Universidade de Brasília

Faculdade de Economia, Administração e Contabilidade

Departamento de Economia

High Dimensional Models for Forecasting Power Electricity Consumption

Pedro Caiua Campelo Albuquerque

Brasília

2019

Pedro Caiua Campelo Albuquerque

High Dimensional Models for Forecasting Power Electricity Consumption

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade de Brasília como requisito à obtenção do título de Mestre em Ciências Econômicas.

Universidade de Brasília

Faculdade de Economia, Administração e Contabilidade

Departamento de Economia

Orientador: Prof. Marina Delmondes de Carvalho Rossi, PhD

Brasília

2019

Pedro Caiua Campelo Albuquerque

High Dimensional Models for Forecasting Power Electricity Consumption

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade de Brasília como requisito à obtenção do título de Mestre em Ciências Econômicas.

Trabalho aprovado. Brasília, 17 de Junho de 2019:

Prof. Marina Delmondes de Carvalho Rossi, PhD
Universidade de Brasília
Orientador

Prof. Daniel Oliveira Cajueiro, PhD
Universidade de Brasília

Prof. Ivan Marques de Toledo Camargo, PhD
Universidade de Brasília

Brasília
2019

Resumo

Este trabalho usa técnicas de aprendizado de máquina para prever o consumo de energia elétrica (CEE) do Brasil no curto e médio prazo. Os modelos são comparados com modelos referenciais, como o *Random Walk* e a *ARIMA*. Nossos resultados mostram que os métodos de aprendizado de máquina, especialmente *Random Forest* e *Lasso Lars*, têm precisão superior para todos os horizontes de previsão, removendo o sobreajuste presente nos modelos tradicionais. *Random Forest* e *Lasso Lars* conseguiram acompanhar a tendência e a sazonalidade nos diferentes horizontes temporais. Ainda, o ganho em prever CEE utilizando modelos de aprendizado de máquina em relação aos tradicionais é muito maior no curtíssimo prazo. A seleção de variáveis dos modelos de aprendizado de máquina mostra ainda que os valores defasados de CEE são extremamente importantes para a previsão de curtíssimo prazo, devido à sua alta autocorrelação. As demais variáveis são importantes para horizontes temporais mais longos.

Palavras-chave: Energia elétrica, Atividade econômica, Brasil, Economia emergente, Previsão, Random Forest, Lasso, Lars, Seleçao de modelo.

Abstract

This work uses machine learning techniques to predict Brazilian power electricity consumption (PEC) for short and medium term. The models are compared to benchmark specifications such as Random Walk and autoregressive integrated moving average (ARIMA). Our results show that machine learning methods, especially Random Forest and Lasso Lars, have superior accuracy for all forecasts horizons by removing the overfitting present in traditional models. Random Forest and Lasso Lars managed to keep up with the trend and the seasonality in different time horizons. The gain in predicting PEC using machine learning models compared to traditional ones is much higher in very-short term. Machine learning variable selection further shows that lagged consumption values are extremely important for very short-term forecasting due to its high autocorrelation. Other variables are important for longer time horizons.

Key-words: Power electricity, Economic activity, Brazil, Emerging economies, Forecasting, Random Forest, Lasso, Lars, Model selection.

List of Figures

List of Tables

List of abbreviations and acronyms

ANEEL	Electric Energy National Agency
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BNDES	Brazilian Bank for Economic and Social Development
GDP	Gross Domestic Product
INMET	Brazilian Institute of Metereology
LASSO	Least Absolute Shrinkage and Selection Operator
LARS	Least Angle Regression
MTFG	Medium-Term Forecast Group
MSE	Mean Square Error
OLS	Ordinary Least Squares
ONS	National Electricity System Operator
PEC	Power Electricity Consumption
RMSE	Root Mean Square Error
SRFG	Short-Term Forecast Group
VSTFG	Vert Short-Term Forecast Group

Contents

1 Introduction

Forecasting economic activity is essential to the development of the economy. Accurate forecasts allows the government to better organize its budget and also help economic agents to set their expectations. We usually follow GDP as the main representation of economic activity, but it is released only quarterly and usually these data are adjusted afterwards. Thus, it is interesting to find some new way to forecast short-term economic activity. This helps to surpass the problems of the usual measurements, such as a considerable lapse of time between the analyzed intervals.

Power electricity consumption (PEC) is released hourly and it has a great causal relation with economic activity (MAZA; VILLAVERDE, 2007). Developed economy requires a large amount of electricity to satisfy the needs of industries, families, farmers, and government. In this sense, satellite data on lights at night are a strong and useful indicator of economic activity (HENDERSON et al., 2012; BUNDERVOET et al., 2015; NASA, 2000). The Federal Reserve Board's monthly index of industrial production (until 2005) is partially based on a survey that measures delivered electricity.

Moreover, forecasting PEC is crucial to the planning of electricity industry and the operation of electric power systems. Accurate forecasts maintain the balance between power electricity supply and demand since we cannot store electrical energy. Consequently, these forecasts play an important role in future decisions on energy management and for saving in operation and maintenance costs.

The purpose of this paper is to make accurate forecasts of Brazilian PEC for 1 day and up to 3 months ahead. We use traditional econometric models as well as machine learning techniques in order to get an accurate forecast. Our results show that machine learning models perform much better than the traditional ones. Also, we have bigger gain in forecasting very short term with machine learning models relative to regular ones. In a 1-day forecast, our best model has a forecast error that is 40 times smaller than traditional ones and about 20 times for the other horizons.

PEC is a highly seasonal and cyclical variable. Its consumption level varies with

the day of the week or season of the year, and is greatly correlated with its past values. Hence, it is commom to forecast short-term PEC by using just lagged variables via ARIMA model (ALMESHAIEI; SOLTAN, 2011; HUANG; SHIH, 2003; RODRIGUES et al., 2014). Calendar variables is also an important set of variables that captures the serie's seasonality. Lebotsa et al. (2018) and Fan and Hyndman (2011) include calendar variables in order to catch the complex nonlinear relationship between electricity demand and its drivers.

Weather variables are relevant to predict a variety of economic variables. Dell et al. (2014) exhibit that shocks in temperature affect many economic outcomes, such as economic growth, industrial output, and energy demand. The authors also note that poor countries, such as Brazil, appear to be much more sensitive to these shocks for many outcomes. Including weather variables manages to control PEC highly seasonality in shor-term forecasts (LEBOTSA et al., 2018; FAN; HYNDMAN, 2011).

Hydroelectric plants are responsible for about 90% of all electricity consumed in Brazil (ANEEL, 2002). This type of energy is highly seasonal and influenced by weather factors (ADEGBEHIN et al., 2016). River flow, rain incidence, and temperature increase directly affect Brazil's energy production. This corroborates the inclusion of weather variables for PEC forecasts.

Electric energy price variables and other economic variables capture the whole economic environment. These sets of variables are able to follow the trend of PEC, given the positive relationship between them. El-Shazly (2013) include economic variables to build a dynamic econometric model to forecast PEC and produce reliable ex-post forecasts.

We consider an approach based on a high dimensional data. We build a database with more than 1500 variables, including lagged demand, calendar variables, weather variables, electric energy price variables, and other economic variables. We estimate our models for 6 different forecasts horizons using a rolling window of 546 observations.

The outline of the paper is as follows. We present our model in Section 2 describing the structure and all models used to forecast power electricity consumption. In Section 3 we detail our dataset and analyze all variable sets used. We then show our general

results and examine the best models in Section 4. Also, we observe which variable are more important to predict PEC for each forecast horizon. Lastly, in Section 5, we present a brief conclusion concerning our findings.

