INDUSTRIALES
ETSII | UPM

POLITÉCNICA

END OF DEGREE PROJECT

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES

Forecasting Model for Electricity Demand in Spain

Juan Hugo Pol Moncada

Mentors: Jesús Juan Ruiz

Carlos González Guillen

June 2022

Madrid

ACKNOWLEDGMENTS

I would like to convey my most sincere thanks to all those who have accompanied me not only in the completion of the work, but also throughout my stay at the ETSII.

First, to my tutors Jesús Juan and Carlos González, who have always been available to guide me in the realization of this Final Degree Project.

Secondly, to my classmates, especially Paula Pons Forteza, for her invaluable help throughout these last four years.

Finally, to my family, for their unconditional support.

EXECUTIVE OVERVIEW

The objective of this work is to deepen in the development of time series models (Reg-ARIMA) for the prediction of electricity demand in Spain.

ARIMA models try to express a certain value as a linear function of previous data and errors, being very popular in the field of engineering and statistics when predicting time series. In this work we use Reg-ARIMA models, which also make use of a certain number of regressors to explain the values of the time series.

Therefore, prior to the development of the prediction models, the relationship between two factors such as temperature and the festive nature of the different days of the year with the variation in energy consumption is studied, and the appropriateness of including them as regressors in the models. It seems reasonable to think that the amount of electricity demanded on a summer day, or a working day differs greatly from that demanded on a winter day or a holiday, respectively, thus justifying the study of both factors.

As for temperature, we have data on the maximum and minimum temperatures of 10 autonomous communities distributed throughout the peninsular territory. The study of the (non-linear) relationship between energy consumed and temperature corroborates the importance of including temperature as another regressor. In addition, it is observed that the effect of temperature is not entirely immediate, since the temperature on previous days significantly influences the current day's electricity demand.

On the other hand, the greatest complexity of the work lies in the treatment of holidays or non-working days, since it is on days with such characteristics where the greatest error is made when predicting. In fact, the analysis of previous Reg-ARIMA models, which do not have any regressor referring to holidays, show very high errors on very specific days, mainly coinciding with important national holidays. In addition, a comparison of the energy demanded on a working day and a public holiday shows that the amount of energy consumed on working days is generally much higher than that consumed on public holidays, especially if the public holiday coincides with a normal working day. For these reasons, it is essential to identify holidays using various regressors.

Once the influence of temperature and holidays on electricity demand has been studied, and whether or not to include them as regressors, the model building phase can begin. The construction of a Reg-ARIMA model basically requires two phases: training and validation. On the one hand, during training, data on electricity demand, maximum and minimum temperatures and different holidays in Spain between 2011 and 2018 are used. From these data the model is adjusted, identifying the number of parameters to be used and the value of the respective coefficients. On the other hand, in the validation, the predictions corresponding to the year 2019 are made, comparing the results obtained with the known data, in order to verify that the model has "learned" and not "memorized", and that it has therefore the capacity to generalize.

Throughout the paper it is observed that obtaining the optimal model is a trial and error process, consisting of the progressive elaboration of several previous Reg-ARIMA models with different structures or regressors and the subsequent comparison of the results obtained.

This comparison between errors helps to decide on the need to include or not some additional parameter or regressor, taking into account not only the error made, but also how optimal the model is. Between two models with similar errors, the one with the smaller number of parameters will always be chosen.

The prediction of electric power demand is approached by developing 24 univariate seasonal Reg-ARIMA models, one for each hour of the day. The electric power demand data show a high daily seasonality, i.e., the demand of hour h of day t is highly correlated with the demand of the same hour of the contiguous days. Therefore, one of the reasons for the decision to build 24 models is to take advantage of this correlation. For simplicity, these models will always have the same structure, although the coefficients will take different values.

However, it is impossible to accurately predict the demand for electrical energy, since it is subject to the variability of consumers' daily actions. For this reason, once the optimal model has been obtained, a technique called "hourly refreshment" is developed with the aim of reducing the errors made. The technique consists of studying the correlation between the errors made in predicting the demand for an hour h with the errors made in the demand predictions for the immediately preceding hours. This high correlation invites the construction of 24 linear regression models for the prediction of the errors for all the hours of the day, using as regressors the prediction errors of the previous hours. In this way, two different models are obtained for each hour of the day: a Reg-ARIMA model for the prediction of demand and a linear regression model for the prediction of the error made.

Finally, based on the percentage root mean square error, the errors obtained by using only the Reg-ARIMA models and the errors obtained after applying the hourly refreshing technique are compared. The ECM is calculated for the 24 hours of the day, the 12 months of the year and the 7 days of the week, showing that the errors obtained after applying the hourly refreshment are significantly better than those obtained directly from the Reg-ARIMA models.

Key words: Electricity demand, Reg-ARIMA models, temperature, holidays, hourly refreshment, mean square error.

INDEX

INTRODUCTION

Se It has become an almost daily occurrence for the television news to open the morning newscast with despairing news related to the incessant growth in the price of electricity. A growth that very few consumers understand, as the electricity market is undoubtedly one of the most complex. The rising cost of raw materials, largely caused by the war; the low weight of renewables in the final price or government interventionism based on taxes or "tolls" are just some of the factors that try to explain the various historical records that have been broken this year 2022 in Spain as far as the price of MWh is concerned (545 Eu/MWh this March, specifically). Factors that have little or nothing to do with the ordinary citizen, who keeps wondering what to do to lower the electricity bill. [1]

The Spanish electricity market consists, in a very simplified form, of a series of daily and intraday auctions, where producers and traders launch their offers to sell and buy, respectively. However, how do the large power plants sense the amount of energy they need to produce?

This is where Red Eléctrica Española plays a fundamental role, as they are in charge of predicting the demand for electricity for each hour of the day in Spain. Based on this information, the generators submit their sales offers, and the large retailers submit their respective purchase offers. This process is carried out daily, as the impossibility of storing electrical energy on a large scale means that what is going to be consumed must be produced almost instantaneously. This instantaneousness makes the margin of error minimal and shows the importance of correctly predicting the total amount of electrical energy that needs to be generated to cover the population's demand. On the one hand, there must always be an overproduction to avoid any type of shortage, but at the same time this must be as light as possible, reducing as much as possible the energy produced and wasted or other aspects such as, for example, emissions of polluting gases into the atmosphere. [2]

Therefore, to carry out actions as simple as turning on the light or charging the cell phone, the work of REE to predict the demand for electricity is essential. Throughout the paper we will go into the development of time series models, very similar to those used by REE to carry out this task, in addition to discussing how factors related to temperature or the festive or non-festive nature of the days for which the prediction is made affect.

However, it is impossible to predict electricity consumption with complete accuracy, since it is conditioned by the daily actions of consumers (the time at which they put the dishwasher on, for example). For this reason, it is important not only to develop prediction models that are as accurate as possible, but also to develop techniques that are capable of updating the predictions and minimizing the errors made.

Predicting electricity demand is therefore a task that is as necessary as it is complicated, and this will be reflected in the complexity of the models developed, the number of regressors used and the correction techniques employed.

1. ELECTRICITY DEMAND ANALYSIS

1.1. TIME SERIES

There are mainly two types of time series: univariate and multivariate. However, in this paper only the former are discussed in depth, since the prediction of electricity demand is carried out on the basis of 24 univariate time series models.

The electric power demand data constitute a discrete univariate series, i.e., they are a set of values of a single variable, the demand; obtained at equidistant instants of time and ordered chronologically. In this case, we have a database consisting of hourly electricity demand values from 2011 to 2019. After dividing such data by hours, the 24 univariate time series around which the 24 Reg-ARIMA models are built are obtained. [3]

The fact that the data are collected chronologically indicates that the data are not independent, i.e., what happens at instant t depends on what happened at instant t-1. Therefore, the use of time series models, which are characterized by their ability to detect the dynamics of the data series, is justified.

1.2. STOCHASTIC PROCESS

In spite of the above, throughout this work, time-dependent and time-independent data series, characterized mainly by their randomness, appear recurrently. These are the so-called stochastic processes, whose future values depend on chance.

White noise is a type of stochastic process that is very important in ARIMA models, and is characterized by being a series whose values have zero mean, constant variance and are also uncorrelated. [4]

An example of white noise process is the generation of random numbers from a normal distribution with mean equal to 0 and variance equal to 1.

Figure 1 shows the 1000 random numbers computed.

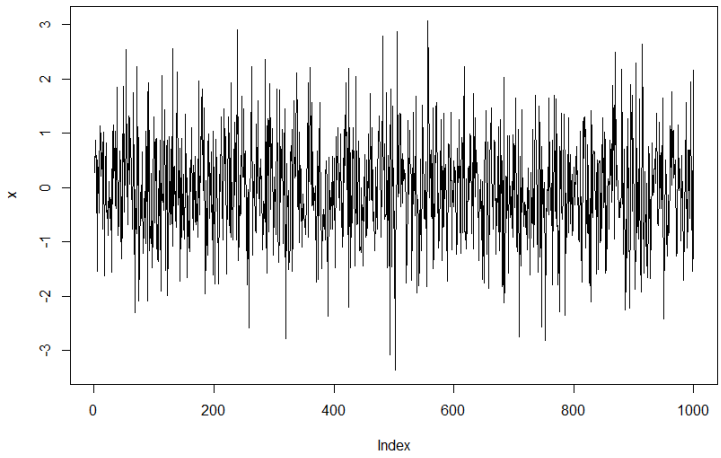

Figure 1- WHITE NOISE

The correlation between the values of a series can be obtained from the autocorrelation function. In this case, the correlation is practically zero. In fact, if it is not zero, it is because the calculated numbers are pseudorandom, the scope of which is beyond the scope of this work. Therefore, this series of pseudo-random numbers fulfills the conditions of having a mean equal to 0, constant variance and zero correlation between their values, thus being an example of white noise.

Figure 2 shows the null correlation between random numbers.

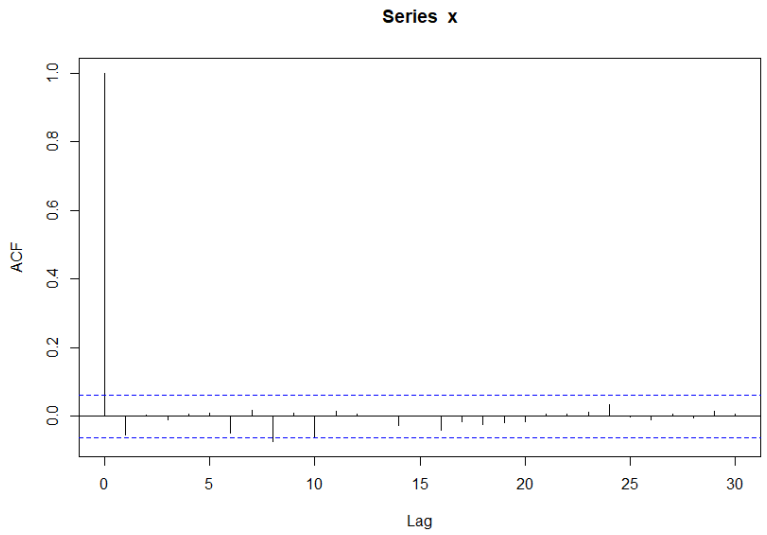

Figure 2- ACF Function

1.3. STATIONARITY OF A SERIE

A process is said to be stationary when the properties of a stochastic process do not vary over time, i.e. when the mean and variance are constant over time and also the correlation between two time instants only depends on the distance between them, not on the instant at which they are measured. [5]

Therefore, the correlation function satisfies that.

$$corr(s,t) = corr(s+h, t+h), \qquad for\ any\ \ s,t\ and\ h$$

Electricity demand is a clear example of a non-stationary series. From Figure 3, which plots the values of electricity demand in Spain between 2011 and 2019, it can be seen that neither the mean nor the variance are constant.

Moreover, the demand series does not meet the third condition of stationarity either, since the correlation between the demand of two instants s and t depends on the time at which they are measured. Figure 4 shows this phenomenon. For example, the correlation between the demand at 01:00 and 07:00 is very different from the correlation between 07:00 and 13:00.

The time dependence of the mean and variance can be solved from differencing and logarithmic transformation techniques, thus eliminating the trend of the series. Figure 5 shows the result of applying these techniques to the demand series.

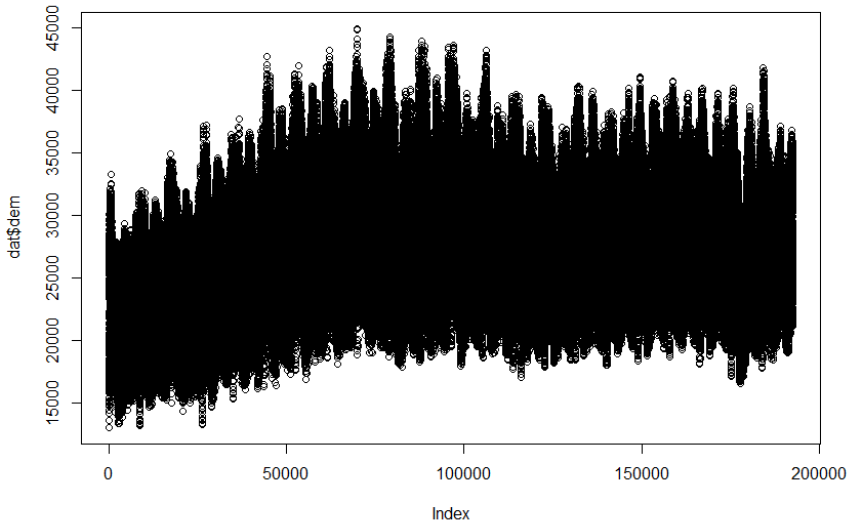

Figure 3- ELECTRICITY DEMAND

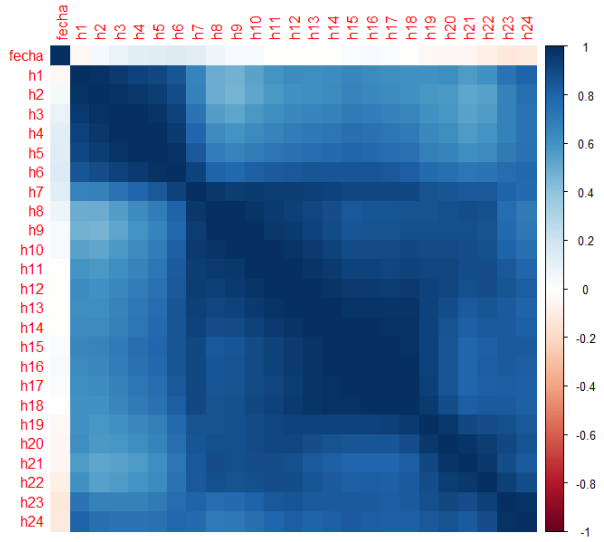

Figure 4- CORRELATION MATRIX

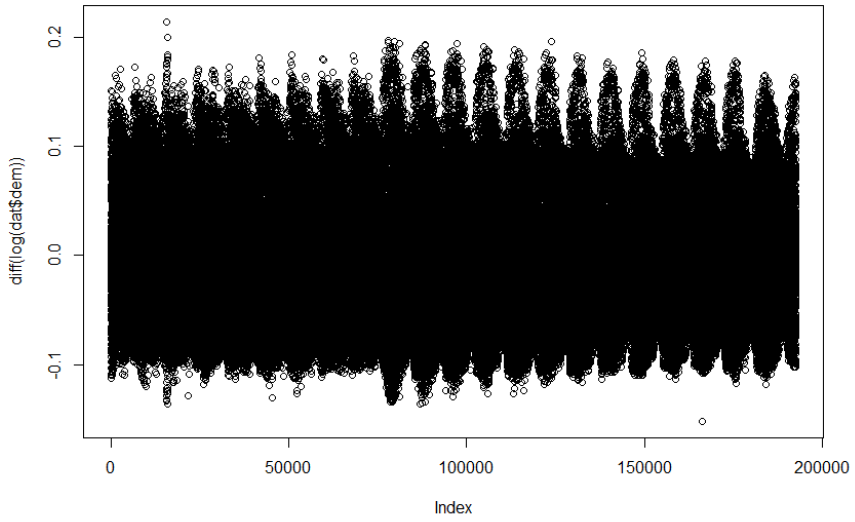

Figure 5- LOG(Xt/Xt-1)

