A model to predict client's phone calls to Iberdrola Call Centre

UNIVERSIDAD COMPLUTENSE
MADRID

Participants:

Cazallas Piqueras, Rosa
Gil Franco, Dolores M
Gouveia de Miranda, Vinicius
Herrera de la Cruz, Jorge
Inoñan Valdera, Danny Javier

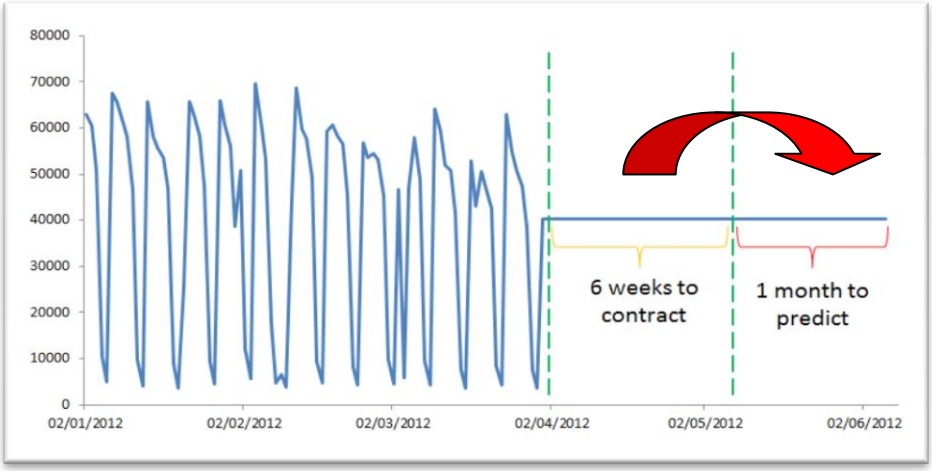
Directed by: Benjamin Ivorra

Problem description

Iberdrola is a company that provides electricity and/or gas to the customers. Customers can contact Iberdrola by letters, internet and **Phone calls (call center)**. A good quality of this service is essential to ensure the customers' satisfaction. To do so, a requirement is to have a sufficient number of operators in order to avoid large waiting time. In call center we need to avoid **over-staff** and **under-staff** because Iberdrola needs a good prediction about the number of operators. Obviously, keeping a high number of operators at each moment is relatively expensive for the company.

The **objective** of this research is to developing mathematical models to predict the volume of customer calls.

Considering historical data we need to predict the volume of calls that Iberdrola's call center would receive for next month six weeks in .



Data analysis

In this section, we are going to analyze the provided data. IBERDROLA gives us a data base with two types of variables: weekly data and daily data.

On weekly data we have the following variables:

VARIABLE PORTFOLIO:

- Portfolio1: this variable indicates the regulated market customers. The cost of the light or electricity is calculated with the cost of other competitive factories.
- Portfolio2: free market customers. In this case, IBERDROLA fix his own cost.
- Portfolio3: gas customers (since 01/2007).

VARIABLE BILL TYPE:

- Bill type 1: estimated consumption bill. This type of bill is calculated with the information of some previous months (for example with the mean).
- Bill type 2: real consumption bill (you have to pay what you have to spent).
- Bill type 3: fixed quota bill.
- Bill type 4: other, for instance, bill mistakes, supplementary material...

On daily data we have:

- Calls: number of phone calls received every day (**TARGET VARIABLE**). This is the most important one, because it's the variable that we have to predict.

To describe the evolution of the calls we have to convert weekly data into daily data in the variable bill type. To do that, we assign different weights to the different days of the week. The weights are calculated to minimize square error.

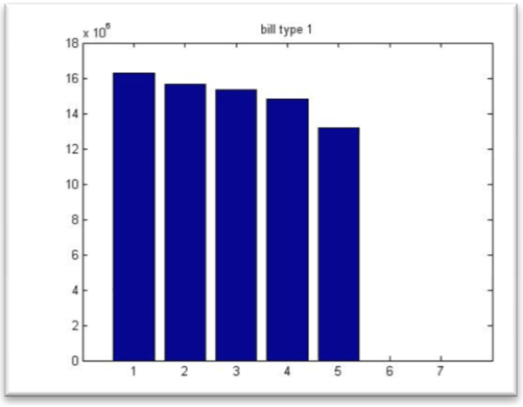
We create binary variables to explain days off and weekends. If we have a day off in the middle of the week we will recalculate the weight of the rest of the weekdays.