2 Data

We use Brazilian data from various sources. The initial date from our dataset is February 1, 2017, when PEC hourly data starts to be disclosed, and ends on July 31, 2018. Although PEC data are released hourly, most other explanatory variables data are released only daily or monthly. Therefore, we use data at a daily frequency. Our sample consists of 546 days (one and a half year). We use one year for the training set and six months for the holdout set.

The first set of variables that we include are calendar variables. This set includes day of the week, day of the month, month, season of the year, year and a dummy for holidays, totaling 6 calendar variables.

2.1 Power Electricity Consumption

We use the hourly energy load by subsystem (MW/h) for Brazil, which is released by the National Electricity System Operator (ONS) to measure Power Electricity Consumption (PEC). Figure 1 shows the behavior, trend, seasonality, and residue of the total daily electricity consumption in Brazil. We can see that PEC is a highly seasonal variable.

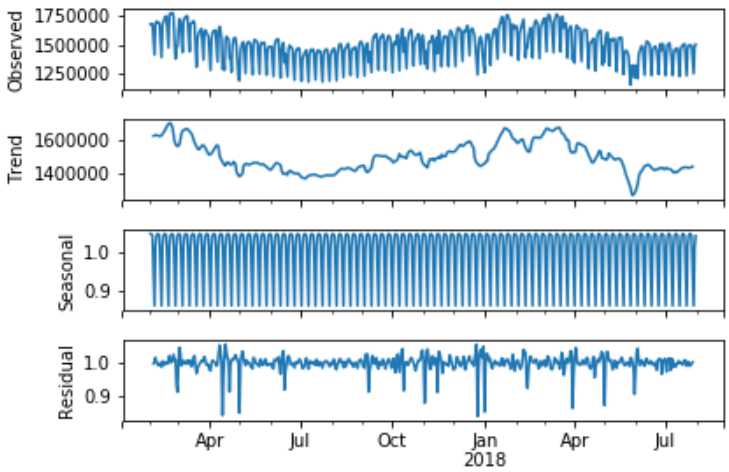


Figure 1 – Seasonal Plot

We can see in Figure 1 the effect on PEC of the Brazil's truckers strike. At the end of May 2018, truck drivers blocked roads across Brazil demanding a reduction in diesel

prices that had risen more than 50% in the last 12 months. With trucks stalled, partially blocking roads, fuels were no longer delivered to several gas stations and to other activities that expected raw materials and essential products, such as food, also ran out of supplies. We note that PEC level fell by about 30% during the Brazil's truckers strike. This relation between the two variables reinforces the great causal relation among economic activity and electric energy.

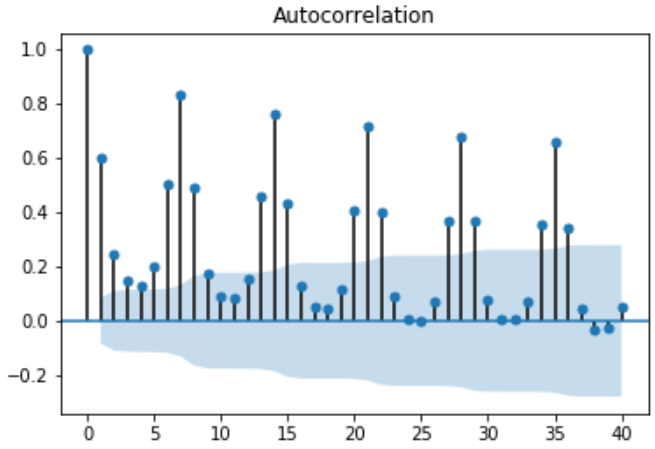


Figure 2 – Autocorrelation Plot

Figure 2 shows how PEC is correlated with its lagged values. We observe a great correlation with its value from the previous day. However, the correlation is even greater with the consumption of a week ago, corroborating the cyclical behavior of the PEC present in Figure 1. This figure indicates the importance of including lagged demand in forecast models.

We select the hourly energy load by subsystem (KW/h) for Brazil and by Region in the ONS website [1]. However, we also take PEC daily data separated by consumption class for each region from Electric Energy National Agency (ANEEL) website [2]. Consumption classes set includes PEC variables divided into groups such as industrial, commercial and services, own consumption, residential, rural, street lighting, public service and others [3].

This division by consumption class not only improves the performance of the

[1] <http://sdro.ons.org.br/SDRO/DIARIO/index.htm>
[2] <http://www.aneel.gov.br>
[3] We list all consumption classes in the appendix.

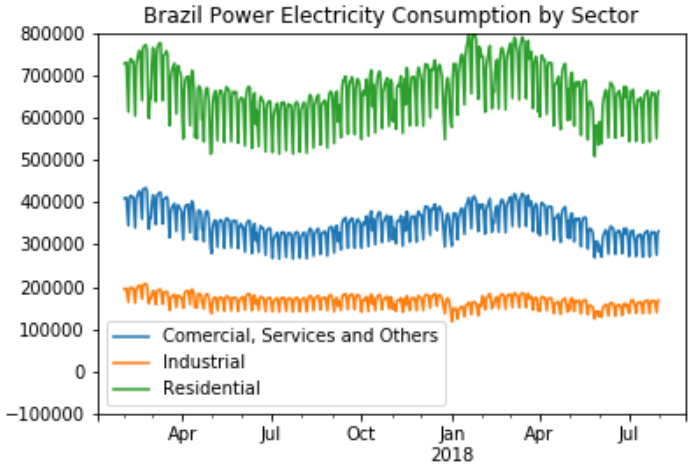


Figure 3 – PEC by sector

model's predictions, but is also crucial to understand how the consumption of electricity is distributed across several economic activities or classes. Most of the electricity consumption is concentraded in commercial and services (23%), industrial (12%), and residential consumption (44%). The first two are directly linked to country's economic activity, while the third one is indirectly related. Figure 3 shows these variables and we can see that the sum of these variables is close to 80% of all PEC in Brazil. This fact reinforces the link between of economic activity and PEC.

We have a total of 185 PEC variables after taking into account PEC's division by hour and by consumption class for each region.

2.2 Weather Variables

We collect this set of variable from the Brazilian Institute of Meteorology (INMET) website [4] that releases the historical series issued by stations scattered throughout Brazil. This set of variables contains air nebulosity, atmospheric pressure (mbar), dry bulb temperature (°C), humidity bulb temperature (°C), relative humidity (%), wind direction and wind speed (m/s) disclosed for 9 a.m., 3 p.m., and 9 p.m. for each station.

The most important weather variable for forecasting PEC is the the dry bulb temperature, which is also a very seasonal variable with a large variance. Inspired by

[4] <http://www.inmet.gov.br>

Boldin and Wright (2015), we adjust the temperature by subtracting the average of the entire training set, if the day belongs to the training set; and over the entire holdout set, if the day belongs to the holdout sample. Adjusted-temperature controls both for the seasonal effect and for the variance of temperature. This methodology is aligned with the World Meteorological Organization guidelines for the calculation of climate normals (ORGANIZATION, 2017).

