UNIVERSIDADE DE BRASÍLIA
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
DEPARTAMENTO DE ECONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

High Dimensional Models for Forecast Power Electricity Consumption

Pedro Caiua Campelo Albuquerque

Orientador: Marina Rossi

Brasília
2019

Abstract

This present work uses machine learning models to predict power electricity consumption, for 6 different time horizons. We observed that the machine learning models have superior results than the original econometric models, by removing the overfitting present in these models. The best predictions was from Random Forest model, that managed to keep up with the trend and the seasonality of the different time horizons. With this model, we are able to capture the trend of power electricity consumption and make inferences about economic activity, due to the positive relationship between the two variables. Our work further concludes that lagged consumption values are extremely important for very short-term forecasting (1 and 7 days) and calendar and meteorological variables become more important for forecasting as it increases the forecast horizon. Price and economic variables are more relevant for 7 and 15 days ahead forecast.

Key-words: Power Electricity. Machine Learning. Forecast.

Sumário

1 Introduction

Maza and Villaverde (2008) review the bibliography to understand the causality of electric energy and economic growth. The relationship can be divided into 4 categories: causality from electricity consumption to GDP growth, causality from GDP growth to electricity consumption, bidirectional causality and no causality at all. Literature does not necessarily converge to a specific result. Although a vast majority finds a relation between these two variables, the direction of this causality is not so clear. The authors argue that the newer and better controlled models point causality from electricity consumption to GDP growth.

Still, most of the electricity consumption today is consumed by industrial and commercial activities. Therefore, the main objective of this work is to forecast electricity consumption in the short term and in the medium term, and thus, be able to measure a tendency for economic activity. With GDP released quarterly, we want to structure the forecast models of electricity consumption to capture the trend of economic activity.

Huang and Shih (2003) and Almeshaiei and Hassan Soltan (2011) uses just lagged demand to forecast short-term of energy load via ARMA Model. Rodrigues *et. al.* (2014) also uses just past values to forecast the daily and hourly energy consumption via Artificial Neural Networks. Lebotsa *et. al.* (2018) and Fan and Hyndman (2012) includes calendar and meteorological variables to forecast PEC.

Hence, its a fact that the variable selection for power electricity consumption forecast is not trivial. So, we prefer to build a high dimensional database that includes all the main variables and let the machine learning models choose the most relevant variables for each horizon prediction.

We construct a high dimensional database with more than 500 variables, including lagged demand, calendar variables, meteorological variables, price of electric energy variables and economic variables. We use we use as many traditional models as the machine learning models, expecting better performance for the machine learning models, by removing the overfitting present in traditional econometric models. We run our models for 6

differents forecast horizons.

We also want to analyze which variables are most relevant for each forecast horizon, and thus, enable an improvement in the future variable selection to predict the power electricity consumption.

In Section 2, we present our model – describing in-depth the structure and all models used to forecast power electricy consumption. In Section 3, we exhibit our data set and analyze all sets of variables used. In Section 4, we show our general results and examine the best models. Also, we observe witch variable are more importantes to predict PEC in differents forecasts horizons. Lastly, in Section 5, we present a brief conclusion concerning our findings.

2 Econometric Models

2.1 Model Structure

This section delineates the empirical methods used in this paper for forecasting future power eletricity consumption (PEC). As we said, electricity consumption is a high seazonal variable, so we consider that the forecast of PEC h, y_{t+h}, periods ahead is modeled as a function of blocks of predictors in differents lag times. Accordingly, we used three lags of the candidate variables divided into five groups aligned with the literature of economic activity and power electricy forecast. The models were estimated using a rolling window of 546 observations.

The Brazilian GDP is release quarterly, hence the main ideia os this paper, is to meansure economic activity in the period that the data is not released yet. Finally, we estimated the models for 6 differents forecast horizons, 1, 7, 15, 30, 60 and 90 days ahead. Moreover, we have divided the forecasts horizons into two sets: (i) short-term forecast, which contains 1, 7 and 15 days ahead and (ii) medium-term forecast, which contains 30, 60 and 90 days ahead. Our estimated equation is given by:

$$y_{t+h} = \alpha_0 + \sum_{i=0}^{2} \gamma_i y_{t-(h*i)} + \sum_{i=0}^{2} \beta_i X_{t-(h*i)} + u_{t+h} \tag{2.1}$$

where y_t is the power electricity consumption (KM/h) in Brazil at time t, α_0 is a constant term, X_t is the vector of the candidates containing all candidate variables and u_t is an error term.

The lags structure means that if we make a forecast for $t + h$ periods ahead, we would use the variables from t, $t - h$ and $t - 2h$. From example, if it is March 21th and we want forecast the power electricity consumption for March 28th, that is, a 7 day ahead forecast. In this case, we would date from today, from March 14 (past 7 days) and March 7 (past 14 days). This structure was the first hyper-parameters defined. We had tested for a several other lags structures and this one had the minor forecast error. Which is quite

intuitive, because PEC is a seasonal variable and if we define a lag structure very close, like 1, 2 and 3 days past, we would lost forecast accuracy for picking close days.

Our explaining variables are divided into five sets aligned to the literature of economic activity and power electricity consumption forecast. The first set, represented by the vector yt_h, are the lagged demand of power electricity. The next four sets are in the X_t vector, that is, $x_t^1, x_t^2, x_t^3, x_t^4 \in X_t$. These sets are, respectively, calendar variables, meteorological variables, price of electric energy and economic variables. Section 3 will better explain each set of variable.

2.2 Models

This section exhibit and explain all the models used to predict PEC, with the structure showed in last section. Shyh-Jier Huang *et al.* (2003*) forecast short-term load using of autoregressive moving average (ARMA). Medeiros *et al.* (2017) use high-dimensional and machine learning models to forecast inflation in real-time, and concluded that these models perform better for big data than traditional models*. Following this line, we use several models to forecast PEC and analyse which perform better. Our initial predict models, ARIMA and Random Walk, use only lagged comsumption variables. Then, we begin to introduce the explanatory variables and to predict with OLS. After that, we use some machine learning models, such as Lasso, Lars, Lasso Lars, Ridge, Elastic Net and Random Forest. All models are explained below.

2.2.1 ARIMA

Autoregressive integrated moving average (ARIMA) uses just lagged demand of the variable of interest. AR *(p)* indicate that the evolving variable of interest is regressed on its own p lagged values. I(i) is used when the series shows evidence of non-stationarity and an initial differencing step can be applied non-stationary . Finally, MA (q) specify that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The ARIMA (p,i,q) equation can be

represented by

$$y_t^* = \gamma + \Phi_0 + \Phi_1 y_{t-1}^* + \cdots + \Phi_p y_{t-p}^* + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} \tag{2.2}$$

where $y_t^* = y_t - y_{t-1}$, because our serial data is non-stationary (but the first difference is). Both p and q values are determined running ARIMA model for several different values for each and the one that gives the smallest mean squared error are choosen.

2.2.2 Random Walk

Our random walk model (without drift) represents the last value we have available for forecast. That is, the best value for forecast k periods ahead is the value available today. The model can be represented by

$$y_t = y_{t-k} + u_t \tag{2.3}$$

This model is a good measure of sensitivity, because it is reasonable to accept, at least in the short term, that the last available value is a good prediction. So, if other models present better results than the random walk, it can be indicative of a good predictor.

2.2.3 OLS

From now on, the models use the sets of explanatory variables, $x_t^1, x_t^2, x_t^3, x_t^4 \in X_t$, in addition to the lagged values of PEC. OLS is one of the most traditional econometric models. It minimizes the mean square error (MSE) and their parameters are defined by

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 \right] \tag{2.4}$$

By minimizing MSE, OLS has difficult to eliminate irrelevant variables for forecasting. In this way, OLS ends up overfitting the model, which is bad for forecasting values

outside the sample.

