PROCEDURE FOR CREATING CUSTOM MLR-BASED STLF MODELS BY USING GA OPTIMIZATION

Slobodan ILIĆ [1], Aleksandar SELAKOV [1], Srđan VUKMIROVIĆ [*]

[*1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

[*] Corresponding author; E-mail: srdjanvu@uns.ac.rs

This paper presents a novel procedure for short-term load forecasting (STLF) in distribution management systems (DMS). The load is forecasted for feeders that can be of a primarily residential, commercial, industrial or combined type. Each feeder has various amounts of distributed energy resources (DER) installed, which accounts for multiple different load patterns. Hence, the DMS cannot use a single STLF model for all forecasts. The proposed procedure addresses the specificity of each particular feeder type, by creating customized STLF models. It uses a genetic algorithm (GA) to select the best inputs for different multiple linear regression (MLR) models. The GA chooses variables from a dataset constructed using load and temperature measurements. The dataset is extended by adding nonlinear transformations and mutual interaction effects of the measurements, as well as calendar variables. This extension enables for the modelling of nonlinear influences and extracts the nonlinearity to the domain of input variables. The models' performance is assessed by the mean absolute percentage error (MAPE). The proposed procedure is applied to a set of measurements collected from a US electric power utility, which operates in the city of Burbank, CA. The obtained MLR model is compared with a previously proposed naïve benchmark, and a special comparison model, developed by correlation analysis. The proposed method is extendable to suit DMS systems with different types of electricity consumers.

Key words*: Short-term load forecasting; Genetic algorithm; Input variable selection; Multiple linear regression.*

1. Introduction

Short-term load forecasting (STLF) presents an important functionality for any electric power system. STLF is defined as the prediction of the load shape for a given set of consumers in some future period. The usual horizon of the forecasting is the next 24 hours (i.e., the next day), observed from the moment it is implemented. Forecasts are sometimes generated up to seven days ahead [1], depending on the specific needs of the system operator. Many important decisions are based on STLF, such as unit commitment, generator scheduling, maintenance plans, and economic dispatch. Accomplishing STLF is difficult, due to the nonlinearity of the load series, its dependence on many different factors (environmental, social, and economic), and their random-like behavior. The trend of applying new methodologies to tackle the challenges of power systems regarding STLF resulted in the development of numerous forecasting models. These models can roughly be categorised into the following groups: the classical methods [2],

[3]; the artificial intelligence methods [4], [5]; and the hybrid methods [6], [7], [8], [9]. Artificial intelligence and hybrid methods are often used for forecasting electricity price [10], [11], [12], which is closely related to STLF, as well as wind power [13], which proves the generality and practicality of such methods. However, a large number of procedures proposed in the literature produce dedicated models that cannot be generalized easily, and thus applied to a particular new real system. This inadequacy is mainly because most STLF models have been developed to suit the specific needs of the load process to which they were applied. To address the problem of generality in the field of STLF, [14] proposes a multivariate meta-learning system that is dedicated to finding a framework for STLF. While the mentioned approach produces promising results, the primary focus of the authors is the learning-based selection of an existing algorithm, which produces the best results for a given situation, rather than the customisation of a particular algorithm to suit the characteristics of a specific electric consumer. A framework for an intelligent energy management system in industry is proposed in [15], in which the authors analyse the input variables in order to choose the best model configuration. However, the primary data analysis of the potential STLF models' inputs is mainly based on statistical methods, the examination of which we expand in this paper. The field of STLF, in general, lacks the well-established methodology to produce benchmarking models for comparative assessment [16], and determining which inputs correlate to the predicted variables the most is a tedious task. Many papers have been written to address this problem, regardless of the technique used to model the actual prediction process. In [17], the authors proposed the method to determine the input space of the STLF Neural Network (NN) that is based on phase-space embedding of a load time series and results in a more parsimonious layout of the NN. In [18], the authors proposed a technique for reducing attributes, based on variable precision with a rough set.

This paper proposes a new method for input variable selection when designing a STLF model that must suit a particular consumer type for a given electric utility. The method uses a GA to select the variables that most influence the predicted load. The paper is organised as follows. The models for STLF that use MLR are described in Section 2. Section 3 presents the proposed methodology in detail, to enable its reproduction for the purposes of comparison. The obtained model has been applied to recorded data, the results of which are described in Section 4, together with an appropriate discussion. Section 5 concludes the paper while presenting the analysis of the solution and possible future development.

