MULTIPLE LINEAR REGRESSION APPROACH FOR SHORT-TERM FORECASTING OF ELECTRIC ENERGY CONSUMPTION IN TOGO

Comlanvi Adjamagbo, Akim Adekunle Salami, Yao Bokovi, Djamil Gado and Ayite Senah Akoda Ajavon
Research Laboratory in Engineering Sciences (Larsi), Regional Center of Excellence for the Control of Electricity (CERME), Department of Electrical Engineering of the National Superior School of Engineers (ENSI), University of Lome (UL), BP 1515, Lome, Togo.

Abstract
A Linear Multiple Regression approach is used to model the energy consumption of electricity in Togo. This model is developed from the load data recorded at the electric power source stations in Togo during the period from 2016 to 2017. This model predicts four input parameters (Day of the week, the type of day (working day) or not, Hours in the day and Load data of the same time of the previous day) is used to predict the electrical energy consumption data for the period of 2018 with a MAPE of 4.4964% and a correlation coefficient R² equal to 95.5889%.

Introduction:
The fundamental functions of modern energy management systems are based on an accurate short-term prediction model of the electrical load [1]. The precision of the prediction model leads to savings and increased security measures in the operation of systems for generating and transmission of electrical energy [2]. Large prediction errors can lead either to too careful or too risky planning, which can also lead to heavy economic losses [3].

Statistical approaches require an explicit mathematical model which gives the relationship between the load and several input factors [4]. Several classical models are applied for load predictions, such as regression-based methods for example [5] [6] and time series methods [7].

To predict electrical load, regression methods are usually used to model the relationship between load consumption and other factors such as weather conditions [8], type of day, and customer category. Engle et al. [9] presented several regression models for predicting the next day's load.

This paper describes the experiences we have gained during the development of a short-term prediction model of the electric charge of the next 48 half-hours per day for all year 2018 in Togo with a Linear Regression method. Multiple.

Our goal is to predict the load data with various combinations of explanatory variables to determine which configuration case gives the best results.

The problems encountered and the solutions proposed are discussed. The developed model should provide daily load profile forecasts for the next seven days. The forecast results for all year 2018 are also presented.
Data presentation:
Electricity load or consumption data is taken at the various source substations in Togo during the period from 2016 to 2018. The readings are made in 30-minute steps over the day, which makes 48 data per day or 52,608 data. Figure 1 gives an overview of the load statements in Excel.

![Figure 1: Presentation of load statements in Excel.](image1)

The evolution of electricity consumption in Togo from 2016 to 2018, shown in Figure 2, shows an increasing trend in electricity consumption. Extrapolating a linear trend estimates the annual rate of increase in electrical energy consumption to be around 6.5183%.

![Figure 2: Evolution of Togo's electrical load from 2016 to 2018 (MW).](image2)

Figure 3 shows the electrical energy consumption over a period of one (01) week (from Monday January 25 to Sunday January 31, 2016). We see a daily profile appear (7 patterns per week) reflecting a daily cyclical variation.

![Figure 3: Evolution of Togo's electrical load over a period of one week (MW).](image3)
**Multiple linear regression method:**

Multiple linear regression analysis relies on descriptive analysis of data to observe the relationships between a quantitative dependent variable and n quantitative independent variables. Any method using regressions is based on the acceptance of the founding assumptions of parametric statistics and the notion of least squares fit. The concept of least squares consists in minimizing the sum of the residuals raised to the power of two between the observed value and the extrapolated one [10].

The descriptive equation for multiple linear regression is as follows (Equation (1)) [11], [12], [13]:

\[ y = X\beta + \epsilon \]  

where:
- \( y \) is the vector of responses;
- \( X \) is the matrix of explanatory variables;
- \( \beta \) is the vector of the model parameters;
- \( \epsilon \) is the vector of errors.

It is therefore a question of calculating the vector of the estimators \( \hat{\beta} \) which is the solution, in the "least squares" sense. This vector of estimators \( \hat{\beta} \) is defined by the equation (2):

\[ \hat{\beta} = (X^*X)^{-1} * X^*y \]  

This model can be used to make predictions. It is therefore a question of applying the relation defined by the equation (3):

\[ \hat{y} = X * \hat{\beta} \]  

where \( \hat{y} \) is the vector of predicted responses.

