Load Forecasting using Combination Model of Multiple Linear Regression with Neural Network for Malaysian City

(Peramalan Beban Menggunakan Model Gabungan bagi Regresi Linear Berganda dengan Rangkaian Neuron untuk Bandaraya di Malaysia)

NUR ARINA BAZILAH KAMISAN, MUHAMMAD HİSYAM LEE*, SUHARTONO SUHARTONO, ABDUL GHAPOR HUSSIN & YONG ZULINA ZUBAIRI

ABSTRACT

Forecasting a multiple seasonal data is differ from a usual seasonal data since it contains more than one cycle in a data. Multiple linear regression (MLR) models have been used widely in load forecasting because of its usefulness in the forecast a linear relationship with other factors but MLR has a disadvantage of having difficulties in modelling a nonlinear relationship between the variables and influencing factors. Neural network (NN) model, on the other hand, is a good model for modelling a nonlinear data. Therefore, in this study, a combination of MLR and NN models has proposed this combination to overcome the problem. This hybrid model is then compared with MLR and NN models to see the performance of the hybrid model. RMSE is used as a performance indicator and a proposed graphical error plot is introduce to see the error graphically. From the result obtained this model gives a better forecast compare to the other two models.

Keywords: Error plot; hybrid model; neural network; regression model; residuals

INTRODUCTION

Modelling a load forecasting is a very important task for electricity to function in a more efficient and safety way, give the most optimum cost for cost saving applications relying on operating reconstruction (Hahn et al. 2009; Soares & Medeiros 2008). By predicting the future load demands, power system planners and demand controllers could ensure that they would be enough supply of electricity to cope with increasing demands (Mastorocostas et al. 2000). For these reasons, load forecasting has attracted not only researchers but also organisation with same interest to forecast energy usage by using various methods from classical methods to the advanced methods.

Forecasting seasonal load demand has become increasingly challenging in recent period. Classical methods such as regression, holt winter’s and exponential smoothing are suitable for a large number of series, for analyst with limited skill and also for a norm of comparison. It follows certain pattern which was determined by the parameter of the model. But it is common for load data to contain a seasonal pattern. Because of these reasons, new models such as artificial neural network and fuzzy time series have been developed in order to find a better and accurate forecast for load forecasting (Chatfield 2005).

Modeling multiple seasonal loads is no longer optimal with standard seasonal methods. Multiple seasonal cycles is differ from the usual seasonal time series. In multiple seasonal cycles, it can contain two or more cycles. Malaysian data contain both daily and weekly cycles. Malaysia has many public holidays because of the ethnicity variation. The cycles from Monday to Friday are similar and Saturday and Sunday are quite discrete. And the patterns for public holidays are quite similar to weekends compare to weekdays. And as stated by Gould et al. (2008), the levels of the daily cycles may change from one week
to the next and yet it still highly correlated with the prior levels of the next day.

A simple or singular method can lead to a bias or inaccurate forecast (Basheer et al. 2014). Thus, a few research teams such as Basheer Shukur et al. (2014) and Mohamed and Ahmad (2010) have developed a hybrid model to deal with this double cycle problem. Some have also used a double seasonal ARIMA to overcome double seasonal difficulties (Mohamed et al. 2011, 2010). Dudek (2016) proposed a linear regression model with patterns of daily cycle. The daily cycle is used as both input and output data. The linear regression model employ more than one independent variable. In regression model, load data will be represented as a linear combination of variables related to other factors and in this study it is the day type (Kyrriakides & Polycarpou 2007). But regression model often suffer from a number of difficulty due to a nonlinear relationship between load demand and the influencing factors (Kyrriakides & Polycarpou 2007).

Since neural network is a flexible model in modeling a nonlinear data, it will be used to analyse the residuals that will be obtained from the MLR model. Neural network has the competence and accessibility in estimated nonlinear data. It captures some of the key properties which replicate the models of biological neurons works (Bates & Granger 1969; Jang 1993; Kyrriakides & Polycarpou 2007; Ringwood et al. 2001; Taylor, 2003; Xiaojuan et al. 2010; Ying & Pan 2008; Zhang 2003; Zhang et al. 2008). Due to this issue, a combination with neural network (NN) is considered in order to enhance the forecast.