2.2.4 Lasso

Working with high dimensional data, *shrinkage* methods form a prosperous alternative to traditional models. The main ideia is to shrink irrelevante variable to zero.

As well as the OLS, the least absolute shrinkage and selection operator (LASSO), also minimizes the MSE. However, lasso has a shrinkage coefficient , λ, that forces irrelevant variables to zero. Their parameters are determined by

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right] \tag{2.5}$$

LASSO can be seen as a generical OLS model, because if $\lambda = 0$, the paremeters $\hat{\beta}$ and $\hat{\beta}_{OLS}$ are the same. Still, if the coefficient of shrinkage are equal to some λ_c, all β_i will be equal to zero. As well, $\lambda_{ols} \leq \lambda \leq \lambda_c$. Our choice of this hyper-parameter was made using a list of possible values to λ and the chosen parameter was the one with the lowest MSE.

2.2.5 Lars

Least angle regression (LARS) is also a good model to use when working with high dimensional data. The main ideia of his algorithm is to start with all β equal to zero and to increase the coefficient of the most correlated x_i with y. This growth stops when finds some x_j that own as much correlation with the residual. Again, the model rises (β_i, β_j) in their joint least squares direction, until other predictor x_k, and so on.

Efrom *et. al.* (2004) concluded that the LARS algorithm and his variations works gracefully for the case where 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.

2.2.6 Lasso Lars

Efrom *et. al.* (2004) proof that a simple modification of the LARS algorithm can results in the simulation of all LASSO models for all possible values of λ. So, in the traditional LARS method, if a coefficient changes the signal, the direction remains the same. However, in the LASSO model, when a coefficient reaches the value 0, it is discarded from the set of active variables. For that reason, LASSO LARS algorithm make a simple modification in the original LARS: if any coefficient becomes zero this variable is discarded and the model recalculates the search direction.

2.2.7 Ridge

As well as LASSO, Ridge Regression, is a generic version of OLS that also has a shrink coefficient λ. However, the difference is that the Ridge's restriction is squared and their parameters are determined below

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right] \tag{2.6}$$

By adding a squared constraints, the loss function is strictly convex, and therefore has a unique minimum. Consequently. Ridge algorithm make it harder to zero the coefficients and can't eliminate completely some irrelevants variables. This can result in several variables with close values to zero and a lower interpretability of the results.

Again, the λ parameter is chosen after running several Ridge models for various λ*'s* values and picks which hyper-parameter has the lowest MSE.

2.2.8 Elastic Net

Elastic Net is also a regularized regression model that combines the constraints of Lasso and Ridge Regression. Their parameters are given by

$$\hat{\beta} = \underset{\hat{\beta}}{argmin} \left[\|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right] \tag{2.7}$$

Elastic Net becomes a general case of LASSO and Ridge Regression. And so, as in Ridge model, the Elastic Net method makes the loss function strictly convex, forcing it to have a unique minimum. Its hyper-parameters, λ_1 *and* λ_2 are also calculated after several regressions tests until finding the lowest MSE.

2.2.9 Random Forest

Random Forest is a method of machine learning for classification and regression of variables. This model creates several random decision trees in the training period and mixes them to obtain the most accurate prediction possible, resulting in the node of the variables that appear most frequently

Random Forest seeks to reduce the overfiting of traditional models when making a bagging (or Botstrap), combining learning models. By making this average of the various decision nodes possible, the model decreases its variance in exchange for an increase in bias.

3 Data

Power Electricity Consumption (PEC) is measured using the hourly energy load by subsystem (KW/h) released by the National Electricity System Operator (ONS). The initial date from ou dataset is February 1, 2017, when the hourly data of electricity consumption started to be disclosed, and ends July 31, 2018, totaling one and a half years of hourly data of PEC. However, the time lag analyzed is daily, due to the other variables, totaling 546 days observed.

ESTOU AQUI FALANDO DA DIVISÃO DE TESTE E TREINO E O QUE EU FACO EM CADA BASE.

We used February 2017 to Abril 2018 for the training set. The holdout set is from May 2018 to July 2018.

All the models delineated in the previous section are trained until the previous period of its forecast horizon. That is, if it is for a 30 day forecast horizon, we will train the models until June 30, 2018, and test the models trained for the next 30 days and compare them with the original values of electric power consumption.

This comparison is made via the mean square error (MSE). Then, the model that presents the lowest MSE, arrives closer to the real value of PEC and is considered the best model.

Figure 1 shows how the total daily electricity consumption in Brazil has behaved, its trend, seasonal and residual. We observe PEC is a variable with a great seasonality. However, besides being a cyclical variable, this graph showed that the average consumption of electric energy showed a slight decrease in its average, during the analyzed period.

Figure 2 exhibits the behavior of the PEC by region. It is easy to observe that the consumption of electric energy has different conduct by region of Brazil and the division of the variable by region better the perfomances of the forecast models.

