MODELING THE ELECTRICITY CONSUMPTION OF THE BRAZILIAN FREE TRADING ENVIRONMENT WITH GENETIC PROGRAMMING

Bruno Quaresma Bastos
Departamento de Engenharia Industrial, PUC-Rio
Marquês de São Vicente Street, 225, Rio de Janeiro, Brazil, 22430-060
brunoq.b@aluno.puc-rio.br

Gheisa Roberta Telles Esteves
Departamento de Engenharia Industrial, PUC-Rio
Marquês de São Vicente Street, 225, Rio de Janeiro, Brazil, 22430-060
gheisa@esp.puc-rio.br

Fernando Luiz Cyrino Oliveira
Departamento de Engenharia Industrial, PUC-Rio
Marquês de São Vicente Street, 225, Rio de Janeiro, Brazil, 22430-060
cyrino@puc-rio.br

ABSTRACT

It is important for companies interested in trading at the Free Trading Environment (FTE) to understand the development of the market over time, considering macroeconomic variables and specific indicators. In this context, this article proposes the use of genetic programming (GP) to build a multivariate model for the electricity consumption of the Brazilian FTE. Variables such as industrial production, number of clients in the FTE, and industrial electricity tariff, are considered as candidate variables in the GP framework. Different models are built with GP, and the best one is selected through the evaluation of the forecasts on a validation set. To evaluate the model's performance, forecasts are made on a test set, and are compared with forecasts of other methods, such as artificial neural networks, multiple linear regression, SARIMA and exponential smoothing. The results show that the symbolic regression model built via GP provides the best forecasts for the FTE.

KEYWORDS. Genetic Programming, Forecasting, Electricity Consumption.

OR in Energy, Metaheuristics.

1. Introduction

In Brazil, there are two different electrical energy trading environments, the Regulated Trading Environment (RTE) and the Free Trading Environment (FTE). In the RTE, distribution companies (DISCOs) purchase energy through public auctions by means of medium and long term contracts (Street, et al., 2012), whereas, in the FTE, net consumers, generation companies and trading companies freely negotiate energy through bilateral contracts (Street, et al., 2012).

The RTE has always been responsible for the major share of the Brazilian electricity consumption (see Figure 1). The FTE, however, is gaining ever more attention, especially due to increasing electricity tariffs of DISCOs (ANEEL) and to special benefits for those who trade renewable energy in the FTE (Bruno, Ahmed, Shapiro, & Street, 2016). In this context, it is important that companies, when assessing the possibility of trading electrical energy at the FTE, understand the development of the market. One way of doing this is by producing valuable medium and long term forecasts of the market's electrical energy consumption.

Several techniques can be applied to forecast electricity consumption. Multivariate models can consider explanatory variables in their modeling, as opposed to univariate models, which cannot. When considering explanatory variables, one can take into account the many factors that affect the electricity consumption; furthermore, one is capable of making what-if analysis (Hong & Fan, 2016) by producing scenarios based on pre-established (long-term) assumptions for one or more explanatory variables. Both aspects are of great importance for decision making in the long-term.

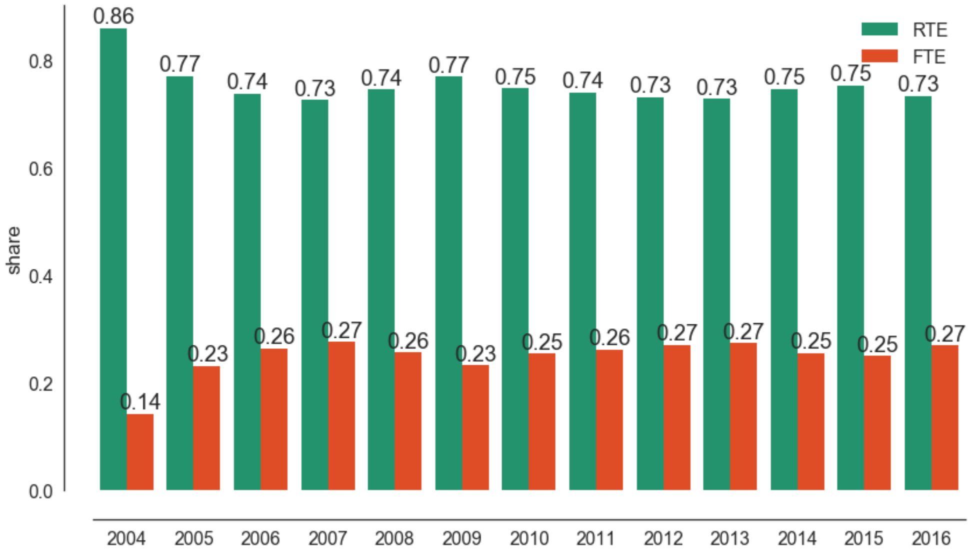

Figure 1 – Share of electrical energy consumption in Brazil. Source: (EPE)