1.4. SEASONALITY OF A SERIES

A series is said to have a seasonal component when it behaves similarly every certain period of time. Analyzing the autocorrelation function of the time series of the demand at 10.00 h we can observe the presence of a weekly seasonal component, i.e., the demand at 10.00 h on day t is highly correlated with the demand at the same time on day t-7. In addition, the autocorrelation function also indicates that there is some correlation between the demand at 10.00 h on nearby days. For example, the demand on a Monday at 10:00 a.m. is correlated with the demand on a

Tuesday at the same time. However, this correlation between nearby days is much lower than the one mentioned above. [6]

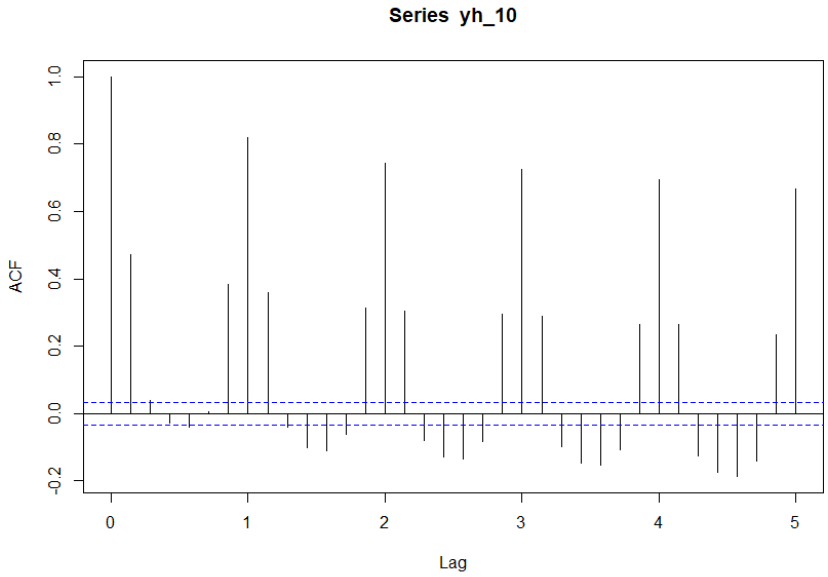

Figure 6- SEASONALITY OF DEMAND

The treatment of seasonality is very important because the presence of a seasonal component leads to the loss of stationarity. As will be seen later, it is necessary to eliminate the seasonal component using differencing techniques in order to work with ARIMA models.

2. INTRODUCTION TO ARIMA MODELS

2.1. INTRODUCTION

The use of linear regression models is often insufficient to detect the dynamics of a time series. Therefore, the introduction of correlation between data as a phenomenon generated from linear lagged relationships induces the use of autoregressive (AR) models, moving average (MA) models and a combination of the two, the autoregressive moving average (ARMA) models. ARMA models can only work with stationary time series, and as explained above, the electricity demand series shows a strong seasonal component, which causes the loss of stationarity. The presence of stationarity justifies the choice of working with seasonal integrated autoregressive moving average models (S-ARIMA), which use differencing techniques in order to eliminate the seasonal component.

2.2. AR(p) MODELS

Autoregressive models are based on the idea that the current value of a time series can be expressed from a linear regression of past values plus an error term.

Thus, an autoregressive model of order p is of the type

$$x_t = \emptyset_1 * x_{t-1} + \emptyset_2 * x_{t-2} + \ldots + \emptyset_p * x_{t-p} + w_t \,,$$

where the series x_t is stationary. The w_t term corresponds to the error term and is assumed to be white noise.

Another way to represent AR models is by means of the B operator, defined as follows,

$$x_{t-1} = B * x_t$$

$$x_{t-2} = B * x_{t-1} = B^2 * x_t$$

$$\ldots$$

$$x_{t-k} = B^k * x_t$$

Therefore, the following expression is obtained,

$$\left(1 - \emptyset_1 * B - \emptyset_2 * B^2 - \ldots - \emptyset_p * B^p\right) * x_t = w_t$$

$$\emptyset(B) * x_t = w_t$$

The use of this notation will appear recurrently throughout the work and, in addition, an important property emerges from it. Analyzing an AR model of order 1, we have that

$$x_t = \emptyset * x_{t-1} + w_t = \emptyset * (\emptyset * x_{t-2} + w_{t-1}) + w_t = \emptyset^2 * x_{t-2} + \emptyset * x_{t-1} + \emptyset^2 * w_{t-1} + w_t$$

Continuing with backward iteration, one can express x_t as follows.

$$x_t = \emptyset^k * x_{t-k} + \sum_{i=0}^{k-1} \emptyset^i * w_{t-i}$$

Therefore, for the process to be stationary it must be fulfilled that |∅|<1, i.e., that the roots of the polynomial ∅(B) are outside the unit circle. Otherwise, the so-called explosive processes appear, where the values of ∅^i grow very rapidly in magnitude and therefore so do those of the series. [7]

2.3. MA(q) MODELS

On the other hand, moving average models consist of defining the values of the time series x_t from the error w_t and the previous errors, assuming that the error is white noise. Thus, MA models are of the type

$$x_t = w_t + \theta_1 * w_{t-1} + \theta_2 * w_{t-2} + \cdots \theta_q * w_{t-q}$$

Using operator B, we obtain

$$x_t = \theta(B) * w_t,$$

where

$$\theta(B) = 1 + \theta_1 * B + \theta_2 * B^2 + \cdots + \theta_q * B^q$$

In this case the moving average processes are stationary for any value of the parameters θ_i, since the error is white noise, the MA processes are the result of adding q+1 stationary processes. [8]

2.4. ARMA(p, q) MODELS

Therefore, the ARMA models consist of a combination of the AR and MA models, defining the stationary series x_t from previous values x_(t-i), with 1≤ i ≤ p; from the current error w_t and previous errors w_(t-j), with 1≤ j ≤ q. The parameter p is the order of the autoregressive part, while parameter q is the order of the moving average part.

Thus, ARMA models are of the type.

$$x_t = \emptyset_1 * x_{t-1} + \emptyset_2 * x_{t-2} + \ldots + \emptyset_p * x_{t-p} + w_t + \theta_1 * w_{t-1} + \theta_2 * w_{t-2} + \cdots \theta_q * w_{t-q}$$

Again using the operator B, we obtain

$$\emptyset(B) * x_t = \theta(B) * w_t$$

The ARMA process will be stationary if the stationarity condition of AR models is met, since the moving average part is always stationary.

It is important to emphasize that ARMA models can only be applied to stationary series. The presence of stationarity in hourly power demand series therefore makes it impossible to use ARMA models directly for their treatment. [9]

2.5. ARIMA(p, d, q) MODELS

As seen above, the use of differencing techniques serves to eliminate the trend of a series, which leads to the loss of stationarity. The idea of ARIMA models consists of differencing the time series at order d, thus obtaining a stationary series, and then applying an ARMA model to the differenced series.

Formally, a process x_t is said to be ARIMA if it is satisfied that the differenced series ⟦(1-B)⟧^d*x_t is an ARMA process.

Therefore, ARIMA models are of the type.

$$\emptyset(B) * (1 - B)^d * x_t = \theta(B) * w_t ,$$

where the term d is the order of differentiation. [10]

2.6. S-ARIMA (p,d,q)x(P,D,Q) MODELS

The treatment of seasonal series requires the use of S-ARIMA models, which introduce a series of modifications to take into account this seasonal component. In seasonal time series it happens that the dependence of x_t on past values tends to intensify for values multiples of a certain lag s. For example, in series with annual seasonality there is a strong dependence on past values multiples of a factor s = 12. For this reason two new operators are introduced, Φ_P (B^s) and Θ_Q (B^s), of order P and Q respectively and of the form.

$$\Phi_P(B^s) = 1 - \Phi_1 * B^s - \Phi_2 * B^{2s} - \cdots - \Phi_P * B^{Ps}$$

$$\Theta_Q(B^s) = 1 + \Theta_1 * B^s + \Theta_2 * B^{2s} + \cdots + \Theta_Q * B^{Qs}$$

On the other hand, the concept of seasonal differentiation is introduced in order to eliminate the seasonal tendency of the hourly electricity demand series. Since the seasonality is weekly, we have a factor s = 7, leaving the differentiation as follows

$$y_t = (1 - B^s)^D * x_t,$$

where y_t is the differentiated series. [11]

Finally, we define the seasonal moving average integrated autoregressive model (SARIMA) of the form.

$$\Phi_P(B^s) * \emptyset(B) * (1 - B^s)^D * (1 - B)^d * x_t = \Theta_Q(B^s) * \theta(B) * w_t$$

2.7. Reg-ARIMA MODELS

In this work, Reg-ARIMA models are used, which also use regressors such as temperature or the holiday nature of the different days of the year to explain the values of the demand for electricity.

A Reg-ARIMA model consists of a linear regression of the series, whose errors follow an S-ARIMA model. Therefore, its structure is as follows

$$y_t = c + \alpha_1 * X_1 + \alpha_2 * X_2 + \cdots + \alpha_n * X_n + error\ ,$$

where

$$\Phi_P(B^s) * \emptyset(B) * (1 - B^s)^D * (1 - B)^d * error = \Theta_Q(B^s) * \ \theta(B) * w_t$$

The parameter c is the constant of the model, while the parameters X_i are the regressors. Therefore, the α_i are the constants that multiply the regressors and whose values must be calculated.

On the other hand, the error follows an S-ARIMA model. The different polynomials that appear have been explained previously, and w_t is the error of the S-ARIMA model and is considered white noise. [12]

3. INFLUENCE OF TEMPERATURE ON ENERGY DEMAND

3.1. QUADRATIC RELATIONSHIP BETWEEN DEMAND AND TEMPERATURE

In order to analyze the relationship between electricity demand and temperature, a database of daily maximum and minimum temperatures for 10 Spanish autonomous communities spread throughout the peninsula is available. Figure 7 shows the logarithm of the 10:00 h demand and the average of the normalized maximum and minimum temperatures, showing how the relationship between both variables is not linear.

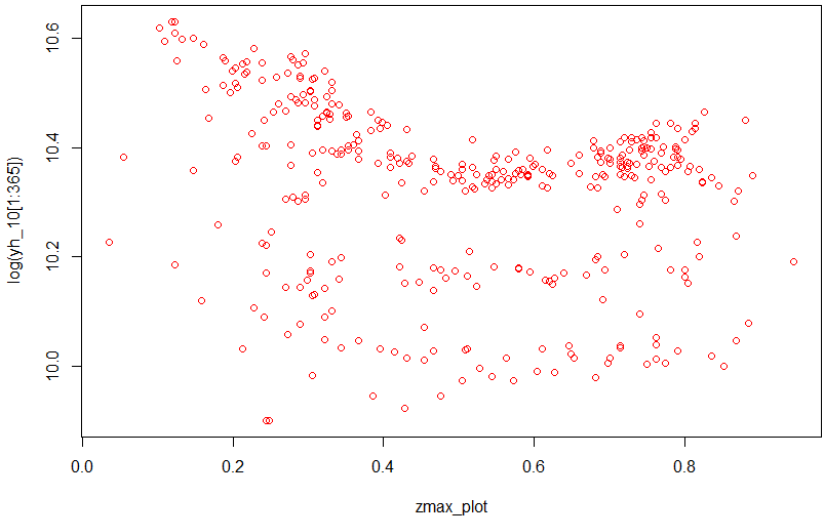

Figura 7- DEMAND ~ MAX TEMP

From the graph it is intuited that the relationship between the electrical energy demand and the temperature can be quadratic. To check this, a regression is performed using a quadratic polynomial, obtaining the following result

$$y_{10} = 10{,}31 - 0.76 * z_{max} + 0.65 * z_{max}^2 + error$$

All the coefficients of the model are significant. Nevertheless, a fairly low R^2 is obtained, around 10%, which is reasonable considering that temperature alone can never explain most of the variability in electricity demand. Assuming that the relationship between temperature and demand is quadratic is a simplification, although as shown in Figure 8 the curve obtained fits the data relatively well.

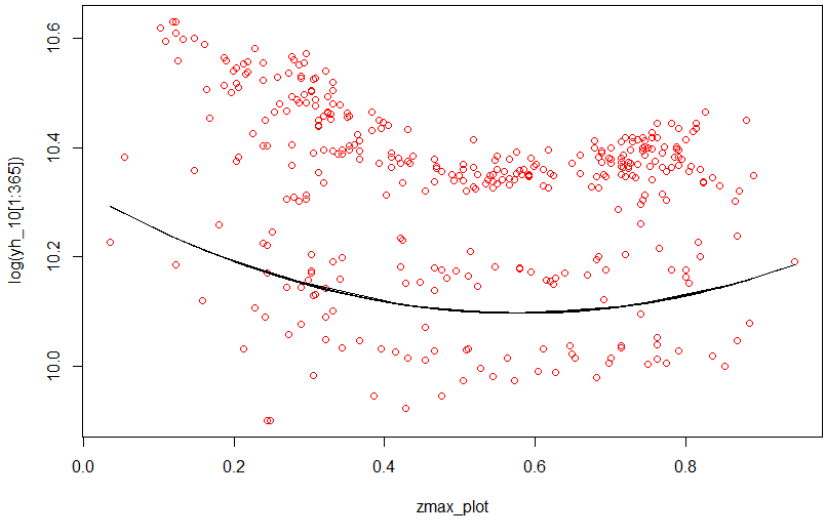

Figure 8- QUADRATIC POLYNOMIAL MAX TEMPERATURE

The same analysis can be performed for the minimum temperature, confirming that the relationship between minimum temperature and electricity demand continues to be quadratic. The results of a regression using a quadratic polynomial between the logarithm of the demand at 10:00 h and the normalized average minimum temperature are very similar to the previous ones. Figure 9 shows how the quadratic polynomial reasonably fits the data.

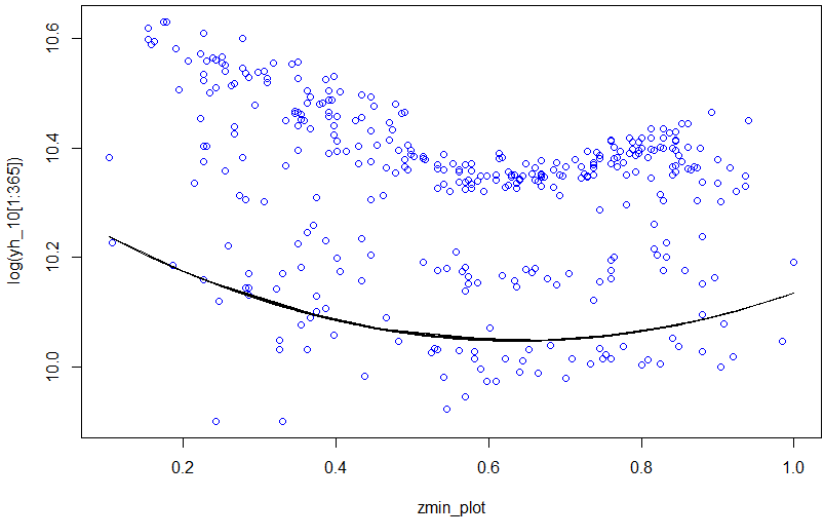

Figure 9- QUADRATIC POLYNOMIAL MIN TEMPERATURE

Therefore, when including temperature as a regressor, not only its value but also its square will be included.