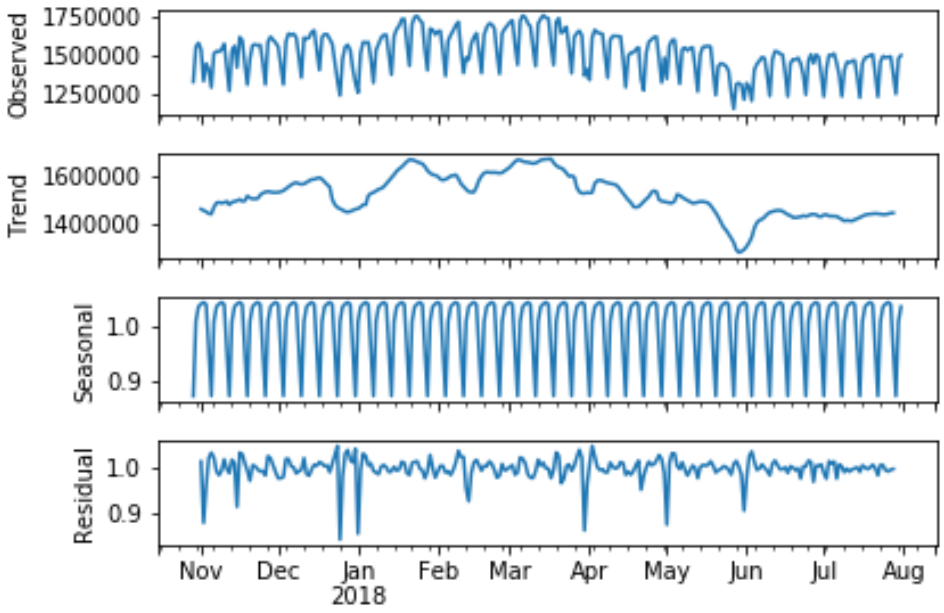

Figure 1 – Seasonal Plot

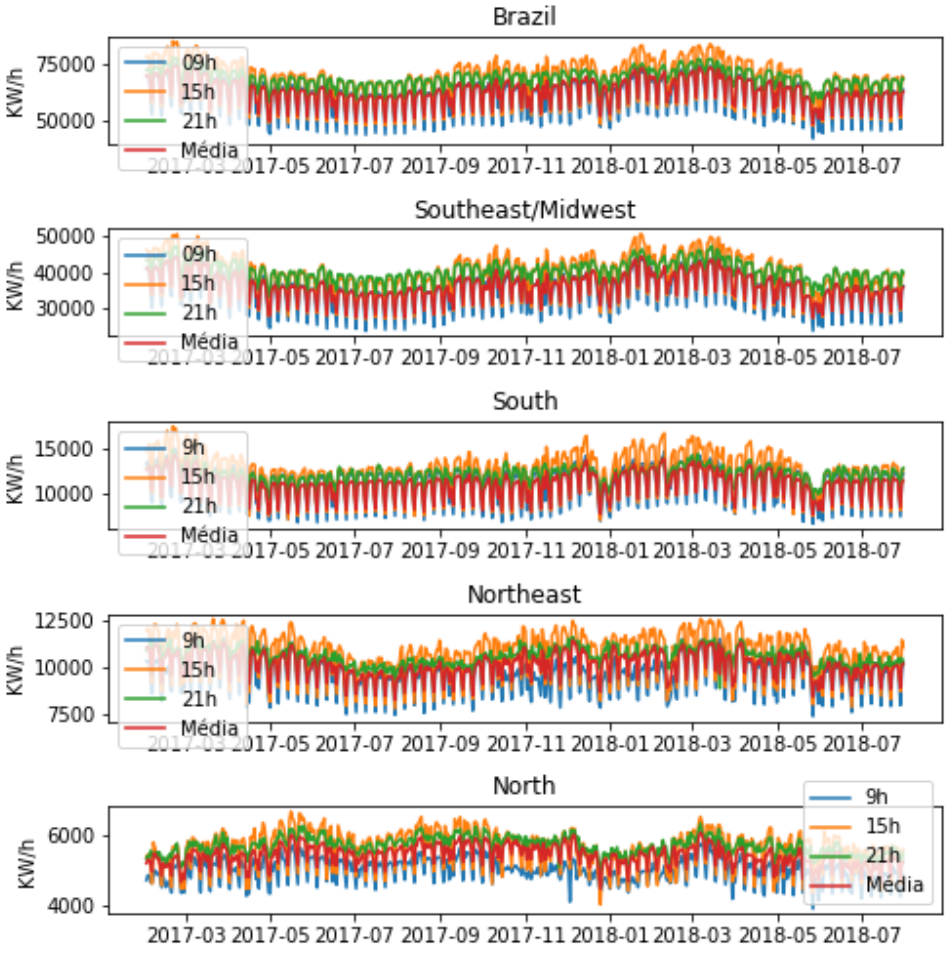

Figure 2 – PEC by Hour

3.1 Power Eletricity Consumption

Huang and Shih (2003) use just lagged demand to forecast short-term of energy load via ARMA Model. The authors also use a variation of this model including non-Gaussian process considerations and concluded that the performance of ARMA model is better ensured, improving the load forecast accuracy significantly. Almeshaiei and Hassan Soltan (2011) also forecasts eletric power load using ARMA/ARIMA models. The authors propose a practical guide for forecasting future electricity consumption, arguing that every electric network needs special forecasting method because each country is indifferent in the factors that affect the electricity demand.

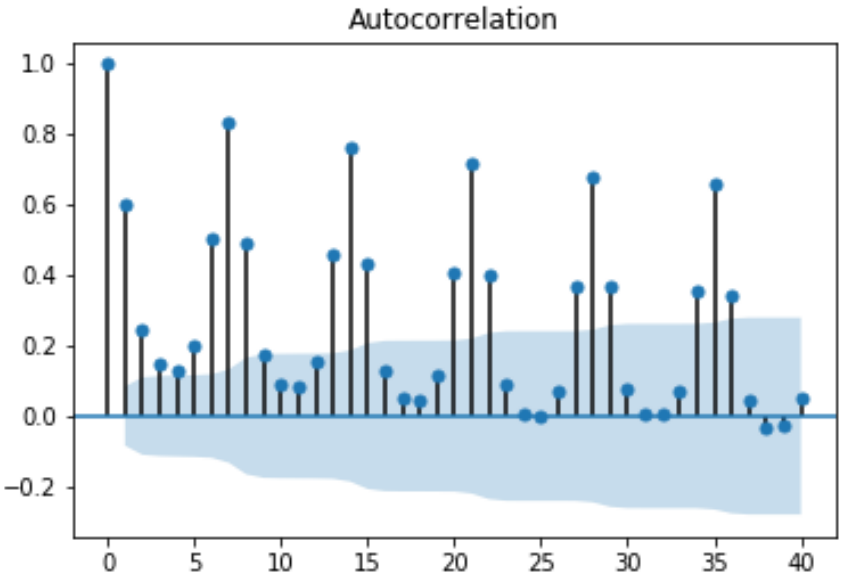

Figure 3 – Autocorrelation Plot

Figure 3 shows how the electric energy consumption is correlated with its lagged demand. 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 with the cyclicity of the PEC. These figures clarify the importance of including lagged demand in forecast models.

Therefore, we selected the hourly energy load by subsystem (KW/h) to Brazil and by Region on the National Electricity System Operator (ONS) website. In addition, we

took the daily data of PEC divided by consumption class afor each region in Electric Energy National Agency (ANEEL) website. The set of consumption classes includes PEC variables divided in groups such as industrial, comercial and services, own consumption, residential, rural, street lighting, public service and others.

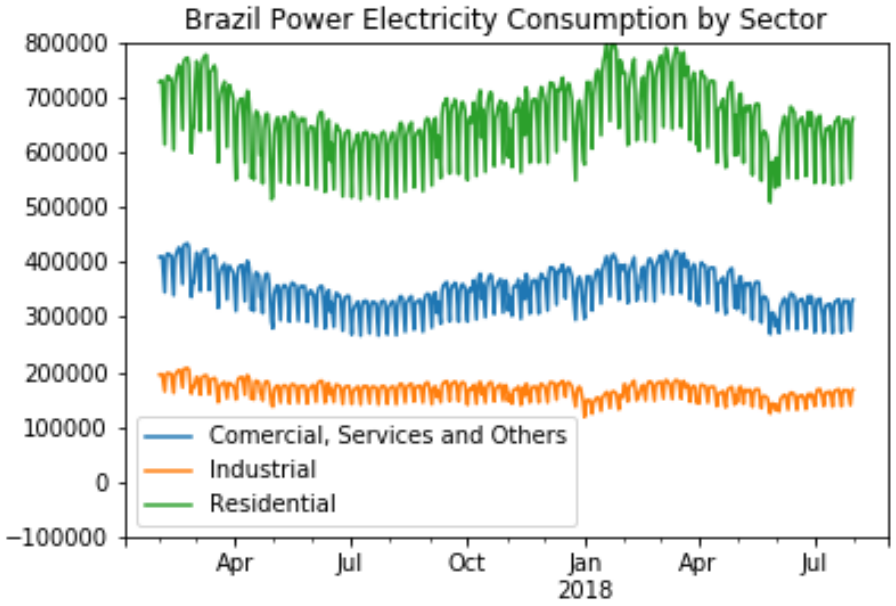

Figure 4 – PEC by sector

This division by consumption class not only improves the performance of prediction models, but also is fundamental to understand where the greatest demand for electricity consumption is. The vast majority of PEC are divided into commercial and services, industrial and residential consumption. The first two are directly linked to economic activity of the country, while the third is indirectly related. Figure 5 shows the trend of these variables, and we can see that the sum of the first two variables exceeds residential consumption, consequently, the largest share of Brazil's electricity consumption is for commercial and industrial activities. This fact reinforces the relation of economic activity and electric energy consumption.

With the division of the variables by hour and by consumption class for each region, a total of 185 past PEC variables has been detected.

3.2 Calendar Variables

Lebotsa *et. al.* (2018) included lagged demand, calendar variables and meteorological variables to construct a partially linear model to predict short term electricity demand. These variables and the use of nonlinear trend variables made the short term electricity demand of perform better.