According to (Hong & Fan, 2016), the most common multivariate models for load forecasting are multiple linear regression (MLR) models (Chatfield, 2000), artificial neural networks (ANNs) (Haykin, 1999) and support vector regression (SVR) models (Smola & Schölkopf, 2004). MLR models are based on regression analysis, a statistical technique for estimating the relationships among variables (Hong & Fan, 2016). ANNs and SVR are, on the other hand, machine learning models. ANNs are inspired by biologic neurons and by the massive parallel structure of the brain (Haykin, 1999). SVR models are based on statistical learning theory and on the principle of structural risk minimization (Chou & Ngo, 2016).

When modeling a problem with MLR, one must specify the functional form that relates input and output variables. This relationship, however, is not always known. When modeling a problem with ANN and SVR, one does not need to specify the functional form. The ANNs are trained to learn and to generalize the input-output relationship from the historical data. The SVR, on the other hand, maps the input-output relationship in a high-dimensional space and finds a function that best fits the data (Kong, Liu, Shi, & Lee, 2015). Both ANN and SVR models offer no insights to the input-output relationships. However, the insight may be desirable in order to understand the process that is being modeled.

The variables that affect the electricity consumption of the Brazilian FTE and the relations that these variables have with it are not quite known. However, knowing which are the variables, and what are their relations with the Brazilian FTE, is valuable to understand the dynamics of the FTE. In this context, genetic programming (GP) (Koza, 1992) could be used to build a multivariate model for the electricity consumption of the Brazilian FTE. Genetic programming is an evolutionary computation technique that aims at learning computer programs (models) by a process that mimics theory of evolution (e.g., survival of the fittest) and genetics (e.g., gene mutation). GP uses this process to search the best computer program (model) for a specific problem, in a space of possible computer programs (models) (Lee, Lee, & Chang, 1997). GP could, therefore, be used to construct a model that best relates input and output, without having prior knowledge about their relationship (Yang, Li, Wang, Lian, & Ma, 2015). As with ANN and SVR, GP does not require the specification of a functional form. The form will be discovered by the evolutionary algorithm. GP offers, thus, interpretability, as it constructs a model with defined functional form and coefficients.

Genetic programming has already been applied in electricity consumption and load forecasting. Amber, Aslam, & Hussain (2015) developed an MLR model and a GP model to forecast the daily electricity consumption of a building; the input variables presented to the model were: temperature, weekday index, solar radiation, relative humidity and wind speed. Forouzanfar, et al. (2012) applied multi-level GP to forecast the transport energy demand of Iran; they used as explanatory variables: population, gross domestic product (GDP) and number of vehicles. Lee, Lee, & Chang (1997) applied genetic programming to create a long-term load forecasting model; the inputs included GDP and population (with time lags).

This work aims at building a representative multivariate model for the monthly electricity consumption of the Brazilian FTE. In order to do that, genetic programming technique is employed to produce a symbolic regression model. Industrial production, industrial electricity tariff, number of FTE clients and FTE consumption (with time lags) are considered as candidates for explanatory variables. The GP procedure is applied several times, so that different models are built. The best one is selected by evaluating the (*ex-post*) forecasts on a validation set. In order to assess the performance of the selected model, its forecasts are evaluated in a test set and compared with forecasts produced using MLR, ANN, SARIMA (Box & Jenkins, 1976) and Exponential Smoothing (Hyndman, Koehler, Snyder, & Grose, 2002).

This paper contributes to the modeling of the electricity consumption of the Brazilian FTE. It proposes use of genetic programming to discover a model for the FTE electricity consumption. The genetic programming process maps different candidate variables to find a nonlinear multivariate model for the FTE consumption. The model could then be used to produce what-if analysis or to forecast the electricity consumption of the FTE, assessing decision-makers in long-term decisions related to trading in the FTE.

This paper is structure as follows: after the introduction, a brief overview on genetic programming is made in the second section. In the third section, the modeling of the problem is described. In the fourth section, the results are presented, and in the fifth section the conclusions are made.