3.2. RELATIONSHIP OF ELECTRIC POWER DEMAND WITH TEMPERATURE IN PREVIOUS DAYS

Next, we study how the electric power demand of day t is not only influenced by the temperature of the day itself, but also by the temperature of the previous days. In fact, the influence of the temperature in the previous days is significantly greater than that of the day itself. One of the reasons for this phenomenon is the thermal inertia of many buildings, which consists of the capacity of certain materials to retain heat and release it gradually. Therefore, it can happen that the internal temperature of certain buildings is highly influenced by the temperatures of previous days. [13]

In order to check that indeed the temperature of previous days influences the electrical energy consumption again, a quadratic regression model between both variables is performed.

First, the influence of the maximum temperature in the 4 previous days is studied, so 4 other models are built. In all cases, the explained variability is higher than that of the previous model. In the case of the maximum temperature on day t-1, the explained variability is 13.5 %, while the explained variability of the maximum temperature on days t-2, t-3 and t-4 is 15.81 %, 14.55 % and 11.3 %, respectively. From these results, the influence of the maximum temperature of the previous days on the energy demand, especially that of days t-2 and t-3, is confirmed.

Figure 10 shows the quadratic relationship between the maximum temperature of the previous days and the demand. It can be seen that the flattest curve, and therefore the one with the least influence on electric energy consumption, is the one corresponding to the maximum temperature on day t.

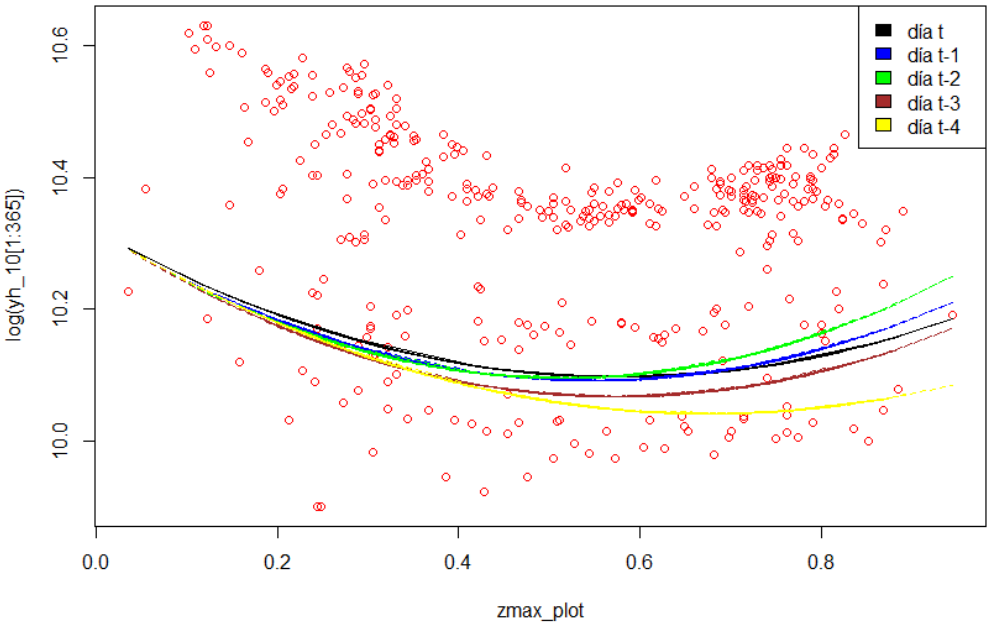

Figure 10- MAX TEMP PREVIOUS DAYS

The same analysis is performed for the influence of the minimum temperature of the previous days, again obtaining similar results. In this case, the variability explained by the minimum temperature on day t-1 is 14.8%, while that of days t-2, t-3 and t-4 is 15%, 13.8% and 13%, respectively. Figure 11 shows these results.

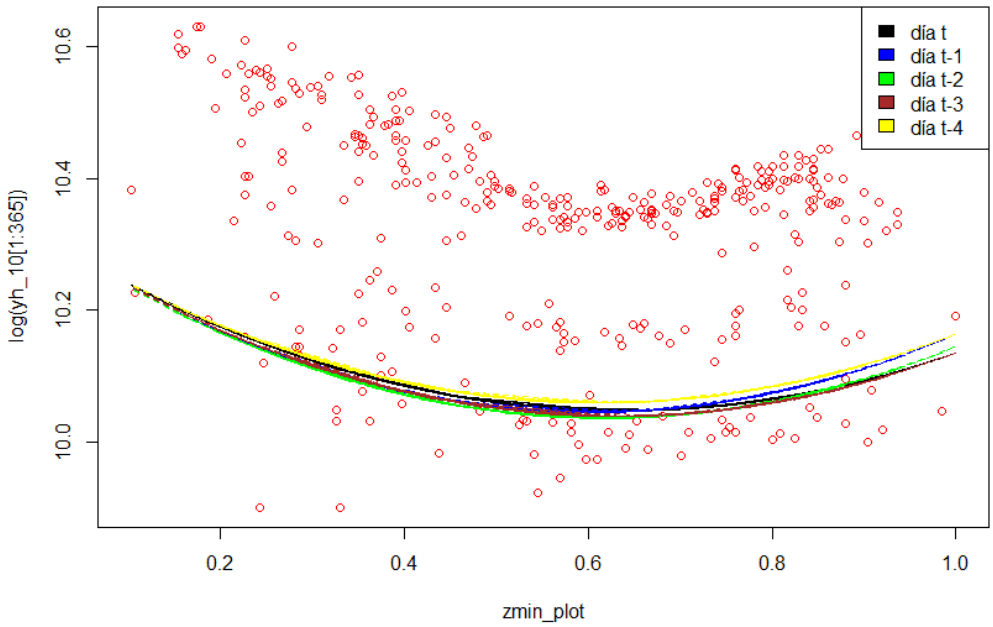

Figure 11- MIN TEMP PREVIOUS DAYS

The influence of temperature on electricity demand and the introduction of the maximum and minimum temperatures and their squares on days t, t-1, t-2, t-3 and t-4 as regressors in the 24 Reg-ARIMA models, totaling 20 regressors for each hourly model, are justified. Since there are 24 models, there will be 480 parameters to estimate related to temperature.

4. INFLUENCE OF HOLIDAYS

4.1 INTRODUCTION

The treatment of holidays is one of the most complex aspects when it comes to predicting electricity demand, since it is on this type of day that electricity consumption shows the greatest variability. The electricity demand shows very different values depending on the day of the week in question. For example, there are significant differences between the energy demanded on Mondays and Fridays.

This effect is intensified on special days or holidays, where the energy demanded generally decreases significantly. In addition, the festive nature of a given day usually affects the demand on the preceding and following days. Holidays or special days range from weekends to national or local holidays.

Therefore, obtaining good predictions for public holidays helps not only to improve the prediction of the day itself, but also to reduce the error made on subsequent days, since, as will be seen later, the error made on the current day is partially carried over to the following days.

In order to evaluate the influence of public holidays, a database of national and local holidays is available. In the case of national holidays, 100% of the Spanish population does not work, while in the case of local holidays only a percentage of the total population enjoys this condition.

Figure 11 shows the average electricity demand at 10:00 am for holidays (national and local) and non-holidays. The average at 10.0 h on holidays is approximately 27,000 MWh, while that of non-holidays is about 33,000 MWh, which represents a decrease of more than 16% in the energy demanded on holidays.

It is important to highlight that this graph compares working days with national and local holidays. The decrease in electricity demand when comparing working days only with national holidays decreases even more, reaching up to 24%, while if compared with local holidays it decreases on average by only 5.7%, something logical considering the lesser influence of local holidays in terms of people working.

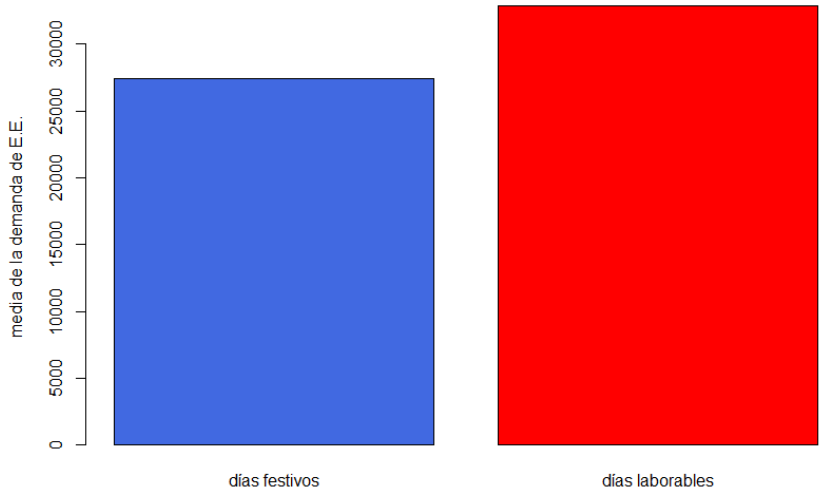

Figure 12- AVERAGE DEMAND ON HOLIDAYS AND WORKING DAYS

In addition, the day of the week on which a holiday falls also modifies the demand profile. On the one hand, if the holiday coincides with a normal working day (any Wednesday, for example), it causes a much lower amount of energy demanded than expected, breaking the weekly seasonality of demand. On the other hand, if the holiday coincides with a Saturday or Sunday, its influence is significantly lower.

Figure 13 shows the average electricity demand at 10:00 am on a working day versus the average energy demanded at the same time on national holidays that coincide with a supposedly working day. The decrease in energy demanded on public holidays is almost 29%.

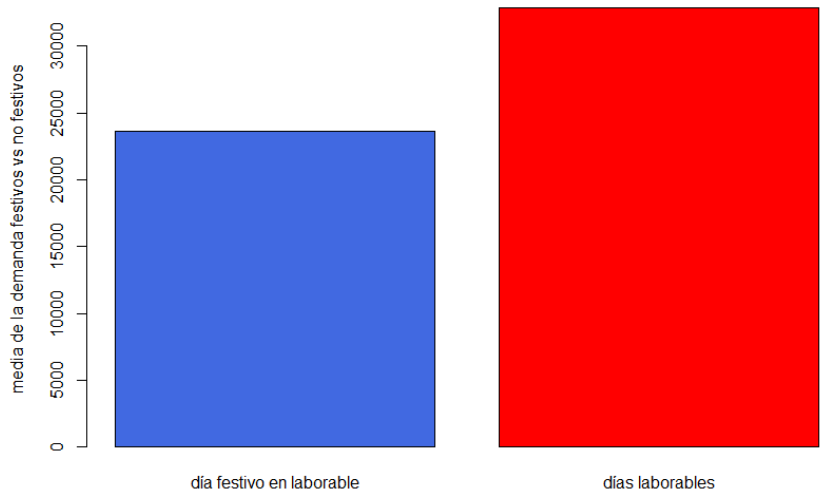

Figure 13- AVERAGE DEMAND HOLIDAYS COINCIDENT WITH WORKING DAYS vs WORKING DAYS

On the other hand, Figure 14 shows the average energy demanded on normal weekends and weekends on which there is a national holiday, showing the lower influence on energy

demand of holidays when they coincide with a weekend day. In this case, demand decreases on average by only 9.4%.

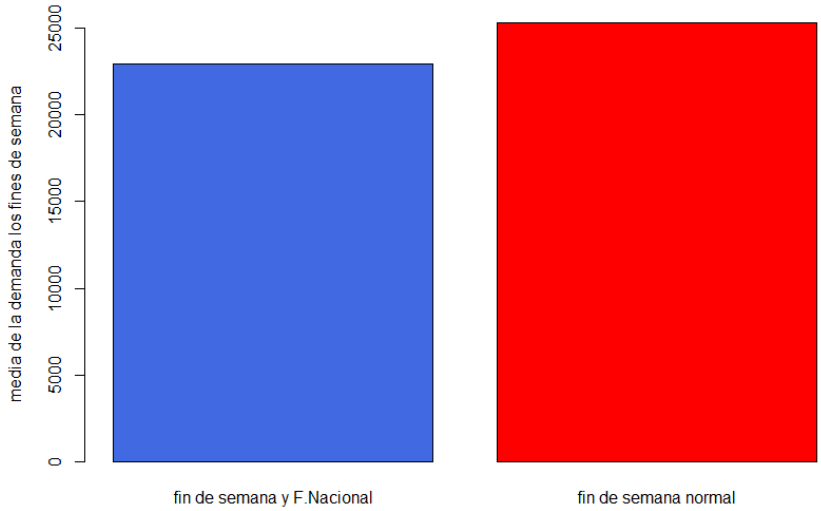

Figura 14- AVERAGE DEMAND HOLIDAYS COINCIDING WITH WEEKEND vs WEEKEND

In this work, January 1 and 6, May 1, August 15, October 12, November 1 and December 6, 25 and 31 are identified as national holidays.

In addition, Thursday and Good Friday share the same national holiday characteristic, although the treatment of these is different, as they never fall on a weekend. The fact that these last two holidays do not fall on a weekend means that the amount of electricity demanded on these days is much lower than that demanded on any Thursday or Friday. Figure 15 shows this difference, which can be as much as 27%.

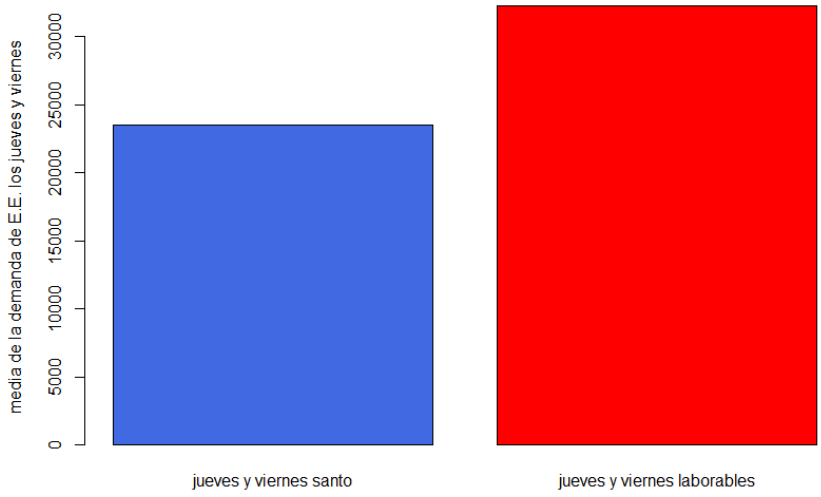

Figure 15- AVERAGE DEMAND ON THURSDAYS AND FRIDAYS

In this section we have studied and verified the influence that national holidays have on electricity demand, especially those that coincide with a working day. Therefore, the inclusion of 9 regressors is justified, each one corresponding to one of the national holidays previously identified. These regressors consist of 9 vectors that take a value equal to 1 on those national holidays that coincide with a working day and a value equal to 0 on the remaining days. The high influence on demand of this type of days motivates us to treat each of them as a single regressor. We also include a tenth regressor referring to national holidays falling on a weekend, which is constructed in the same way as the previous one.

On the other hand, two more regressors are included that try to identify the influence on demand of Holy Thursday and Good Friday. Again, the structure is identical to the previous regressors.

Finally, a last regressor is added referring to local holidays, which takes values between 0 and 1 depending on the percentage of the population affected by the holiday. In this case they are not treated individually because of their lesser influence and because of the large increase in the number of coefficients to be estimated.

Therefore, to model the effect of public holidays, 13 different regressors are used for each model, and therefore 312 regressors in total.

5. CONSTRUCTION OF THE Reg-ARIMA MODEL

5.1. DATA PRE-PROCESSING

Before building any model, it is essential to study the data to be worked with and to fix any outliers that may appear. Moreover, this is a phase of the work in which a lot of time is usually invested, because without consistent data an accurate model can never be obtained.

First of all, several outliers appear in the electricity demand data, mainly related to those days when the time in Spain is changed.

Figure 16 shows the electricity demand data before being modified. The presence of outliers is clearly observed.