2. Multiple linear regression models for STLF

General linear regression models that are used for STLF and have normal error terms can be defined with the following equation:

$$L_i = \alpha_0 + \alpha_1 X_{i,1} + \alpha_2 X_{i,2} + \cdots + \alpha_{p-1} X_{i,p-1} + e_i \tag{1}$$

where $\alpha_0 \dots \alpha_{p-1}$ are the model parameters, $X_{i,1} \dots X_{i,p-1}$ are the known constants, and e_i is the independent, normally distributed random variable $N(0, \sigma^2)$. Then, the response function is:

$$E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{p-1} X_{p-1} \tag{2}$$

where $X_1 \dots X_{p-1}$ represent the predictor variables. Definition (1) thus implies that the observations L_i are independent normal variables, with a mean value of $E[L_i]$, as given by (2), and a constant variance σ^2.

2.1. Quantitative and qualitative prediction variables

Prediction variables that constitute the prediction function are often quantitative. For example, if we model the consumption in some area as a linear function of the number consumers in that area, the load time series will exhibit an increasing pattern if the number of consumers increases. The count of consumers in this case can be regarded as a quantitative variable.

Definition (1) does not strictly imply the use of variables that are quantitative. The use of qualitative variables, sometimes called class or dummy variables, is also possible. These variables represent such information as the type of the day, i.e., weekday or weekend, and can be included in the model. Indicators with values of 0 or 1 are used to identify the classes of a quantitative variable. For example, if the load (L) prediction is implemented based on the information about the type of the day for which the prediction is being performed (i.e., whether it falls on a weekend or on a workday), we define a qualitative prediction variable X_i in the following way:

$$\begin{cases} X_1 = 1, & \text{if the day is a weekday} \\ X_1 = 0, & \text{if the day is a weekend} \end{cases} \tag{3}$$

The predicting function is then defined as $E[L] = \alpha_0 + \alpha_1 X_1$, which becomes $E[L] = \alpha_0 + \alpha_1$, for the working day, and $E[L] = \alpha_0$, for the weekend.

2.2. Polynomial regression

Polynomial regression models can contain polynomials of the predictor variables, which make the response function curvilinear. For example, if one of the variables used to predict the load (L) is the temperature (Ti), and the order of the polynomial is three [3], then the prediction model can be written as follows:

$$L_i = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_1^2 + \alpha_3 T_1^3 + e_i \tag{4}$$

Although equation (4) has nonlinear terms, it is in fact a special case of (1) because the prediction variables T_1, T_1^2, and T_1^3 are independent, which we can present as X_{i1}, X_{i2}, and X_{i3}, while the model proposed in (4) becomes:

$$L_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} \tag{5}$$

2.3. Transformed variables

In addition to the basic variables that form the prediction model, their transformed shape can also be considered, as the model is still linear in its parameters. The transformations of the input variables can be nonlinear functions, such as *ln* and *log*. For example, we can write the following equation:

$$L_i = \alpha_0 + \alpha_1 T_1 + \alpha_2 \ln T_2 + \alpha_3 T_3 \tag{6}$$

The basic temperature measurement is transformed by a natural logarithm, and the variable enters the model as such. This is still a special case of (1) because, when $\ln T_2 = X_{i2}$, (6) can be rewritten as:

$$L_i = \alpha_0 + \alpha_1 T_1 + \alpha_2 X_{i2} + \alpha_3 T_3, \ln T_2 = X_{i2} \quad (7)$$

2.4. Interaction effects

In general, when two or more variables influence the process to be predicted, the behaviour of one variable can influence the behaviour of another one, and such a combined influence is called the interaction effect. For example, the same environmental temperature of 12 degrees Celsius will be perceived differently in January versus in August and thus cause different usage of heating/air-conditioning devices and, consequently, a different system load. Modelling such effects can be accomplished by multiplying two variables, for example, the current month and the current temperature. The model obtained in such a way is presented with the following equation:

$$L_i = \alpha_0 + \alpha_1 T_1 + \alpha_2 * m * T_2 + \alpha_3 T_3 \quad (8)$$

where *m* stands for the month in which the temperature T_2 occurred. We can see that this equation is still a special case of (1) because, when $X_{i2} = m * T_2$, (8) can be rewritten as:

$$L_i = \alpha_0 + \alpha_1 T_1 + \alpha_2 X_{i2} + \alpha_3 T_3 \quad (9)$$

3. The proposed methodology

This paper suggests a novel method for determining the inputs to a STLF model through the use of a GA. For previous examples of GAs used in electric power systems, see [19] and [20]. The proposed algorithm analyses the influence of the input variables to the predicted electric load. The variables that are candidates for input model selection are based on the recorded measurements of the load and the temperature. However, apart from the raw load and temperature measurements, their nonlinear transformations are also added as candidates to form the prediction model. The idea behind this approach is that, when combined, two or more variables, which are not strongly correlated to the targeting variable themselves, may have a significant impact on the prediction, as described in Section 2. Once the input variables are selected, a prediction model is constructed using classical MLR, for which the parameters are calculated by the ordinary LSE method. For the purpose of benchmarking, the performance of the proposed model was compared with two different models formed by the traditional approach. The first approach is a naïve MLR model proposed in the literature [16], which had also initiated the idea used herein. The second one was formed using correlation analysis, and it takes into account only those input variables that most correlate with the predicted load. The proposed methodology was applied to a set of load series data representing the energy consumption in a US utility from the city of Burbank. The differences in the behaviour of different models are analysed, and appropriate conclusions are thereupon formulated.

When proposing a conceptually new method to tackle the problem of STLF, it is necessary to validate its performance. Hundreds of different models have been proposed in the literature over the past decades, often claiming the improvement in precision in comparison with the other models. However, it is not always possible to validate the performance of those models, partly because they were developed for a

very narrow field of application and partly because many of them are still on theoretical grounds and lack significant practical value [16].

In this paper, we describe the application of the benchmark model proposed in [16] to our model. Based on the performance of the model and the correlation analysis of the load data time series, we constructed several MLR models for STLF that are customised for this particular load time series. The proposed process can serve as a starting point in constructing a real-life commercial application of STLF functionality, mainly as a benchmark but also as a component of more advanced, hybrid methods. Finally, we propose a GA optimising procedure to construct an efficient MLR model and apply it to the model at hand.

3.1. Research data

The models that are proposed in this paper require two sets of data for their construction. In the first case, we design the model to take into account only the temperature variables, while in the second case, the sole information that composes the model involves the past behaviour of the load process at the system level. Most electric utilities have access to these two sets of data; some even have their own meteorological facilities, while others use weather services. To determine the parameters of the model, a training period ranging from 1 January 2009 to 21 August 2011 was used. To validate the performance of the model, a period from 22 August 2011 to 28 August 2011 was used. We made this choice based on a previous analysis of the entire recorded dataset and the fact that the validation period chosen represents the week with the maximum consumption of electrical energy for the city of Burbank, CA. The tables and figures in Section 4 are all constructed with respect to this choice of the training/validation period. Some of the temperature variables, which were used to form the prediction model, would not be available in a real-life situation, and in this paper, the recorded measurements were used. This situation results in slightly altered errors in the actual exploitation phase of a model because temperature predictions are used instead of the recorded values.

3.2. Correlation analysis

One of the methods to select the input variables that constitute a STLF model is through a correlation analysis. In this paper, we shall examine the characteristics of several time series to determine the impact on the targeting load series. The first model is constructed using only temperature data, and the components of the signal that influence the targeting data series the most are chosen to enter the model. Apart from basic temperature measurements, there are also variables that represent its polynomial or nonlinear transformations. The reason for this choice can be seen from the fact that the exponent to a degree of a recorded temperature has higher correlation to the predicted load value (e.g., 15 hours ahead) than its basic recorded measurement. The detailed table of correlation-selected variables will be presented in Section 4.2. The second model is constructed using only the load data, and the components that correlate the most with the process to be predicted are taken into account. The approach of adding nonlinear transformations of the basic measurements, similar to the one mentioned above, is also applied here. The time period in which the temperature measurements influence the resulting load series was observed to be one week before, including the forecasting day. This choice was made based on common sense and the assumption that the temperature in buildings, ground, water, etc., shall only accumulate for a couple of days. However, the time period that was used to assess the correlation with the past load behaviour was

one week but without the prediction day because these data are not available at the moment of prediction (i.e., the prediction is performed for the period 24 hours ahead).

3.3. GA Optimisation of the STLF model

A GA is used to determine the best set of input variables for STLF functionality. A highlight of the proposed methodology is its applicability to different types of consumers in different electrical utilities. Furthermore, the algorithm does not have to be limited only to the application in MLR models but could also be used in the different STLF techniques already proposed in the literature.

3.3.1 Creating the extended input variable candidates set

Most utilities have access to load and temperature measurements. However, the dependency of the future load, if based only on the basic values of these recorded measurements, can rarely be determined with a satisfying accuracy. For this reason, we propose an extended data set that will contain the candidate variables for creating the input of a STLF model. The extensions of the basic recorded measurements data set are created in several ways:

- Creating nonlinear transformations of the basic measurements, including exponential, logarithmic and integral transformations,
- Including calendar variables (section 2.4),
- Modelling interaction effects by combining calendar variables with recorded measurements.

