparison. A classical ANFIS is firstly developed based on the work presented in [5], as a benchmark. This is then followed by CS-optimized-ANFIS.

6. Optimized ANFIS model

In ANFIS the advantages of fuzzy system and Neural Network are utilized to develop a layer-by-layer network with a number of nodes. For first order Sugeno-type (type-3) FIS with only two inputs, the following two rules hold [16];

1 *If x is A_1 and y is B_1, then* $f_1 = p_1 x + q_1 y + r_1$

2 *If x is A_2 and y is B_2, then* $f_2 = p_2 x + q_2 y + r_2$

Where p_i, q_i and r_i are the consequent parameters. In Figure 1, the circular nodes are fixed and the square nodes require update of their parameters and are called adaptive nodes. If $o_{j,i}$ is the output of node *i* in layer, the operation taken place in each layer is explained in [17], and for *i = 1,2*, the output *f* is given by:

$$f = \overline{w_1} f_1 + \overline{w_2} f_2 \tag{7}$$

$$f = \sum_i \overline{w_i} f_i \tag{8}$$

Where w_i is the data firing strength in node *i*. If the estimated output is f_k and the actual output is b_k the error difference for one iteration is;

$$E_k = \left(b_k - f_k\right)^2 \tag{9}$$

To improve the searching ability of CS it is combined with Levy flight as presented in [18], and the objective function to be minimized is presented in equation (10), subject to constraint stated in equation (11);

$$E = \sum_k^K E_k \tag{10}$$

$$e_{\min} \le e_k \le e_{\max} \tag{11}$$

The procedure followed in optimizing the ANFIS network is fully described in [19].

7. Results and discussion

Because of the nature of the data, average results will be considered in this work. From table 1 it can be observed that temperature and dew points recorded highest values of *R* for all the four seasons. This is because temperature plays a vital role in energy consumption. With a lower temperature use of room heaters is necessary, and with high temperature use of air conditioners is also necessary, this will, therefore increase the energy consumption. The dew point, on the other hand is affected by minimum daily temperature [20] experimentally.

Nova Scotia is characterized by cold. Such temperature of below 0°C will results to more energy consumption through the use of heating devices. In all the four seasons it can be seen that the benchmark (HC-ANFIS) curve fluctuates around the actual load, NCNHNW curve follow the same pattern, but with positive error since it is above the actual load curve. The other curve that represents NCNHW data presents negative error by being below the actual load curve. The presented CHW

algorithm follows the actual load curve with negligible variation. It is worth saying that introducing such algorithm that will improve the data selection before forecasting, irrespective of the region or season is necessary for forecasting future energy demand.

Table 1. Correlation coefficients and P-values for the forecasting variable

Season	**Winter**		**Spring**		**Summer**		**Autumn**	
Variables	**R**	**P-value**	**R**	**P-value**	**R**	**P-value**	**R**	**P-value**
Training data								
Temperature	-0.4241	0.0000	-0.4017	0.0000	0.4527	0.0000	-0.5443	0.0000
Dew Point	-0.3876	0.0000	-0.4742	0.0000	0.4511	0.0000	-0.5726	0.0000
Relative Humidity	-0.1397	0.0000	-0.5033	0.0000	-0.3677	0.0000	-0.3704	0.0000
Wind Direction	0.0207	0.0963	0.0263	0.0324	-0.0510	0.0000	0.0516	0.0000
Wind Speed	0.0557	0.0000	0.2962	0.0000	0.1218	0.0000	0.127	0.0000
Testing data								
Temperature	-0.554	0.0000	-0.6575	0.0000	0.4301	0.0000	-0.5893	0.0000
Dew Point	-0.538	0.0000	-0.6477	0.0000	0.3357	0.0000	-0.5771	0.0000
Relative Humidity	-0.2741	0.0000	-0.4056	0.0000	-0.4773	0.0000	-0.2625	0.0000
Wind Direction	0.0011	0.9607	0.1278	0.0000	-0.0763	0.0003	0.0206	0.3353
Wind Speed	0.0745	0.0005	0.2642	0.0000	-0.0052	0.806	0.1427	0.0000

To validate the proposed CHW data selection method comparison is made with NCNHW and NCNHNW data selection methods using the developed CS-ANFIS. From Table 2 it can be seen that CHW approach produces lowest MAPE of 0.63, NCNHW and NCNHNW produce 2.43 and 1.55 respectively. This shows that the proposed data selection approach can select the actual forecasting variables.