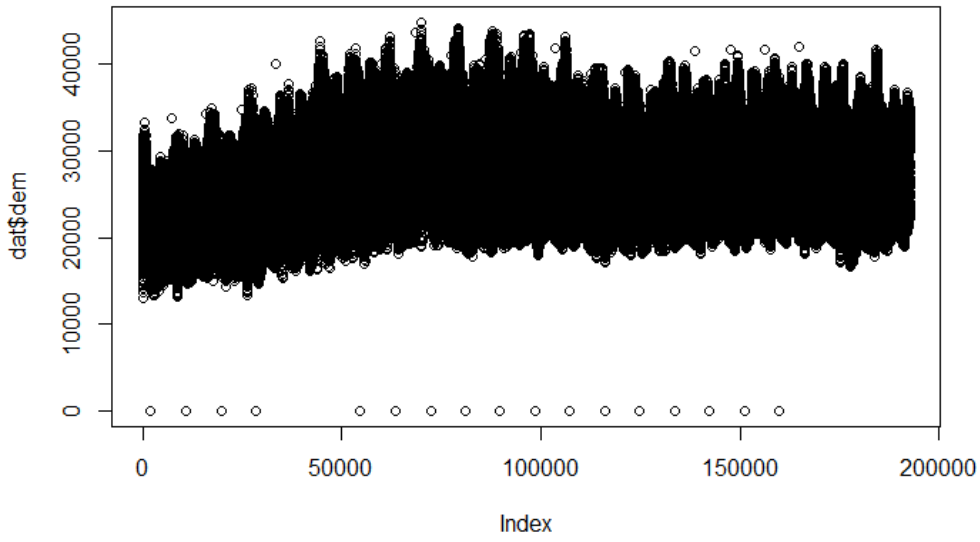

Figure 16- ELECTRICITY DEMAND SPAIN

On the one hand, on the dates generally between March 25 and 31, the time in Spain is advanced to 03.00 h when it is actually 02.00 h. On the other hand, on the days when the time in Spain is advanced to 03.00 h, the time in Spain is 02.00 h. It has been identified that in the 03.00 h time series, the electricity demand data take a value equal to 0 on the days when the time is advanced. The solution taken has been to change this value to that of the demand of the following hour.

On the other hand, between the end of October and the beginning of November, the time in Spain is delayed, being 02.00 h when it would really be 03.00 h. Very high demand values have been detected in the 03.00 h time series, doubling the value of the neighboring hours. In this case, it has been decided to divide its value by two.

Figure 3, which belongs to the chapter on the time series of electricity demand, shows the modified demand for electricity in Spain.

Secondly, the data for maximum and minimum temperatures have been normalized so that the values of this regressor are of the same order as the values of the other regressors related to holidays or outliers, which, as already explained, consist of vectors that take values between 0 and 1.

For example, the normalization of the maximum temperature has been done as follows,

$$z_{max} = \frac{t_{max} - b}{a - b},$$

where

$$a = \max(t_{max})$$

$$b = \min(t_{max})$$

The normalization of the minimum temperature has been done in the same way.

5.2. PRE-MODEL BUILDING

Reaching the final Reg-ARIMA model is a process consisting of several previous models on which different modifications are made in order to improve the errors obtained. In this section we present only some of the previous models that have been used for the prediction of the electricity demand at 10.00 h. The structure of the other 23 models is the same. The structure of the other 23 models is the same, although logically the calculated coefficients take different values.

As explained in previous chapters, the demand data are taken in logarithms to eliminate heteroscedasticity.

One of the parameters used to check the improvement of successive models is the Akaike information criterion (AIC), which evaluates the different models according to their goodness of fit and also their complexity. This criterion is interesting because although the addition of parameters to the model always means an improvement in its standard deviation, however small, it is as important to obtain a model with a low error as an optimal model for performance purposes, and the addition of parameters slows it down significantly. This is why the AIC is useful to know to what extent parameters should be added.

5.2.1. MODEL 1

Model 1 uses the "Auto.Arima" function, which chooses the number of parameters that correctly fit the model. In addition, several regressors are introduced in this model to represent the influence of temperature, such as normalized maximum and minimum temperatures and

their squares of day t. Neither holidays nor temperatures of previous days are included as regressors in this model.

First, the function "Auto.Arima" results in a model of the type (1,0,0,0)x(0,1,1), in which the weekly seasonal difference (s=7) is used to eliminate the trend of the series.

Secondly, we obtain a standard deviation very close to 6% and an AICc equal to -8159.78. Figure 17 shows the presence of predictions in which an error of more than 40% has been made. These predictions coincide mostly with national holidays or neighboring days, which shows the importance of including holidays as a regressor.

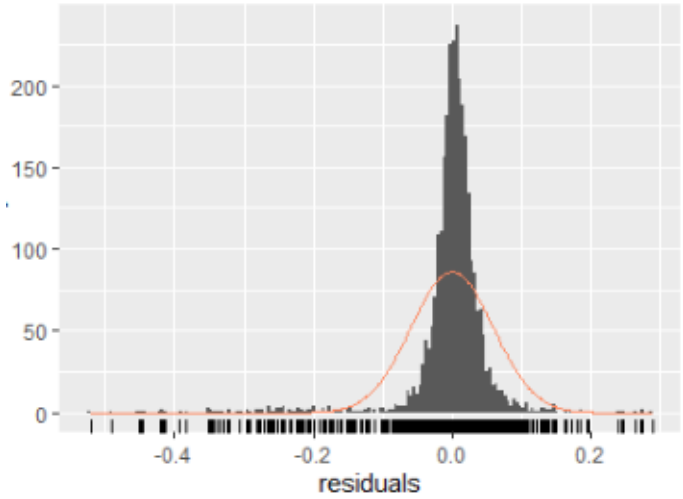

Figure 17- MODEL 1 errors

Table 1 shows the days in 2011 on which the error in predicting electricity demand at 10:00 am is greater than 10%. It can be seen that there are many national holidays, such as October 12, November 1 and December 25. In addition, there are also many days close to national holidays, such as the last days of March, May 2 or October 31.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
1	20110421	20110422	20110425	20110502	20110725	20110815	20111012	20111031	20111101	20111206	20111208	20111225

Table 1 - DAYS WITH ERRORS GREATER THAN 10% IN 2011

The error committed in this model is too high, especially that of the special days, so it is decided to discard the model.

5.2.2. MODEL 2

In model 1 the need to introduce holidays as regressors has been proven. For this reason, model 2 includes a regressor for each national holiday coinciding with a supposedly working

day, a regressor for national holidays coinciding with a weekend day, a regressor for local holidays, and two more regressors for Holy Thursday and Good Friday.

Figure 18 shows that there are practically no predicted values in the training with an error greater than 20%, which indicates a clear improvement with respect to the previous model.

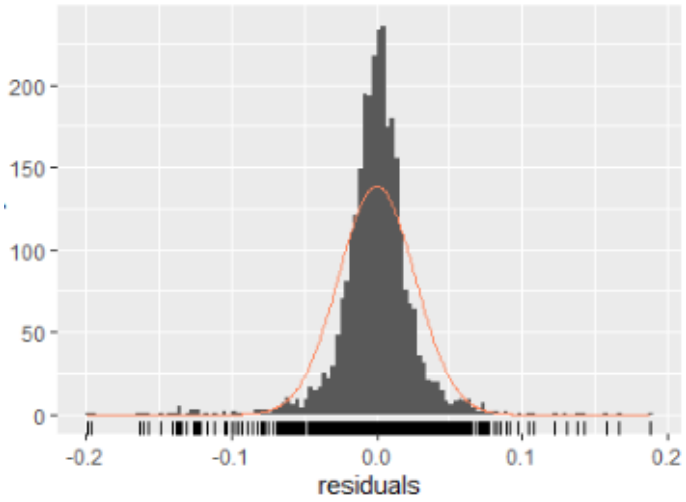

Figura 18- ERRORES MODELO 2

In addition, both the AICc and the standard deviation decrease significantly. An AICc equal to -12895.31 and a standard deviation of 2.63 % are obtained.

Table 2 shows the days in 2011 in which the error in predicting the demand for electricity at 10.00 h is greater than 10 %. It can be seen that they coincide in all cases with national holidays or with days close to them, which may be surprising since in this model special days are included as regressors.

Nevertheless, Table 3 shows that the error committed on these specific dates in model 2 is much lower than that committed in model 1. From this analysis it can be concluded that the inclusion of holidays as regressors significantly reduces the error committed on this type of day, although this error is still clearly higher than that committed on any normal day.

	V1	V2	V3	V4
1	20110815	20111031	20111208	20111225

Table 2- DAYS WITH ERRORS GREATER THAN 10% IN 2011 MODEL 2

	2011/08/15	2011/10/31	2011/12/08	2011/12/25
MODELO 1	34.28674	16.05631	26.55655	16.27392
MODELO 2	11.65887	14.66019	13.42369	12.24932

Table 3 - COMPARISON OF ERRORS BETWEEN MODELS 1 and 2

Therefore, model 2 significantly improves model 1, thus justifying the inclusion of holidays as regressors. The analysis done for the errors made in 2011 can be extended to the rest of the years used to train the model, obtaining similar results.

5.2.3. MODEL 3

Previously, the influence of the temperature of the previous days on the current day's electricity demand has been studied. Therefore, the maximum and minimum temperatures and their squares of days t-1, t-2, t-3 and t-4 are included in model 3, with the objective of verifying that such inclusion results in a better model.

In addition, the autocorrelation function of the previous model is plotted in Figure 19. The presence of a high correlation with lags t-1 motivates the addition of a moving average coefficient, while the high correlation with lags t-7 also justifies the introduction of a seasonal moving average coefficient.

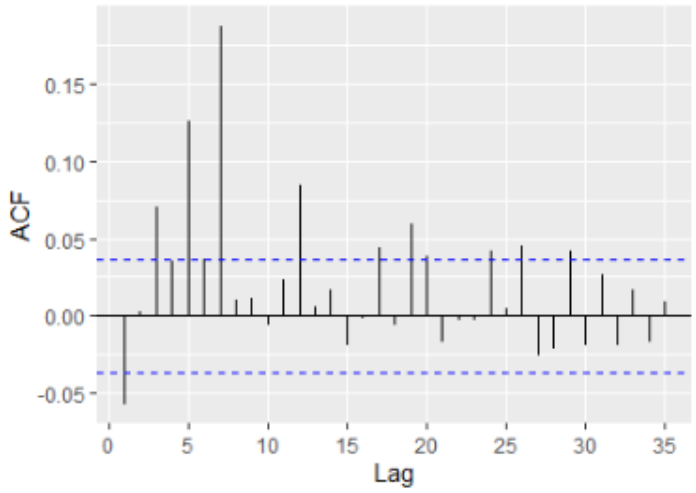

Figure 19- ACF MODEL 3

This model gives an AICc equal to -13298.04 and a standard deviation of 2.44%. Although the AICc has decreased considerably, as can be seen in Figure 20, there are still several predicted values in the training whose error exceeds 10%.

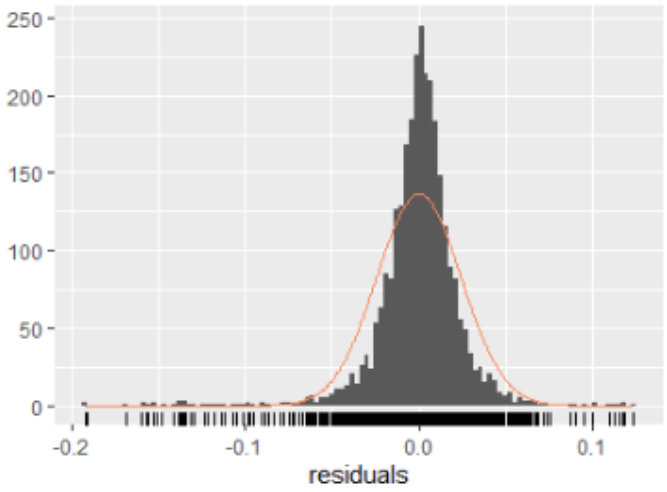

Figure 20- MODEL 3 errors

Studying again the days in 2011 when the predicted values of 10.00 h electricity demand during training exceed 10%, it is revealed that these coincide with those of the previous model. However, Table 4 shows that the errors made on these dates have decreased.

	2011/08/15	2011/10/31	2011/12/08	2011/12/25
MODELO 1	34.28674	16.05631	26.55655	16.27392
MODELO 2	11.65887	14.66019	13.42369	12.24932
MODELO 3	11.40029	13.32157	13.14786	12.33213

Table 4 - COMPARISON OF ERRORS BETWEEN MODELS 1, 2 AND 3

Finally, Figure 21 shows the autocorrelation function of model 3 and shows that the correlation previously existing with lags t-1 and t-7 is no longer significant.

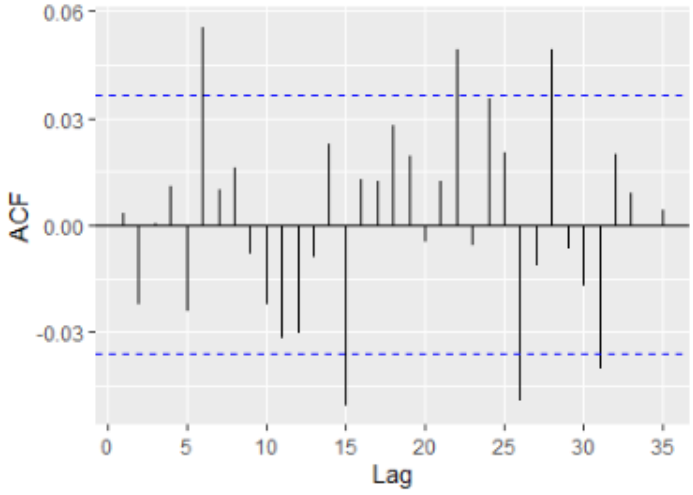

Figure 21- ACF MODEL 3

A significant correlation with the t-6 delay is observed in the ACF. To decrease this correlation, a total of 6 moving average coefficients would have to be used, which is considered

excessive for the small improvement. In fact, the AICc obtained with an ARIMA(1,0,6)x(0,1,2) model is equal to -13303.67, i.e. identical to the previous one, indicating that the model has not improved. When faced with two models giving similar results, the one using fewer parameters is always chosen.

Thus, it is concluded that the addition of the maximum and minimum temperatures and their squares of the previous days and the introduction of a moving average coefficient and a seasonal moving average coefficient result in a better model.

5.2.4. MODEL 4

During pre-modeling and analysis of the errors committed, high errors appear on specific days, many of which do not have any particular characteristics at first sight. Figure 22 shows the errors committed during the training of the 10.00 h Reg-ARIMA model, which has all the regressors explained above. This figure shows the presence of several predicted values with an error between 10 and 20%, which is considered too high. Specifically, there are 26 predicted values out of a total of 2922 whose error is in this range.

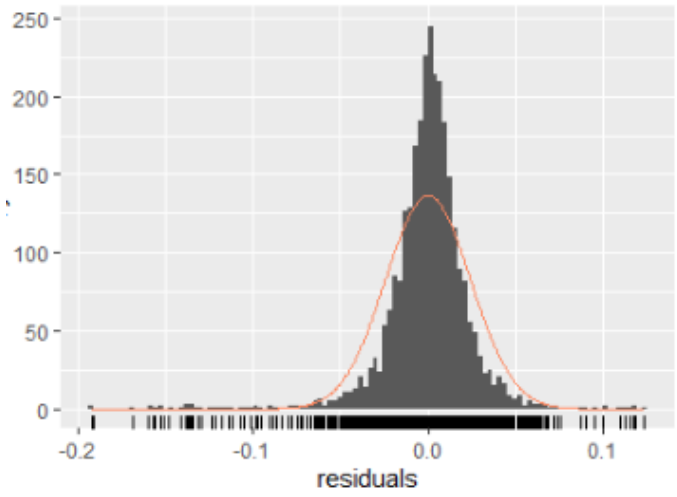

Figure 22- MODEL 4 errors

Table 5 shows those days in which the error committed during the training of model 4 exceeds 10 %. Analyzing the outlier days in 2012, several conclusions can be drawn.