The entire extended data set of input variable candidates is presented in the Table 1. The parameters i and k represent the hour and the day lag, respectively. The parameter n represents the degree to which a variable is exponentiated. This is an important step, as the correlation analysis above demonstrates better correlation with such variables.

Table 1 Components of the extended data set of input variable candidates

Candidates	Description
$T^n{}_{d-k}(h-i)$	nth degree of the recorded temperature
$\ln T_{d-k}(h-i)$	Natural logarithm of the recorded temperature
$\int_{h-i-23}^{h-i} T_{d-k}(t)dt$	Integral of the recorded temperature, over the 24 hours
$m * T^n_{d-k}(h-i)$	Interaction effect between the current month and the nth degree of the temperature
$d * T^n_{d-k}(h-i)$	Interaction effect between the current day and the nth degree of the temperature
$h * T^n_{d-k}(h-i)$	Interaction effect between the current hour and the nth degree of the temperature
$L^2_{d-k}(h-i)$	nth degree of the recorded load
$d * L_{d-k}(h-i)$	Interaction effect between the current day and the load
$\ln L_{d-k}(h-i)$	Natural logarithm of the recorded load
$\int_{h-i-23}^{h-i} T_{d-k}(t)dt$	Integral of the recorded temperature, over the 24 hours

Some variables that are added to the dataset do not have significant correlation with the predicted load time series but can still affect the predicted load time series. These variables are mainly qualitative calendar and interaction variables described in Section 2.

3.3.2 Population

The population of the GA consists of a fixed number of genes, each representing a subset of the extended variable set. A fixed number of elements (variables) exists in one gene, and it represents the final number of the MLR model parameters. This number is modified in different GA optimisations, and the best choice for this particular approach is presented in Section 4, and is represented by the number of model components. The initial elements that will form a particular gene are selected randomly from the extended dataset, with uniform distribution. The population size that provided the best results in terms of accuracy and training time was 100 genes.

3.3.3 Selection

For each gene in a population, the appropriate variables are extracted from the dataset. The matrices that represent inputs and outputs for each day of the training period and of the prediction period are then formed. The MLR model can be described with the matrix equation (10), where S stands for the all-selected variables matrix, *y* represents all respective outputs for those variables, and the vector *b* stands for the unknown coefficients. The vector *y* contains the recorded loads for each hour of each day of the training period. The matrix *S* contains the same number of instances (i.e., rows) as *y*, with each instance representing a particular set of variables chosen by the GA. Once the GA forms the matrices, it determines the model parameters with the LSE method, namely by solving the matrix equation (11).

$$S * b = y \tag{10}$$

$$b = (S^T * S)^{-1} * S^T * y \tag{11}$$

Once the coefficients of the model are calculated, the GA also calculates the load predictions for the prediction part of the dataset, using the first matrix equation (10). When predictions are available, the assessment of the model's performance is calculated through the mean average percentage error (MAPE) in the following way:

$$MAPE = \sum_{i=1}^{n} \left| \frac{y_i - p_i}{y_i} \right| * 100\% \tag{12}$$

where *n* stands for the number of hours in the forecasting period, and *y* and *p* stand for the actual and predicted values, respectively. Finally, the MAPE is used as a fitness function. This procedure is repeated for each individual of the current population, and from that point, the GA continues its operation. Selection of the best individuals is performed as a classical roulette selection, since this option provided the best results.

3.3.4 Crossover and mutation

The in-memory representation of a GA gene that is used throughout these experiments is by indices of the selected variables in the extended dataset. For this kind of representation, the best results were obtained by using the scattered crossover, and the uniform mutation.

4. RESULTS AND DISCUSSION

4.1. Comparison with a naïve model

The model proposed in [16] consists of several variables that represent the constant value, the linear trend, the combined influence of the current day and the temperature, the different interaction effects between an exponent of the current temperature, and the current month/hour. The model is constructed for the purposes of benchmarking and can be stated with the following equation:

$$L_d(h) = \alpha_0 + \alpha_1 * Trend + \alpha_2 * d * h + \alpha_3 * m + \alpha_4 * m * T_d(h) + \alpha_5 * m * T_d^2(h) + \alpha_6 * m * T_d^3(h) + \alpha_7 * h * T_d(h) + \alpha_8 * h * T_d^2(h) + \alpha_9 * h * T_d^3(h) \quad (13)$$

The authors have reported the acceptable performance of this model when used on their dataset. The mean average percentage error (MAPE) is reported to be oscillating around 5%, depending on the prediction preferences. The lowest MAPE for the hourly load is reported to be 4.89%, for the 24 hours ahead prediction. When applied to the dataset used in this paper, the model did not perform with a satisfactory accuracy. The performance of the benchmark model is presented in Figure 1, together with the performances of the models developed by correlation analysis. The horizontal axis represents the forecasting period from 22 August 2011 to 28 August 2011. The vertical axis represents the electrical load of the system in Megawatts. To consider the benchmark model plausible, the MAPE would have to be in the order of magnitude reported by the authors, which is around 5%. The MAPE for the testing period in our repeated experiments (with the naïve model) is 16.18%, which is not acceptable, neither for benchmarking purposes nor for commercial application.