METHODS

The idea of the hybrid model came from Zhang (2003). In this study, a hybrid model combination of two models together. The first model will be used to forecast the output data and the second model will be used to forecast the residuals output from the first model. MLR model is selected because of the suitability in modelling a consistent parameter over time. In this case, the parameters will be the days since the data have been arranged into 24 sets of independent series of data. Neural network will be combined with MLR model because of its benefit in modelling a nonlinear data. As being stated by Zhang (2003), the relationship between an actual value and a forecasting can be written as:

\[ y_t = \hat{y}_t = \epsilon_t, \]  

where \( y_t \) is the actual value, \( \hat{y}_t \) is the forecasted value and \( \epsilon_t \) is the error or residual. From this equation, the \( \hat{y}_t \) will be calculated by using MLR model and the error term or residual, \( \epsilon_t \) will be forecast by using NN model.

MLR model employ more than one independent variable. The linear regression model relating \( y \) to \( x_1, x_2, K, x_7 \) is:

\[ y = \beta_{0} + \beta_{1}x_{1} + L + \beta_{k}x_{k} + \epsilon, \]  

where \( \beta_{0}, \beta_{1}, K, \beta_{k} \) is the mean value of the dependent variable \( y \) when the values of the independent variables are \( x_1, x_2, K, x_7 \). \( \epsilon \) is (unknown) regression parameters relating the mean value of \( y \) to \( x_1, x_2, K, x_7 \). \( \epsilon \) is the error term that describes the effects on \( y \) of all factors other than the values of the independent variables \( x_1, x_2, K, x_7 \).

In this study the linear regression model can be written as:

\[ y = \beta_{0} + \beta_{1}x_{1} + L + \beta_{k}x_{7} + \epsilon, \]  

where \( x_1, x_2, K, x_7 \) are dummy variables where \( x_1 \) is Monday, \( x_2 \) is Tuesday and so on since the data used in this study is a daily recorded data. After data is analysed with
MLR, the residuals obtained from the in-sample forecast of MLR will be analysed by using neural network in Matlab software. A two-layer feed-forward network with sigmoid hidden neurons and linear output neurons will be used in the neural network process. The network will be trained with Levenberg-Marquardt backpropagation algorithm and if there is not enough memory, scaled conjugate gradient backpropagation will be used to train the network. The step by step process of the process can be written as follows:

**First step:** Data is modelled by using the MLR model.

**Second step:** Residual obtained from the in-sample forecast is calculated and the ACF and PACF of the residuals are plotted to find the lags. If the lags are not stationary, then the residuals will be difference to 1 to make the ACF and PACF lags stationary. The lags obtained will be used as the input nodes that will be used in the next process.

**Third step:** Neural network model is then used to model the out-sample forecast of the residual.

**Forth step:** The final out-sample forecast is obtained by adding the out-sample forecast from the MLR model with the residuals obtained from the neural network modelling process.

The results of the point estimated for the parameters are listed in the Table 1 below. The parameters \(x_1\), \(x_2\), \(x_3\), \(x_4\), \(x_5\), \(x_6\) and \(x_7\) represent Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday, respectively. The parameters estimation of the MLR model can be written as:

For example, we could write the equation of the MLR model of 6 am as:

\[
y = 25098 + 8.0x_1 + 8.5x_2 + 8.5x_3 + 8.2x_4 + 4.3x_5 + 2.4x_6 + 6.8x_7
\]

(4)

The input lags in neural networks are determined based on autoregressive (AR) term in Box-Jenkins model. Park et al. (1996), Tang and Fishwick (1993) and Zhang and Qi (2005) suggested that lags can be determined by the AR terms and ignoring the MA terms. Based on the ACF and PACF plots of the residuals, the best Box-Jenkins model is either ARIMA(0,1,1) or ARIMA(2,1,1) or ARIMA(2,1,2). The model for ARIMA(0,1,1) can be written as:

\[
(1 - B)y_t = (1 - \theta B)a_t
\]

\[
y_t - y_{t-1} = a_t - \theta a_{t-1}
\]

(5)

and for ARIMA(2,1,1) can be written as:

\[
(1 - B)y_t = (1 - \theta B - \theta^2 B^2)a_t
\]

\[
y_t - y_{t-1} - y_{t-2} = a_t - \theta a_{t-1} - \theta^2 a_{t-2}
\]

(5)

and for ARIMA(2,1,2) can be written as:

\[
(1 - B - B^2)y_t = (1 - \theta B - \theta^2 B^2)a_t
\]

\[
y_t - y_{t-1} - y_{t-2} = a_t - \theta a_{t-1} - \theta^2 a_{t-2}
\]

(5)

Since only the AR terms need to be considered to choose the input lags, by ignoring the term in (4), the input lags to be considered are lag 1 and for (5) and (6), the input lags are 1 and 2. The input lags will be used to forecast the out-sample of the residuals of the NN model.