First, there are days corresponding to national holidays, such as January 1 or December 25, which indicates how difficult it is to correctly predict the demand for electricity on this type of special days. Secondly, there are days that are very close to national holidays, such as April 30 or December 24, which shows the effect of national holidays on the electricity demanded on neighboring days. Finally, there are days that do not seem to have any special characteristics, such as November 14.

	2011	2012	2013	2014	2015	2016	2017	2018
1	20110815	20120101	20131224	20140502	20151208	20161031	20170101	20180430
2	20111031	20120329	0	20141208	20151224	20161208	20171013	20181207
3	20111208	20120430	0	20141224	0	20161225	20171225	20181224
4	20111225	20121114	0	0	0	0	0	0
5	0	20121206	0	0	0	0	0	0
6	0	20121224	0	0	0	0	0	0
7	0	20121225	0	0	0	0	0	0

Table 5 - ATYPICAL DAYS ON DEMAND FOR 10.00 a.m.

The presence of atypical days occurs in a similar way for the rest of the hours of the day, so the inclusion of a new regressor that helps to improve its prediction is justified. The error committed in the predictions of electricity demand on the dates of national holidays and the upcoming days decreases thanks to the inclusion of the regressor for atypical days, which consists of a vector that takes a value equal to 1 on those days where the error is greater than 10% and a value equal to 0 on the rest of the days.

Figure 23 shows the errors committed during the training of a model equal to the previous one, to which the regressor for outlier days has been added. It can be seen that the presence of predicted values with an error greater than 10% is much lower.

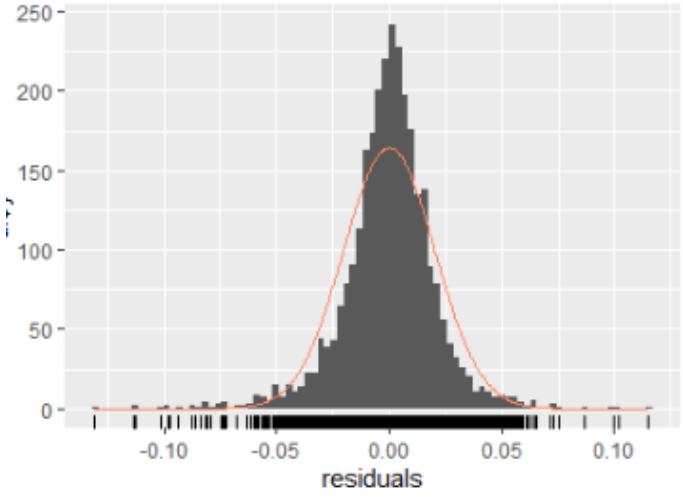

Figure 23- ERRORS in the TRAINING in the DEFINITIVE MODEL at 10.00 a.m.

Table 6 shows a table with the days on which the error is greater than 10%. From the table, it can be concluded that the day on which it is most difficult to predict the demand for electricity at 10:00 am is December 26, probably because it is so close to such important dates as December 24 and 25. Once again, the influence of national holidays on the demand of neighboring days is observed.

	2011	2012	2013	2014	2015	2016	2017	2018
1	0	20121226	20131225	0	0	0	20171226	20181226

Table 6 - ATYPICAL DAYS ON DEMAND FOR 10.00 a.m.

On the other hand, the inclusion of the outlier days as a regressor provides a model with an AICc equal to -14357.26 and with a standard deviation of 2%, significantly improving any of the previous models.

Due to the decrease in the number of days where the error exceeds 10% and the obtaining of a very low standard deviation and AICc, model 4 is adopted as definitive.

Therefore, a Reg-ARIMA(1,0,2)x(0,1,2) model is obtained with 34 regressors: 20 to represent the maximum and minimum temperatures and their squares for days t, t-1, t-2, t-3 and t-4; 13 more regressors to represent the national and local holidays and a last regressor of outlier vectors. In the next chapter, its structure is discussed in more detail.

6. STRUCTURE OF THE REG-ARIMA MODEL

Therefore, the final hourly model consists of a Reg-ARIMA (1,0,2)x(0,1,2)(s=7) model, with 34 regressors. The structure of any hourly model is of the following type,

$$y_h = c + * \alpha_h^T * Z_t + \beta_h^T * F_t + \gamma_h^T * A_t + error_h \,,$$

$$\emptyset(B) * (1 - B^7)^1 * error_h = \Theta_Q(B^7) * \theta(B) * w_t \,,$$

where

- The parameter c is the constant of the hourly model, which takes a different value for each of the models.
- Z_t is a vector that includes the 20 regressors related to temperature and is identical for all 24 models.

$$Z_t = (z_{máx}^t, z_{min}^t, z_{máx}^{t-1}, z_{min}^{t-1}, z_{máx}^{t-2}, z_{min}^{t-2}, z_{máx}^{t-3}, z_{min}^{t-3}, z_{máx}^{t-4}, z_{min}^{t-4},$$
$$z_{máx}^{t\ 2}, z_{min}^{t\ 2}, z_{máx}^{t-1\ 2}, z_{min}^{t-1\ 2}, z_{máx}^{t-2\ 2}, z_{min}^{t-2\ 2}, z_{máx}^{t-3\ 2}, z_{min}^{t-3\ 2}, z_{máx}^{t-4\ 2}, z_{min}^{t-4\ 2})$$

- α_h^T is a vector of the same dimension as Z_t, which consists of the parameters to be estimated that multiply the temperature regressors. They take different values for each model.

$$\alpha_h^T = (\alpha_1^h, \alpha_2^h, \alpha_3^h, \alpha_4^h, \alpha_5^h, \alpha_6^h, \alpha_7^h, \alpha_8^h, \alpha_9^h, \alpha_{10}^h, \alpha_{11}^h, \alpha_{12}^h, \alpha_{13}^h, \alpha_{14}^h, \alpha_{15}^h, \alpha_{16}^h, \alpha_{17}^h, \alpha_{18}^h, \alpha_{19}^h, \alpha_{20}^h)$$

- F_t is a vector that includes the 13 regressors related to the special days and is again identical for all 24 models.

$$F_t = (I_{01Enero}, I_{06Enero}, I_{01May}, I_{15Ag}, I_{12Oct}, I_{01Nov}, I_{06Dic}, I_{25Dic}, I_{31Dic},$$
$$FestNacFinde, FestLocal, JuevesSS, ViernesSS)$$

- β_h^T is a vector of equal dimension to Ft, which includes the parameters to be computed that multiply the different regressors of the holidays and is different for each model

$$\beta_h^T = (\beta_1^h, \beta_2^h, \beta_3^h, \beta_4^h, \beta_5^h, \beta_6^h, \beta_7^h, \beta_8^h, \beta_9^h, \beta_{10}^h, \beta_{11}^h,$$
$$\beta_{12}^h, \beta_{13}^h, \beta_{14}^h, \beta_{15}^h, \beta_{16}^h, \beta_{17}^h, \beta_{18}^h, \beta_{19}^h, \beta_{20}^h)$$

- A_t represents the regressor for outlier days and is the same for each model.

- γ_h^T is the parameter that multiplies the regressor of the outlier days and is different for each model.

- $error_h$ is the error of the linear regression of the demand following a model S-ARIMA

- $\emptyset(B) = (1 - \emptyset_1 * B)$ and is the autoregressive polynomial of the model

- $\theta(B) = 1 + \theta_1 * B + \theta_2 * B^2$ and is the moving average polynomial of the model
- $\Theta_Q(B^7) = 1 + \Theta_1 * B^7 + \Theta_2 * B^{14}$ y is the seasonal moving average polynomial (s=7)
- $(1 - B^7)^1$ indicates the weekly seasonal difference applied to the data.
- w_t S-ARIMA model error considered as white noise

The result is a Reg-ARIMA model (1,0,2)x$(0,1,2)_{s=7}$ with 6 parameters and 34 regressors, so there are a total of 39 coefficients to be estimated per model (the seasonal difference is not accompanied by any coefficient). Considering that there are 24 models, the total number of coefficients to be estimated amounts to 936.

Table 7 shows the estimated coefficients of the seasonal ARIMA model.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24
AR 1	0.915	0.912	0.906	0.902	0.901	0.899	0.897	0.891	0.886	0.905	0.907	0.885	0.876	0.873	0.870	0.884	0.878	0.896	0.941	0.948	0.939	0.933	0.949	0.950
MA 1	-0.049	-0.074	-0.040	-0.009	0.034	0.067	-0.014	-0.061	-0.141	-0.161	-0.187	-0.216	-0.222	-0.228	-0.211	-0.231	-0.213	-0.229	-0.233	-0.177	-0.108	-0.166	-0.162	-0.155
MA 2	-0.219	-0.197	-0.193	-0.193	-0.194	-0.170	-0.121	-0.090	-0.041	-0.049	-0.056	-0.041	-0.043	-0.039	-0.066	-0.064	-0.071	-0.079	-0.101	-0.096	-0.104	-0.096	-0.124	-0.119
S-MA 1	-0.819	-0.842	-0.858	-0.854	-0.856	-0.846	-0.771	-0.592	-0.619	-0.798	-0.828	-0.773	-0.778	-0.775	-0.768	-0.830	-0.820	-0.808	-0.738	-0.709	-0.703	-0.786	-0.918	-0.853
S-MA 2	-0.157	-0.134	-0.120	-0.125	-0.124	-0.131	-0.208	-0.206	-0.138	-0.156	-0.125	-0.132	-0.102	-0.081	-0.104	-0.132	-0.141	-0.151	-0.053	-0.056	-0.061	-0.089	-0.066	-0.125

Table 7 – COEFFICIENTS OF THE SEASONAL ARIMA MODEL

Table 8 shows the coefficients that multiply public holidays.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24
l_01Ene	-0.001	0.051	0.053	0.027	0.003	-0.037	-0.134	-0.237	-0.323	-0.371	-0.340	-0.290	-0.253	-0.231	-0.202	-0.217	-0.240	-0.231	-0.204	-0.180	-0.155	-0.132	-0.107	-0.099
l_06Ene	0.010	0.008	-0.007	-0.021	-0.027	-0.023	-0.101	-0.172	-0.233	-0.232	-0.201	-0.175	-0.173	-0.170	-0.156	-0.171	-0.186	-0.180	-0.152	-0.120	-0.109	-0.091	-0.045	-0.043
l_01May	0.018	0.014	0.013	0.001	-0.003	0.007	-0.122	-0.228	-0.249	-0.221	-0.182	-0.161	-0.161	-0.150	-0.128	-0.153	-0.181	-0.184	-0.181	-0.171	-0.153	-0.099	-0.073	-0.065
l_15Ago	0.006	0.005	0.008	0.003	0.003	-0.007	-0.036	-0.076	-0.092	-0.094	-0.089	-0.101	-0.106	-0.105	-0.084	-0.071	-0.087	-0.088	-0.095	-0.087	-0.077	-0.053	-0.033	-0.021
l_12Oct	0.026	0.022	0.020	0.015	0.013	-0.013	-0.083	-0.162	-0.186	-0.157	-0.119	-0.101	-0.109	-0.103	-0.088	-0.104	-0.131	-0.135	-0.131	-0.118	-0.093	-0.069	-0.049	-0.043
l_01Nov	0.028	0.025	0.017	0.007	0.002	-0.018	-0.072	-0.183	-0.202	-0.151	-0.122	-0.113	-0.125	-0.119	-0.084	-0.093	-0.120	-0.123	-0.105	-0.090	-0.085	-0.060	-0.045	-0.038
l_06Dic	0.031	0.026	0.020	0.011	0.009	-0.010	-0.053	-0.118	-0.141	-0.109	-0.059	-0.038	-0.037	-0.051	-0.037	-0.050	-0.072	-0.076	-0.061	-0.055	-0.041	-0.031	-0.020	-0.017
l_25Dic	-0.044	-0.005	-0.006	-0.032	-0.044	-0.022	-0.164	-0.238	-0.300	-0.302	-0.259	-0.207	-0.197	-0.190	-0.199	-0.235	-0.250	-0.216	-0.222	-0.192	-0.149	-0.128	-0.109	-0.098
l_31Dic	0.024	0.011	0.006	0.001	-0.005	-0.020	-0.078	-0.119	-0.141	-0.119	-0.102	-0.103	-0.110	-0.117	-0.119	-0.130	-0.139	-0.139	-0.110	-0.089	-0.113	-0.157	-0.185	-0.138
Viernes_SS	-0.022	-0.028	-0.036	-0.034	-0.035	-0.053	-0.111	-0.183	-0.212	-0.187	-0.153	-0.137	-0.143	-0.139	-0.126	-0.130	-0.151	-0.154	-0.150	-0.140	-0.125	-0.106	-0.084	-0.069
Jueves_SS	0.015	0.021	0.018	0.018	0.011	0.004	0.011	0.013	0.025	0.034	0.038	0.029	0.028	0.015	0.003	0.001	-0.009	-0.020	-0.017	-0.023	-0.038	-0.050	-0.044	-0.029
FiestasNaciFinde	0.007	0.009	0.004	0.001	-0.008	-0.017	-0.046	-0.082	-0.116	-0.108	-0.106	-0.102	-0.094	-0.086	-0.083	-0.095	-0.090	-0.096	-0.082	-0.068	-0.053	-0.043	-0.024	-0.021
FiestasLocales	-0.010	-0.023	-0.031	-0.041	-0.047	-0.063	-0.141	-0.234	-0.296	-0.277	-0.246	-0.214	-0.212	-0.190	-0.167	-0.196	-0.209	-0.208	-0.199	-0.169	-0.128	-0.095	-0.060	-0.056
Atipicos	-0.093	-0.063	-0.065	-0.051	-0.050	-0.048	-0.095	-0.112	-0.130	-0.120	-0.116	-0.111	-0.105	-0.102	-0.128	-0.121	-0.126	-0.118	-0.111	-0.100	-0.111	-0.142	-0.187	-0.140