4.2. Correlation analysis of the temperature series influences

The temperature influence was independently tested from any other variable. The temperature variables that constitute the prediction model were chosen based on their correlation to the load series. The time span that was included for the possible choice of temperature variables was seven days, including the recorded measurements of the forecasting day. This choice was based upon trial and error, keeping in mind the heat capacity of the environment. Apart from the basic temperature measurements, the other variables also served as candidates to form the model, such as their nonlinear transformations and the interaction variables. Based on the selected variables, we can construct a model that is described by the following equation:

$$L_d(h) = \alpha_1 T_d^4(h-15) + \alpha_2 T_d^3(h-15) + \alpha_3 * T_d^4(h-16) + \alpha_4 * T_d^4(h-14) + \alpha_5 * T_d^3(h-16) + \alpha_6 * T_d^2(h-15) + \alpha_7 * T_d^3(h-14) + \alpha_7 * T_d^2(h-14) \quad (14)$$

The proposed model was tested with the recorded temperature measurements, and the obtained results are shown in Figure 1. The MAPE of this model for the testing period is 14.08%, which is better than the first model but is still not acceptable, since greater precision can be achieved by a simple correlation analysis of the load series (explained in detail in 4.3). We can see, however, that the prediction curve fits the actual load time-series better than the benchmark model. This potentially means that the generated model could still provide some useful information (i.e. not contained in the load time series, hence useful when combined), regardless of the large prediction error.

4.3. Correlation analysis of the load series influences

The influence of the past behaviour of the load time-series on the prediction accuracy was tested in isolation from the other variables. The load variables that enter the model were chosen based on the correlation analysis to the prediction signal. At the time of the forecasting, only the previously recorded data were available; therefore, the most recent variables start from one day before the forecasting day and look back seven days. In addition to the basic load measurements, other variables can be chosen as model candidates, namely nonlinear transformations and interaction effects. Based on the presented variable choice, we can form the model described by the following equation:

$$\begin{aligned} L_d(h) = {} & \alpha_1 * L_{d-1}(h) + \alpha_2 * \ln\big(L_{d-1}(h)\big) + \alpha_3 * L_{d-1}^2(h) + \alpha_4 \\ & * L_{d-1}(h-1) + \alpha_5 * \ln\big(L_{d-7}(h)\big) + \alpha_6 * L_{d-1}^2(h-1) \\ & + \alpha_7 * \ln\big(L_{d-1}(h-1)\big) + \alpha_8 * L_{d-7}(h) \end{aligned} \tag{15}$$

The proposed model was tested with the recorded load measurements, and the obtained results are shown in Figure 1. The MAPE of this model for the testing period is 6.65%, which is better than both previous models and could be used for comparison purposes with the other models. The fitting of the prediction curve to the actual load time series is the best in comparison with the previous two models.

4.4. Combined temperature and load model

We proceed by combining two previously obtained models to form a model of satisfactory accuracy for comparison purposes. The combined model is formed by integrating the influence of the temperature and the load variables. The model can be defined with the following equation:

$$\begin{aligned} L_d(h) = {} & \alpha_1 T_d^4(h-15) + \alpha_2 T_d^3(h-15) + \alpha_3 * T_d^4(h-16) + \alpha_4 * T_d^4(h-14) \\ & + \alpha_5 * T_d^3(h-16) + \alpha_6 * T_d^2(h-15) + \alpha_7 * T_d^3(h-14) + \alpha_8 \\ & * T_d^2(h-14) + \alpha_9 * L_{d-1}(h) + \alpha_{10} * \ln\big(L_{d-1}(h)\big) + \alpha_{11} \\ & * L_{d-1}^2(h) + \alpha_{12} * L_{d-1}(h-1) + \alpha_{13} * \ln\big(L_{d-7}(h)\big) + \alpha_{14} \\ & * L_{d-1}^2(h-1) + \alpha_{15} * \ln\big(L_{d-1}(h-1)\big) + \alpha_{16} * L_{d-7}(h) \end{aligned} \tag{16}$$

with a total of 16 parameters. The results of the combined model are shown in Figure 1. The MAPE for the displayed period is 5.95%, which can be useful for comparison purposes. We can see that the fitting of the curve of the prediction to the actual recorded data is best in comparison with the previous three models.