Fan and Hyndman (2012) propose a new statistical method to forecast the short term electricity demand, allowing nonlinear and nonparametric terms within the regression framework. Their model inputs are calendar variables, lagged demand and temperature variables, and try to catch the complex nonlinear relationship between electricity demand and its drivers.

As we saw above, PEC is an extremely seasonal variable. At least in the short term, the inclusion of calendar variables controls part of this cyclicity. This group of variables includes day of the week, day of the month, month, season of the year, year and a dummy for holidays, totaling 6 calendar variables.

3.3 Meteorological Variables

In the same line of calendar variables, the meteorological variables serve to control a little of the seasonality of the consumption of electric energy. Thus, as seen in the two studies cited above, it is highly recommendable to place this set of variables in the forecast models of electricity consumption in the short term.

These data were collected from the National Institute of Meteorology (INMET) website that release the historical series by stations scattered throughout Brazil. This set of variables contains air nebulosity, atmospheric pressure (mbar), dry bulb temperature (oC), humidity bulb temperature (oC), relative humidity (%), wind direction and wind speed (m/s) disclosed for the hours 9, 15 and 21 for each station.

After aggregating the data by region and for brazil, we have a total of 180 meteorological variables for all regions and all hours analyzed by INMET.

The most important meteorological variable for the forecast of electric energy consumption is the temperature of the dry bulb, which is also a very seasonal variable with a large variance. Boldin and Wright (2015) proposes an adjusted temperature that controls both the seasonal effect and the variation of the normal temperature per year.

$$temp\_adj_i = temp_i - \frac{1}{n}\sum_{i=1}^{n} temp_i$$

where $temp_i$ is the temperature on day i.

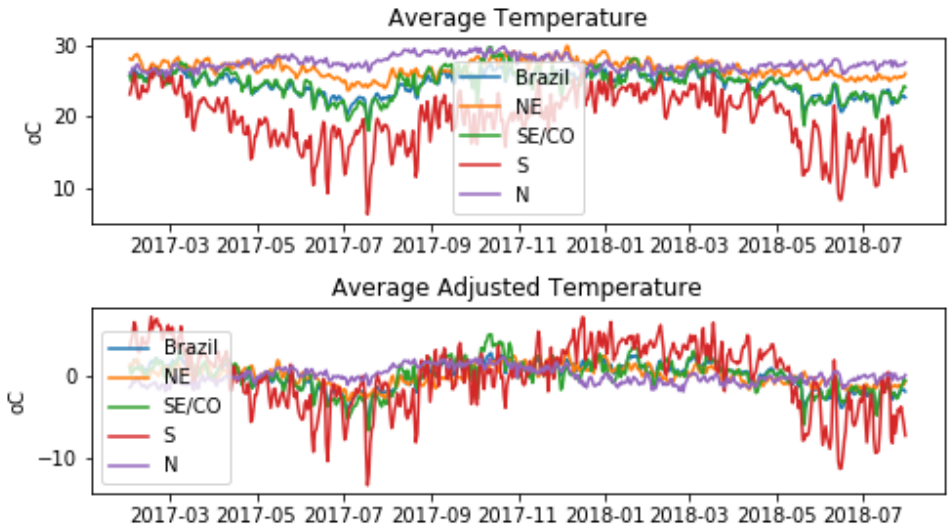

Figure 5 – Average Temperature

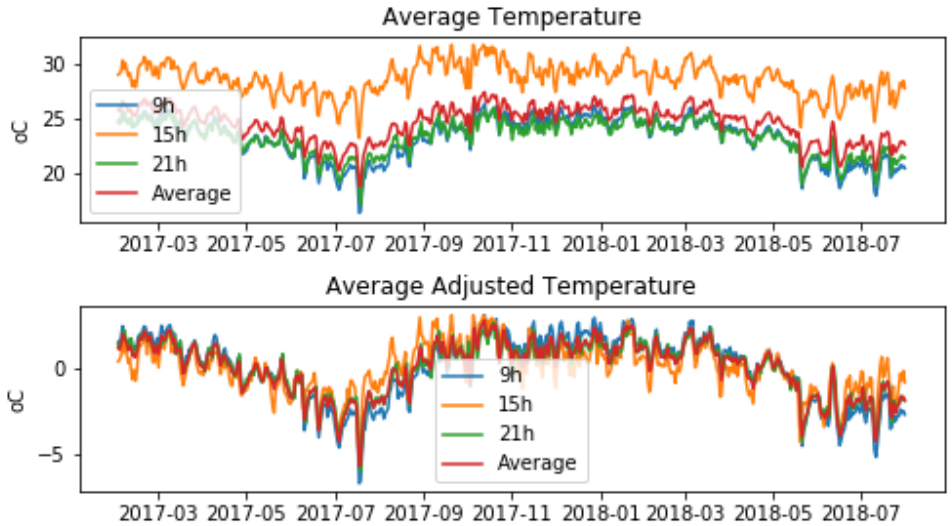

Figure 6 – Adjusted Average Temperature

Figures 6 and 7 show how the adjusted variability loses part of its variance and seasonality. Besides that, comparing figures 2 and 7, it is possible to observe that in the times of higher temperature also occurs a greater consumption of electric energy in the day. This is a signal that the meteorological variables are important to predict PEC.

3.4 Price of Electric Energy

Since our interest variable is a demand variable, we include variables that measure the cost of electric energy. We pick the price set from Electric Energy National Agency (ANEEL) and its also divided by consumption class each region. Aggregating all variables by region and by consumption class, we have a total of 60 price variables. Yet, cost of electricity is a monthly variable and as our dataset is a daily variable, we set the same price for the whole month.

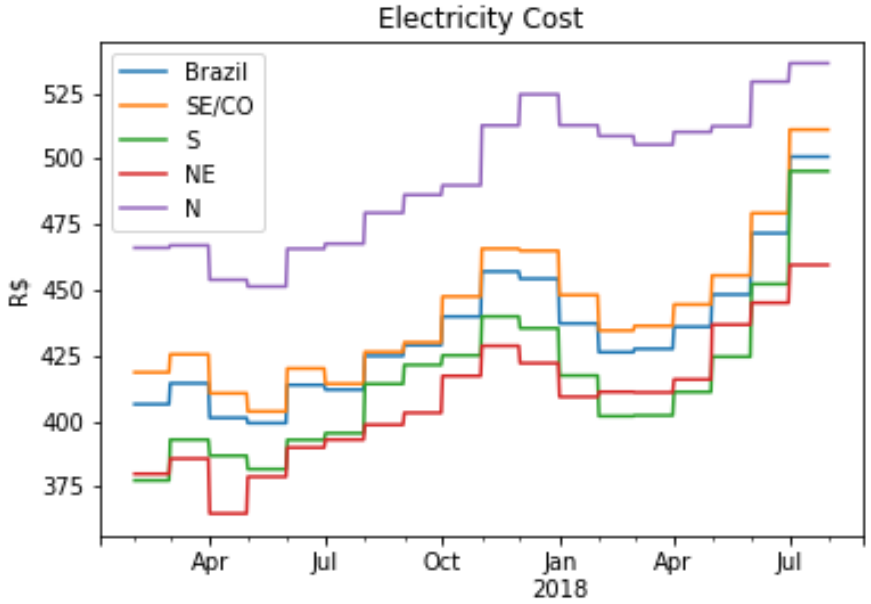

Figure 7 – Average Price

When comparing figures 2 and 8, we can observe that the negative relation between price and demand, in this case, is not so strong. The Northeast region, for example, has the lowest tariff and yet does not have the highest consumption of electricity. It's all about the region's structural capacity. Hence, this set of variables may be indicating that the

price elasticity of demand of electricity might be low in some time horizons.