Table 8 – COEFFICIENTS OF THE REGRESSORS FOR THE HOLIDAYS

Figure 9 shows the coefficients that multiply the regressors related to temperature.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24
Zmax t	-0.085	-0.067	-0.067	-0.056	-0.055	-0.057	-0.045	-0.043	-0.079	-0.165	-0.290	-0.406	-0.497	-0.572	-0.611	-0.628	-0.641	-0.649	-0.590	-0.527	-0.466	-0.428	-0.413	-0.406
(Zmax t)^2	0.092	0.069	0.063	0.051	0.049	0.046	0.025	0.017	0.047	0.115	0.222	0.325	0.419	0.502	0.565	0.599	0.625	0.635	0.595	0.535	0.471	0.427	0.426	0.425
Zmax t-1	-0.291	-0.252	-0.218	-0.198	-0.181	-0.172	-0.162	-0.147	-0.177	-0.252	-0.289	-0.269	-0.250	-0.220	-0.219	-0.230	-0.217	-0.190	-0.163	-0.137	-0.146	-0.148	-0.145	-0.137
(Zmax t-1)^2	0.310	0.280	0.243	0.220	0.203	0.190	0.172	0.157	0.171	0.229	0.263	0.248	0.239	0.216	0.210	0.223	0.219	0.198	0.173	0.147	0.145	0.125	0.107	0.101
Zmax t-2	-0.146	-0.122	-0.113	-0.111	-0.109	-0.125	-0.123	-0.114	-0.122	-0.120	-0.101	-0.096	-0.092	-0.096	-0.100	-0.097	-0.116	-0.107	-0.088	-0.103	-0.126	-0.119	-0.121	-0.118
(Zmax t-2)^2	0.108	0.091	0.092	0.096	0.094	0.101	0.106	0.102	0.102	0.101	0.077	0.075	0.073	0.081	0.090	0.091	0.106	0.102	0.091	0.100	0.111	0.109	0.106	0.105
Zmax t-3	-0.134	-0.116	-0.111	-0.107	-0.112	-0.108	-0.105	-0.112	-0.140	-0.164	-0.148	-0.112	-0.100	-0.105	-0.109	-0.112	-0.104	-0.099	-0.088	-0.083	-0.064	-0.058	-0.070	-0.069
(Zmax t-3)^2	0.115	0.100	0.095	0.093	0.094	0.093	0.085	0.091	0.116	0.143	0.134	0.095	0.086	0.088	0.092	0.091	0.084	0.082	0.076	0.071	0.051	0.039	0.045	0.046
Zmax t-4	-0.060	-0.053	-0.057	-0.056	-0.057	-0.054	-0.036	-0.012	-0.012	-0.029	-0.016	-0.009	0.004	0.009	0.009	-0.007	-0.017	-0.012	-0.007	-0.004	-0.023	-0.028	-0.018	-0.024
(Zmax t-4)^2	0.047	0.052	0.057	0.051	0.055	0.052	0.036	0.007	0.004	0.019	0.011	0.009	-0.002	-0.004	-0.001	0.011	0.022	0.018	0.019	0.013	0.030	0.037	0.026	0.029
Zmin t	-0.098	-0.122	-0.130	-0.150	-0.155	-0.150	-0.146	-0.185	-0.208	-0.192	-0.127	-0.064	-0.005	0.029	0.044	0.033	0.028	0.036	-0.026	-0.061	-0.088	-0.088	-0.109	-0.101
(Zmin t)^2	0.103	0.127	0.145	0.158	0.163	0.161	0.152	0.177	0.208	0.210	0.182	0.153	0.113	0.091	0.072	0.070	0.061	0.049	0.076	0.092	0.100	0.093	0.109	0.106
Zmin t-1	-0.093	-0.079	-0.073	-0.053	-0.049	-0.046	-0.059	-0.061	-0.057	-0.047	-0.056	-0.088	-0.096	-0.112	-0.139	-0.141	-0.149	-0.160	-0.119	-0.097	-0.041	-0.035	-0.046	-0.036
(Zmin t-1)^2	0.108	0.098	0.082	0.066	0.057	0.052	0.055	0.050	0.040	0.031	0.044	0.065	0.072	0.079	0.105	0.102	0.108	0.104	0.074	0.050	0.009	0.011	0.031	0.025
Zmin t-2	-0.010	-0.028	-0.024	-0.029	-0.017	-0.010	-0.010	-0.004	-0.010	0.007	-0.017	-0.027	-0.035	-0.047	-0.030	-0.039	-0.029	-0.026	-0.035	-0.023	-0.032	-0.027	-0.023	-0.022
(Zmin t-2)^2	0.014	0.026	0.027	0.031	0.022	0.012	0.005	-0.008	0.002	-0.024	0.000	0.017	0.028	0.045	0.030	0.037	0.030	0.030	0.034	0.031	0.038	0.027	0.023	0.025
Zmin t-3	-0.021	-0.016	-0.008	-0.005	-0.005	0.001	-0.008	-0.013	0.014	0.029	0.022	-0.009	-0.010	0.003	-0.010	-0.007	-0.018	-0.019	-0.009	-0.029	-0.028	-0.017	-0.023	-0.025
(Zmin t-3)^2	0.017	0.013	0.000	-0.005	-0.006	-0.012	-0.007	-0.007	-0.030	-0.046	-0.044	-0.017	-0.012	-0.026	-0.013	-0.018	-0.010	-0.013	-0.026	-0.012	-0.013	-0.009	0.003	0.001
Zmin t-4	-0.051	-0.035	-0.025	-0.020	-0.006	-0.003	-0.002	-0.001	-0.032	-0.036	-0.053	-0.064	-0.064	-0.080	-0.095	-0.080	-0.074	-0.080	-0.069	-0.056	-0.041	-0.056	-0.061	-0.058
(Zmin t-4)^2	0.035	0.022	0.021	0.024	0.011	0.005	0.001	0.004	0.035	0.033	0.041	0.053	0.057	0.076	0.083	0.077	0.073	0.077	0.068	0.059	0.047	0.038	0.043	0.050

Table 9 – COEFFICIENTS OF THE TEMPERATURE REGRESSORS

7. HOURLY REFRESHMENT

As seen during the analysis of the demand series in Chapter 1, the h-hour power demand is highly correlated with the demand of the nearby hours. Therefore, it seems logical to think that the errors made in predicting power demand between nearby hours will also be highly correlated. That is, if the error made in predicting energy demand at 10:00 h is 5% positive, the error made at 11:00 h should be similar, both in percentage and sign. This reasoning is shown in Figure 24, which shows the correlation matrix between the errors made on day *t*.

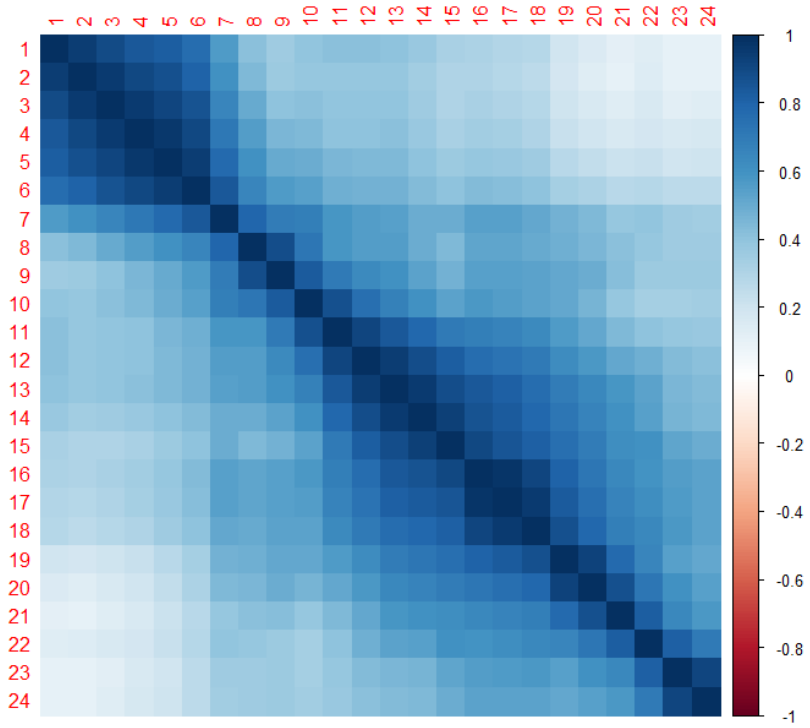

Figure 24 – T-DAY ERROR CORRELATION MATRIX

Therefore, the idea of the hourly refreshment consists of constructing 24 linear regression models, by means of which the error of hour h is predicted from the errors of the previous hours h-1, h-2 and h-3. The three previous hours are chosen as regressors because the correlation between the h-hour error and the h-4 hour error is significantly lower.

The starting point for the hourly refreshment is 11.00 am. In other words, the exact errors at 10:00 h, 09:00 h and 08:00 h are known, since the real electricity demand data are known for those hours and, of course, the prediction made by the hourly model in question. Based on these known errors, the error that will be made in predicting the demand at 11:00 am is estimated by means of a linear regression model. In the same way, the error to be made at 12:00 h is estimated using a different linear regression model, which uses the predicted error at 11:00 h, calculated from the previous linear regression model, and the known errors at 10:00 h and 09:00 h. This process is repeated for the 24-hour period. This process is repeated for the next 24 hours, until 10:00 h on day t+1.

This procedure tries to emulate the work done by REE to adjust the electricity demand forecasts for day t+1. The process is similar, except that the Spanish REE completes the

adjustment of the forecasts for the whole day t+1 and does not stop at 10.00 h, as it does in this work. The task of adjusting the electricity demand beyond 10.00 h requires a two-step error prediction that is beyond the scope of this paper, but that undoubtedly forms a future line of study.

The actual value of the electricity demand at 11:00 a.m. can be expressed as follows

$$y_{11,t} = \hat{y}_{11,t} + \varepsilon_{11,t}\,,$$

where

$$\varepsilon_{11,t} = \underbrace{\vartheta_1^{11} * error_{10,t}^{real} + \vartheta_2^{11} * error_{09,t}^{real} + \vartheta_3^{11} * error_{08,t}^{real}}_{= \hat{\varepsilon}_{11,t}} + \epsilon_{11,t}\,,$$

and therefore the predicted value to which the hourly refreshment is applied is

$$\hat{\hat{y}}_{11,t} = \hat{y}_{11,t} + \hat{\varepsilon}_{11,t}$$

where

- $y_{11,t}$ is the actual value of the electric power demand at 11:00 a.m.
- $\hat{y}_{11,t}$ is the value predicted by the Reg-ARIMA model for 11:00 a.m.
- $\varepsilon_{11,t}$ is the error committed by the Reg-ARIMA model at 11.00 a.m.
- $\hat{\varepsilon}_{11,t}$ is the error estimated from the linear regression, which uses the actual errors of 10.00 h, 09.00 h and 08.00 h as regressors.
- $\epsilon_{11,t}$ is the error made in estimating the 11.00 h error.
- $\hat{\hat{y}}_{11,t}$ is the value predicted by the Reg-ARIMA model at 11.00 h, to which the hourly refreshing is applied.

The same applies to all other hours of the day t

$$\hat{\hat{y}}_{12,t} = \hat{y}_{12,t} + \hat{\varepsilon}_{12,t}$$

$$\hat{\varepsilon}_{12,t} = \vartheta_1^{12} * \hat{\varepsilon}_{11,t} + \vartheta_2^{12} * error_{10,t}^{real} + \vartheta_3^{12} * error_{09,t}^{real}\,,$$

$$\hat{\hat{y}}_{13,t} = \hat{y}_{13,t} + \hat{\varepsilon}_{13,t}$$

$$\hat{\varepsilon}_{13,t} = \vartheta_1^{13} * \hat{\varepsilon}_{12,t} + \vartheta_2^{13} * \hat{\varepsilon}_{11,t} + \vartheta_3^{13} * error_{10,t}^{real}\,,$$

.

.

.

$$\hat{\hat{y}}_{24,t} = \hat{y}_{24,t} + \hat{\varepsilon}_{24,t}$$

$$\hat{\varepsilon}_{24,t} = \vartheta_1^{24} * \hat{\varepsilon}_{23,t} + \vartheta_2^{24} * \hat{\varepsilon}_{22,t} + \vartheta_3^{24} * \hat{\varepsilon}_{21,t}$$

The hourly refreshment of day t+1 has the particularity that the estimated errors of 01.00 h, 02.00 h and 03.00 h depend on the predicted errors of the previous day. As was the case with the errors of day t, the correlation between the errors of the first hours of day t+1 and the last hours of day t is high.

Figure 25 shows the correlation between 01.00 h , 02.00 h and 03.00 h of day t+1 and 24.00 h , 23.00 h and 22, 00 h of day t.

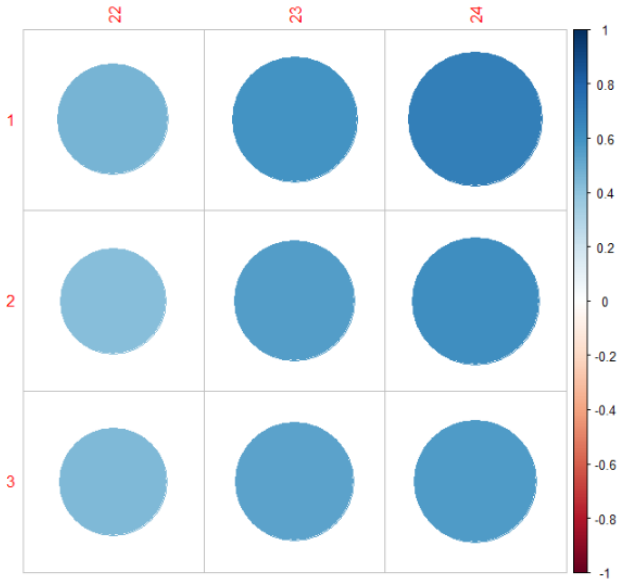

Figure 25 – CORRELATIONSHIPS between LAST HOURS DAY t and FIRST HOURS DAY t+1

Therefore, the hourly refreshment for day t+1 is of the type

$$\hat{\hat{y}}_{01,t+1} = \hat{y}_{01,t+1} + \hat{\varepsilon}_{01,t+1}$$

$$\hat{\varepsilon}_{01,t+1} = \vartheta_1^{01} * \hat{\varepsilon}_{24,t} + \vartheta_2^{01} * \hat{\varepsilon}_{23,t} + \vartheta_3^{01} * \hat{\varepsilon}_{22,t}\,,$$

$$\hat{\hat{y}}_{02,t+1} = \hat{y}_{02,t+1} + \hat{\varepsilon}_{02,t+1}$$

$$\hat{\varepsilon}_{02,t+1} = \vartheta_1^{02} * \hat{\varepsilon}_{01,t+1} + \vartheta_2^{02} * \hat{\varepsilon}_{24,t} + \vartheta_3^{01} * \hat{\varepsilon}_{23,t}\,,$$

.

.

.

$$\hat{\hat{y}}_{10,t+1} = \hat{y}_{10,t+1} + \hat{\varepsilon}_{10,t+1}$$

$$\hat{\varepsilon}_{10,t+1} = \vartheta_1^{10} * \hat{\varepsilon}_{09,t+1} + \vartheta_2^{10} * \hat{\varepsilon}_{09,t} + \vartheta_3^{10} * \hat{\varepsilon}_{08,t}$$

The mean of the regression models does not appear because the white noise error is equal to 0.

Table 10 shows a table with the coefficients, the standard deviation and the explained variability of the different linear regression models.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12	m13	m14	m15	m16	m17	m18	m19	m20	m21	m22	m23	m24
Omega 1	0.937	0.989	1.031	1.031	0.919	0.985	1.138	1.016	0.946	0.769	0.889	1.142	1.041	1.072	1.071	0.819	1.101	1.189	0.792	0.933	1.009	0.726	0.810	0.993
Omega 2	-0.161	-0.205	-0.075	-0.064	0.064	-0.047	-0.017	-0.141	0.013	-0.291	0.015	-0.292	-0.026	-0.032	-0.180	-0.089	-0.180	-0.304	0.063	0.040	-0.290	-0.055	0.403	-0.177
Omega 3	0.080	0.152	-0.021	-0.008	-0.044	-0.044	-0.264	0.061	-0.049	0.331	-0.092	0.089	-0.058	-0.116	0.030	0.227	0.063	0.083	0.049	-0.106	0.106	0.060	0.114	-0.071
Desv. Tipica	1.513	0.662	0.584	0.605	0.540	0.604	0.993	1.374	1.118	1.038	0.938	0.771	0.691	0.567	0.688	0.864	0.508	0.668	1.110	0.823	0.945	0.983	0.896	0.598
Var. Explicada	42.100	89.300	91.590	91.140	92.580	89.710	73.250	63.750	78.230	73.800	76.350	85.230	88.900	92.430	88.420	83.010	94.440	90.740	75.820	85.430	77.720	68.700	67.860	84.490

Table 10– COEFFICIENTS LINEAR REGRESSION MODELS

In conclusion, the hourly refreshment technique tries to take advantage of the correlation between the errors committed during training to establish linear relationships between them and thus be able to predict them, with the aim of improving the predictions of the future hours' electricity demand.

8. RESULTS

8.1. INTRODUCTION

The objective of this section is to evaluate the errors committed by the 24 Reg-ARIMA models during the validation stage and to study the effect of the application of the hourly refreshment on the predictions.

First, Figure 26 shows the standard deviation in percent of the 24 Reg-ARIMA models (in black) and of the 24 linear regression models that predict the error (in red).