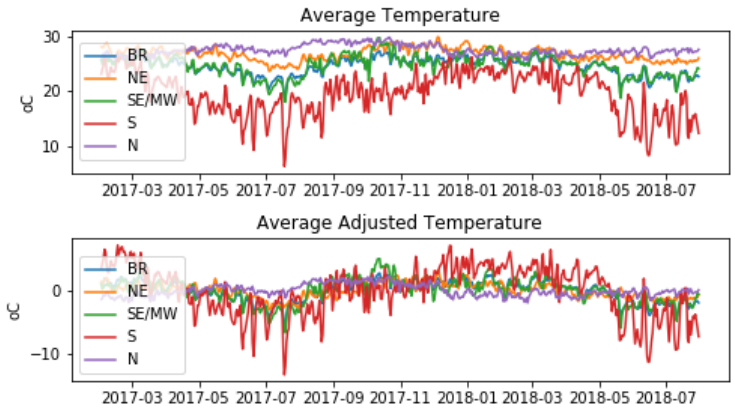


Figure 4 – Average Temperature

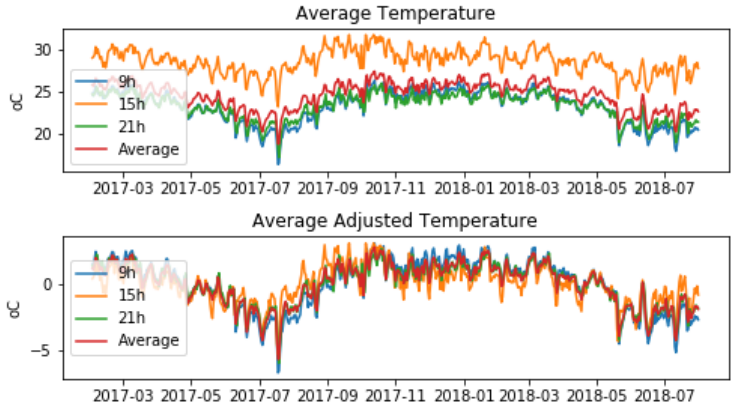


Figure 5 – Adjusted Average Temperature

Figures 4 and 5 show how the adjusted temperature loses part of its variance and seasonality. When comparing Figures 1 and 5 we observe that PEC and temperature have positive correlation. This is a further indication that weather variables are potentially important in predicting PEC.

We have a total of 180 weather variables for all regions and all hours analyzed by INMET after aggregate the data by region and including the adjusted temperature.

2.3 Electric Energy Price

We retrieve the price dataset from Electric Energy National Agency (ANEEL) website [5] and divide by consumption class for each region. However, price of electricity is a monthly variable and as our variable of interest has daily frequency, we set the same price for all days in a month. Figure 6 shows price of power electricity for Brazil and its regions.

After aggregating all variables by region and by consumption class, we have a total of 60 price variables.

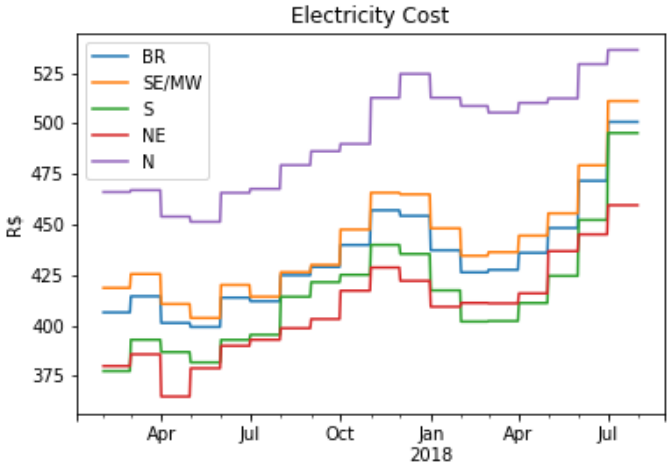


Figure 6 – Average Price by Region

2.4 Economic Variables

We use a database from Brazilian Central Bank (BCB) website [6] that includes several indices of price, traffic of vehicles, unemployment, wage, industrial production variables, monetary variables, consultations with credit bureaus, Brazilian Bank for Economic and Social Development (BNDES) disbursements and several others. We include a wide list of economic variables, some at a daily and others at a monthly frequency. For those at a monthly frequency, we also set the same value for all days within a month.

[5] <http://www.aneel.gov.br>
[6] <https://www.bcb.gov.br/estatisticas/indicadoresconsolidados>

We have a total of 79 economic variables. After merging all sets of explanatory variables listed in this section, we have a total of 510 variables [7].

[7] We list all sets of variable in the appendix.

3 Econometric Models

3.1 Estimation

Since PEC is a highly seasonal variable, we consider that its forecast h periods ahead is a function of three blocks of predictors. Each block of predictors is the merge of all our explanatory variables in a different time period, which depends on our forecast horizon h. The first block uses the data from the last available date, the second one uses data from the past h days, and the third block uses data from the past $2 * h$ days. After aggregating all predictor's blocks we have a total of 1530 candidate variables for 546 observations.

The lags structure means that a forecast for $t + h$ periods ahead uses variables from t, $t - h$ and $t - 2h$. That is, in a 7 days ahead forecast, we use data from today and from past 7 and 14 days. We test several other lags structures and this one presents the smallest forecast error. We see from Figure 2 that PEC has a large autocorrelation with different past weeks and this structure builds a database that captures information from past data. With a nearby lag structure [1], our database loses information that better capture PEC's seasonality. The lag structure is the first hyper-parameter that we define.

We take the period from February 2017 to Abril 2018 as the training set. The holdout set is from May 2018 to July 2018. We split the training set in two in order to choose all hyper-parameters of the models for each forecast horizon. We run several hyper-parameters values for each model, such as λ from Lasso, which we present below. We select the parameters with the lowest root mean square error (RMSE) for each forecast horizon.

After we select the hyper-parameters, we run the models for the entire training set and obtain the model's parameters ($\boldsymbol{B} = \left[\gamma, \beta\right]$). Next, we use these parameters in our holdout set and we obtain the predictions for the models and forecasts horizons. We use the RMSE to evaluate our models. Thus, the model that presents the lowest RMSE is

[1] Example of a nearby lag structure: forecast $t + h$ use variables from t, $t - 1$ and $t - 2$. A 7 day ahead forecast use data from today, yesterday and the day before yesterday.

considered the best one because it is closer to the real value of PEC.

We make predictions for 6 different forecast horizons: 1, 7, 15, 30, 60 and 90 days ahead, divided into three groups: (1) very short-term forecast group (VSTFG) that includes 1 and 7 days, (2) short-term forecast group (STFG) containing 15 and 30 days and (3) medium-term forecast group (MTFG) including 60 and 90 days. Our estimated equation is given by:

$$y_{t+h} = \alpha_0 + \sum_{i=0}^{2} \gamma_i y_{t-(h*i)} + \sum_{i=0}^{2} \beta_i X_{t-(h*i)} + u_{t+h},$$

in which y_t is the power electricity consumption (MW/h) in Brazil at time t, α_0 is a constant term, X_t is a matrix containing all candidate variables, $\boldsymbol{B} = \Big[\gamma, \beta\Big]$ is a vector with all the linear parameters, and u_t is an error term.

3.2 Models

Our benchmarks models, ARIMA and Random Walk, use only lagged PEC. We build a high-dimensional dataset with all candidate variables. We then use machine learning models to predict PEC. We expect these models to perform better than traditional ones when working with big data. All models are explained below.

3.2.1 ARIMA

Autoregressive integrated moving average (ARIMA) uses just lagged variable of interest. The ARIMA (p,i,q) equation can be represented by:

$$y_t^* = \gamma + \Phi_0 + \Phi_1 y_{t-1}^* + \cdots + \Phi_p y_{t-p}^* + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q},$$

in which y_t is the PEC at time t, $y_t^* = y_t - y_{t-1}$ and u_t is an error term. We use Augmented Dickey–Fuller test and find that our series has an unit root and therefore we use its first difference in the ARIMA.