The number of neuron could influence the performance of MLP forecast performance. However, using minimum number of neuron is most recommended (Masters 1993). Each neuron is processing unit that used logistic function to calculate the linear combination of inputs. The number of the nodes in the hidden layer could be decrease or increase based on the performance of the network training. In this study, the number of neurons that used is one up to five neurons. But in this study only the one that show the best performance will be discussed and only out-sample forecast is consider since the proposed method only appropriate for the out-sample forecast. For the proposed model, the step by step diagram on how the model working is shown in Figure 1 below.

This method is done by using Minitab and Matlab program. The MLR modeling part is done by using the Minitab software while the NN modeling part is done by using Matlab software.

**Table 1. Parameters estimation for multiple linear regression model**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Constant</th>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(x_4)</th>
<th>(x_5)</th>
<th>(x_6)</th>
<th>(x_7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 am</td>
<td>25098</td>
<td>8.0</td>
<td>8.5</td>
<td>8.5</td>
<td>8.2</td>
<td>4.3</td>
<td>2.4</td>
<td>6.8</td>
</tr>
<tr>
<td>12 pm</td>
<td>51470</td>
<td>17.7</td>
<td>18.4</td>
<td>18.1</td>
<td>15.5</td>
<td>-1.9</td>
<td>-10.3</td>
<td>16.7</td>
</tr>
<tr>
<td>6 pm</td>
<td>40033</td>
<td>15.4</td>
<td>15.4</td>
<td>15.2</td>
<td>14.3</td>
<td>10.2</td>
<td>6.8</td>
<td>13.9</td>
</tr>
<tr>
<td>12 am</td>
<td>25402</td>
<td>7.8</td>
<td>7.9</td>
<td>7.8</td>
<td>7.7</td>
<td>7.7</td>
<td>5.2</td>
<td>6.9</td>
</tr>
</tbody>
</table>

**Table 2. Input lags for number of nodes in NN**

<table>
<thead>
<tr>
<th>Hour</th>
<th>ARIMA model</th>
<th>Lags</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 am</td>
<td>(2,1,2)</td>
<td>1.2</td>
</tr>
<tr>
<td>12 pm</td>
<td>(0,1,1)</td>
<td>1</td>
</tr>
<tr>
<td>6 pm</td>
<td>(2,1,1)</td>
<td>1</td>
</tr>
<tr>
<td>12 am</td>
<td>(0,1,1)</td>
<td>1</td>
</tr>
</tbody>
</table>
GOODNESS-OF-FIT TESTS

Root mean square error (RMSE) is used as the performance indicator in order to see the performance of the model other than the comparison plot to observe the goodness-of-fit of the hybrid model. The smaller the number of the RMSE indicates the better the model is. RMSE is preferred compared to MSE because of their theoretical relevance in statistical modelling. Often, the RMSE is preferred to the MSE as it is on the same scale as the data (Hyndman & Koehler 2006). The formula of the RMSE can be seen as:

\[\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \quad (8)\]

where \(\hat{y}_i\) is the predicted value and \(y_i\) is the observed value.

Other than RMSE, we also suggested an error plot in order to compare the results and to determine whether a hybrid model should be considered or not. Usually, plot of error or residual plot is used to see the error graphically. But the problem with residual plot is that it depends on the value from the data. If the data has large number, then the number error could also give a big value which will affect the point in the residual plot. The proposed plot will transform the value of data into percentage by using the formula in order to standardize the value.

\[x_t = \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \quad (9)\]

where \(y_t\) is the actual value and \(\hat{y}_t\) is the forecast value. The value of \(x_t\) will be in the range of 0 to 1, with 0 being the best and 1 being the worst.

\[\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \quad (10)\]

where \(y_i\) is the actual value and \(\hat{y}_i\) is the predicted value. MAPE gives the accuracy of the prediction in percentage. The smaller the MAPE, the better the model.