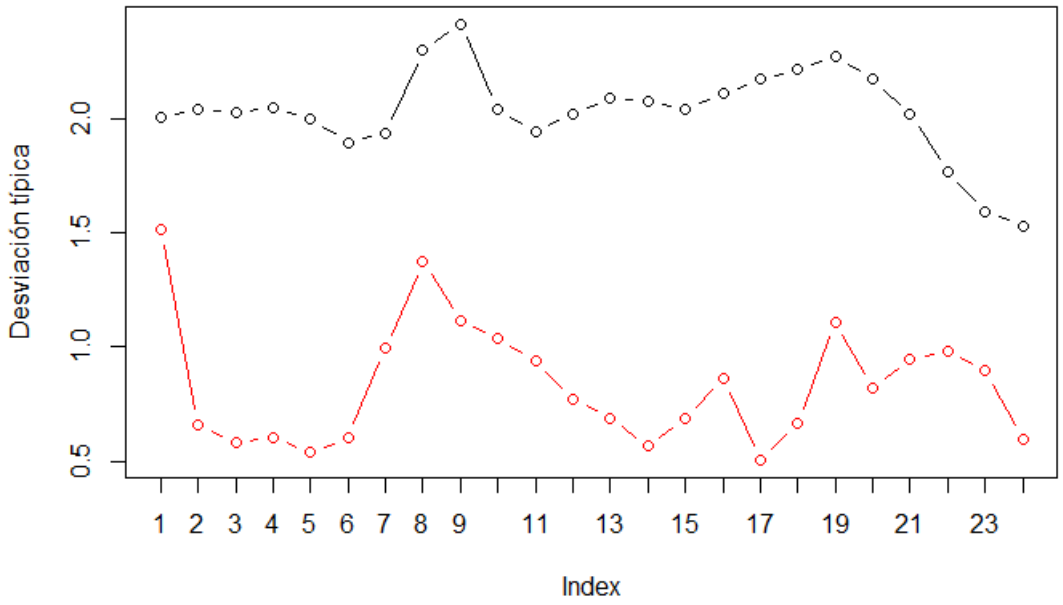

Figura 26 – DESVIACIÓN TÍPICA DE LOS MODELOS REG-ARIMA Y DE LOS MODELOS DE REGRESIÓN LINEAL

It is found that the linear regression models that estimate the error committed have a much lower standard deviation than the Reg-ARIMA models. It follows from this result that the preiction of the error is more accurate than the prediction of the electricity demand itself. For this reason, the application of the a priori hourly refresh should help to improve the predictions of the Reg-ARIMA models.

Although the decrease of the standard deviation is evident in all hours, there are some hours with higher values than the average.

First of all, the value of the standard deviation of the linear regression model of the 01.00 h error is surprising. In fact, the explained variability of this model is the lowest of the 24, as shown in Table 10. It is believed that both the low explained variability and the high standard deviation are due to the influence of the regressors.

On the one hand, the temperature of the last hours of day t should be similar to the temperature at 01:00 h on day t+1, but since they belong to different days, very different values of the temperature regressor are used during the prediction. Specifically, the averages of the maximum and minimum temperatures and their squares of days t, t-1, t-2, t-3 and t-4 are used to predict the demand of the last hours of day t, while those of days t+1, t, t-1, t-2 and t-3 are used to predict the demand of 01.00 h of day t+1. These values can be very disparate and an explanation for the standard deviation and explained variability obtained, since values of maximum and minimum temperatures for the last hours of day t and 01.00 h of day t+1 are used that are much more different than they really are.

Secondly, the standard deviation of the 09.00 h model is also higher than normal. One of the reasons that may explain this situation is that at 09.00 h the population usually starts working, which makes electricity consumption more variable and therefore more difficult to predict.

8.2. THE PERCENT ROOT MEAN SQUARE ERROR

On the other hand, the Mean Squared Error (MSE) is used to analyze the errors made in predicting the demand for electric power. The MSE is expressed as follows

$$MSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}\,,$$

where

- n is the total number of observations
- y_i is the real value
- $\hat{y}_i$ is the predicted value

However, since the errors increase with increasing demand, the percentage Mean Squared Error, which is independent of the scale of the data, is used. The percent MSE is defined as follows [12].

$$MSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} \left(\frac{(y_i - \hat{y}_i)}{y_i} * 100 \right)^2}$$

8.3. Percent MSE OF Reg-ARIMA MODELS

Figure 27 shows the percent MSE as a percentage of the predicted values of electricity demand for days t and t+1, i.e., from 11:00 a.m. on day t to 10:00 a.m. on day t+1.

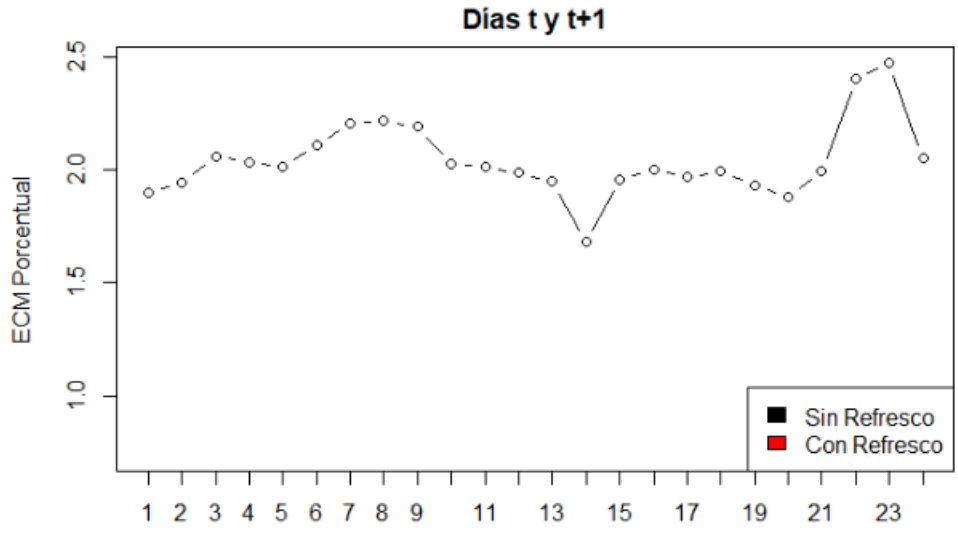

Figure 27 – Percent MSE of the REG-ARIMA MODELS

The indices on the x-axis can be misleading. The first observation corresponds to the percent MSE at 11.00 h on day t, not 01.00 h, and therefore the last one corresponds to 10.00 h on day t+1.

A number of conclusions can be drawn from the above figure. First, the Reg-ARIMA models that commit a higher percent MSE are those of 08.00 h and 09.00 h, being close to 2.5%, and therefore it is concluded that the electricity demand of hours 08.00 and 09.00 are the most difficult to predict. It has already been seen in Figure 26 that the standard deviations of the models for these two hours are the highest of all, so the result obtained here is logical. One of the reasons that may explain the difficulty of predicting the demand for electricity during these hours is precisely that it is between 08:00 h and 09:00 h when the city "wakes up" and the population's work activity begins, which can mean a highly variable consumption of electricity.

Secondly, the hour with the lowest percent MSE is 24.00 h, which is the simplest hour to predict according to this criterion. In this case, the percent MSE is approximately 1.5%.

Finally, the rest of the hours present a similar percent MSE, ranging between 1.8% and 2.3%, approximately. Within this third group, the hours with the highest percent MSE are 17.00 h , 18.00 h and 19.00 h.

8.4. Percent MSE OF THE DAY T

Figure 28 shows the percent MSE as a percentage of the predicted values of electric power demand by hours of day t with and without hourly refreshment. That is, the errors made between 11.00 h and 24.00 h are plotted.

Figure 28 – Percent MSE REFRESHMENT SCHEDULE DAY T

Several conclusions can be drawn from the previous figure. First, the quality of the hourly refreshment decreases as the hours progress. It is important to remember that the hourly refreshment consists of an estimation of the error based on known and/or predicted errors from previous hours, and since it is an estimation, it always makes an error. Therefore, this error made when predicting the error is carried forward as the hours advance. For this reason, among the predictions to which the hourly refreshment is applied, the 11:00 a.m. prediction is the one with the lowest percent MSE, since it uses three known errors as regressors. The 12:00 h prediction will have a higher present MSE, since it uses two known errors and one estimated error, and so on. However, it is true that there are specific hours, such as 23.00 h or 24.00 h, whose percent MSE decreases, when following the above reasoning it should increase. This is due to the fact that these are hours in which it is easier to predict the demand for electricity.

Secondly, it can be seen that the application of the hourly refreshment improves the percent MSE in all the hours of day t. The hours with the lowest error are in the hours of the day t. The percent MSE of the day t is the one with the lowest error. The hours with the lowest error are between 11.00 h and 14.00 h, whose ECM does not exceed 1.5%.

Thirdly, the percent MSE plot of the predicted values to which the hourly refreshment has been applied is increasing until 19:00 h, so that the improvement in the error is steadily decreasing. From then on, between 19:00 h and 22:00 h, the percent MSE remains more or less constant. Finally, the percent MSE decreases again for the last two hours of the day.

Fourth, the hours of the day t when the percent MSE is highest are between 19:00 h and 22:00 h. This is explained by the decrease in the percent MSE between 19:00 h and 22:00 h.

The percent MSE decreases in the last two hours of the day. This is explained by the decrease in the quality of the hourly refreshment and also by the fact that these are hours in which the variability of the energy demanded is higher. This last aspect is one of the reasons why the percent MSE at 23:00 h and 24:00 h is lower. Despite the fact that the hourly refreshment of these hours is of lower quality due to a higher previous error, the greater ease in predicting the demand for electric power in these hours compensates for this, resulting in a lower percent MSE.

Table 11 shows the percent MSE for day t.

	11,00	12,00	13,00	14,00	15,00	16,00	17,00	18,00	19,00	20,00	21,00	22,00	23,00	24,00
sin_ref	1.902	1.943	2.061	2.036	2.014	2.108	2.205	2.219	2.190	2.025	2.013	1.986	1.949	1.681
con_ref	0.738	1.077	1.318	1.414	1.521	1.564	1.652	1.710	1.794	1.704	1.751	1.767	1.708	1.506

Table 11 – Percent MSE VALUES DAY T

8.5. Percent MSE OF DAY T+1

As will be seen below, the percent MSE of the hourly predictions of day t+1 is conditioned by the low explained variability of the linear regression model for the prediction of the 01.00 h error. In addition to the low quality of this model, the continuous decrease in the improvement caused by the hourly refreshment explains the similarity between the percent MSE of the predicted values of day t+1 with and without hourly refreshment.

Figure 29 shows the percent MSE of the predicted values with and without hourly refreshment.

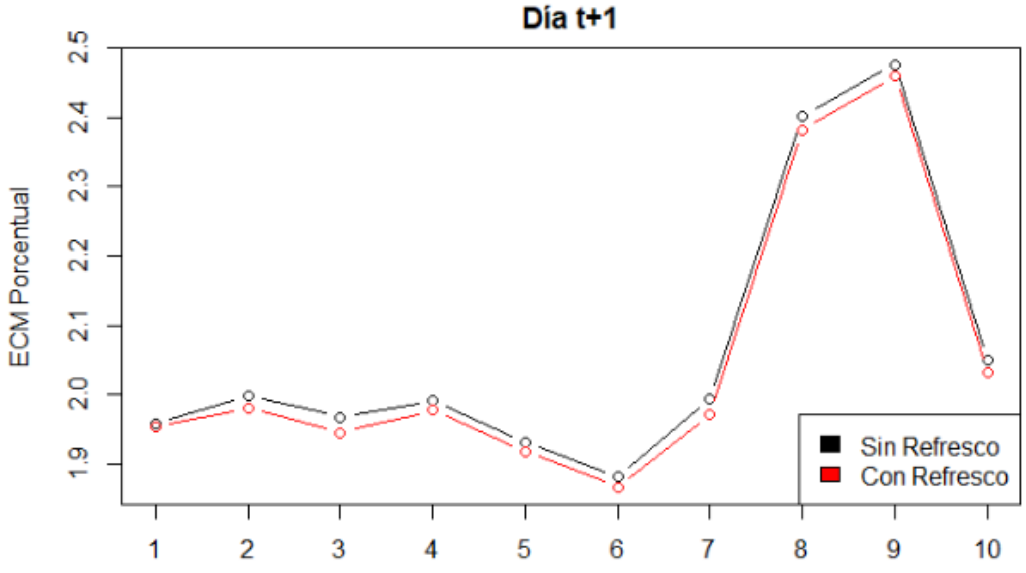

Figure 29 – Percent MSE DAY t+1

Two conclusions can be drawn from Figure 29. On the one hand, it is observed that the improvement due to the application of the hourly refreshment is not significant in any of the

hours of the day t+1. Moreover, at specific hours such as 01.00 h, the improvement is not even appreciable.

On the other hand, the hours where the mean square error is highest are 08.00 h and 09.00 h, being almost 2.5%. The high standard deviation of the Reg-ARIMA and linear regression models of the 09:00 h error has already been mentioned, which is one of the reasons for obtaining such a high percent MSE.

The percent MSE for the hours of day t+1 are shown in Table 12..

	01,00	02,00	03,00	04,00	05,00	06,00	07,00	08,00	09,00	10,00
sin_ref	1.959	1.998	1.968	1.992	1.931	1.882	1.993	2.402	2.476	2.050
con_ref	1.954	1.981	1.946	1.977	1.919	1.867	1.971	2.382	2.460	2.033

Table 12– Percent MSE DAY t+1

Figure 30 shows the ECMs for 24 hours, i.e., from 11:00 h on day t to 10:00 h on day t+1.

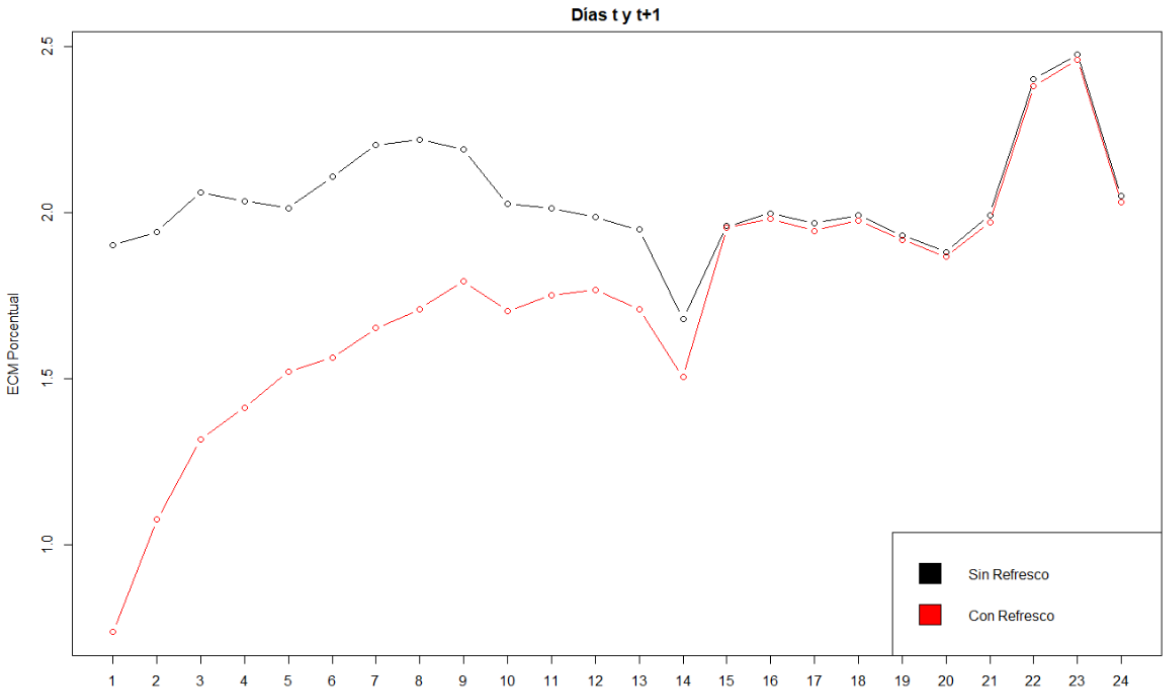

Figure 30 – Percent MSE of 24 Hours with REFRESCO

From this overview, it is clear that the quality of the hourly refreshment decreases as time progresses, being this decrease very significant for the hours of the day t+1.

8.6. Percent MSE BY DAYS OF THE WEEK

The calculation of the mean square error obtained in the validation stage for the different days of the week can be useful to understand on which days it is more difficult to predict the electricity demand. Figure 31 shows the percent MSE by days of the week..