2. Genetic Programming Overview

Genetic programming (GP) was first introduced by Koza in 1992 (Koza, 1992). Since then, it has been used in different types of problems. According to Amber, Aslam, & Hussain (2015), GP is a branch of machine learning algorithms in which a population of computer programs is evolved. The computer programs are also called GP individuals or GP solutions. Each individual is a computer program and, thus, a potential solution of the problem presented to the GP. Individuals are represented as trees whose nodes are procedures, functions, variables and constants (Lee, Lee, & Chang, 1997). Figure 2 illustrates the representation of the program $x^2 + y$. The trees are evolved in the GP process. Individuals can be multigene, i.e., can contain more than one gene. In this case, each gene receives a weight, and the multigene individual is the weighted sum of all of its genes (see (Searson, 2015)). In general, the restriction on the number of genes (for multigene individuals) and on the depth of trees allows one to produce less complex and more compact models. In GP modeling, the user defines all the possible functions, variables and constants that can be used as nodes (Lee, Lee, & Chang, 1997).

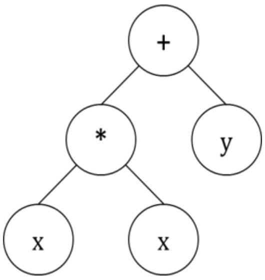

Figure 2 – Tree representation of the program $\mathrm{x}^2 + \mathrm{y}$

The evolution of programs with GP follows six steps (see (Forouzanfar, Doustmohammadi, Hasanzadeh, & Shakouri G, 2012)). (1) First, the algorithm creates an initial random population of programs and sets current population as this initial population. Following this, (2) the algorithm selects programs of the current population for breeding. The selection is usually made considering the fitness of the programs: the best ones, which have higher fitness, have more probability of being selected by the algorithm. It is possible, however, to consider other criteria in the selection process, e.g., fitness and model complexity (see, for example, Pareto tournament selection in (Searson, 2015)). After the selection, (3) genetic operators such as crossover, mutation and reproduction are applied. A new population of offspring is then created. (4) Each offspring in this new population is evaluated (fitness). (5) This new population is made to be the current population, and steps (2)-(5) are repeated until a stopping criteria is achieved (e.g., until the procedure achieves a given number of generations). (6) Select the best program created by the procedure. For an in depth read on GP, please refer to (Koza, 1992).

In the context of this work, the GP method is applied to find a model that estimates the electricity consumption of the Brazilian FTE based on macroeconomic variables and variables related to the Brazilian FTE and RTE. GP is then used to produce a model that has the form $\hat{y}_{gp} = f(x_1, x_2, \dots, x_n)$.

3. Methodology

3.1 Explicative Variables

In order to develop a multivariate model for the electricity consumption of the FTE, one needs first to identify the variables that could help explain it. Considering that approximately 90% of the total electricity consumption of the FTE is due to industrial clients (CCEE, 2016), the variables selected for the GP modeling were related to the Industry sector; indicators of the FTE were also selected as candidate variables. The variables considered to explain the FTE consumption were the following:

- Total electricity consumption of the FTE;
- Total number of clients participating of the FTE;
- Total industrial production in Brazil;
- Electricity tariff for the industrial class in the RTE.

In this work's modeling approach, it is considered that the consumption of the FTE (i.e., the variable of interest), at a given time t, could be explained by all the above-listed variables at given time t and at times $t - k$, where $k = 1{,}2{,}3$. Considering this, the final candidate variables for the problem were those presented in Table I.

Table I – Candidate variables for the model

FTE Consumption	Number of Clients	Industrial Production	Industrial Electricity Tariff
$\text{consumption}_{(t-1)}$	$\text{clients}_{(t)}$	$\text{industrial\_production}_{(t)}$	$\text{industrial\_tariff}_{(t)}$
$\text{consumption}_{(t-2)}$	$\text{clients}_{(t-1)}$	$\text{industrial\_production}_{(t-1)}$	$\text{industrial\_tariff}_{(t-1)}$
$\text{consumption}_{(t-3)}$	$\text{clients}_{(t-2)}$	$\text{industrial\_production}_{(t-2)}$	$\text{industrial\_tariff}_{(t-2)}$
	$\text{clients}_{(t-3)}$	$\text{industrial\_production}_{(t-3)}$	$\text{industrial\_tariff}_{(t-3)}$

3.2 Dataset

The data of the variables used in the modelling were obtained from different sources. Data of the electricity consumption and number of clients of the FTE were collected from the Electrical Energy Trading Council (CCEE, in portuguese). Data regarding the industrial production and industrial electricity tariffs were provided by Tendências Consultoria Integrada. The historic data consists of monthly observations, and was obtained from January 2008 until December 2016, totalizing 105 observations for each of the variables listed above.