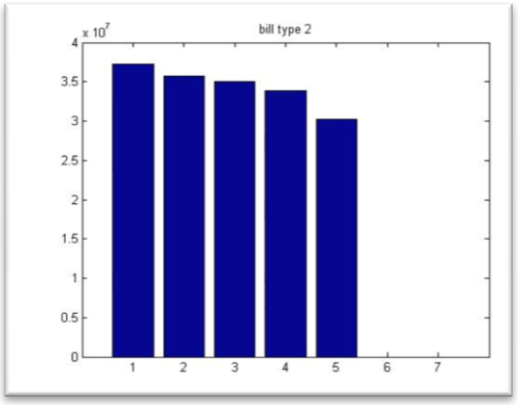
The weights that we have assign are the following:

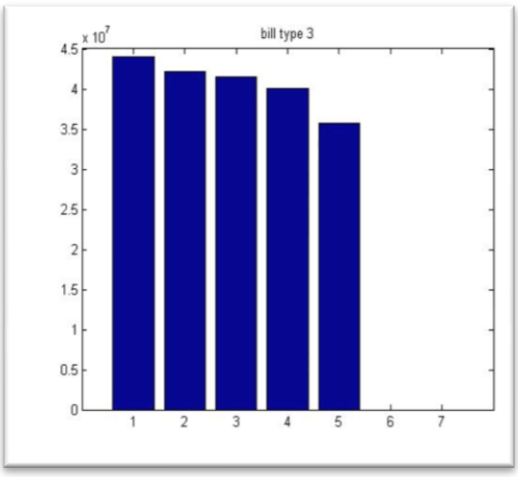
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
0.215	0.208	0.201	0.197	0.179	0	0

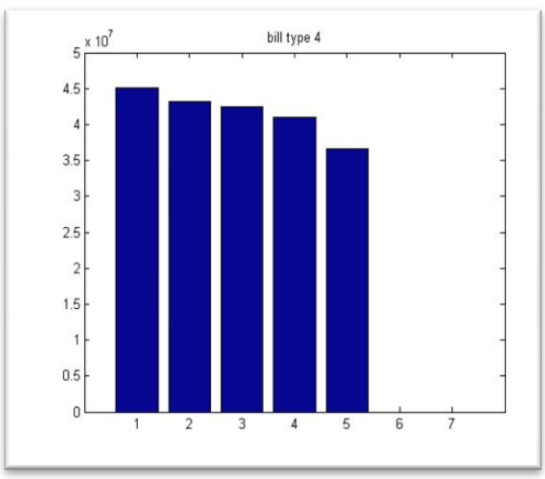
On Saturday and Sunday, we have put the value 0 because the company told us that they don't send bills in weekends. Furthermore, we have put a mayor value on Monday and Tuesday because in these days, the company send more bills.

Now, we are going to represent the distribution of the different types of the variable bill type:

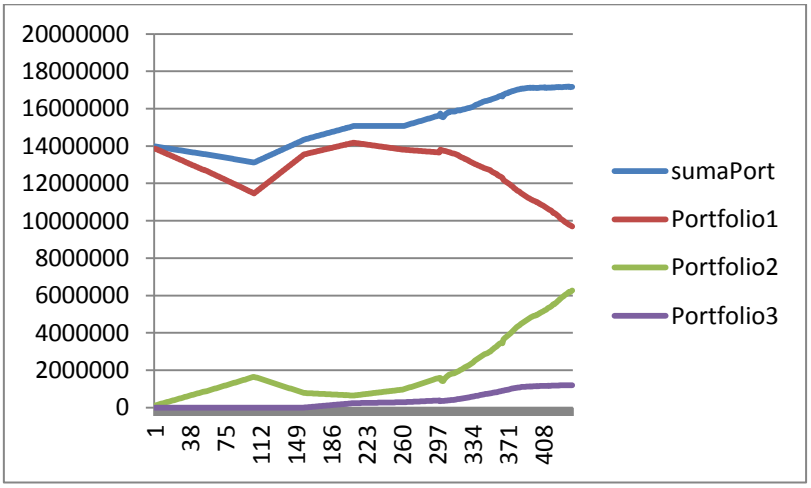






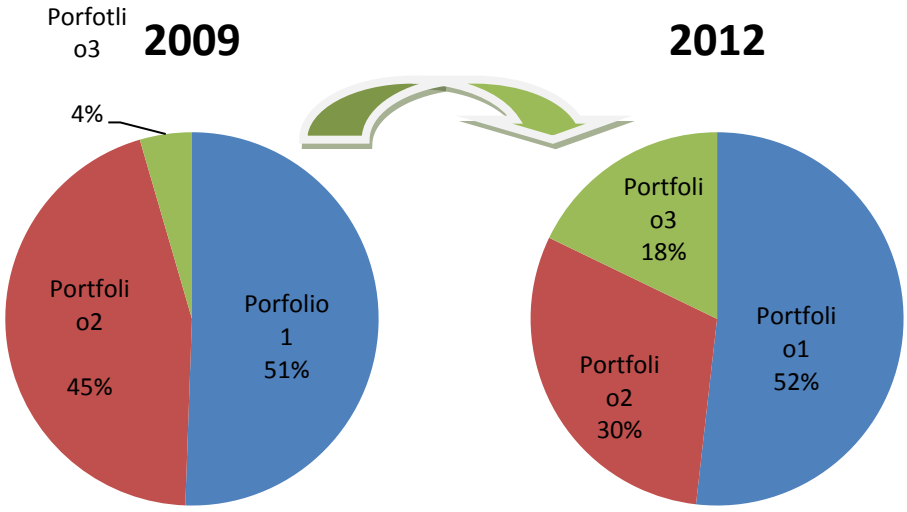


After that, we are going to study the distribution and the mean characteristics of the variables available. First, we start with the variable portfolio.

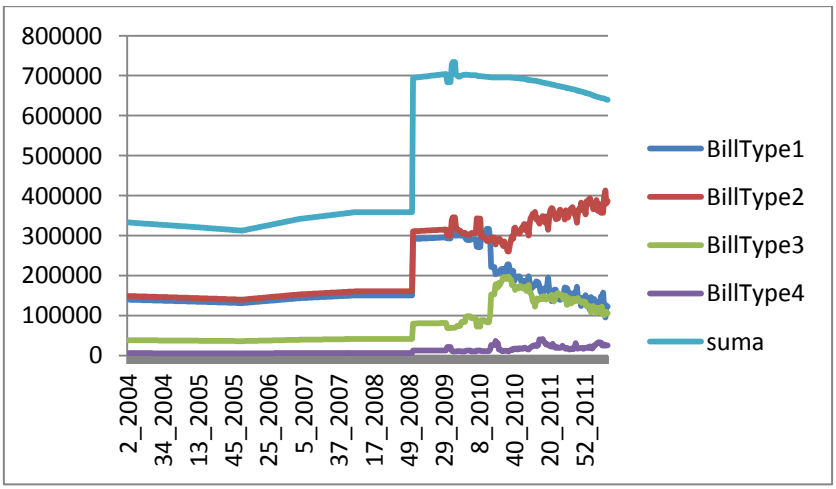


In this graphic we have represented the three types of portfolio and the sum of them. We can observe that the regulated market (PortFolio1) is decreasing while the free market (PortFolio2) is increasing. Regarding to the gas customers (PortFolio3), we see that in the first years, this variable doesn't have value but have been increasing since the year 2007. We can suppose that in the future, the customers will prefer gas than electricity.

In addition to this, we can represent de evolution of this variable with two histograms. In the first one, we represent the year 2009 and in the second one we represent the year 2012. The results can be seen in the following graphics:



It is very important to say that after some studies, we have decided to ommit this variable in our model. The main reason is because data aren't relevant for our study. Now, we are going to study the variable bill type:

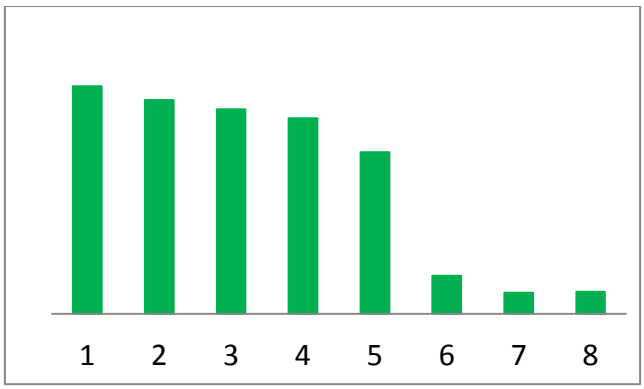


As we can see, there are an important structural break in the year 2009, where the behavior of this variable start to change.