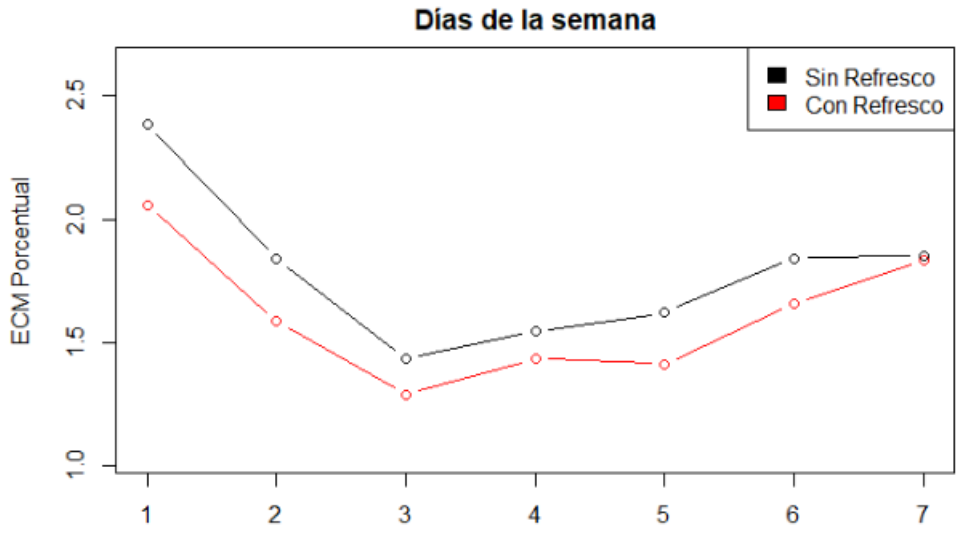

Figure 31- Percent MSE of the days of the week

First, it is observed that the application of the hourly refreshment clearly decreases the percent MSE of all days of the week. The change from an percent MSE on Mondays of 2.4% to 2.1% is particularly noteworthy.

From the figure it can be concluded that by far the most difficult day to predict energy demand is Monday. The other days of the week have a similar percent MSE, so there is no significant difference in the difficulty of predicting demand between any of these days. Despite this, the low error on Wednesdays, Thursdays and Fridays stands out.

Generally speaking, and specifically in 2019, many long weekends and local holidays fell on Mondays, which is logical considering their proximity to the weekend. This may be the main reason for their high percent MSE.

However, if the same reasoning is followed, it is surprising that the percent MSE of Fridays is so low. As with Mondays, there are also many holidays that, in order to "extend" the weekend, fall on Fridays. In addition, national holidays such as November 1, December 6 or evidently Good Friday coincided with this day of the week in 2019. Therefore, it is easier to predict the electricity demand of the last day of the working week than that of the first day of the working week.

On the other hand, Tuesdays, Wednesdays and Thursdays, especially the last two, also show a low percent MSE. The reason for this low error is the opposite of that of Mondays. The fact that these three days make up the middle part of the workweek means that fewer local holidays or long weekends coincide with these three days, and therefore predicting demand is easier, since, as we have seen throughout the paper, the holiday nature of the days is the aspect that makes this task most difficult.

Finally, the percent MSE on weekends is slightly higher than on Tuesdays, Wednesdays, Thursdays and Fridays. It has been previously verified that the fact that some holidays coincide on weekends does not significantly modify the demand for electrical energy, so the existence of a certain similarity between the percent MSE of these days is not surprising.

8.7. ECM PER MONTH

The same analysis of the previous section can be performed for the months of 2019. Figure 32 shows the percent MSE for each month.

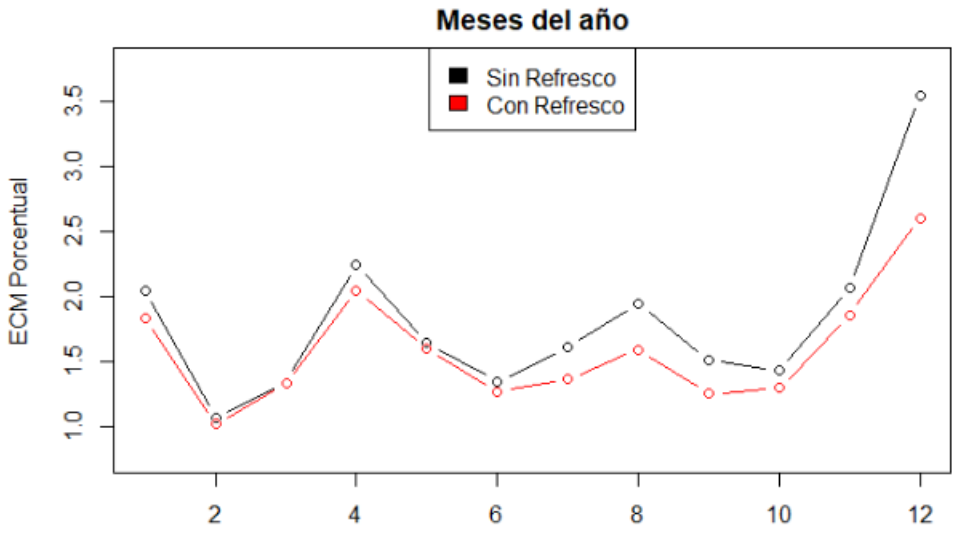

Figura 32 – Percent MSE by month

The analysis of the percent MSE committed by months yields some striking results. First, as with the percent MSE committed by weeks, the application of the hourly refreshment decreases the error committed in all months. However, in this case the improvement is very different depending on the month. It can be seen that the hourly refreshment decreases the error in months such as December from more than 3.5% to 2.6%, while in other months, such as February, March, May or June, the improvement is practically not appreciable. In other months, such as January, April, July, August, September, October and November, the improvement is significant, although not as high as in December.

On the other hand, the high error in the month of December stands out, especially for the Reg-ARIMA model without hourly refreshment, being more than 3.5%. This is due to the number of holidays that this month hosts.

Also noteworthy are the high percent MSE of months such as January, April, August or November. All these months have important national holidays, which may be a reason for the high percent MSE.

Finally, the rest of the months have a similar percent MSE, which oscillates around 1.5% for the predictions with hourly refreshment, although months such as February or September stand out, which are the ones with the lowest percent MSE.

9. CONCLUSIONS

Throughout this work we have studied the time series of electricity demand in Spain from the construction of 24 univariate Reg-ARIMA models.

The first important aspect to take into account is the stationarity of the demand series, since the ARIMA models, which are very expanded in the treatment of time series, can only work with stationary series, and as we have seen, the stationarity of the series entails the loss of stationarity. Although the complete demand series presents annual, weekly and even daily stationarity, the division of this series into 24 series, one for each hour of the day, has revealed the presence of a strong weekly stationarity. In other words, the demand for hour h of day t is highly correlated with the demand for hour h of day t-7. This characteristic of the series justifies the use of S-ARIMA models, which use autoregressive and seasonal moving average polynomials, in addition to seasonal differencing techniques, in order to work with seasonal time series.

Secondly, the influence of several regressors on the amount of electricity demanded was studied. On the one hand, the high influence of the demand for electrical energy on day t with the average maximum and minimum temperatures of the same day and especially of the 4 previous days has been observed. As a simplification, the relationship between demand and temperature was considered to be quadratic. To corroborate the influence of maximum and minimum temperatures on demand, a series of quadratic regression models were run, and it was found that the explained variability of the models using the temperatures of the previous days as regressors was always higher than that using the temperature of day t as regressor. Thus, the inclusion in each model of the mean maximum and minimum temperatures and their squared values for days t, t-1, t-2, t-3 and t-4 is justified.

On the other hand, the relationship between demand and holidays has been studied in depth. From the elaboration of several graphs, it has been seen how the energy demanded decreases significantly on holidays, especially when these coincide with a working day. In addition, the development of previous models that do not include any regressor referring to holidays reveals the presence of several days of the year in which the error committed during training is very high. These days mostly coincide with national holidays. It is therefore of vital importance to identify this type of special days by including several regressors. These regressors consist of a vector that takes a value equal to 1 on public holidays and a value equal to 0 on other days. We have chosen to include 13 regressors: 9 regressors representing the 9 national holidays, taking a value equal to 1 on those working days coinciding with a holiday; one regressor for national holidays falling on a weekend, two other regressors identifying Thursday and Good Friday, and a last regressor representing local holidays, taking in this case a value between 0 and 1 that represents the percentage of the population that was affected by the local holiday as a percentage of one.

In addition, a last regressor has been included to represent the outlier days, which are those days that, after training the model with all the regressors already mentioned, continue to present an error greater than 10%. It has been seen that these days mostly coincide with days close to national holidays.

days close to national holidays, which leads us to think that these days also condition the demand for the days close to them.

Thirdly, the process to obtain the final prediction model consisted in the development of several previous models and the evaluation of the errors made with each one of them to check that each model improves the previous one. In particular, the Akaike information criterion (AIC), which penalizes the inclusion of parameters, has been followed. In this process, the importance of including the regressors explained above has been corroborated, since the inclusion of each of them has led to significant improvements.

Fourthly, an hourly correction technique, called hourly refreshment, has been applied, which tries to take advantage of the correlation between the errors made when predicting the demand of hour h and the errors made when predicting the demand of the immediately preceding hours in order to improve the predictions. This technique consists of the development of 24 linear regression models, which attempt to explain the h-hour prediction error from the errors of hours h-1, h-2 and h-3. In all cases, except for the 01.00 h error regression model, very high explained variabilities are obtained. It is believed that the temperature regressor is the main responsible for the value of the explained variability of the 01.00 h regression model to be lower than expected, because although it is expected that the temperature at 01.00 h and 24.00 h, 23.00 h and 22.00 h should be similar, they are used to predict the mean temperature of day t in the first case and the mean temperature of day t-1 in the other three, which can be very different.

The Mean Squared Error study is used to check how the application of the hourly refreshment significantly improves the results obtained. The error committed with and without hourly refreshment has been analyzed by months, days of the week and hours.

On the one hand, it has been shown that the application of the time refreshment significantly reduces the ECM in all months of the year, especially in December, since this is the month with the greatest number of national holidays and therefore the most difficult month to predict, so that the time refreshment has a greater effect. On the other hand, in months such as January, April or August, significant improvements are also obtained, while in other months, such as February or March, the improvements are practically not appreciable.

On the other hand, the ECM also decreases significantly on all days of the week. It is concluded that the most difficult day to predict is Monday, probably because of its proximity to the weekend, while the easiest days to predict are Wednesday, Thursday and Friday. The difference in the errors committed on Mondays and Fridays is surprising, although it is thought that this is due to the fact that in 2019, the year in which the validation of the models was carried out, many special days coincided with Mondays, so the difficulty of predicting that day increases significantly.

Finally, it is also corroborated that the time correction improves the error committed at all times of the day. However, it is observed that the time refreshment is losing effectiveness. This is logical, since the known errors of the previous three hours are used to predict the error committed at 11:00 h, the starting point of the refreshing. However, since it is a prediction, an error is always made. This error will grow and increase as the hours progress. For example, for the 12:00 h prediction, the predicted error of 11:00 h and the known errors of 10:00 h and 09:00 h are used, and so on. Therefore, it is reasonable that the time of day at which the hourly

refreshment has the greatest effect is 11.00 h, and that this effect decreases progressively. Furthermore, the low explained variability of the linear regression model at 01.00 h favors the decrease in the quality of the hourly refreshment. It is concluded that the most difficult hours to predict are 09.00 h and 08.00 h, probably because they are the hours when the city "wakes up" and the population starts to work, so the variability of the electric power demanded increases. On the other hand, the hours where the least error is made are mostly between 22:00 h and 06:00 h, which is reasonable considering that these are hours when the population activity is minimal, and therefore so is the variability of the energy demanded.

Thus, 24 models are obtained Reg-ARIMA(1,0,2)x$(0,1,2)_{s=7}$ univariate models, one for each hour of the day, with 6 parameters and 34 regressors. As the seasonal difference is not accompanied by any coefficient, there are 39 coefficients to be estimated per model, and therefore 936 coefficients to be estimated in total. There are also 24 linear regression models to predict the error, whose application clearly improves the results obtained using only the Reg-ARIMA models.

10. FUTURE LINES

In this work there are three aspects that, either because they are simplifications or because they are assumptions, would merit further study.

First, it has been determined that the nonlinear relationship between electric power demand and maximum and minimum temperatures can be approximated by a quadratic relationship. As already explained in the corresponding chapter, this consideration is no more than a simplification. Nevertheless, it has been seen how such a simplification has helped considerably to improve the prediction models. Nevertheless, more time could be devoted to think of another way of modeling the demand-temperature relationship. Perhaps the use of spline models would be a good option.

Secondly, it has been observed that the linear regression model used for the prediction of the 01.00 a.m. errors has a much lower explained variability than the rest of the models. It has been reasoned that the influence of the temperature factor could be one of the main causes, although this theory has not been explored too deeply. One of the future lines of work could be to improve this model, as this would considerably improve the quality of the hourly refreshment.

Thirdly, in this work only 1-step predictions have been made, stopping at the prediction of 10.00 h on day t+1. It would be very interesting to continue until 24.00 h of the same day, applying techniques such as seasonal momentum.

Finally, it has been seen that many of the days close to national holidays present high errors. Perhaps we could try to use regressors that represent this type of days, in order to reduce the error committed on these days.

BIBLIOGRAPHY

[1] V. Cordero, «Kelisto,» 12 Mayo 2022. [En línea]. Available: https://www.kelisto.es/electricidad/consejos-y-analisis/por-que-sube-el-precio-de-la-luz-6181.

[2] S. F. Munguía, «Xetaka,» 8 Enero 2021. [En línea]. Available: https://www.xataka.com/energia/como-funciona-mercado-electrico-que-a-pesar-que-precio-a-veces-llegue-a-cero-apenas-va-a-repercutir-nuestra-factura-1.

[3] J. A. Mauricio, de *Introducción al Análisis de Series Temporales*, Madrid, 2005, pp. 1-9.

[4] J. F. López, «Economipedia,» 2019. [En línea]. Available: https://economipedia.com/definiciones/proceso-estocastico.html.

[5] R. H. S. a. D. S. Stoffer, «Stationary Time Series,» de *Time Series Analysis and Its Apllications*, Springer, 2010, p. 23.

[6] A. M. Alonso, «Clasificación de Series Temporales,» de *Introducción al Análisis de Series Temporales*, p. 11.

[7] R. H. S. a. D. D. Stoffer, «Introduction to Autoregressive Models,» de *Time Series Analysis and Its Applications*, Springer, 2014, p. 84.

[8] R. H. S. a. D. D. Stoffer, «Introduction to Moving Average Models,» de *Time Series Analysis and Its Applications*, Springer, 2014, p. 91.

[9] R. H. S. a. D. D. Stoffer, «Autoregressive Moving Average Models,» de *Time Series Analysis and Its Applications*, Springer, 2014, p. 93.

[10] R. H. S. a. D. D. Stuffer, « Integrated Models for Nonstationary Data,» de *Time Series Analysis and Its Applications*, Springer, 2014, p. 141.

[11] R. H. S. a. D. D. Stuffer, « Multiplicative Seasonal ARIMA Models,» de *Time Series Analysis and Its Applications*, Springer, 2014, p. 154.

[12] L. E. E. UPM, «Modelo de Predicción Demanda de Energía Eléctrica,» Madrid.

[13] B. González, «Ecología Verde,» 4 Diciembre 2018. [En línea]. Available: https://www.ecologiaverde.com/como-afecta-la-produccion-electrica-al-medio-ambiente-1745.html.

[14] «Rincón Educativo,» [En línea]. Available: https://www.rinconeducativo.org/es/recursos-educativos/la-planificacion-energetica-en-espana.

[15] L. E. E. UPM, «Modelo Predicción Demanda de Energía Eléctrica,» Madrid.