3.5 Economic Variables

Shazly (2013) build a dynamic econometric model of nonstationary heterogeneous panels including economic variables. Their models provides a satisfactory technique to analyze energy demand at a disaggregated sectoral level for policy purposes. Maza and Villaverde (2008) argues that existis a causality from electricity consumption to GDP growth.

How the main idea of this paper is measure economic activity with consumption of electric energy, we include a wide list of economic variables, some daily and other monthly. As in the price variables, we set the same value to economic variable for the whole month. We pick the economic variables from a database of Central Bank, and include several indices of price, traffic of vehicles, unemployment, wage, industrial production variables (consumer goods, capital goods, automobile), monetary variables (monetary base and currency issued), consults to SPC and Serasa, BNDES expenses and several others. This set is composed by a total of 79 economic variables.

Adding all the explanatory variables listed in this section, we have a total of 510 variables. As I explained in the previous section, these variables are divided into 3 blocks of lags that depend on the forecast horizon, k. Thus, our model has 1530 variables and its estimated using a rolling window of 546 observations. We expact that our machine learning models select the most relevant variables for forecast PEC for each forecast horizon.

4 Results

MUDAR OS RESULTADOS E CONCLUSOES. MUDAR MEHTHODS EMPIRICOS TB.

In this session we show the results of the four-best performed models for each time horizon and make an analysis of these models and their parameters. First of all, we check which are the best models.

Model Rank	1 dia (MSE)	7 dias (MSE)	15 dias (MSE)	30 dias (MSE)	60 dias (MSE)	90 dias (MSE)
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)

Table 1 – Models Results

Table 1 exibiths the four models that presented the smallest prediction error (MSE) for each forecast horizon. These results corroborate to what we delineat in section 2. The machine learning models present much better perfomance than the models that use only lags demand and OLS, which ends up being overfitted and presents poor results in all time horizons analyzed.

Rank	Model	
	Short-term	Medium-term
1st	Random Forest	Lasso Lars
2nd	Lasso Lars	Lars
3rd	Lars	Ridge
4th	Ridge	Lasso
5th	Lasso	Random Forest

Table 2 – Models Rank

In table 2, we classified the 5 best models to forecast via analyzing analyzing the results presents in Table. We can observe that the 5-best models of forecast electric power consumption are the regularized models that select better the set of relevant variables. Our the best prediction model was Random Forest, which presented the lowest MSE in most of the forecast horizons analyzed.

Paremeters	1 Day				7 Days			
	Random Forest	Lasso Lars	Lars	Lasso	Random Forest	Lars	Ridge	Lasso Lars
∑(PED Variables)	0.74	-203.06	-251.15	80.12	0.9	266.75	361.89	246.78
∑(Calendar Variables)	0.2	-55575.01	-54343.69	-122161.95	0.01	116088.1	29363.31	117827.2
∑(Meteorological Variables)	0.05	2615.98	2633.04	22853.45	0.08	25852.79	18584.8	30527.85
∑(Price Variables)	0.01	-16332.11	-20921.88	-77237.63	0.01	-61841.9	-20269.3	-53178.2
∑(Economic Variables)	0.02	1502.11	1450.26	5352.6	0.03	2758.27	3560.09	3226.2

Table 3 – (a) Paremeters Analysis

Paremeters	15 Days				30 Days			
	Lasso Lars	Random Forest	Lars	Ridge	Lasso	Lasso Lars	Ridge	Lars
∑(PED Variables)	51.24	0.43	37.55	129.07	58.31	-101.15	196.96	-87.53
∑(Calendar Variables)	-50502.52	0.33	-45871.43	-36740.47	-177819.21	-154571.91	-80242.57	-80369.78
∑(Meteorological Variables)	-12791.22	0.13	13409.61	-12422.43	56999.25	54648.9	57121.47	54308.08
∑(Price Variables)	-14790.79	0.04	-7991.82	-18047.1	-18402.41	10405.24	3155.14	47408.99
∑(Economic Variables)	-7523.73	0.09	-2645.92	-3731.25	-9692.37	-4650.39	-20018.51	-2452.44

Table 4 – (b) Paremeters Analysis

Paremeters	60 Days				90 Days			
	Elastic Net	Ridge	Lars	Lasso Lars	Lasso Lars	Lars	Random Forest	Ridge
∑(PED Variables)	-47.56	-47.82	17.01	-34.51	-165.46	78.88	0.18	-400.96
∑(Calendar Variables)	-18938.03	-16871.96	21533.29	-8004.23	42786.93	42731.1	0.51	110.25
∑(Meteorological Variables)	-20711.24	-19325.94	-8132.87	-38750.11	-9214.65	-4171.4	0.28	-17230.67
∑(Price Variables)	-84044.65	-76144.52	-22053.61	-121277.64	-82185.74	-93558.98	0.04	-16851.77
∑(Economic Variables)	-7660.18	-6923.76	-50566.99	-50581.41	-42507.03	-42128.09	0.01	-30923.86

Table 5 – (c) Paremeters Analysis

To make a better analysis of the most important variables for the model, we added the coefficients of each set of variables described in the previous session. Tables 3, 4 and 5 (a,b,c) shows this sum for the 4 best models in each forecast horizon.

Since each set of variables has a different size, comparing the sum between them does not give us much. What we did in this work was to compare the sum of the same model with the forecast horizon shown in Tables 3, 4 and 5 (a,b,c). That is, we basically compared how the sum of coeficientes for each set varied in every time horizon and infer the importance of this set of variables for each one.

In order to clarify our analysis, we will compare here the sum of the coefficients of the Random Forest model, which is our main model. However, we also checked with the coefficients sum of the other models that also presented good results, such as LARS, Ridge, LASSO LARS and LASSO.

4.1 Short-term

4.1.1 1 Day

As can be observed in figures 3 and 4, the electricity comsumption is highly correlated with the consumption of the previous day. The LASSO, for example, zeroed all coefficients of all variables except for the power electric consumption for Brazil at 11 p.m. of the previous day.