3.2.2 Random Walk

Random Walk (without drift) is a process in which the variable current value is constituted by its past value plus an error term. That is, the best value for forecasting k periods ahead is the value available today.

Random Walk is commonly used to trade exchange rates in short-term. Random Walk without drift can outperform the Random Walk with drift in terms of the RMSE (MOOSA; BURNS, 2016). The model can be represented by:

$$y_t = y_{t-k} + u_t,$$

in which y_t is PEC at time t and u_t is an error term. This model is a good measure of sensitivity in a autocorrelated series, since the last available value is probably correlated with the forecast. Thus, Random Walk, is commonly used as a benchmark for forecasts.

3.2.3 Lasso

Shrinkage methods present a prosperous alternative to traditional models when working with high dimensional data (GARCIA et al., 2017). The main idea of these methods is to shrink irrelevant variables to zero.

The least absolute shrinkage and selection operator (Lasso) minimizes the mean squared error (MSE), just as OLS. However, Lasso has a shrinkage coefficient, λ, that forces irrelevant variables to zero. The parameters are determined by:

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right].$$

Lasso can be seen as a generical OLS model because if $\lambda = 0$ the parameters, then $\hat{\beta}$ and $\hat{\beta}_{OLS}$ are the same. On the other hand, if the shrinkage coefficient is equal to some $\lambda_c > 0$ large enough, then all the variables are considered irrelevant and all β_i are equal to zero. Generically we have, $\lambda_c \geq \lambda \geq \lambda_{OLS} = 0$.

3.2.4 Lars

Least angle regression (Lars) is also a good model to use when working with high dimensional data because it provides a way of producing an evaluation of which variables to include. The main idea of his algorithm is to start with all values of β equal to zero and to increase all the coefficient associated with the x_i that is most correlated with y. In other words, the algorithm increases the parameters values in an equiangular direction to each one's correlations with the residual.

Lars algorithm and its variations work gracefully for the case in which there are many more variables than observations. Lars is easy to modify to produce efficient algorithms, like Lasso Lars, and is useful in cross-validation or similar attempts (EFRON et al., 2004).

3.2.5 Lasso Lars

A simple modification in Lars algorithm can result in the simulation of all Lasso models for all possible λ values (EFRON et al., 2004). Thus, in the traditional Lars method, if a coefficient changes the signal, the direction remains the same. However, in Lasso model, when a coefficient reaches the value 0, it is discarded from the active variable set. For that reason, Lasso Lars algorithm makes a simple modification in the original Lars. That is, if any coefficient becomes zero this variable is discarded and the model recalculates the search direction.

3.2.6 Ridge

Ridge Regression, like Lasso, is a generic version of OLS that also has a shrinkage coefficient λ. However, the difference is that the Ridge's restriction is squared and its parameters are determined by:

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right].$$

By adding a squared constraint, the loss function becomes strictly convex, and therefore has a unique minimum. Consequently, this model also reduces overfitting presented in traditional models. However, Ridge algorithm makes it harder to zero the coefficients and can not completely eliminate some irrelevant variables.

3.2.7 Elastic Net

Elastic Net is also a regularized model regression that combines the restrics in Lasso and Ridge Regression. The parameters are given by:

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right].$$

Elastic Net becomes a general case of Lasso and Ridge Regression. Therefore, as in the Ridge model, the Elastic Net method makes the loss function strictly convex, forcing it to have a unique minimum.

3.2.8 Random Forest

So far, all machine learning models present a method to shrink less relevant variables. However, Random Forest is a machine learning method thats combine a tree of predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (BREIMAN, 2001). Briefly, this model grows a forest of trees and let them vote for the most popular class (if in a classification problem) or averaging the forecasts (if in a regression problem).

Random Forest seeks to reduce the overfitting of traditional models when making a bagging (or bootstrap), combining learning models. Moreover, this machine learning algorithm can also model arbitrarily complex relations between inputs and outputs and intrinsically implement feature selection. Medeiros and Freitas (2016) argue that Random Forest has a great ability to select variables and a powerful potential to identify nonlinearities between macroeconomic variables.

4 Results

We list our best models for each forecast horizon by analyzing their root mean square error (RMSE). We normalize all variables before running the models in order to make the comparison between them easier.

Table 1 – Models results

Model Rank	1 dia (RMSE)	7 dias (RMSE)	15 dias (RMSE)	30 dias (RMSE)	60 dias (RMSE)	90 dias (RMSE)
1st	Random Forest (35753.39)	Random Forest (66621.08)	Lasso Lars (75882.53)	Lasso (72545.07)	Elastic Net (80127.27)	Lasso Lars (84758.67)
2nd	Lasso Lars (48300.59)	Lars (68400.7)	Random Forest (79375.37)	Lasso Lars (74151.69)	Ridge (80876.71)	Lars (85059.85)
3rd	Lars (50234.63)	Ridge (68846.63)	Lars (86543.14)	Ridge (79611.33)	Lars (81952.96)	Random Forest (86887.45)
4th	Lasso (61572.13)	Lasso Lars (68898.78)	Ridge (93536.83)	Lars (108984.56)	Lasso Lars (91794.13)	Ridge (92784.78)

This table shows the models ranking by the forecast error. The number between parentheses is the RMSE.

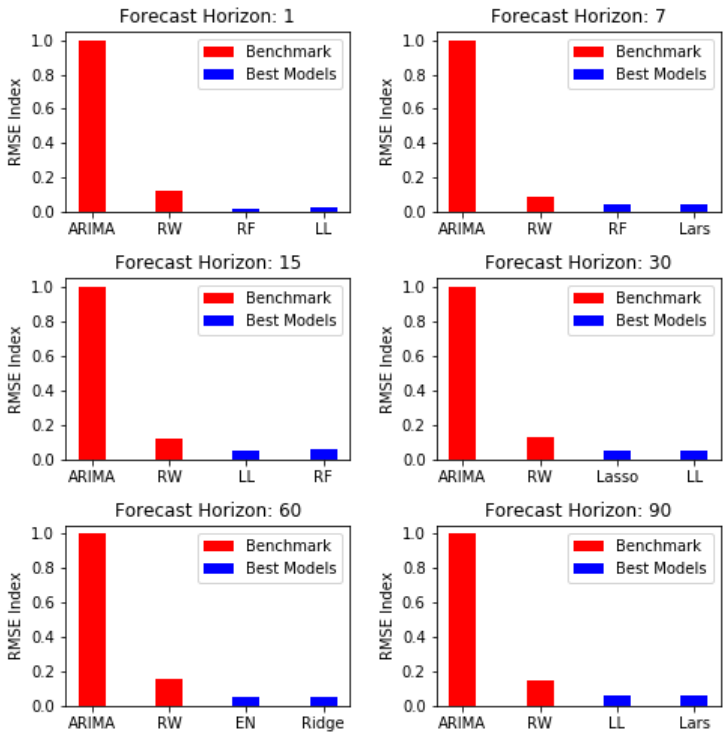
Table 1 exhibits the four models that present the smallest RMSE for each forecast horizon. These results corroborate what we delineate in section 2: machine learning models present much better performance than the traditional ones. We also analyze the gain in predicting PEC using machine learning models instead of our benchmarks models. We normalize the error by dividing the RMSE of the best models and of the Random Walk by ARIMA's RMSE for each time horizon. Figure 7 shows this gain.