THE DATA SET

We used data from Tenaga Nasional Berhad Johor Bahru. Data were recorded every day for 3 years from a station.
named Pusat Bandar Johor Bahru. Pusat Bandar Johor Bahru (PBJB) is a commercial area in Johor Bahru. Four selected hours used in this paper are 6 am, 12 pm, 6 pm and 12 am. These four selected hours represent the peak hours and the dip hour in Johor Bahru and generally it can represent the daily scenario in Malaysia too.

For casual hours like 6 am and 12 am, there are not much differences of load usage between weekdays and weekends but for peak hours such as 12 pm and 6 pm the load usage during weekdays is higher compare to the weekend’s usage. These four hours are selected not only because of the high and low usage, but also to see the pattern and differences of the load usage between these hours.

RESULT AND DISCUSSION

A comparison between residual plot and percentage error plot is discussed here. As can be seen from Figure 2, residual plot do not have a fix scale on the y-axis. The scale on the y-axis depends on the value of the residual. Therefore, there is not much information gained from this plot. Compare to percentage error plot, the scale on the y-axis is fixed from 0 to 1. It is easier to determine whether the forecast is good or not. And one can set its own benchmark when making a decision to determine a good or bad forecast.

Other than using the percentage error plot as a comparison between models, it also could be used as a benchmark plot for certain purpose. In this study, we used this plot to determine whether a data need to be forecasted with hybrid or not. We set the benchmark at 0.4. If there is any point exceeds 0.4, a hybrid model will be considered. As can be seen from the figures below, 12 pm and 6 pm contain points which is greater than 0.4. To ensure that this rule could be applied, we test all of the data used in this study with the proposed hybrid model and compare it with the result in RMSE.

FIGURE 2. A comparisons between residual plot (a) and fractional residual plot (b) for selected times
Here is the result of the hybrid model. 6 am, 12 pm, 6 pm and 12 am from Pusat Bandar Johor Bahru was selected for this purpose. Comparing the plots of percentage error above and the error measurement in Table 3 below, it is agreed that for 12 pm and 6 pm where there are points that exceed 0.4, hybrid model gives smaller value of RMSE and for 12 am where all the points are below 0.4 in the plot, NN shows smaller value of RMSE compare to hybrid model which suggest that hybrid model is not necessary.

As for 6 am, although all the points in the plot are below 0.4, hybrid model shows a smaller value of RMSE. This is understood since 6 am represent the transition time when most people in Malaysia start their day. At this hour, they usually busy preparing their self to working and other activities. Therefore, transition hour should be considered to be tested with hybrid model. The result of the out-sample forecast can be seen in the figures and tables below. Other than that, absolute error plot also shows that all forecast by using NN gives small error which could also suggest that it is a good model if hybrid model is not used for this data.

The result of the out-sample for selected hours can be seen in Table 3. Both tests show consistent results for the models of selected hours. For 6 am, 12 pm and 6 pm, the hybrid model gives a better result compare to other models. The RMSE for 6 am was reduce to almost 41.3% while when data is forecasted with the hybrid model compare to MLR and 46.1% was reduce for RMSE compare to NN model. The RMSE for 12 pm reduce to almost 35.5% compare to MLR and reduce 17.1% compared to NN model. As for 6 pm, the hybrid model also fit the data well compare to MLR and NN. A reduction in RMSE result can be seen from Table 3. Other than using the absolute error plot as a comparison between models, it also could be used as a benchmark plot for certain purpose. In this study, we used this plot to determine whether a data need to be forecasted with hybrid or not. We set the benchmark at 0.4. If there is any point exceeds 0.4, a hybrid model will be considered. As can be seen from the figures below, 12 pm and 6 pm contain points which is greater than 0.4. To make sure that this rule could be applied, we test all of the data used in this study with the proposed hybrid model and compare it with the result in RMSE.

Here is the result of the hybrid model. 6 am, 12 pm 6 pm and 12 am from Pusat Bandar Johor Bahru was selected for this purpose. Comparing the plots of percentage error above and the error measurement in Table 2 below, it is agreed that for 12 pm and 6 pm where there are points that exceed 0.4, hybrid model gives smaller value of RMSE and for 12 am where all the points are below 0.4 in the plot, NN shows smaller value of RMSE compare to hybrid model which suggest that hybrid model is not necessary.

As for 6 am, although all the points in the plot are below 0.4, hybrid model shows a smaller value of RMSE. This is understood since 6 am represent the transition time when most people in Malaysia start their day. At this hour, they usually busy preparing their self to working and other activities. Therefore, transition hour should be considered to be tested with hybrid model. The result of the out-sample forecast can be seen in the figures and tables below. Other than that, absolute error plot also shows that all forecast by using NN gives small error which could also suggest that it is a good model if hybrid model is not used for this data.