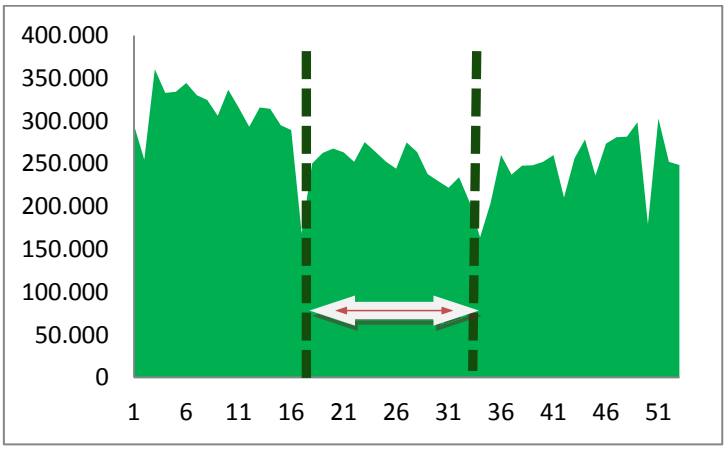
This important change is due to an European regulation: Iberdrola had to send the bills monthly instead of twice-monthly, so this fact made that the values of the four variables have been duplicated since this year.

Finally, we are going to explain the most important variable: calls. To do that, we represent different characteristics in the following graphics:
In the first one, we have represented calls distribution during weeks and holidays. From 1 to 7 represent the days of the week (from Monday to Sunday) and the number 8 correspond to holidays.

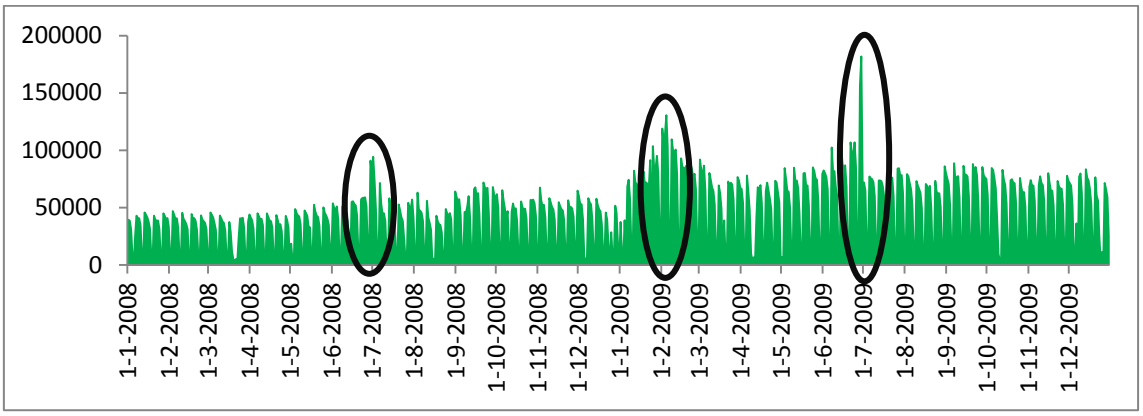
There are a mayor amount of calls on Monday and Tuesday in spite of Sunday, when the number of calls is very low.



In the next graphic, we represent the call seasonality pattern, where an important chance is produced between holidays (from July to September). It is due to people are on holidays and they are go out, so the number of calls decrease.



The last one is the representation of the variable calls for two years (from 1/1/2008 to 1/1/2010).



We can see three especial events, due to:

- 7/2008: technical problems
- 2/2009: European regulation
- 7/2009: technical problems

In next studies, we are not going to consider this events, because this fact makes more difficult to find the correct model.

Modelling

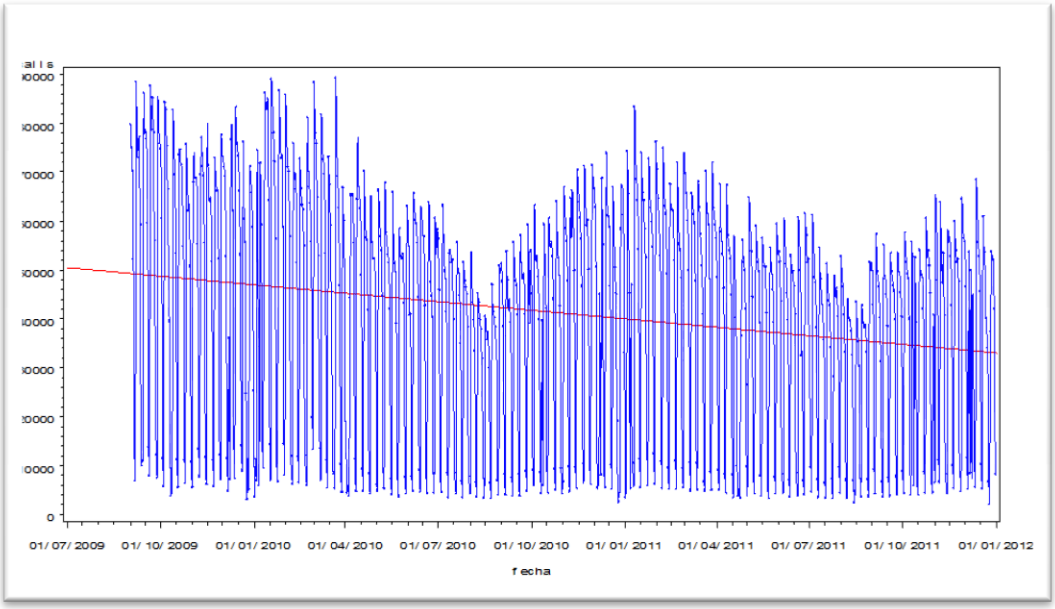
As we have been speaking previously, we are going to consider the data restricted to dates after 2009 because new regulations.