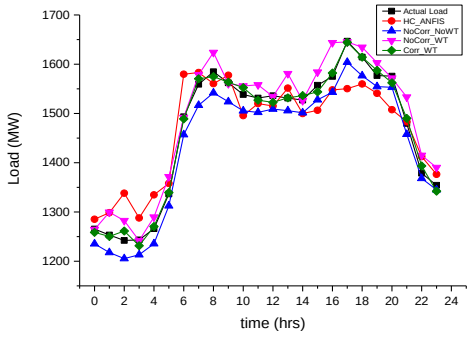

Figure 1. Actual and forecasted load based for Monday 20/01/2014 in winter season

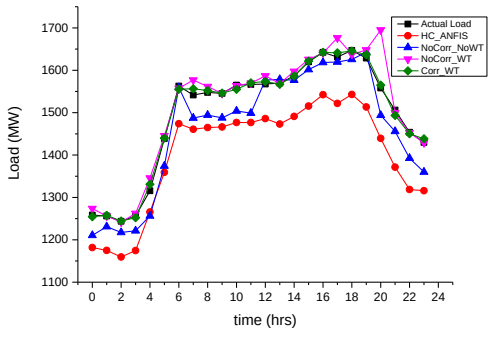

Figure 2. Actual and forecasted load based for Monday 31/03/2014 in spring season

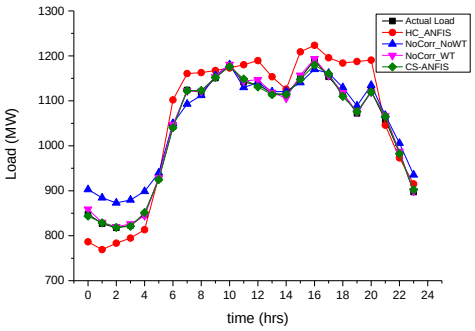

Figure 3. Actual and forecasted load based for Monday 07/07/2014 in summer season

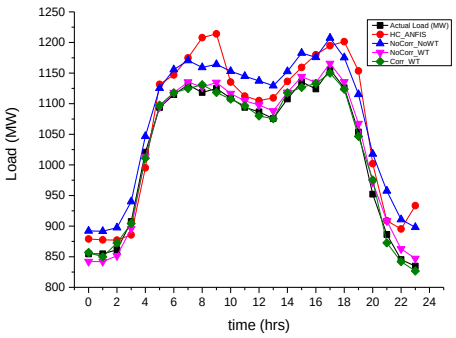

Figure 4. Actual and forecasted load based for Monday 29/09/2014 in autumn season

Table 2. Percentage reduction in the average forecasting error using the proposed algorithm

Data Selection method	**MSE**	**RMSE**	**MAPE**
CHW	73.37	8.50	0.63
NCNHNW	1275.96	35.37	2.43
NCNHW	778.04	27.11	1.55
HC-ANFIS	2383.93	48.74	3.55

8. Conclusion

This paper addresses two main forecasting issues. Firstly, data selection and analysis framework are formulated to handle the variability in the order of the forecasting variables. Some of these variables are in the order of thousandth, some hundredth, some tenth, and some are even on the order of units. The proposed algorithm is able to combine these variables within a minimum range that can accommodate each variable. Three different case studies based on three data selection approaches were tested. All three data groups were used in the forecasting differently, using an optimized ANFIS model. The optimization is by replacing the GD in classical ANFIS with CSA. It was found that selecting the data according to each season using correlation analysis, hypothesis test and then applying wavelet transform to decompose the data, improve the forecasting accuracy and enhance the forecasting speed.

Acknowledgements

the authors wish to acknowledged the contribution of TETFUND, Nigeria, Kano University of Science and Technology (KUST), Wudil, Nigeria and Universiti Teknologi Malaysia (UTM) for supporting this work.