In order to evaluate the performance of the models in forecasting the FTE consumption and generalization capability, the historic data was divided into three sets: training set, validation set and test set. The data in the training set is used by the GP to obtain the model for the FTE consumption through evolutionary process described in previous sections. The training set contained 87 examples of the historic data, with time *t* ranging from April 2008 until June 2015).

The data in the validation set is used to evaluate the performance of the models and to identify the best one. The set used in this work contained 9 examples of the historic data, with time *t* from July 2015 until March 2016. The test set, finally, is used to evaluate if the selected model can generalize well the data, i.e., if it does provide good forecasts. The set also contained 9 examples of the historic data, with *t* ranging from April 2016 until December 2016. The split is implemented in order to select and evaluate the model obtained by the methodology. In Figure 3, the time series of the variable of interest is illustrated.

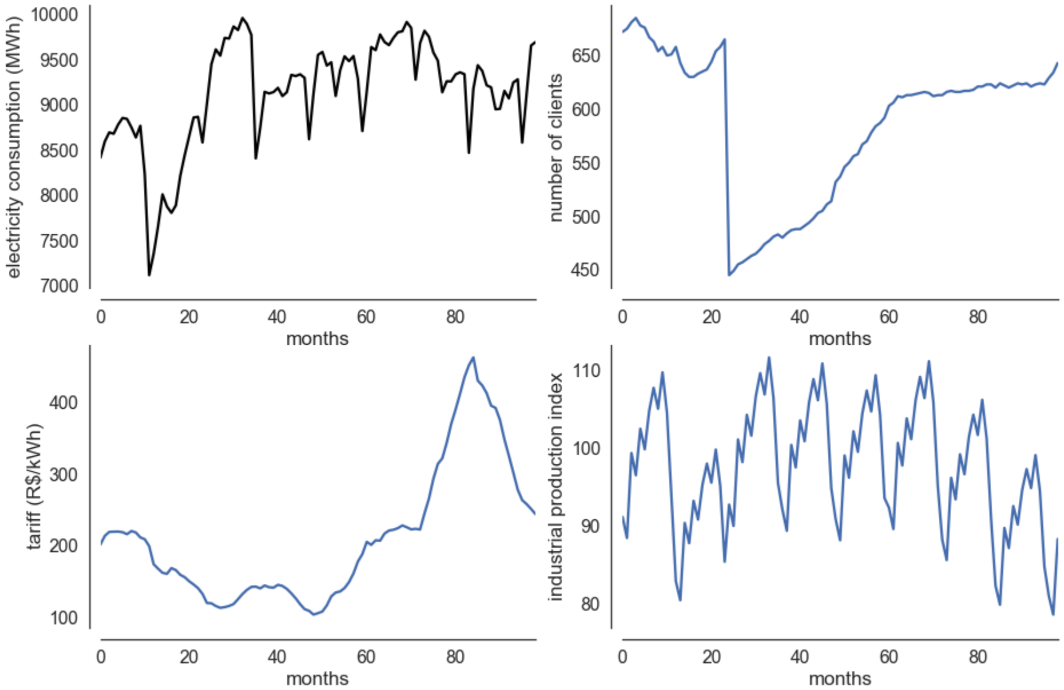

Figure 3 – Time series of independent variable (black) and candidate explanatory variables (blue)

An excerpt of the database is shown in Table II. The variable of interest is listed in the second column, and the explicative variables in the subsequent columns. The first example, observed at time $t = 1$ (April 2008), is listed in first row that follows the table's heading.

Table II – Database structure

t	cons(t)	cons(t-1)	cons(t-2)	cons(t-3)	clients(t)	clients(t-1)	...
1	8678	8693	8592	8413	685	681	...
2	8779	8678	8693	8592	678	685	...
3	8852	8779	8678	8693	676	678	...
4	8842	8852	8779	8678	667	676	...
5	8748	8842	8852	8779	663	667	...
6	8637	8748	8842	8852	654	663	...
7	8766	8637	8748	8842	658	654	...
8	8231	8766	8637	8748	650	658	...
9	7117	8231	8766	8637	651	650	...
10	7352	7117	8231	8766	658	651	...
11	7658	7352	7117	8231	643	658	...
12	8009	7658	7352	7117	634	643	...
13	7875	8009	7658	7352	630	634	...
14	7805	7875	8009	7658	630	630	...
15	7887	7805	7875	8009	633	630	...