ARIMA model

In this section we analyze univariate model for our time series using SAS software. We choose SAS for his reliability and high capacity to work with this kind of models.

Time series analysis consists on several phases. First, the dynamic structure of the model is selected and then the parameters of the model are estimated. Diagnostic tests are performed to test the model assumptions, and the outcomes may suggest alternative specifications of the model. When an acceptable model has been obtained it can be used to forecast future values of the variable, in our case the number of calls in a call center. To do the analysis we will follow Box-Jenkins method that allows modeling time series.

First, we are going to analyze the evolution of data. If we graph it we will see time evolution data and data fluctuations that may be due to seasonal effects.



Time series evolution

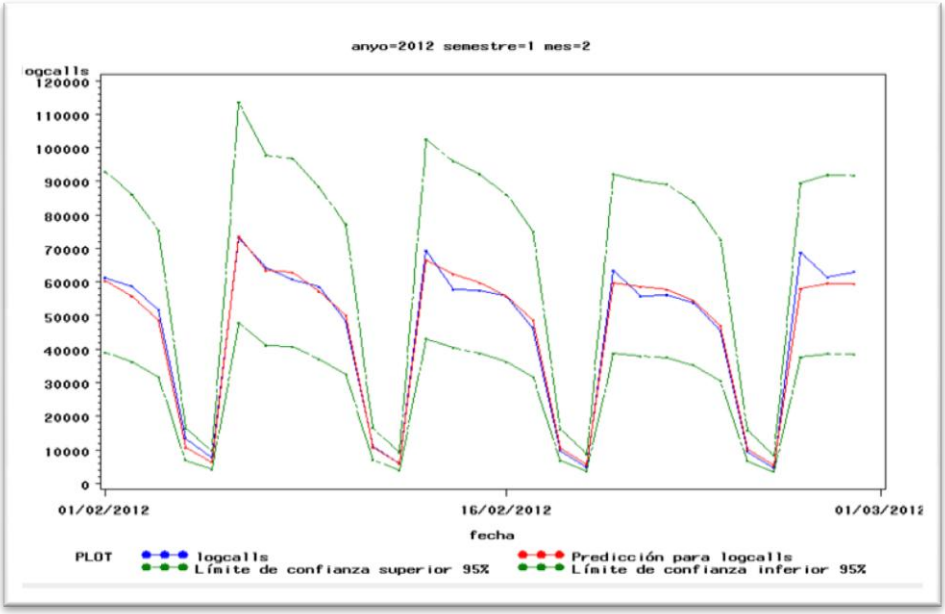
The series shows negative growth over time, therefore we will consider models for the logarithm of this series.

The model that we propose is an $ARIMA(1,1,1)x(0,1,1)_7$ without constant term, namely