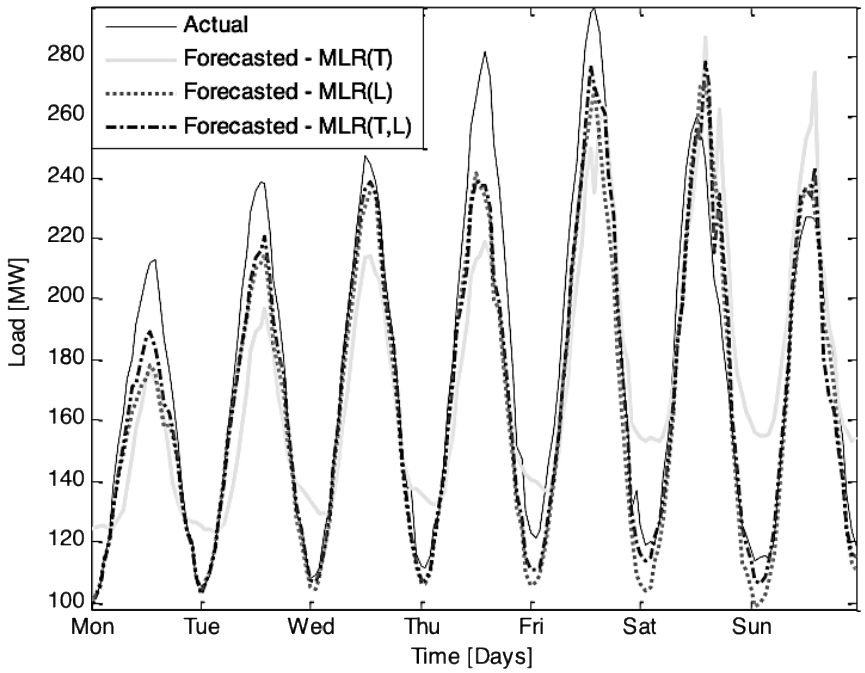

Figure 1 Performances of different MLR models determined by a correlation analysis. The MAPE for each of the models is presented in Table 2

4.5. Model obtained by use of a GA

The customised model is obtained by the procedure described in Section 3.3. The optimisation was performed for the same time period as in previous examples, with the prediction period being the peak week for the year 2011. The number of variables that constitute the model is set to 16, which resulted in 16 coefficients to be determined. This choice ensures the fair comparison between methods obtained by purely correlational analysis because the combined method, described in the previous section, is also composed of 16 input variables. The performance of the model obtained by GA is shown in Figure 2, together with the performances of the comparison models.

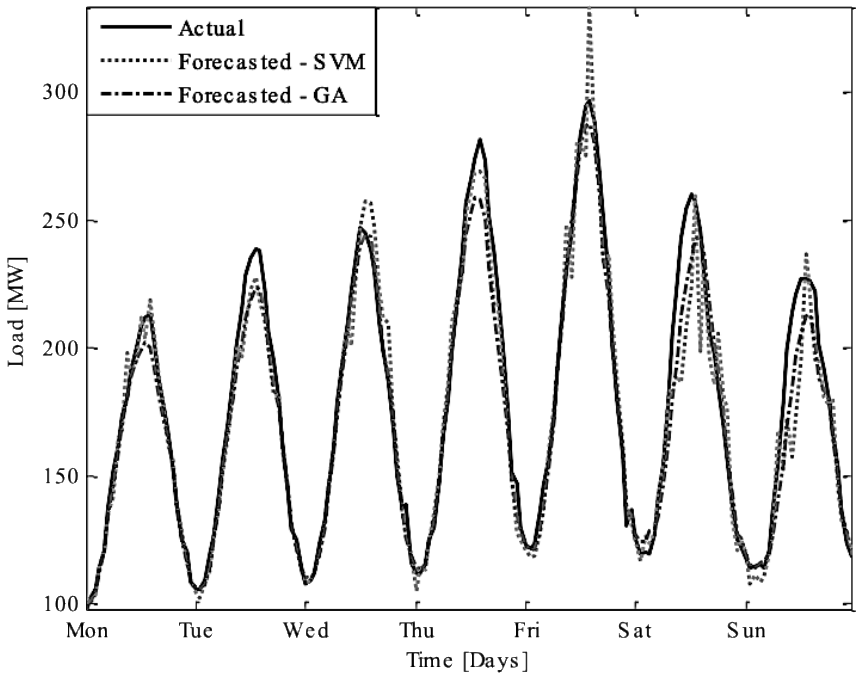

Figure 2 Performance of the model obtained by the GA, in comparison with model obtained by SVM.