3.3 Genetic Programming Application

In this work, the individuals in a population are models for the FTE consumption. The GP method receives the input and output data contained in the training set. GP uses its evolutionary process to learn how to relate the explicative variables in a way that it fits the output values. In order to implement the GP to find a model, the user has to specify several parameters that affect the evolutionary process. Aiming to find the best model for the FTE consumption, different parametrization of GP were tested. The number of individuals in a population, the maximum depth of the program, the maximum number of genes, and others, were varied.

In Table III[1] the different GP configurations tested in the work are presented. The parameters of maximum number of genes and maximum depth of trees were chosen so that it would be possible to search in the space of compact models (reduced number of genes and depth). Furthermore, the population size was made to vary from 500 to 900 individuals, and the number generations from 100 to 300. These values were chosen so that the total number of individuals searched during GP process would not surpass 150,000.

After implementing all experiments, each of them will have found a symbolic regression model for the FTE consumption. The model for the FTE is the one that presents the best forecasting accuracy when evaluated on the validation dataset.

3.4 Model Selection and Forecasting

In order to select the best symbolic regression model for the FTE, all models produced by the GP experiments are implemented to forecast 9 values that follow the training set, which are the target values contained in the validation set. The selected model will be the one whose forecast has lowest Mean Absolute Percentage Error (an accuracy measure that is described in Subsection 3.5) on the validation dataset.

[1] The 'Pareto' column in Table III contains the percentage of tournaments that were made to be Pareto tournaments.

Table III – GP parametrizations tested

ID	**Pop. Size**	**Generations**	**Max. Genes**	**Max. Depth**	**Pareto**	**Functions**
1	900	100	2	3	0,3	*, -, +, sqrt, square, add3, mult3
2	400	250	3	3	0,3	*, -, +, sqrt, square, add3, mult3, tanh, cube
3	400	250	3	3	0,2	*, -, +, sqrt, square, add3, mult3, tanh, cube
4	500	300	3	3	0,2	*, -, +, sqrt, square, add3, mult3, tanh, cube, log
5	500	300	3	4	0,2	*, -, +, sqrt, square, add3, mult3, tanh, cube, log
6	500	300	4	4	0,2	*, -, +, sqrt, square, add3, mult3, tanh, cube, log
7	500	300	4	4	0,3	*, -, +, sqrt, square, add3, mult3, tanh, cube, log

In order to produce the forecasts for the model selection phase, the inputs presented to the models will be observations of (i) number of clients, (ii) industrial production and (iii) industrial electricity tariff. The values of electricity consumption, however, will not be those observed, but those estimated by the models.

Let the training set contain N observations, and let T be the time at which the last value, i.e., the N-th value, was observed. The forecast τ-steps ahead of time T is then represented as $\hat{y}_T(\tau)$. The observed value at time $T+\tau$ is represented as $y_{T+\tau}$. Also, let $x_{1,T+\tau}$ be the observed value of clients at time $T+\tau$, $x_{2,T+\tau}$ be the industrial production at time $T+\tau$, and $x_{3,T+\tau}$ be the industrial electricity tariff at time $T+\tau$. The forecast one step ahead of time, two steps ahead of time, and 3 steps ahead of time are then obtained by functions as those presented in Equations (1), (2) and (3), respectively[2].

$$\hat{y}_T(1) = f\left(y_T, y_{T-1}, y_{T-2}, x_{1,T+1}, x_{2,T+1}, x_{3,T+1}, x_{1,T}, x_{2,T}, x_{3,T}, x_{1,T-1}, x_{2,T-1}, x_{3,T-1}\right) \tag{1}$$

$$\hat{y}_T(2) = f\left(\hat{y}_T(1), y_T, y_{T-1}, x_{1,T+2}, x_{2,T+2}, x_{3,T+2}, x_{1,T+1}, x_{2,T+1}, x_{3,T+1}, x_{1,T}, x_{2,T}, x_{3,T}\right) \tag{2}$$

$$\hat{y}_T(3) = f\left(\hat{y}_T(2), \hat{y}_T(1), y_T, x_{1,T+3}, x_{2,T+3}, x_{3,T+3}, x_{1,T+2}, x_{2,T+2}, x_{3,T+2}, x_{1,T+1}, x_{2,T+1}, x_{3,T+1}\right) \tag{3}$$