$$(1-\phi L)\Delta\Delta_7 y_t = (1-\theta L)(1-\Theta L^7)a_t$$

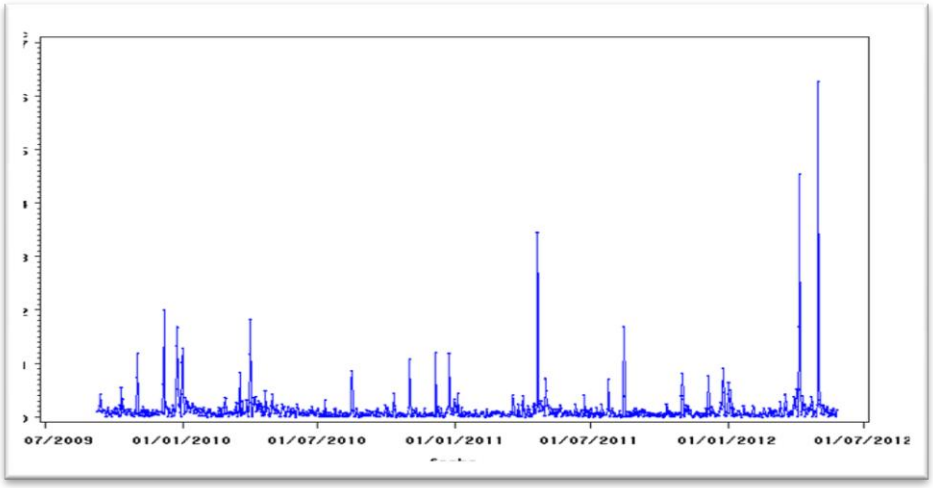
where $y_t = \log(x_t)$ and Δ, Δ_7 are regular and seasonally difference.

Graphically we can see that forecasting for the timeline of one day is a good accuracy for the real data.



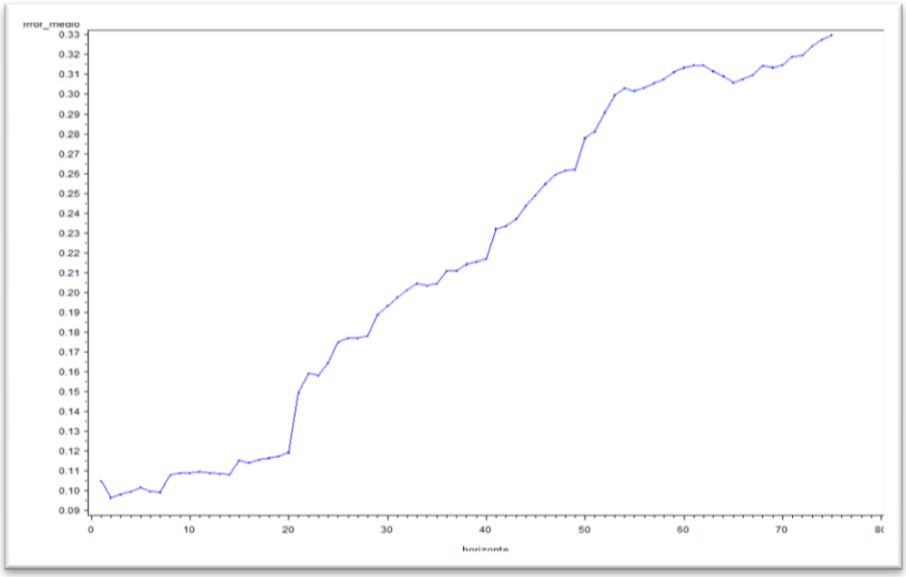
Validation data with forecast and interval confidence 95%

The next figure shows the average error for one day forecast. We can conclude that for one day model makes a good prediction because we have errors between 2% and 12%. The high values in the graph are due to unpredictable events.



Error lever to a one day forecast

This is a reasonable model to predict for a short time, but how shows the next figure for our problem, where we need to predict for a long time, the error grows when the horizon is greater. For the interval of 45-75 days, that is the interval we want to predict, we have average errors between 25% and 33%.



Simulation of the error for different temporal horizon

We can conclude, in agree with the theory, ARIMA models are generally suited for short time forecasts, but it fails for long term predictions. We reject the use of a ARIMA model to this problem.

A structural model

Once we have estimated the ARIMA model, the next step is to get a more realistic model. The main idea, now, is to incorporate some variables to try to include in the model more human behavior content. In such a way, we have deal with "transfer function models". A transfer function is a model that :

- Uses a dynamical model for inputs.
- Adds some seasonal events.
- Estimates an additional model for error term.

For instance, an extended formulation of the model could be as follows:

$$output_t = \alpha + \delta(L)input_t + seasonal + \eta_t$$

Where:

$\boldsymbol{\delta(L)}$ is the polynomial that parameterizes the transfer function. In general, this polynomial uses to be rational. But we have simplified in the way of a linear function

Seasonal, in this case, the long run seasonal cycle is estimated using an harmonic term composed by

$\varpi 1 \times \sin\left(\frac{2\pi t}{365}\right) + \varpi 2 \times \sin\left(\frac{2\pi t}{365}\right)$, where w1 and w2 have to be estimated in the model context and t is a linear trend. And the short run seasonal cycle is modeled using dummy variables as follows:

$D_{it} = 1\ if\ the\ day\ is\ the\ day$ I , 0 in other case.