4.6. Comparison of prediction errors introduced by different models

The models that were constructed for the purposes of the research in the scope of this paper all perform differently when applied to the same testing period with the same training data. This variety of responses is not always accented in the STLF field, partly because a majority of the methods reported in the literature are developed to suit the particular needs of the specific electric utility. In this paper, a procedure for choosing the best model to suit a particular purpose is presented, and the model obtained in such a way clearly demonstrated better performance in comparison with other "ad hoc" models, as well as with the ones constructed by a correlation analysis. The results in terms of the MAPE are shown Table 2. In addition to the MAPE, which is most commonly used to depict the performance of a given model, other errors are also calculated, such as the mean absolute error (MAE) and the mean weighted absolute error (MWAE), which are defined by the following equations:

$$MAE = \sum_{i=1}^{n} |y_i - p_i| \tag{17}$$

$$MWAE = \sum_{i=1}^{n} |(y_i - p_i) * y_i| \tag{18}$$

This last measurement, the MWAE, can serve as an indicator of the system behaviour when it is being fully loaded because the forecasting errors will be treated with greater respect if they occur during the peaking period (high load).

It is not always possible to reproduce the models presented in literature, partly because of the data availability, and partly because of the complexity of the implementation. Furthermore, the selected comparison method does not always have to represent the behaviour of the underlying process very well, especially if it has been designed to suit a specific electric consumer in a completely different utility. The intention of this paper was not the development and fine tuning of a particular model, but rather a proposal of a uniform modelling method, that could be used in various scenarios. Apart from the MLR based models developed in this paper, a comparison with another model obtained by a different technique has also been presented, namely with support-vector machines (SVM). This particular SVM comparison model is currently employed as a production tool in a US electric power utility, from which the data for this research was obtained. The results are presented graphically in the Figure 2, and numerically in the Table 2. This comparison confirms the validity of the STLF customising method proposed in this paper, because the overall precision and stability is better for the obtained MLR model.

Table 2 Performance comparison of the different models, in terms of the MAPE, MAE, and MWAE

Model	MAPE [%]	MAE [MW]	MWAE [-]
Naïve benchmark	16.18	23.76	8784
Purely temperature	13.42	15.94	4769
Purely load	6.65	9.87	2944
Combined	5.96	8.55	2395
GA	3.51	6.88	1429
SVM	4.57	8.85	1824

5. Conclusion

When designing a STLF model, selection of the input variables presents one of the challenges because the performance of different models varies greatly, depending on the process to which it is being applied. One particular forecasting model may perform with exceptional precision on one class of electrical consumers, while it may completely fail on another class. Therefore, different utilities often invest in the design of customised models. In this paper, we investigated the possibility of applying a GA to tackle the mentioned challenge. We proposed a relatively simple but powerful customising procedure that can be applied to different classes of load processes, rather than proposing a single novel STLF model and claiming its advantage over other models. This paper can provide useful insight into the benchmarking process in the field of STLF and can thus serve as a starting point for researchers and practitioners who are designing sophisticated models to suit particular purposes.

An independent model was developed based on a correlation analysis for the purpose of benchmarking. The performance of the benchmark model and the GA-obtained model are directly compared, and the

results clearly demonstrate the advantage of the GA approach. In terms of the MAPE, the GA constructed model is 2.5% more accurate when predicting the load for the peak week of 2011. The GA obtained model has also been compared with an in house STLF commercial tool based on SVM, and it is 1% more precise in terms of MAPE. The performance of the model obtained by the proposed GA optimising procedure may be satisfactory for the purposes of comparison for some utilities, while for others, it may be satisfactory as an actual commercial tool. Note that in the process of comparing STLF models, the recorded temperatures were used. In the actual exploitation phase, the predicted temperatures would be used, which would yield greater forecasting error, the amount of which must be determined by on-site experimentation.

In further development of the concepts proposed in this paper, we will investigate the possibility of applying the mentioned optimising procedure to more sophisticated STLF methods, such as the artificial neural network (ANN) and support vector machine (SVM). Other optimising procedures, when selecting input model variables, such as particle swarm optimisation (PSO), may also be used.

Acknowledgements

The authors thank Burbank Water and Power for providing the data for this research.

References

[1] T. Peng, N. Hubele and G. Karady, "An adaptive neural network approach to one-week ahead load forecasting," *Power Systems, IEEE Transactions on*, vol. 8, no. 3, pp. 1195-1203, Aug 1993.