The MAPE shows a consistent result as the RMSE. For hour 6 am, 12 pm and 6 pm, the hybrid model shows the smallest value of RMSE and MAPE which indicate that the model is the best compare to MLR and NN models. Hour 12 am shows that NN with neuron 5 is slightly better compare to hybrid model but from MAPE we could see that the difference is only 0.001. Thus, from the error measurements results we could see that the hybrid model did improve the forecast of the load data. The more obvious pattern can be seen in figures below.

As can be seen from Figure 3, the outline of hybrid plot is closer to the actual plot outline compare to the other two plots. Although at certain point, the hybrid model could not detect the actual point, but as can be seen at point 31 in x-axis where there is a sudden increment in the load usage reading, the hybrid did follow the actual value pattern compare to the MLR or NN models. While for 12 pm, the pattern of the load is more obvious where a sudden drop-off could be seen from the plot in Figure 4. This is understand as explained above, 12 pm is the peak hour and the sudden drop-off is actually the weekend when people are mostly not working and since this data cover the urban region, the pattern is obvious at 12 pm compare to the other three hours. From the figures above, it is obvious that the hybrid model gives better performance in following the pattern of the actual load compare to MLR and NN.

Overall, as can be seen from Figures 3-5, NN and hybrid give a quite identical plot and could read the pattern of the actual data. Although at certain time such as 12 am the hybrid model does not gives a good result, it can still be concluded that this hybrid model could be used to forecast the load data especially load data from Malaysia. NN is a good model when forecasting a nonlinear data.

<table>
<thead>
<tr>
<th>Hour</th>
<th>Model</th>
<th>MAPE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 am</td>
<td>MLR</td>
<td>0.073</td>
<td>1679</td>
</tr>
<tr>
<td></td>
<td>NN(4)</td>
<td>0.074</td>
<td>1829</td>
</tr>
<tr>
<td></td>
<td>Hybrid(1)</td>
<td>0.029</td>
<td>986</td>
</tr>
<tr>
<td>12 pm</td>
<td>MLR</td>
<td>0.151</td>
<td>6874</td>
</tr>
<tr>
<td></td>
<td>NN(3)</td>
<td>0.115</td>
<td>5346</td>
</tr>
<tr>
<td></td>
<td>Hybrid(5)</td>
<td>0.086</td>
<td>4433</td>
</tr>
<tr>
<td>6 pm</td>
<td>MLR</td>
<td>0.087</td>
<td>3344</td>
</tr>
<tr>
<td></td>
<td>NN(5)</td>
<td>0.073</td>
<td>2840</td>
</tr>
<tr>
<td></td>
<td>Hybrid(4)</td>
<td>0.054</td>
<td>2289</td>
</tr>
<tr>
<td>12 am</td>
<td>MLR</td>
<td>0.080</td>
<td>2332</td>
</tr>
<tr>
<td></td>
<td>NN(5)</td>
<td>0.049</td>
<td>1599</td>
</tr>
<tr>
<td></td>
<td>Hybrid(2)</td>
<td>0.050</td>
<td>1859</td>
</tr>
</tbody>
</table>
be seen from Figure 6, the load at 12 am did not show an obvious seasonal pattern compared to the others. Because of that, NN shows a superior result compared to hybrid model for 12 am. However NN is also a good model in forecasting the load compared to the traditional model, MLR. Therefore, if one does not wish to forecast by using the hybrid model, NN model could be considered as an alternative model.

**CONCLUSION**

The hybrid model has shown to be superior compared to MLR and NN models. The combination of MLR and NN is a good combination especially for a data that has a similar pattern like Malaysian data. Malaysian data has a multiple seasonal pattern since the data contain few cycles in the load and since regression can model a data with variable from other factor and NN is a good model in modeling a nonlinear data, the combination of these two models could enhance the forecasting of the load. The percentage error plot is also a good plot to check the error graphically. It could give brief idea on whether a forecast is good or bad and it also can be used to make a comparison between models.

**REFERENCES**

Basheer Shukur, O., Salem Fadhil, N., Hisyam Lee, M. & Hura Ahmad, M. 2014. Electricity load forecasting using hybrid of multiplicative double seasonal exponential smoothing model