Finally, we have:

$$\eta_t = \frac{\theta(L)\varepsilon_t}{(1-L)(1-L^s)\phi(L)}$$

Where ε_t is a white noise term, $\theta(L)$ is a moving average polynomial term[1], $\phi(L)$ is an autoregressive polynomial term and $(1-L)(1-L^s)$ represent the possibility of regular and seasonal differences if is needed.

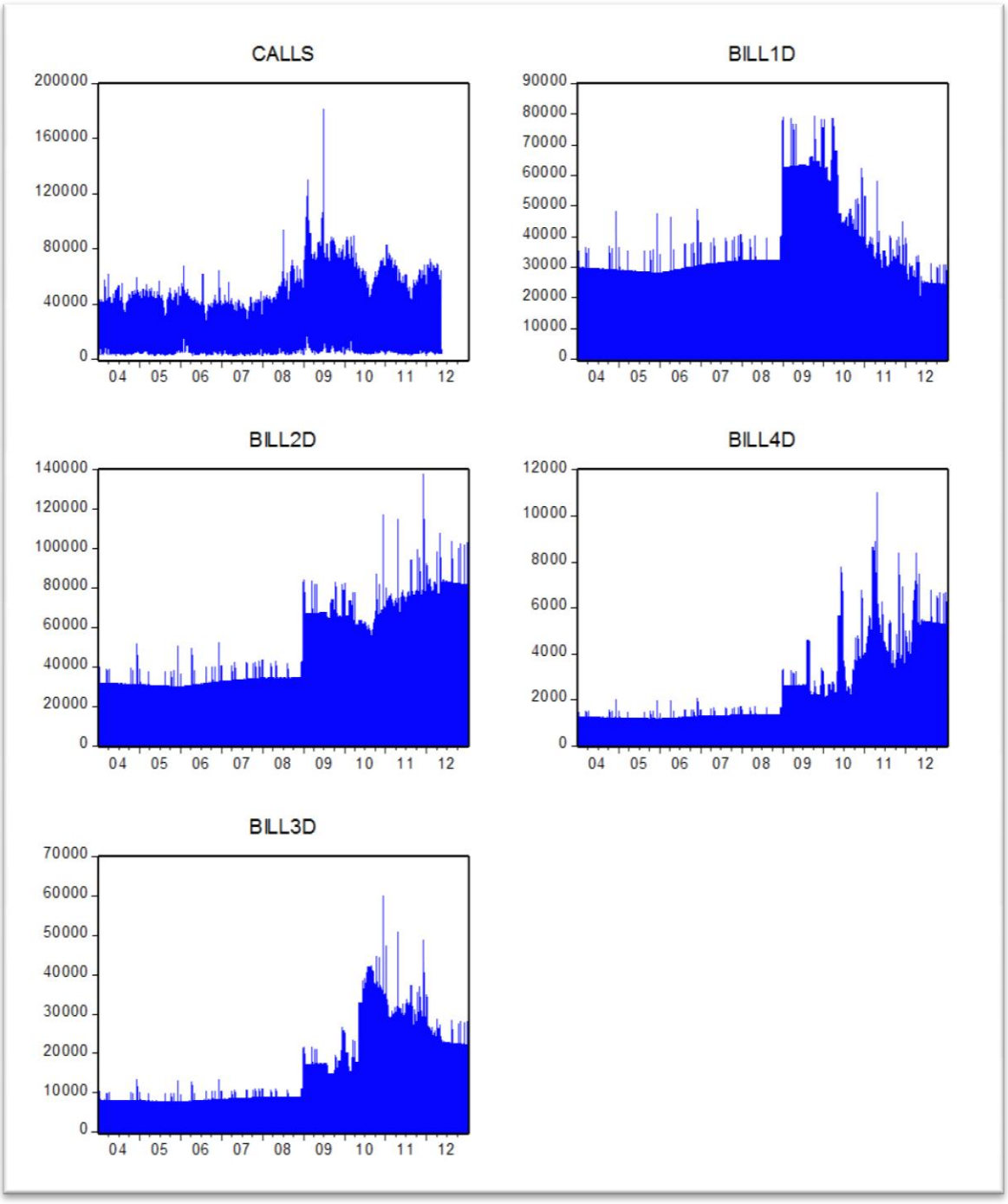
In order to simplify, we have created two variables:

- **NetBill2**: is the difference between Billtype2 and Billtype1
- **Neterror**: is the sum of Billtype3 and Billtype4

The idea is to eliminate the collinearity (it is said, a high correlation among regressors) to gain in estimating efficiency. So, we have summed up the Billing concepts that are closely related.