[2] S. Pappas, L. Ekonomou, P. Karampelas, D. Karamousantas, S. Katsikas, G. Chatzarakis and P. Skafidas, "Electricity demand load forecasting of the Hellenic power system using an ARMA model," *Electric Power Systems Research*, vol. 80, no. 3, pp. 256-264, March 2010.

[3] A. Bracale, G. Carpinelli, P. De Falco and T. Hong, Short-term industrial reactive power forecasting, vol. 107, International Journal of Electrical Power & Energy Systems, 2019, pp. 177-185.

[4] H. Hippert, C. Pedreira and R. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans Power Syst*, vol. 16, no. 1, pp. 44-45, 2001.

[5] M. López, S. Valero, C. Senabre, J. Aparicio and A. Gabaldon, "Application of SOM neural networks to short-term load forecasting: The Spanish electricity market case study," *Electric Power Systems Research*, vol. 91, pp. 18-27, October 2012.

[6] S. Ilić, A. Erdeljan, F. Kulić and S. Vukmirović, "Hybrid artificial neural network system for short-term load forecasting," *Thermal Science*, vol. 16, no. 1, pp. 215-224, 2012.

[7] R.-A. Hooshmand, H. Amooshahi and M. Parastegari, "A hybrid intelligent algorithm based short-term load forecasting approach," *International Journal of Electrical Power & Energy Systems*, vol. 45, no. 1, pp. 313-324, February 2013.

[8] M. Amina, V. Kodogiannis, I. Petrounias and D. Tomtsis, "A hybrid intelligent approach for the prediction of electricity consumption," *International Journal of Electrical Power & Energy Systems*, vol. 43, no. 1, pp. 99-108, December 2012.

[9] P. Singh, P. Dwivedi and V. Kant, A hybrid method based on neural network and improved environmental adaptation method using Controlled Gaussian Mutation with real parameter for short-term load forecasting, vol. 174, 2019, pp. 460-477.

[10] N. Amjady and A. Daraeepour, "Mixed price and load forecasting of electricity markets by a new iterative prediction method," *Electric Power Systems Research,* vol. 79, no. 9, pp. 1329-1336, September 2009.

[11] N. Amjady and F. Keynia, "Electricity market price spike analysis by a hybrid data model and feature selection technique," *Electric Power Systems Research,* vol. 80, no. 3, pp. 318-327, March 2010.

[12] N. Bigdeli, K. Afshar and N. Amjady, "Market data analysis and short-term price forecasting in the Iran electricity market with pay-as-bid payment mechanism," *Electric Power Systems Research,* vol. 79, no. 6, pp. 888-898, June 2009.

[13] N. Amjady, F. Keynia and H. Zareipour, "Short-term wind power forecasting using ridgelet neural network," *Electric Power Systems Research,* vol. 81, no. 12, pp. 2099-2107, December 2011.

[14] M. Matijaš, J. A. Suykens and S. Krajcar, "Load forecasting using a multivariate meta-learning system," *Expert Systems with Applications,* vol. 40, no. 1, p. 4427–4437, September 2013.

[15] J. J. Cárdenas, L. Romeral, A. Garcia and F. Andrade, "Load forecasting framework of electricity consumptions for an Intelligent Energy Management System in the user-side," *Expert Systems with Applications,* vol. 39, no. 5, p. 5557–5565, April 2012.

[16] T. Hong, P. Wang and H. L. Willis, "A Naïve Multiple Linear Regression Benchmark for Short Term Load Forecasting," in *Power and Energy Society General Meeting, 2011 IEEE,* Raleigh, NC, USA, 24-29 July 2011.

[17] I. Drezga and S. Rahman, "Input variable selection for ANN-based short-term load forecasting," *Power Systems, IEEE Transactions on,* vol. 13, no. 4, pp. 1238-1244, Nov 1998.

[18] X. Zhi, Shi-Jie Ye, Z. Bo and S. Cai-Xin, "BP neural network with rough set for short term load forecasting," *Expert Systems with Applications,* vol. 36, no. 1, p. 273–279, January 2009.

[19] Y. Hu, J. Li, M. Hong, J. Ren, R. Lin, Y. Liu, M. Liu and Y. Man, Short term electric load forecasting model and its verification for process industrial enterprises based on hybrid GA-PSO-BPNN algorithm—A case study of papermaking process, vol. 170, Energy, 2019, pp. 1215-1227.

[20] Z. Miljanić, I. Djurović and I. Vujošević, "Optimal placement of PMUs with limited number of channels," *Electric Power Systems Research,* vol. 90, pp. 93-98, September 2012.

Submitted: 05.12.2019.

Revised: 25.01.2020.

Accepted: 07.02.2020.