Power Eletricity Consumption 1 Day Forecast

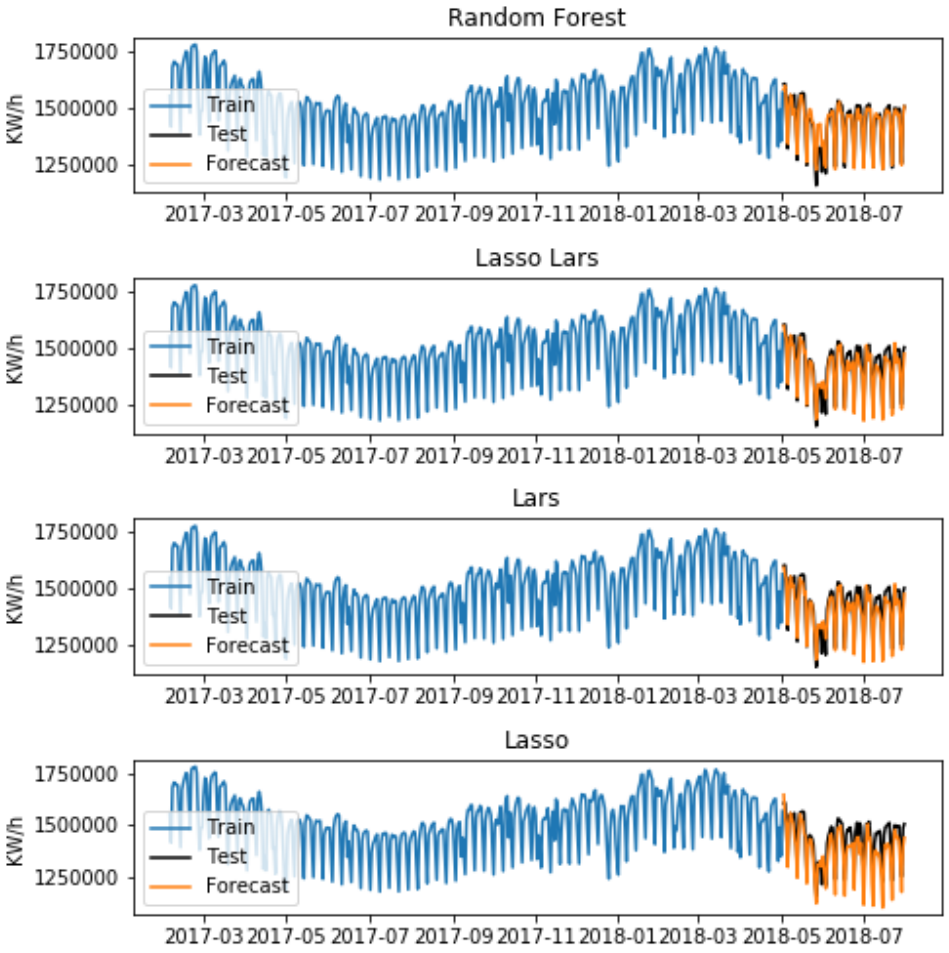

Figure 8 – 30 days forecast

All major models gives greater importance for lagged values of electricity consumption. For example, LARS assigns no value to the coefficients of the 60 electric energy price variables. This makes a lot of sense, since cost of energy consumption in that month will have little or no influence on the electricity consumption of the previous day of the same

month.

4.1.2 7 Days

If energy consumption has a large correlation with the value of consumption on the previous day, the correlation with the value of 7 days ago is even higher, as can be seen at figures 3 and 4. Thus, Table 3 (a) shows that the weight of the past variables of electric power consumption is even higher for this forecast horizon, comparing with day the forecast horizon of 1 day. The Lars model, for example, increased by almost 6 times the sum of lagged values.

The correlation is so strong that this is the only forecast horizon where the Random Walk model appears, picking up exactly the values of electricity consumption from the previous week. It is worth mentioning that this is the only forecast horizon that has a model that is not machine learning.

Figure 9 showed that the predictions of the main models are able to follow the trend and seasonality of the series of electric energy consumption, better until the values of the previous week, exhibited in Random Walk model.

4.1.3 15 Days

For a forecast of 15 days ahead, the correlation with the past values of demand are still present, however, increasingly weak. Comparing with the 7-day forecast, it is observed that the sum of the lag variability parameters decreases for the main models. In random forest, our main model, this figure decline by more than half.

On the other hand, other variables began to gain more importance as we advanced in the forecast horizon. In random forest, all the variable blocks obtained a greater importance in the main forecast models, with a greater significance for the sets of calendar variables, economic variables and even price variables gained a greater importance. Thus, the economic variables were relevant for forecasting energy consumption for 15 days ahead.

Figure 10 shows that, again, the forecast and our best can follow the tendency and

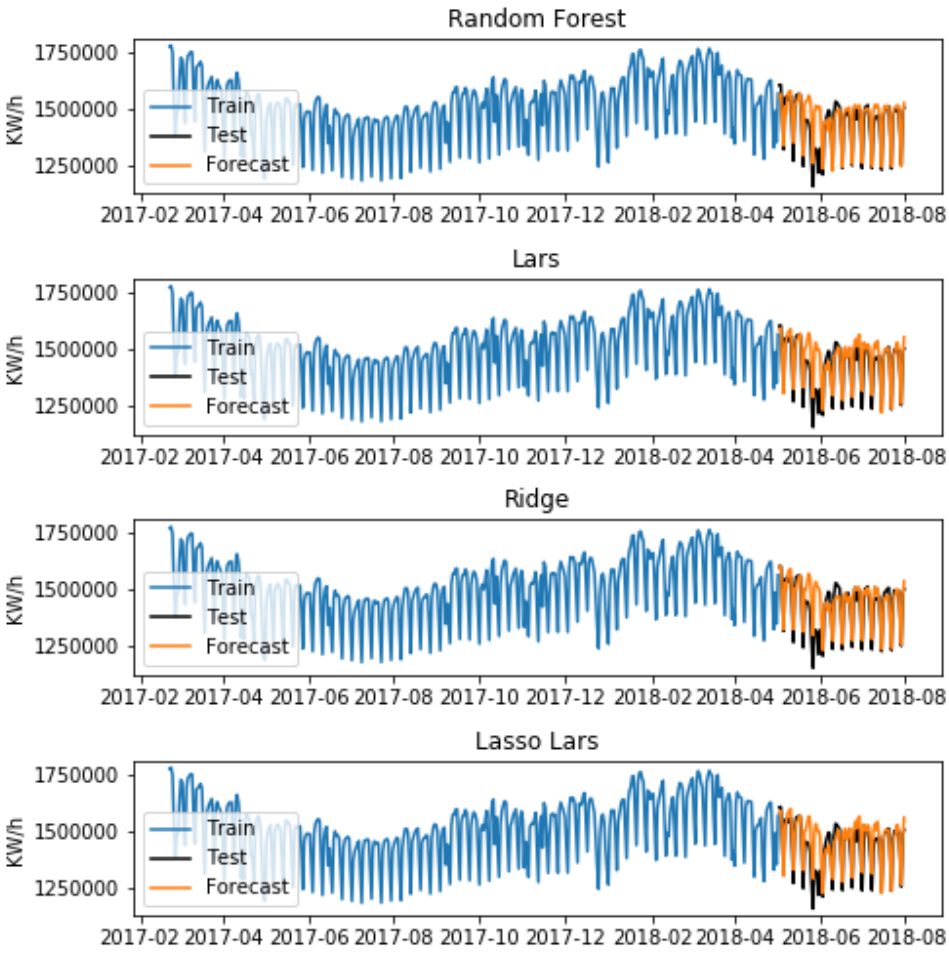

Figure 9 – 7 days forecast

cyclicity of the series of electric energy consumption, with a small difficulty to find the local maximum and minimum of the series.

4.2 Medium-Term

4.2.1 30 Days

Forecasting PEC 30 days ahed, we observe that the coefficientes sum of lagged demand decreases still further. Which makes sense, since the correlation of this series with its past values is decreasing.

On the other hand, the coefficientes sum of the calendar variables increases somewhat and confirms their importance for short and mainly medium-term forecast models.