The relative gain in predicting PEC using machine learning models is larger for short term predictions. The best model has an error that is about 40 times lower than the error of the ARIMA for 1-day forecast and that is about 20 times smaller for the subsequent forecasts horizons. When compared to the Random Walk, our best model has a forecast error that is about 5 times lower for the 1-day forecast and about 3 times lower for the other forecast horizons.

Even when using almost the same variables to predict short-term PEC [1] as benchmark models, machine learning models have a much better performance. Benchmarks models end up being overfitted and present poor results for all time horizons analyzed. Thus, our four best prediction models are always the machine learning ones that better

[1] Benchmark models only use lagged variables to predict and these variables are the main ones for predicting short-term PEC with machine learning models.

select the relevant variables.



RMSE divided by ARIMA's RMSE.

Figure 7 – Relative Gain

In Table 2, we classify the five best models according to the results present in Table 1. For VSTFG, our best prediction model is Random Forest and for STFG and MTFG the best model is Lasso Lars. Both models present the lowest RMSE in most of the forecast horizons analyzed for each group.

Table 2 – Models rank

Rank	Model		
	Very Short-Term	Short-Term	Medium-Term
1st	Random Forest	Lasso Lars	Lasso Lars
2nd	Lars	Lasso	Lars
3rd	Lasso Lars	Random Forest	Random Forest
4th	Ridge	Lars	Elastic Net
5th	Lasso	Ridge	Ridge

This table shows the five best models for each forecast horizon group.

In order to make a better analysis of the most important variables for each model, we sum of the absolute values of the coefficients for each set of variables described in the previous section. Table 3 shows this sum for the four best models in each forecast horizon.

Table 3 – Paremeters analysis

VSTFG

Paremeters	1 Day				7 Days			
	Random Forest	Lasso Lars	Lars	Lasso	Random Forest	Lars	Ridge	Lasso Lars
$\sum$| PED Variables|	0.73	511.94	603.4	529.27	0.89	401.59	475.74	334.08
$\sum$|Calendar Variables|	0.21	55599.06	54616.96	122161.95	0	120141.53	35985.1	123180.74
$\sum$| Weather Variables|	0.04	63548.57	69568.87	66730.31	0.08	97435.28	81025.62	104999.74
$\sum$|Price Variables|	0	18906.59	23374.71	77237.62	0.01	79872.65	43863.99	69113.43
$\sum$|Economic Variables|	0.02	5146.7	6630.49	5617.05	0.02	12243.93	27198.81	13236

STFG

Paremeters	15 Days				30 Days			
	Lasso Lars	Random Forest	Lars	Ridge	Lasso	Lasso Lars	Ridge	Lars
$\sum$|PED Variables|	478.88	0.42	542.32	1147.75	557.29	656.16	1034.1	1223.31
$\sum$|Calendar Variables|	103156.78	0.31	110101.43	89350.27	202878.84	182044.05	93669.63	114530.49
$\sum$|Weather Variables|	116181.17	0.13	143406.53	257546.68	220317.28	178016.79	292604.5	152194.85
$\sum$|Price Variables|	64241.49	0.04	75458.11	137440.58	155713.15	165620.2	165943.6	195263.74
$\sum$|Economic Variables|	22414.62	0.1	27228.64	80774.07	10277.77	7590.1	61500.69	23055.89

MTFG

Paremeters	60 Days				90 Days			
	Elastic Net	Ridge	Lars	Lasso Lars	Lasso Lars	Lars	Random Forest	Ridge
$\sum$|PED Variables|	729.22	689.71	145.98	204.25	456.33	214.95	0.24	813.52
$\sum$|Calendar Variables|	39517.78	36029.92	42947.04	82805.39	42786.92	42731.1	0.49	7113.61
$\sum$|Weather Variables|	131628.7	122329.32	51081.44	79428.17	79564.16	58642.19	0.2	114541.38
$\sum$|Price Variables|	115291.06	106949.99	32461.2	131733.53	107544.62	114738.55	0.05	102818.97
$\sum$|Economic Variables|	46396.04	43915.52	55389.14	50581.4	45921.62	43899.67	0.03	57197.6

Sum of the absolute values of the coefficients for each set of variables and forecast horizon.

Comparing the sum among them does not give us much since each model behaves differently in the valuation of their parameters. Also, each set of variables has a different size and the comparison between them also makes no sense. Therefore, we compare the sum of the parameters for the same model and same variable set between different forecast horizons. This comparison is only for our main models for each forecast horizon group. In other words, we basically analyze how the sum of the coefficients varies for each variable set between each time horizon and thus, infer the importance of this variables set for each horizon.

We note in Figure 2 that PEC is highly correlated with it previous day value. Consequently, all major models attach great importance to lagged values in a 1-day forecast. Other variables have almost no relevance for the prediction in this time horizon. Lasso model, for instance, have all their variables coefficients equal to zero, except for the aggregate PEC at 11 p.m. of the previous day. Random Forest assigns almost no value to all 60 coefficients of electricity price set.

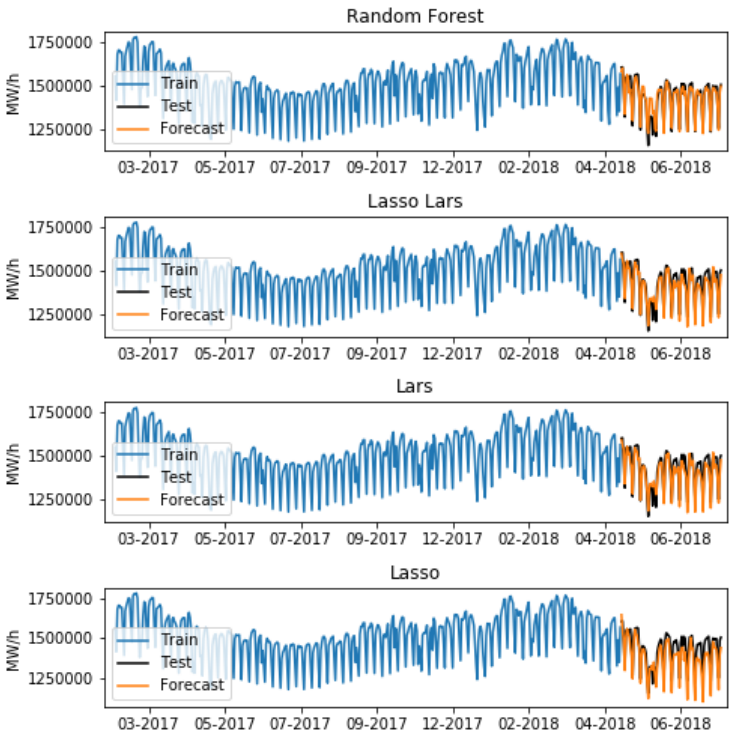


Figure 8 – 1 day forecast

Figure 8 shows that all four best forecasts are able to capture the direction and seasonality of electric power consumption with accuracy.

PEC's autocorrelation of 7 days ago is higher than the previous day autocorrelation, as we note in Figure 2. Table 3 shows that the weight attributed to lagged demand is larger for a 7-day forecast than for 1-day forecast. As lagged values increase their importance in this forecast horizon, other sets of variables reduce their relevance. That is, other explanatory variables are still not very important to make forecasts for the very short-term.

Figure 9 shows that the predictions of the main models are able to follow PEC trend and cyclicity. Additionally, even with a high autocorrelation, our main models perform better than the Random Walk forecasts.