… …

When facing the task of out-of-sample forecasting, the observations of the explicative variables are not available. In this case, rather than providing observations as input to the multivariate model, estimated values (e.g., univariate forecasts) of the input variables are provided. Equations (1)-(3) are then rewritten as (4)-(6). Note that the values that were previously considered as observations at times $T+\tau$, with $\tau > 0$, are now estimated values.

$$\hat{y}_T(1) = f\left(y_T, y_{T-1}, y_{T-2}, \hat{x}_{1,T}(1), \hat{x}_{2,T}(1), \hat{x}_{3,T}(1), x_{1,T}, x_{2,T}, x_{3,T}, x_{1,T-1}, x_{2,T-1}, x_{3,T-1}\right) \tag{4}$$

$$\hat{y}_T(2) = f\left(\hat{y}_T(1), y_T, y_{T-1}, \hat{x}_{1,T}(2), \hat{x}_{2,T}(2), \hat{x}_{3,T}(2), \hat{x}_{1,T}(1), \hat{x}_{2,T}(1), \hat{x}_{3,T}(1), x_{1,T}, x_{2,T}, x_{3,T}\right) \tag{5}$$

$$\hat{y}_T(3) = f\left(\hat{y}_T(2), \hat{y}_T(1), y_T, \hat{x}_{1,T}(3), \hat{x}_{2,T}(3), \hat{x}_{3,T}(3), \hat{x}_{1,T}(2), \hat{x}_{2,T}(2), \hat{x}_{3,T}(2), \hat{x}_{1,T}(1), \hat{x}_{2,T}(1), \hat{x}_{3,T}(1)\right) \tag{6}$$

… …

To evaluate the performance of the selected model in out-of-sample forecasting situations, the model will be applied to forecast 9 of the target values contained in the test set.

[2] It is important to take note that, depending on the model that will be created by GP's evolutionary process, not all variables will appear in its mathematical expression. However, aiming to illustrate the forecasting approach, all variables are explicitly shown in Equations (1)-(3) (and, later, in Equations (4)-(6)).

The forecast will then be compared with forecasts obtained from multivariate models (MLR and ANN) and from univariate models (SARIMA and Exponential Smoothing).

3.5 Accuracy Measure

In order to evaluate the forecasts, the Mean Absolute Percentage Error (MAPE) is adopted. MAPE is recommended as accuracy measure for forecasts by several books, as pointed out in (Hyndman & Koehler, Another look at measures of forecast accuracy, 2006). Its use is not recommended when the observations of the time series have values near zero (Hyndman & Koehler, Another look at measures of forecast accuracy, 2006). As this is not the case in this work, MAPE will be considered as the accuracy measure. MAPE is detailed in Equation (7).

$$MAPE = \frac{1}{h}\sum_{i=1}^{h}\frac{|\hat{y}_T(i) - y_{T+i}|}{x_{T+i}}, \qquad (7)$$

Where $\hat{y}_T(\mathrm{i})$ is the forecasted value of the series for time $T + i$, y_{T+i} is the observed value of the series at time $T + i$, and h is the number of steps ahead of time forecasted.

4. Results

The genetic programming framework was implemented in MATLAB, using the GPTIPS 2 platform (Searson, 2015). Each GP experiment constructed a symbolic regression model for the FTE consumption. The MAPE measures of fit (in the training set) and forecast (in the validation set) for each of the models are presented in Table IV.

Table IV – MAPEs of the symbolic regression models in the training and validation datasets

Model ID	MAPE	
	Training	Validation
1	2.043%	2.975%
2	1.853%	3.027%
3	2.073%	3.253%
4	2.104%	3.513%
5	1.916%	4.011%
6	2.711%	3.208%
7	1.444%	1.561%

Considering the results, the model selected was the one named with ID 7, i.e., which presented the lowest MAPE in validation dataset and lowest ratio between validation MAPE and training MAPE. The model has the following mathematical expression:

$$\begin{aligned}\hat{y}_t = {} & 0.96 * y_{t-1} - 0.96 * x_{1,t-1} - 20.02 * x_{2,t-1} + 19.05 * x_{2,t-2} + 5.90 * x_{2,t-3} \\ & + 24.95\left(x_{2,t} - x_{2,t-1}\right)^2 + \sqrt{x_{1,t-3}} \\ & + 0.0009\left(x_{2,t} - x_{2,t-1}\right)^2\left(x_{2,t} + x_{2,t-1} + 4.054\right)\left(x_{2,t-1} + \sqrt{x_{3,t-3}} + 4.271\right) \\ & + 163.9\end{aligned} \qquad (8)$$

It is possible to see that the model provides relations that are not straightforward. For example, it considers that an increase in the number of clients of the FTE in a previous month will cause a decrease on FTE's electricity consumption in the following month.