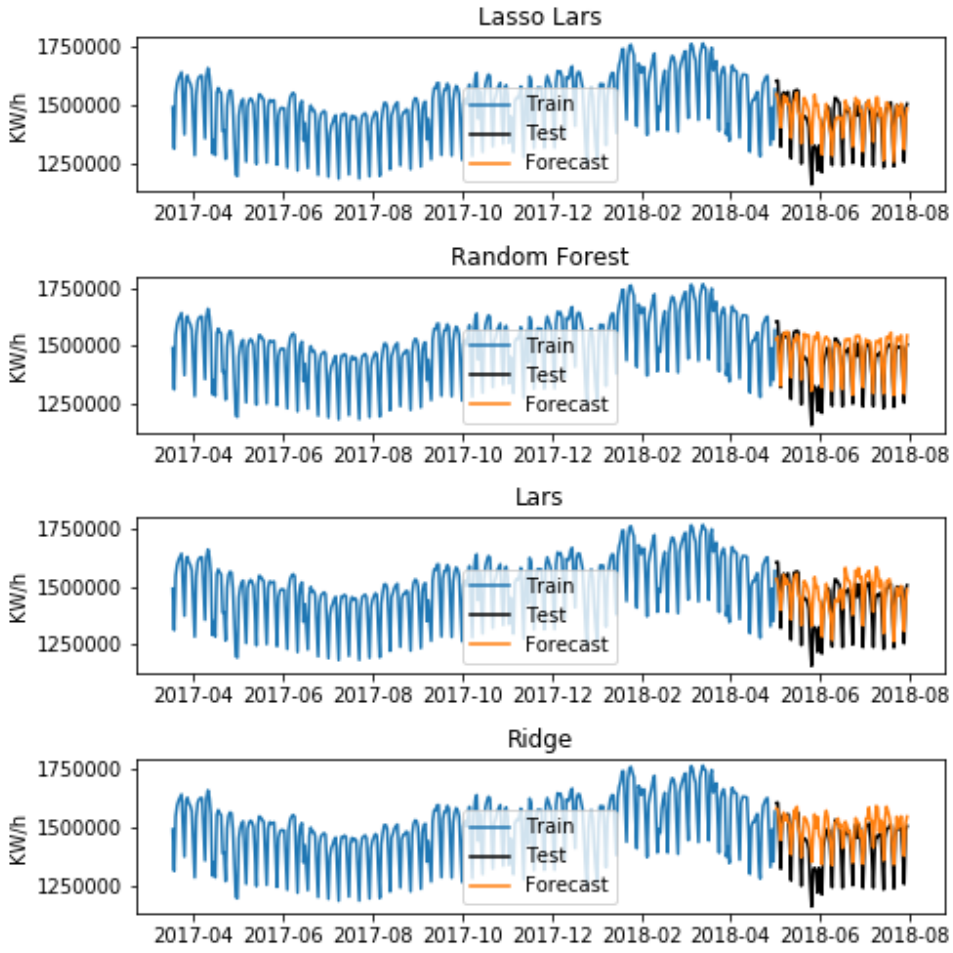

Figure 10 – 15 days forecast

In addition, price variables also increase considerably, showing greater price elasticity of demand of power electricity for longer forecast horizons.

The coefficients sum for meteorological and economic variables has decreased somewhat in the random forest model, but they remain important to predict demand for electric energy

Figure 11 shows that the results of forecasts closely follow the original series of electricity consumption. With small deviations, our main models are able to follow well the variations between local maximum and minimum of the analyzed series.

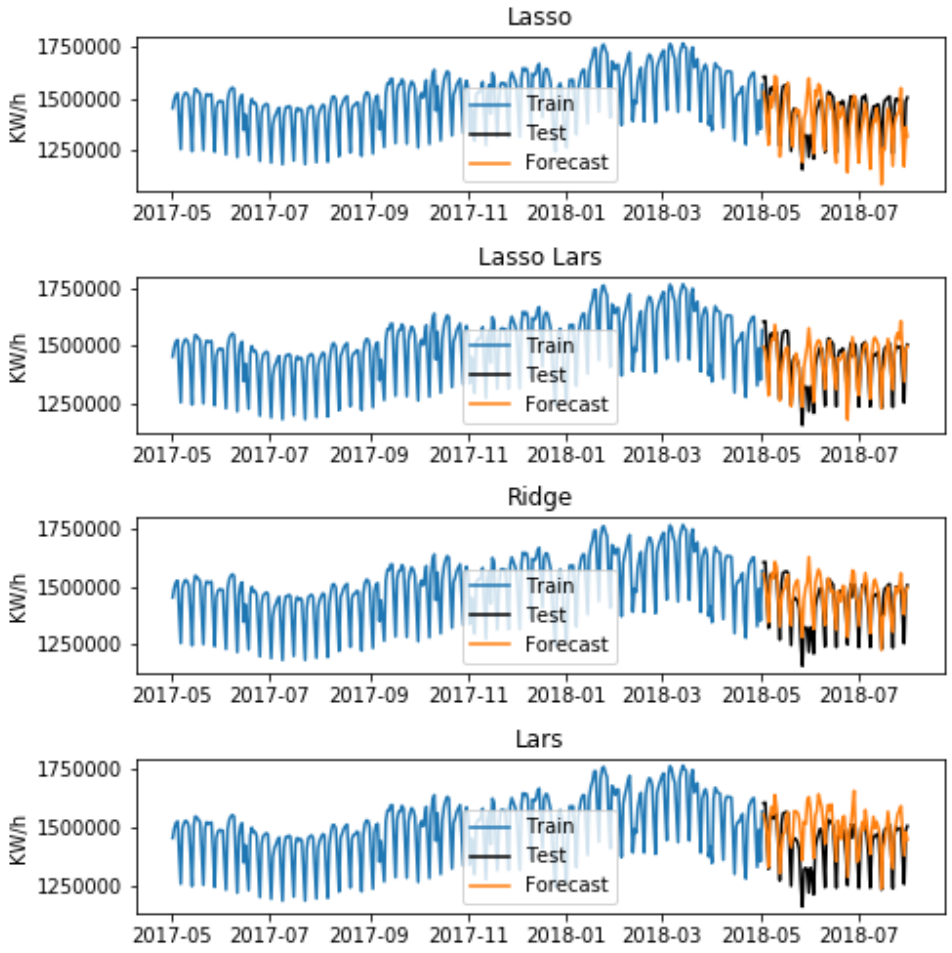

Figure 11 – 30 days forecast

4.2.2 60 Days

Comparing the results of the forecast models from 60-days ahead to 30-days, we observe that the sum of the lagged demand coefficients and calendar variables did not change significantly. That is, the first set decreased significantly until the forecast of 30-days ahead, and remained constant in the next period. The second set increased significantly from 1-day ahead to 30-day head, and also remained constant for the following time horizon.

Also, the meteorological variables that increased a little over the forecast horizons, more than doubling their coefficients sum. This is quite intuitive, because as we increase the time horizon, the variations in meteorological variables increase even more and may become more relevant for the PEC forecast.

On the other hand, prices variables and economic variables were less important for the forecast of 60 days ahead, with emphasis on the large fall in the coefficients sum of economic variables.