**Methodology:**

The choice and methodical analysis of the explanatory variables make it possible to assess the influence of each input parameter on the output of the forecast model. Indeed, it is very important, for the accuracy of the model, to choose adequate input parameters. This step is very useful because it allows you to eliminate some variables that provide very little or no information to describe the output, or to eliminate redundant variables. We took into account the following explanatory variables (Table 1):

**Table 1:** List of explanatory variables used.

<table>
<thead>
<tr>
<th>Data types</th>
<th>Mathematical explanations</th>
<th>Data presentations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day of the week</td>
<td>Monday = 2 ; Tuesday = 3 ; Wednesday = 4 ; Thursday = 5 ; Friday = 6 ; Saturday = 7 ; Sunday = 1</td>
<td>[1 … 7]</td>
</tr>
<tr>
<td>Working day or not</td>
<td>( { \text{if working day then} \ 1 \text{ if not then} \ 0 } )</td>
<td>[0 ou 1]</td>
</tr>
<tr>
<td>Half hour in the day</td>
<td>( \frac{1}{2}h )</td>
<td>[1 … 48]</td>
</tr>
<tr>
<td>Load data for the same time of the previous day</td>
<td>( L_{\frac{1}{2}h-48} )</td>
<td>-</td>
</tr>
<tr>
<td>Load data for the same time of the previous week</td>
<td>( L_{\frac{1}{2}h-336} )</td>
<td>-</td>
</tr>
<tr>
<td>Load data for the same time of the previous year)</td>
<td>( L_{\frac{1}{2}h-17520} )</td>
<td>-</td>
</tr>
</tbody>
</table>
Average charges of the last 24 hours

\[ \text{Mean} \left( \sum_{i=1}^{48} L_{24-i} \right) \]

\( L = \text{data of loads} \)

The data preprocessing is obtained by the MATLAB software.

Either the following nomenclature adopted for the parameters:

A = Day of the week;

B = Working day or not;

C = Hours in the day;

D = Load data for the same hour of the previous day;

E = Load data for the same hour of the previous week;

F = Charging data for the same hour of the previous year;

G = Average charges for the last 24 hours.

Our goal is to predict the load data with various combinations of these explanatory variables in order to determine which configuration case gives the best results. We tested different configuration cases which are summarized in Table 2, for a total of 7 configuration cases.

**Table 2:** Summary of simulation cases in MATLAB.

<table>
<thead>
<tr>
<th>Case</th>
<th>Coefficients ( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>([-0.6273 ; 4.0738 ; 0.2045 ; 0.0882 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
<tr>
<td>Case 2</td>
<td>([-5.8734 ; 0.2045 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
<tr>
<td>Case 3</td>
<td>([-5.8734 ; 0.2045 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
<tr>
<td>Case 4</td>
<td>([-5.8734 ; 0.2045 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
<tr>
<td>Case 5</td>
<td>([-5.8734 ; 0.2045 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
<tr>
<td>Case 6</td>
<td>([-5.8734 ; 0.2045 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
<tr>
<td>Case 7</td>
<td>([-5.8734 ; 0.2045 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
</tbody>
</table>

We have divided our data into two groups. The data for the years 2016 and 2017 are used for learning, that is to say for the determination of the coefficients of the estimators \( \hat{\beta} \) of the model and the data for the year 2018 are used for validation (for the test of the prediction). For each of the configurations adopted previously, we applied the modeling method described in section 3. Thus, we calculated the coefficients of the vector of the estimators \( \hat{\beta} \) from equation (2). Once these coefficients were known, we then performed the prediction of the new load data by equation (3).

To evaluate the performance of each prediction model, we used as measures: the average value of the absolute errors in percentage (%) (MAPE: Mean Absolute Percentage Error, [2]) committed, expressed by equation (4), the histogram of the absolute errors, as well as the correlation coefficient (R²) between the predicted data and the real load data.

\[
\text{MAPE} = \frac{100}{T} \times \sum_{t=1}^{T} \left| \frac{y_t - \hat{y}_t}{y_t} \right|
\]

where

\( y_t \) the real value;  
\( \hat{y}_t \) the predicted value; 
\( T \) the total number of samples.

**Results and Discussions:**

In this part, we will first discuss the choice of the prediction model and then present the results of the prediction for the year 2018 of the electrical load of the energy system of Togo.

**Choice of prediction model:**

Tables 3 and 4 respectively represent the summaries of the coefficients of the estimators \( \hat{\beta} \), the R² and MAPE for the seven case configurations that we have chosen for the input of each model.

**Table 3:** Summary of estimators \( \hat{\beta} \) calculated.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Coefficients ( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>([-0.6273 ; 4.0738 ; 0.2045 ; 0.0882 ; 0.4469 ; 0.1347 ; 0.2832])</td>
</tr>
</tbody>
</table>
We have 7 scenarios. Case 1 composed of all the explanatory variables gives a correlation coefficient of 87.298%. Case 2 composed of only 5 explanatory variables gives the coefficient of 88.5588%, which is the highest value. Case 6 composed of 3 explanatory variables gives the lowest correlation coefficient (61.5796%). However, all the other cases give results of more than 78%. Note also that this is case 2 which quite logically presents the smallest MAPE which is 5.8618% and also case 6 the largest MAPE of 20.0856%. Thus from these results we exclude case 6 of the seven configurations that we have proposed. However, given that the choice of a model is not based only on its precision during its training, but also and especially on its precision during the validation tests, we carried out validation tests for the seven cases of configurations including the results on $R^2$ and MAPE are shown in Table 5.