After selecting the symbolic regression model, we verify if it performs better than other models considering a test dataset.

MLR and ANN models are designed as multivariate models, and SARIMA and Exponential Smoothing are univariate models. To select the input variables of the MLR and ANN models, the standard *backward elimination* procedure (Draper & Smith, 1981) was applied. The candidates for explanatory variables are the ones presented in Table I (i.e., the same candidates as those presented to the GP).

After applying the *backward elimination* procedure, six variables were selected as significant to the FTE consumption: consumption$_{t-1}$, consumption$_{t-3}$, industrial_production$_t$, industrial_production$_{t-1}$, industrial_production$_{t-2}$, industrial_production$_{t-3}$. These variables were considered, then, as input to the MLR and ANN models. MLR was estimated via ordinary least squares (OLS), and resulted in the model presented in Equation (9).

$$\hat{y}_t \cong 0.70 * y_{t-1} + 0.30 * y_{t-3} + 34.27 * x_t - 31.75 * x_{t-1} - 25.05 * x_{t-2} + 22.96 * x_{t-3} \qquad (9)$$

The ANN developed in this study was a fully-connected multilayer ANN, trained with ten different parametrizations. The selected ANN was the one that presented lowest variation between the training and validation MAPE[3]. The chosen ANN had 5 perceptrons in the input layers, 3 in the hidden layers, and 1 in the output layer; the perceptrons had logistic activation function. The parametrizations and results of all ANNs are shown in Table V. It is possible to note that most of the ANN configurations led to an overtraining of the series (very low MAPE in training, and considerably higher MAPE in validation). The configuration with logistic activation function, however, presented a reasonable result, with a ratio of 1.10 between validation MAPE and training MAPE. The training and forecasting of the ANN models were programmed and implemented in MATLAB.

Table V – Results of the ANN modeling

ANN ID	**Config.**	**Activation Function**	**Learning Rate**	**Interval Weight**	**Max Epochs**	**MAPE**	
						Training	**Validation**
1	[5 3 1]	tanh	0,05	[-0,01 0,01]	100.000	0,736%	3,011%
2	[5 3 1]	tanh	0,10	[-0,01 0,01]	100.000	0,569%	1,983%
3	[5 3 1]	tanh	0,15	[-0,01 0,01]	100.000	0,474%	1,559%
4	[5 3 1]	tanh	0,20	[-0,01 0,01]	100.000	0,600%	2,012%
5	[5 3 1]	tanh	0,15	[-0,01 0,01]	1.000.000	0,417%	2,664%
6	[5 3 1]	tanh	0,15	[-0,05 0,05]	1.000.000	0,475%	4,451%
7	[6 3 1]	tanh	0,15	[-0,01 0,01]	1.000.000	0,335%	3,595%
8	[5 3 1]	tanh	0,15	[-0,01 0,01]	10.000	1,123%	2,008%
9	[5 3 1]	tanh	0,10	[-0,01 0,01]	10.000	1,147%	1,935%
10	[5 3 1]	logistic	0,15	[-0,01 0,01]	10.000	2,110%	2,327%

The SARIMA and Exponential Smoothing (ES) models were estimated considering the training and validation sets altogether. Fit and forecast were made using FPW software. The resulting SARIMA for the FTE time series was a SARIMA(0,1,0)x(0,1,1)$_{12}$, with $\Theta_{12} = 0.8844$. The resulting ES model had no trend and multiplicative seasonality, with four seasonal indexes.

Table VI shows the MAPE of the fit (training set) and forecasts (validation and test sets) for all multivariate models implemented. Table VII shows the MAPE of the forecasts (test set) for the univariate models and for the GP model.

[3] This selection criterion was chosen based on the assumption that a low variation on training and validation errors would lead to a model that generalizes well.

Table VI – Comparison of MAPEs on training, validation and test sets (multivariate models vs GP model)

Model	MAPE		
	Training	Validation	Test
MLR	1.937%	2.947%	5.252%
ANN	2.110%	2.327%	4.595%
GP	1.444%	1.561%	2.238%

Table VII – Comparison of MAPEs on the test set (univariate models vs GP model)

Model	MAPE
	Test
SARIMA	3.650%
ES	3.427%
GP	2.238%

It is possible to note that the symbolic regression model, built via genetic programming, produces lowest MAPEs comparing to all other methodologies tested. The model's accuracy on the test set, specifically, is considerably higher than the other models. Even when training the univariate models with training and validation sets altogether, they do not produce better forecast than the GP model. The GP model is, therefore, a relative good model for the electricity consumption of the Brazilian FTE. It could, therefore, be used to forecast the FTE consumption, and, by extension, to infer the electricity wholesale market's development.