The set containing lagged values of PEC loses its relevance for short and medium term forecast.The correlation with the past demand values are still present for a 15-day forecast, however weak. When comparing with the 7-day forecast, we note that the sum

Power Electricity Consumption 7 Days Forecast

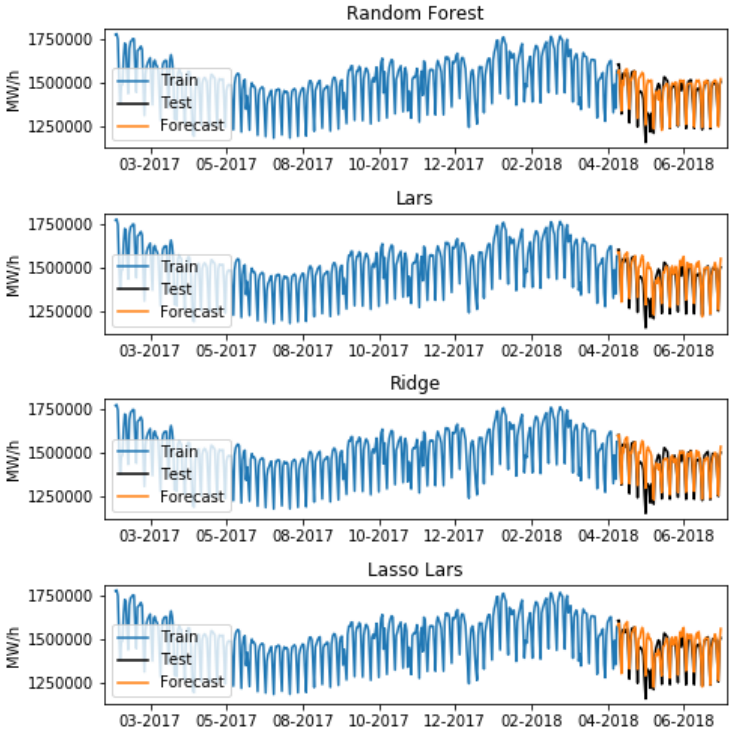


Figure 9 – 7 days forecast

of the parameters associated with the lagged demand decreases in our main models. In Random Forest, this set declines by more than a half.

Consequently, for the following horizons all other variables significantly expand their importance in the major models. The results are intuitive since PEC is highly autocorrelated with its initial lags. As the forecast horizon increases, other variables gain more importance, better capturing the series seasonality and the whole macroeconomic environment.

Figure 10 again shows that the best forecasts are able to follow the serie's trend and cyclicity, with a small difficulty to find the local maximums and minimums.

Moreover, we still have accurate results in a 30-days forecast. Figure 11 shows that the best forecasts can follow the series trend and cyclicity. We note that the predictions made using Lasso model are able to follow the series seasonality even when forecasting a month ahead.

When forecasting 60-day ahead, we see an increse in sum of the coefficients associ-

Power Electricity Consumption 15 Days Forecast

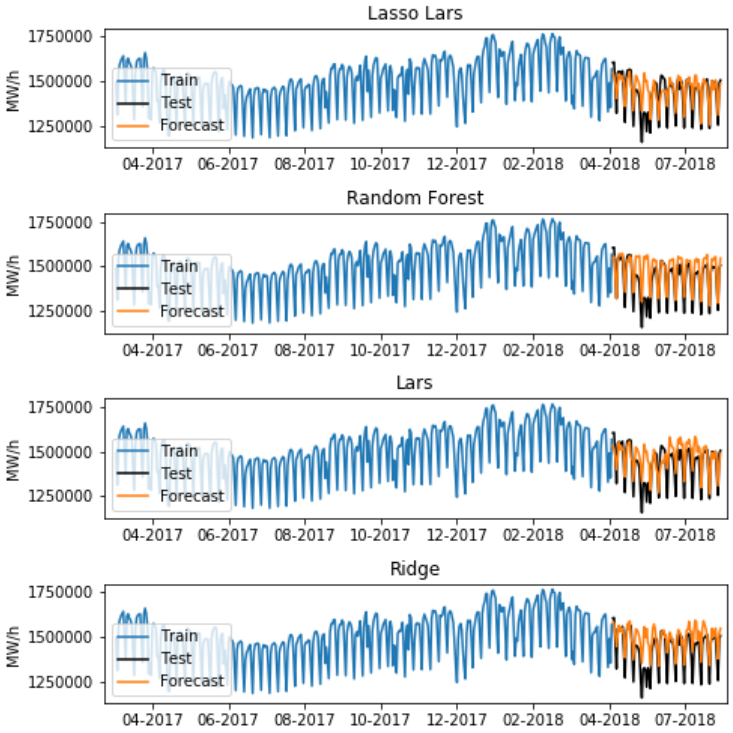


Figure 10 – 15 days forecast

Power Eletricity Consumption 30 Days Forecast

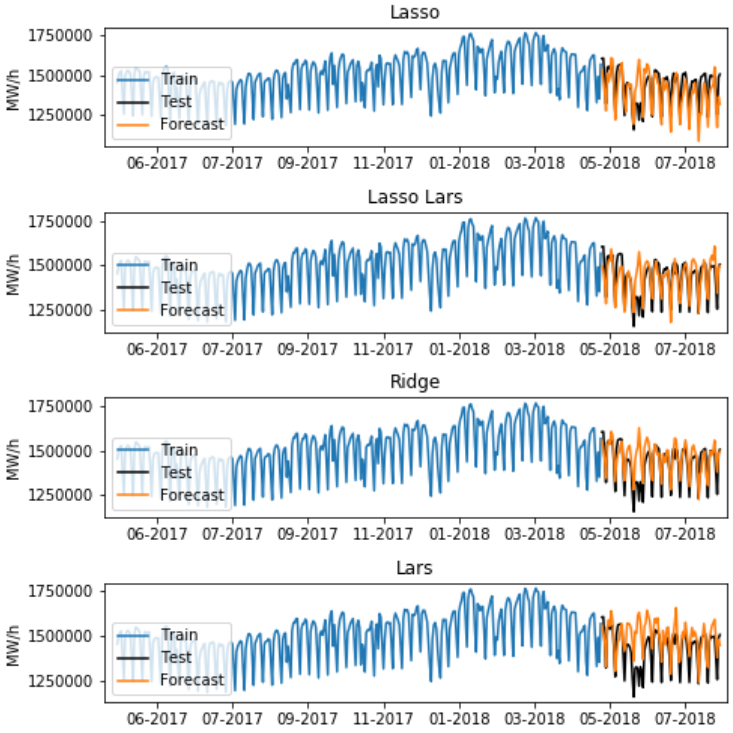


Figure 11 – 30 days forecast

ated with price variables. This is intuitive, since economic agents tend to be more sensitive to price changes for longer horizons, as they have more time to adjust behavior.