### Table 5: Effects of input parameter configurations on the prediction model during validation tests.

<table>
<thead>
<tr>
<th>Cases</th>
<th>MAPE (%)</th>
<th>$R^2$(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>5.6879</td>
<td>90.0178</td>
</tr>
<tr>
<td>Case 2</td>
<td>5.5871</td>
<td>89.8176</td>
</tr>
<tr>
<td>Case 3</td>
<td>5.3032</td>
<td>92.4043</td>
</tr>
<tr>
<td>Case 4</td>
<td>8.6466</td>
<td>77.2865</td>
</tr>
<tr>
<td>Case 5</td>
<td>4.4964</td>
<td>95.5889</td>
</tr>
<tr>
<td>Case 6</td>
<td>22.3584</td>
<td>61.9213</td>
</tr>
<tr>
<td>Case 7</td>
<td>5.7205</td>
<td>89.7302</td>
</tr>
</tbody>
</table>

The results in Table 5 show that all models except the Case 4 model (because its $R^2$ decreased, 77.2865% vs. 78.8005% during training) fit the validation test data. Indeed, the MAPE and $R^2$ measurements were improved during these tests. The model of case 1 allows us to predict the electric charges with an $R^2$ coefficient of 90.0178%. The model of case 2 gives us an $R^2$ coefficient of 89.8176%, which is no longer the highest value, since the model of case 5 gives us an $R^2$ coefficient of 95.5889% which is the largest. The model of case 6 always gives us the lowest correlation coefficient (61.9213%). Note also that this is case 5 which presents the smallest MAPE (4.4964%), which is more logical since its correlation coefficient $R^2$ is the highest (95.5889%). Thus the model of case 5 with the highest coefficient $R^2$ and the lowest MAPE during the validation tests is chosen for the prediction of the electrical load of the energy system of Togo.

Following this work we will present and discuss the prediction of the electric charge for the year 2018 with the model chosen, i.e. the model of case 5.

**Prediction for the year 2018:**

Figure 8 shows the result of the prediction of Togo's electric charge for each half-hour of the year 2018. We can observe a linear correlation between the measured and predicted load data.
This correlation between the measured and predicted load data is well observed if we visualize this result for one week. Figure 9 shows the load prediction result from Sunday 07 to Saturday 13 January 2018. From this figure (Figure 9) we can observe the predicted and measured load data for each day (delimited according to Table 6) from this week.

In Figure 9, we note a strong correlation between the measured and predicted load data for the seven days (from January 07 to 13, 2018), however we observe a very large difference between the curves of the measured and predicted load data for two days. (Sunday 07 and Saturday 13 January 2018). This observation led us to measure the accuracy of the prediction of the electric charge for each day of the week of the year 2018, the results of which are shown in Table 6.

Table 6:- Electric charge prediction precision measurements for each day of the week in 2018.

<table>
<thead>
<tr>
<th>Days of the week</th>
<th>MAPE (%)</th>
<th>R^2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>9.2746</td>
<td>99.1417</td>
</tr>
<tr>
<td>Monday</td>
<td>2.6603</td>
<td>99.6350</td>
</tr>
<tr>
<td>Tuesday</td>
<td>2.0647</td>
<td>99.4225</td>
</tr>
<tr>
<td>Wednesday</td>
<td>1.6415</td>
<td>99.4168</td>
</tr>
<tr>
<td>Thursday</td>
<td>1.6535</td>
<td>99.4155</td>
</tr>
<tr>
<td>Friday</td>
<td>1.3678</td>
<td>99.3635</td>
</tr>
<tr>
<td>Saturday</td>
<td>12.8566</td>
<td>99.0620</td>
</tr>
</tbody>
</table>
Thus the results of Table 6 show that there is a strong correlation between the measured and predicted load data for the seven days of the week in 2018 since the average of the R² for each day is greater than 99%. We also observe a large average of the MAPEs between the curves of the measured and predicted load data for Sundays (9.2746%) and Saturdays (12.8566%). From Monday to Friday, the model of case 5 retained for the prediction of the electric charge of the energy system of Togo presents good performances (MAPE < 2.67% and R² > 99%) on the prediction of the electric charge of each day.

**Conclusion:**

This paper presents the short-term prediction of the electrical load of Togo's energy system by the Multiple Linear Regression method. Among seven models used (each of which differs from the other by the nature of these input parameters), we have chosen for the prediction of the electricity load a model having four input parameters (Day of the week, the type of Day (working or not), Hours in the day and Load data of the same time of the previous day). Our choice was justified thanks to the performances obtained during the validation tests of this model, since the linear multiple regression model allowed us to predict the electrical load of Togo's energy system for the year 2018 with a MAPE of 4.4964% and a correlation coefficient R² equal to 95.5889%.

**References:**

4. Ricco RAKOTOMALALA, (2010) : Pratique de la Regression lineaire Multiple : Diagnostic et sélection de variables. Université Lumière Lyon 2, France