References

[1] Chakravorty A, Rong C, Evensen P, Wlodarczyk, Wiktor T 2014 A Distributed Gaussian-Means Clustering Algorithm for Forecasting Domestic Energy Usage. *In: International Conference on Smart Computing. IEEE*, pp 229–236

[2] Tiong SK, Ahmed SK 2008 Electrical Power Load Forecasting using Hybrid Self-Organizing Maps and Support Vector Machines *The IEEE 2nd Internationsl Power Engineering optimization Conference (PEOCO) IEEE* pp 51–56

[3] Jain A, Satish B (2009) Clustering based Short Term Load Forecasting using Support Vector Machines *Proceeding of 2009 IEEE Buchrest Power Tech.* pp 1–8

[4] Osman ZH, Awad ML, Mahmoud TK 2009 Neural Network Based Approach for Short- Term Load Forecasting *IEEE Power Systems Conference and Exposition* pp 1–8

[5] Huseyin Cevic H, Cunkas M (2015) Short-term load forecasting using fuzzy logic and ANFIS. *Neural Computing and Application* 26 pp 1355–1367
[6] Mandal P, Senjyu T, Urasaki N, Funabashi T 2006 A neural network based several-hour-ahead electric load forecasting using similar days approach *International Journal of Electrical Power and Energy Systems* 28 pp 367–373
[7] Mirasgedis S, Sarafidis Y, Georgopoulou E, et al. 2006 Models for mid-term electricity demand forecasting incorporating weather influences *Energy* 31 pp 208–227
[8] Lin C, Chou L, Chen Y, Tseng L 2014 A hybrid economic indices based short-term load forecasting system *International Journal of Electrical Power and Energy Systems* 54 pp 293–305.
[9] Hong T, Wang P 2013 Fuzzy interaction regression for short term load forecasting *Fuzzy Optimization and Decision Making* 13 pp 91–103
[10] Fattaheian-Dehkordi S, Fereidunian A, Gholami-Dehkordi H, Lesani H 2014 Hour-ahead demand forecasting in smart grid using support vector regression (SVR) *International Transaction on Electrical Energy Systems* 24 pp 1650–1663
[11] Zhu J 2013 The Optimization Selection of Correlative Factors for Long-term power load Forecasting *IEEE Fifth International Conference on Intelligent Human-Machine Systems and Cybernetics* pp 241–244
[12] Bashir ZA, El-Hawary ME 2009 Applying Wavelets to Short-Term Load Forecasting Using PSO-Based Neural Networks *IEEE Transactions on Power Systems* 24 pp 20–27.
[13] Hernandez L, Baladron C, Aguiar JM, et al. 2012 A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework. *SENSORS* 12 pp 11571–11591
[14] Li S, Wang P, Goel L 2016 A Novel Wavelet-Based Ensemble Method for Short-Term Load Forecasting with Hybrid Neural Networks and Feature Selection *IEEE Transaction on Power Systems* 31 pp 1788–1798
[15] Daubechies I 1988 Orthonormal bases of compactly supported wavelets *Communications on Pure and Applied Mathematics* 41 pp 909–996
[16] Salisu S, Mustafa MW and Mustapha M 2017 Predicting Global Solar Radiation in Nigeria using Adaptive Neuro-Fuzzy Approach *Proceedings of the 2nd International Conference of Reliable Information and Communication Technology* pp 513–521
[17] Mustapha M, Mustafa MW, Khalid SN, et al. (2016) Correlation and Wavelet-based Short-Term Load Forecasting using ANFIS. Indian Journal of Science and Technology 9:1–8. doi: 10.17485/ijst/2016/v9i46/107141
[18] Yang X, Deb S 2009 Cuckoo Search via Levy Flights *2009 World Congress on Nature & Biologically Inspired Computing* pp 210–214
[19] Mustafa MW, Abdilahi AM, Mustapha M 2016 Chaos-Enhanced Cuckoo Search for Economic Dispatch with Valve Point Effects *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 14 pp 1220–1227
[20] Mohammadi K, Shamshirband S, Petković D, et al. (2016) Using ANFIS for selection of more relevant parameters to predict dew point temperature *Applied Thermal Engineering* 96 pp311–319