Power Eletricity Consumption 60 Days Forecast

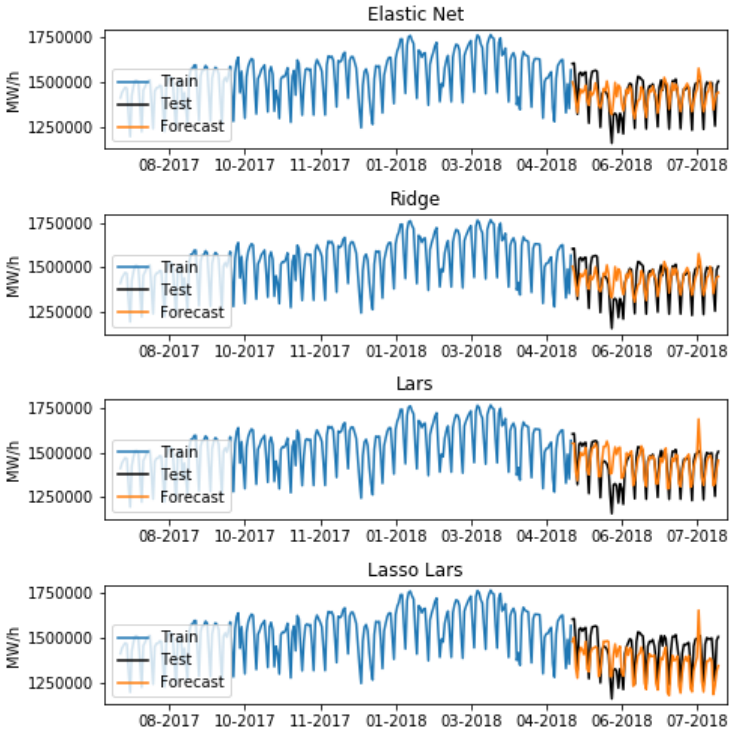


Figure 12 – 60 days forecast

Our predictions for short horizons are very close to the original values of PEC. At this stage, our forecasts are able to catch the direction and cyclicity of the original series, as we can note in Figure 12. That is, as the forecast horizon increases our predictions are less successful in reaching the local maximums and minimums of the series.

When forecasting PEC 90-days ahead we have a smaller accuracy but we still have good results. As we note in Figure 13, the best models have a greater difficulty to follow the series cyclicity, but still, they are able to follow its trend.

Power Eletricity Consumption 90 Days Forecast

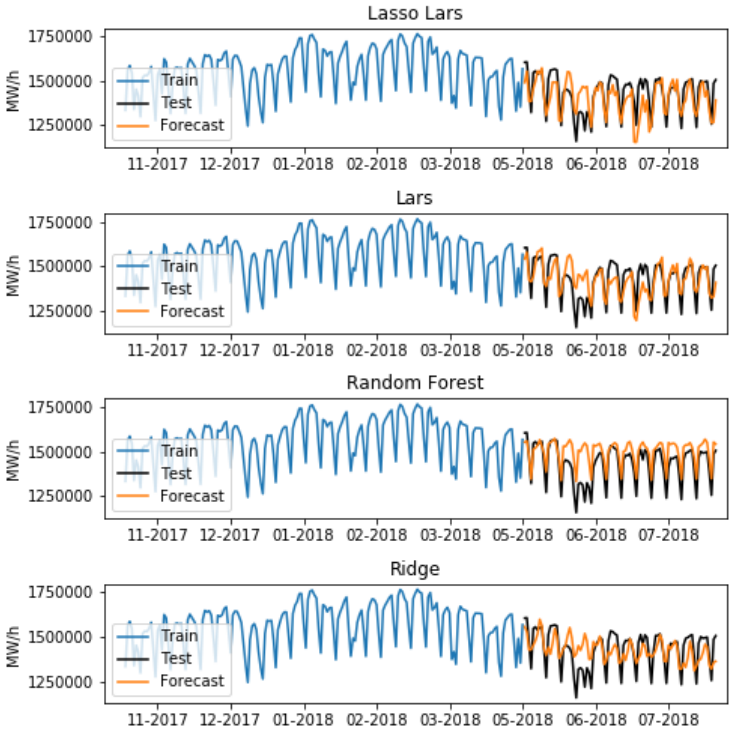


Figure 13 – 90 days forecast

5 Conclusion

This work uses high dimensional data to forecast PEC. We show that machine learning models perform better than traditional ones when forecasting using this kind of database. We forecast PEC for 6 different horizons, which we divide into three groups: very short term, short term, and medium term. The variables in our database can be classified into five sets: lagged demand, calendar variables, weather variables, electric energy price variables, and economic variables. We make predictions using more than ten different models for each forecast horizon, among which are ARIMA, Random Walk, and several machine learning methods.

The results show that the high dimensional models with regularized coefficients, especially Random Forest and Lasso Lars, consistently outperfom the benchmark models. We have a bigger relative gain in the very short term when predicting with machine learning models. In a 1-day forecast, the prediction error of machine learning models is about 40 times smaller than the error of traditional models, while for the subsequent forecast horizons this error is, on average, 20 times smaller. Futhermore, we conclude that lagged demands are the most relevant variables for very short-term forecast and as we consider longer horizon, the other set of variable become more important.

Our paper shows how machine learning models predict PEC better than traditional ones. [Melhorar esse paragrafo a partir daqui - explicar deep learning e selecao de variaveis] Future works may attempt to predict power electricity using deep learning models. We also may facilitate the variable selection for future works that intend to predict PEC.

Reler Texto. Colocar no corretor do word

Bibliography

ADEGBEHIN, A. et al. Effect of weather parameters on hydroelectric power generation in kainji dam niger state, nigeria. In: . [S.l.: s.n.], 2016. 12

ALMESHAIEI, E.; SOLTAN, H. A methodology for electric power load forecasting. *Alexandria Engineering Journal*, v. 50, n. 2, p. 137 – 144, 2011. ISSN 1110-0168. Available at: <http://www.sciencedirect.com/science/article/pii/S1110016811000330>. 12

ANEEL. *Atlas de Energia Elétrica do Brasil*. [S.l.: s.n.], 2002. 12

BOLDIN, M.; WRIGHT, J. H. Weather-adjusting economic data. *Brookings Papers on Economic Activity*, v. 46, n. 2 (Fall), p. 227–278, 2015. Available at: <https://EconPapers.repec.org/RePEc:bin:bpeajo:v:46:y:2015:i:2015-02:p:227-278>. 17

BREIMAN, L. Random forests. In: *Machine Learning*. [S.l.: s.n.], 2001. p. 5–32. 24

BUNDERVOET, T.; MAIYO, L.; SANGHI, A. Bright lights, big cities: measuring national and subnational economic growth in africa from outer space, with an application to kenya and rwanda. *Policy Research Working Paper*, n. 7461. World Bank, Washington, DC., 2015. Available at: <https://openknowledge.worldbank.org/handle/10986/22883>. 11

DELL, M.; JONES, B. F.; OLKEN, B. A. What do we learn from the weather? the new climate-economy literature. *Journal of Economic Literature*, v. 52, n. 3, p. 740–98, September 2014. Available at: <http://www.aeaweb.org/articles?id=10.1257/jel.52.3.740>. 12

EFRON, B. et al. Least angle regression. *Annals of Statistics*, v. 32, p. 407–499, 2004. 23

EL-SHAZLY, A. Electricity demand analysis and forecasting: A panel cointegration approach. *Energy Economics*, v. 40, p. 251 – 258, 2013. ISSN 0140-9883. Available at: <http://www.sciencedirect.com/science/article/pii/S0140988313001485>. 12

FAN, S.; HYNDMAN. Short-term load forecasting using semi-parametric additive models. In: *2011 IEEE PES General Meeting*. United States: IEEE, Institute of Electrical and Electronics Engineers, 2011. p. 1–7. ISBN 9781457710018. 12

GARCIA, M. G.; MEDEIROS, M. C.; VASCONCELOS, G. F. Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, v. 33, n. 3, p. 679 – 693, 2017. ISSN 0169-2070. Available at: <http://www.sciencedirect.com/science/article/pii/S0169207017300262>. 22