5. Conclusions

This paper used genetic programming to find a nonlinear multivariate model for the electricity consumption of the Brazilian Free Trading Environment (FTE). Several genetic programming experiments were made, in order to find a model for the FTE that could generalize well to unseen data (i.e., that could produce forecasts with MAPE that did not vary much from the training to the validation set). The genetic programming experiments searched for models with industrial production, industrial electricity tariffs, number of clients in the FTE and FTE's electricity consumption as explicative variables, all considering time lags.

The forecasts of the best genetic programming model have shown to be more accurate than the forecasts from other models, such as multiple linear regression, artificial neural networks, SARIMA, and exponential smoothing. The genetic programming model allows one to have an understanding about the relations between explicative variables and the variable of interest. Furthermore, the model allows one to produce what-if analysis, assessing in long-term decisions related to the FTE.

Future works include the optimization of the genetic programming framework, aiming to find a more accurate or representative model for the FTE consumption. It is known that the genetic programming finds a suboptimal solution for a problem. An optimal model for the problem could, therefore, be found by enhancing the genetic programming framework. Other future work would be to include other explanatory variables in the modeling framework.

References

Amber, K., Aslam, M., & Hussain, S. (2015). Electricity consumption forecasting models for administration buildings of the UK higher education sector. *Energy and Buildings*, *90*, pp. 127-136.

ANEEL. (n.d.). *Relatórios de Consumo e Receita de Distribuição*. Acesso em 20 de 12 de 2016, disponível em http://www.aneel.gov.br/relatorios-de-consumo-e-receita

Box, G., & Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control.* San Francisco, California, US: Holden-Day Inc.

Bruno, S., Ahmed, S., Shapiro, A., & Street, A. (2016). Risk neutral and risk averse approaches to multistage renewable investment planning under uncertainty. *European Journal of Operational Research, 250*(3), pp. 979-989.

CCEE. (2016). *Market Info - General Data [In Portuguese].*

Chatfield, C. (2000). *Time-series Forecasting.* Boca Raton, Florida, US: Chapman & Hall / CRC.

Chou, J.-S., & Ngo, N.-T. (2016). Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. *Applied Energy, 177*, pp. 751-770.

Draper, N., & Smith, H. (1981). *Applied Regression Analysis* (2ª ed.). John Wiley & Sons.

EPE. (n.d.). Monthly Electrical Energy Consumption per Class (Regions and Subsistems) - 2004-2016 [In Portuguese]. *Energy Market Analysis Bulletin*. Rio de Janeiro, Rio de Janeiro, Brazil. Fonte: http://www.epe.gov.br/

Forouzanfar, M., Doustmohammadi, A., Hasanzadeh, S., & Shakouri G, H. (2012). Transport energy demand forecast using multi-level genetic programming. *Applied Energy, 91*(1), pp. 496-503.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* Prentice Hall.

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting, 32*(3), pp. 914-938.

Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), pp. 679-688.

Hyndman, R., Koehler, A., Snyder, R., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting, 18*(3), pp. 439-454.

Kong, X., Liu, X., Shi, R., & Lee, K. (2015). Wind speed prediction using reduced support vector machines with feature selection. (456, Ed.) *Neurocomputing, 169*, p. 449.

Koza, J. (1992). *Genetic Programming.* The MIT Press.

Lee, D., Lee, B., & Chang, S. (1997). Genetic programming model for long-term forecasting of electric power demand. *Electric Power Systems Research, 40*(1), pp. 17-22.

Searson, D. (2015). GPTIPS2: An open-source software plaatform for symbolic data mining. In: A. e. Gandomi, *Handbook of Genetic Programming Applications.* New York, NY: Springer.

Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*, pp. 199-222.

Street, A., Lima, D., Veiga, A., Fânzeres, B., Freire, L., & Amaral, B. (2012). Fostering wind power penetration into the Brazilian forward-contract market. *2012 IEEE Power and Energy Society General Meeting.* San Diego.

Yang, G., Li, X., Wang, J., Lian, L., & Ma, T. (2015). Modeling oil production based on symbolic regression. *Energy Policy, 82*, pp. 48-61.