Power Eletricity Consumption 60 Days Forecast

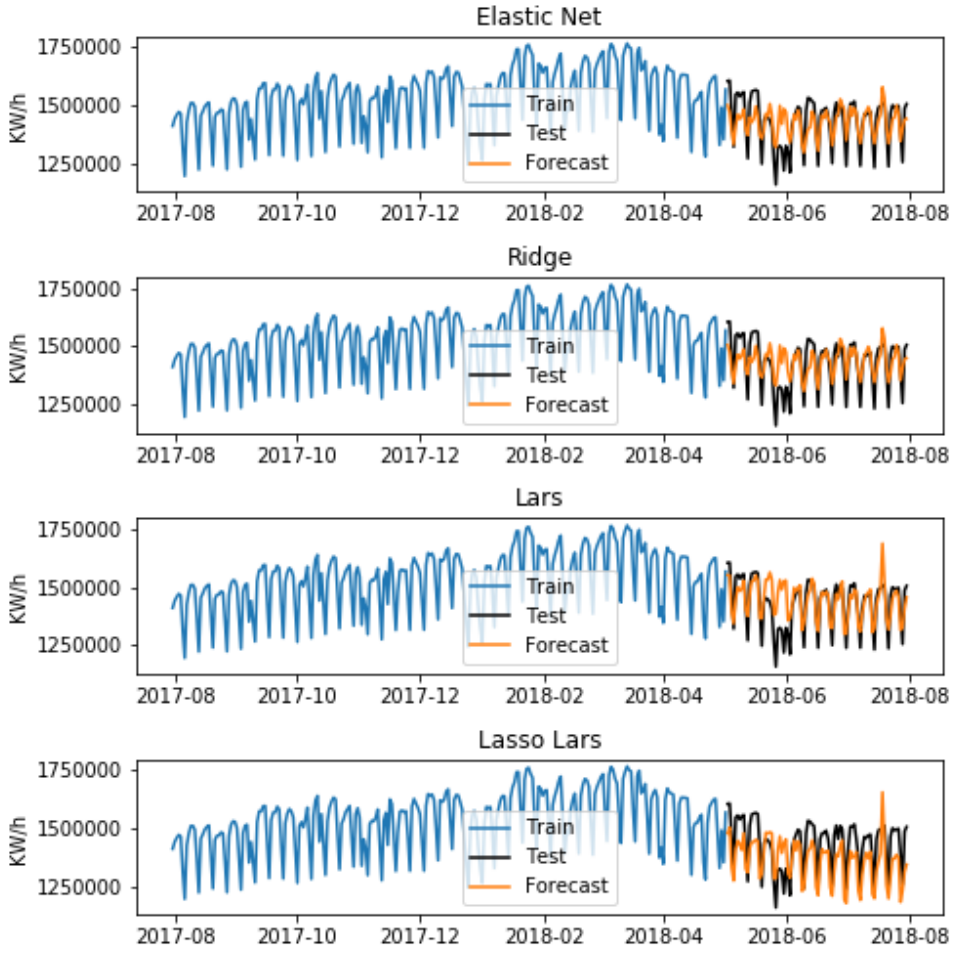

Figure 12 – 60 days forecast

So far, our forecasts have converged almost on values to the original consumption. Now the result are able to catch the trend and cyclicity of the original series. This can be observed in Figure 12. With an excess of Random Forest model, with an increase in the forecast horizon the forecasts of the models are more difficult to reach the local maximums and minimums of the series. This is almost a distribution convergence of our forecasts.

Random Forest stands out from the rest of the models when making predictors that can accompany a wide variation of the consumption of electric energy, even for the longest horizons.

4.2.3 90 Days

Forecasting PEC 90-days ahed, we observe that coefficientes sum did not change significantly, except for a slight decrease in meteorological and price variables, and a modest increase in calendar variables.

Therefore, we infer that the relevance of variables varies more in smaller time horizons and when predicting PEC for 2 and 3 months ahead, they have already stagnated in their relevance.

Power Eletricity Consumption 90 Days Forecast

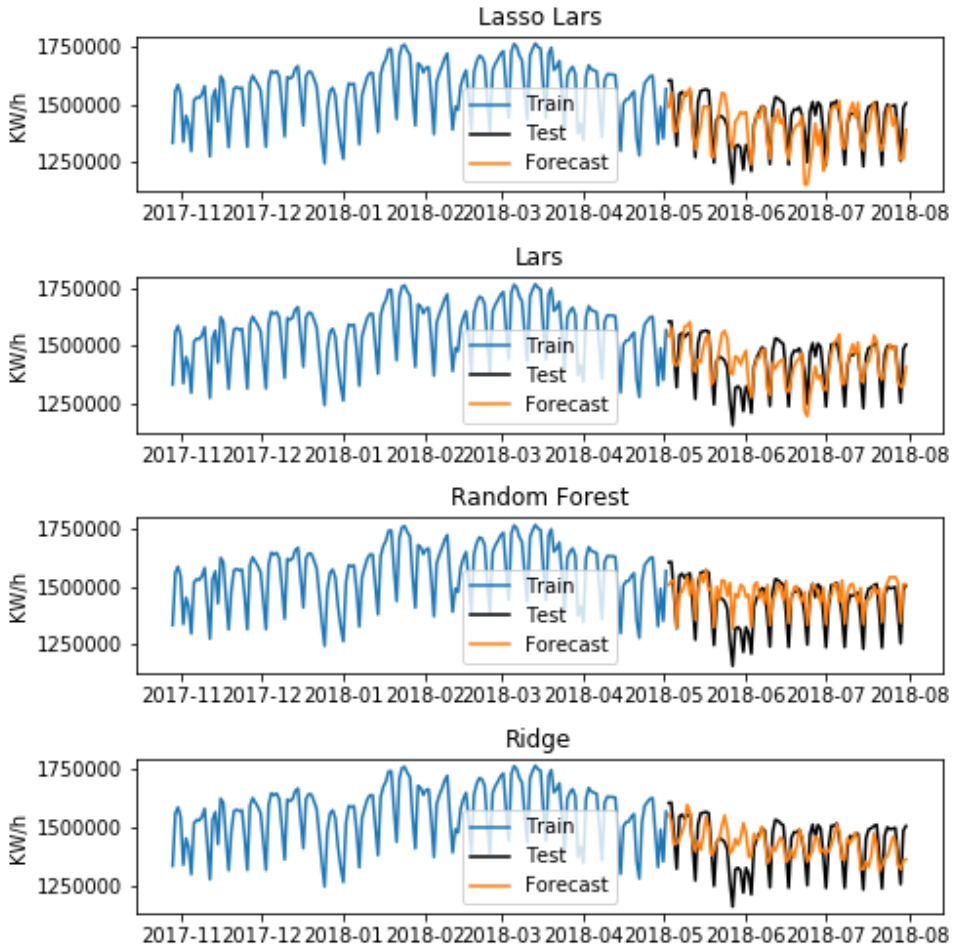

Figure 13 – 90 days forecast

As we observe in figure 13, in the last forecast horizon, the main models have a greater difficulty to follow the series cyclicity, but still, they can follow its trend. Random Forest, although not having the smallest MSE, manages to follow the trend and the seasonality of the series better than the other models.

In this way, if we are making a prediction of 90 days ahead, Random Forest remains the model that best captures the trend and the cyclicity of the series. While the other models can give the trend of electricity consumption for the next 3 months, which is good enough when we want to measure the level of future economic activity.

5 Conclusion

In this work, we build a model to predict power electricity consumption (PEC) for 6 different forecast horizon. These horizons are divided between short term (1, 7 and 15 days) and medium term (30, 60 and 90 days). With GDP released quarterly, we want to structure the forecast models of electricity consumption to capture the trend of economic activity.

We construct a high dimensional database with 5 main sets of variables: (1) lagged demand, (2) calendar variables, (3) meteorological variables, (4) cost of electric energy variables and (5) economic variables. We make predictions for more than 10 different models for all forecast horizons, among them are ARIMA, Random Walk, OLS and machine learning models. We expect the machine learning models perform better than the usual traditional models because they can make a better selection of the relevant variables. The results showed that the five best models are based on machine learning models, especially Random Forest, which presented the lowest mean square error for almost all forecast horizons. This model was able to capture the trend and the seasonality of electricity consumption with good precision.

In addition, we analyze the variation of the sets of variables for all time horizons analyzed. We conclude that for 1-day forecast, lagged demands are the most relevant variables for predicting the consumption of electricity and it loses its importance in the models as we advance in the time horizon until its relevance is stagnated from the medium term horizons. Calendar variables were very relevant for the 7-day forward forecast and have their relevance shows a slight growth until the forecast horizon of 3 months. Yet, the meteorological variables were relevant for medium-term forecast horizons, that is, their relevance grows in the forecast for 30 days and stagnates for the following values.

Finally, price of the power electricity variables and economic variables had a similar trajectory. Both presented a great relevance for the models when forecasting 15 and 30 days ahead, however, they had a drop of relevance in the models for the following horizons of forecast. Our work presents a new model, which uses high dimensional data and machine

learning models to predict power electricity consumption. With the good results of the models, we were able to measure the trend of this consumption and infer about the trend of economic activity in the short term.

In addition, we structure which variables are most relevant to each time horizon, in the short and medium term, which may facilitate the variable selection for future works that predicts power electricity consumption.

Referências