HENDERSON, J. V.; STOREYGARD, A.; WEIL, D. N. Measuring economic growth from outer space. *American Economic Review*, v. 102, n. 2, p. 994–1028, April 2012. Available at: <http://www.aeaweb.org/articles?id=10.1257/aer.102.2.994>. 11

HUANG, S.-J.; SHIH, K. Short-term load forecasting via arma model identification including non-gaussian process considerations. *IEEE Transactions on Power Systems*, Institute of Electrical and Electronics Engineers Inc., v. 18, n. 2, p. 673–679, 5 2003. ISSN 0885-8950. 12

LEBOTSA, M. E. et al. Short term electricity demand forecasting using partially linear additive quantile regression with an application to the unit commitment problem. *Applied Energy*, v. 222, p. 104 – 118, 2018. ISSN 0306-2619. Available at: <http://www.sciencedirect.com/science/article/pii/S030626191830504X>. 12

MAZA, A.; VILLAVERDE, J. A state-space approach to the analysis of economic shocks in spain. *Journal of Policy Modeling*, v. 29, n. 1, p. 55–63, 00 2007. Available at: <http://www.sciencedirect.com/science/article/B6V82-4N1CRSM-7/1/cc6d22223592fe7fbcba46a024018163>. 11

MEDEIROS, G. V. M. C.; FREITAS, E. Forecasting brazilian inflation with high-dimensional models. *Brazilian Review of Econometrics*, v. 36, p. 1980–2447, 2016. ISSN 0306-2619. 24

MOOSA, I.; BURNS, K. The random walk as a forecasting benchmark: drift or no drift? *Applied Economics*, Routledge, v. 48, n. 43, p. 4131–4142, 2016. Available at: <https://doi.org/10.1080/00036846.2016.1153788>. 22

NASA. *Bright Lights, Big City*. 2000. [Online; accessed 10-Abri-2019]. Available at: <https://earthobservatory.nasa.gov/features/Lights>. 11

ORGANIZATION, W. W. M. *Guidelines on the Calculation of Climete Normals*. [S.l.]: Secretariat of the World Meteorological Organization, 2017. 17

RODRIGUES, F.; CARDEIRA, C.; CALADO, J. The daily and hourly energy consumption and load forecasting using artificial neural network method: A case study using a set of 93 households in portugal. *Energy Procedia*, v. 62, p. 220 – 229, 2014. ISSN 1876-6102. 6th International Conference on Sustainability in Energy and Buildings, SEB-14. Available at: <http://www.sciencedirect.com/science/article/pii/S1876610214034146>. 12

Appendix

List of Variables

A. Groups

Consumption Classes (C)	
Index	Class
CSO	Commercial, services and others
OC	Own consumption
SL	Street lighting
IND	Industrial
PP	Public power
RES	Residential
RR	Rural
ARR	Aquicultor rural
IRR	Irrigating rural
PS1	Public service (water, sewage and sanitation)
PS2	Public service (electric traction)
TOT	Total
AVG	Average

Regions (R)	
Index	Region
SE	Southeast
MW	Midwest
S	South
NE	Northeast
N	North
BR	Brazil

B. Variables

Power Electricity Consumption - 185 lagged variables	
Index	Variable
PEC_R_i	Hourly PEC of the hour i and region R (MW/h) - (i = 0, $\cdots$, 24,)
PEC_R_TOT	Sum of daily PEC of region R (MW/h)
PEC_R_AVG	Average daily PEC of Region R (MW/h)
PEC_R_C	Daily PEC of the region R for consumption class C (MW/h)

Economic Variables - 79 variables
Selic target imposed by BCB
SELIC effective rate
CDI
Dollar for purchase
Dollar for purchase variance
Dollar for sale
Dollar for sale variance
Euro for purchase
Euro for purchase variance
Euro for Sale
Euro for sale variance
IBOVESPA quotation
IBOVESPA daily minimum value
IBOVESPA daily maximum value
IBOVESPA daily absolute variance
IBOVESPA daily percentage variance
IBOVESPA daily volatility
INPC monthly variance
INPC cumulative
IPCA monthly variance
IPCA accumulated
IPA-M monthly variance
IPA-M accumulated
IPA-DI monthly variance
IPA-DI accumulated
IGP-M monthly variance
IGP-M accumulated
IGP-DI monthly variance
IGP-DI accumulated
Crude steel production (observed)
Crude steel production (seasonally adjusted)
Heavy vehicles traffic on toll roads (observed)
Heavy vehicles traffic on toll roads (seasonally adjusted)
construction production inputs (observed)
construction production inputs (seasonally adjusted)
CPS and Usecheque consultations (observed)
CPS and Usecheque consultations (seasonally adjusted)
Serasa consultations(observed)
Serasa consultations (seasonally adjusted)
Installed capacity utilization in the manufacturing industry - FGV (observed)
Installed capacity utilization in the manufacturing industry - FGV (seasonally adjusted)
Installed capacity utilization in the processing industry - CNI (observed)
Installed capacity utilization in the manufacturing industry - CNI (seasonally adjusted)
Real Industrial Sales (observed)
Real Industrial Sales (seasonally adjusted)
Production hours worked in the manufacturing industry (observed)
Production hours worked in the processing industry (seasonally adjusted)
Real wage in manufacturing industry (observed)
Real wage in the manufacturing industry (seasonally adjusted)
Oil and gross oil production (monthly production in m3)
Natural gas production (monthly production in m3)
General industrial production (observed)

General industrial production (seasonally adjusted)
Industrial production - capital goods (observed)
Industrial production - capital goods (seasonally adjusted)
Industrial production - intermediate goods (observed)
Industrial production - intermediate goods (seasonally adjusted)
Industrial production - consumer goods (observed)
Industrial production - consumer goods (seasonally adjusted)
Automotive industry production (observed)
Automotive industry production (seasonally adjusted)
Consumer confidence index
National consumer expectations index
Industrial entrepreneur confidence index
BNDES disbursements - accumulated amounts
Total employment index (observed)
Total employment index (seasonally adjusted)
Total employment index- manufacturing industry (observed)
Total employment index - manufacturing industry (seasonally adjusted)
Total employment index- commercial (observed)
Total employment index - commercial (seasonally adjusted)
Total employment index - services (observed)
Total employment index - services (seasonally adjusted)
Total employment index - civil construction (observed)
Total employment index - civil construction (seasonally adjusted)
Employed people (formally)
Unemployment rate (brazil)
Monetary Base
Currency paper issued

Calendar - 6 variables
Day of the week (1-7)
Day of the month (1-30)
Dummy for holiday
Seasons (1-4)

Weather Variables - 180 variables	
Index	Variable
CLD_R_i	Cloudiness of region R and hour i (i = 9,15,24, AVG)
AP_R_i	Atmospheric pressure (mbar) of region R and hour i (i = 9,15,24, AVG)
DBT_R_i	Dry bulb temperature (°C) of region R and hour i (i = 9,15,24, AVG)
HBT_R_i	Humid bulb temperature (°C) of region R and hour i (i = 9,15,24, AVG)
RH_R_i	Relative humidity (%) of region R and hour i (i = 9,15,24, AVG)
WD_R_i	Wind direction of region R and hour i (i = 9,15,24, AVG)
WV_R_i	Wind velocity (m/s) of region R and hour i (i = 9,15,24, AVG)

Electric Energy Price Variables - 60 variables	
Index	Variable
PECR_R_C	PEC price of region R and consumption class C (R$)
PECR_R_AVG	Average PEC price of region R (R$)