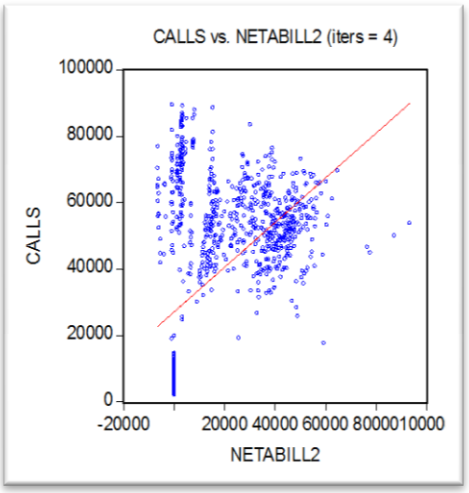
In addition, we use only a data partition. First we have eliminated the sample till sep 09, due to the structural break evidence among inputs and output (as can be seen in the next graph:

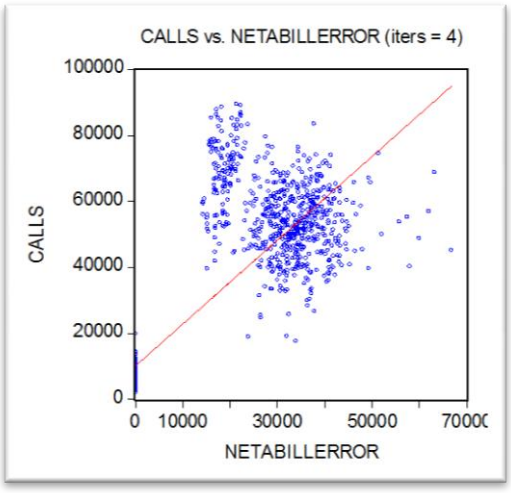
[1] L is a Backstage operator that means: $Y_t\ L^r = Y_{t-r}$



Note that bill variables are completely different previous 2009, so the relationship among then and calls variable could be distorted.

In a first exploration of data, we have obtained some kind of counterintuitive results. As it can be seen in the next graphs, there is a strong correlation in the contemporary moment among input and output:





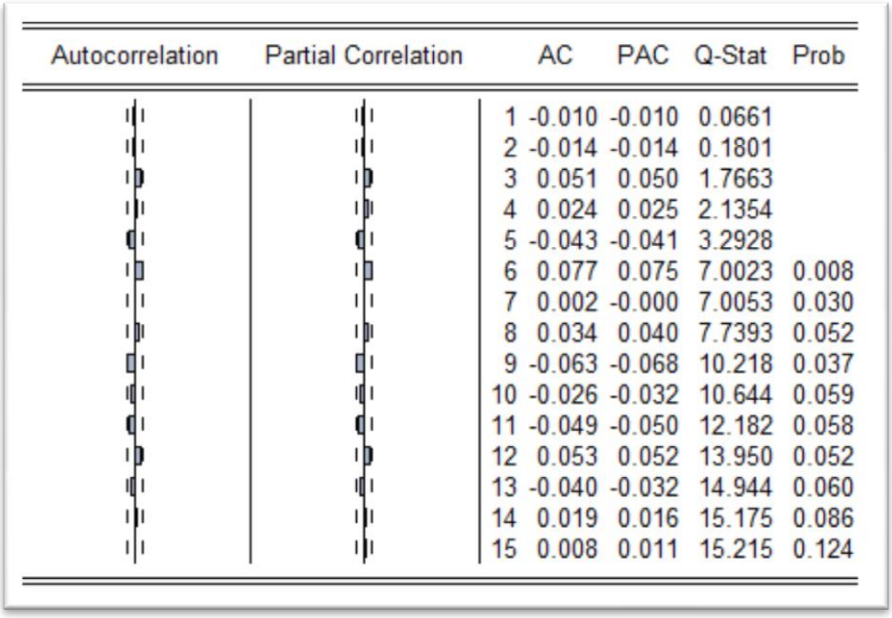
But in fact, we should not expect such a result. We should expect some lag in the response of customers to the bill reception. So, we assume that this model is not able to capture correctly the structural content of billing on calls due to one main reason. Billing data was originally weekly, and we had to transform it daily in a heuristic way. We attribute to this step the main measure error in the variable.

By the way, we will continue as this contemporary correlation has sense and we would like to purpose an improvement in the inputs variables to run new models more realistic.

In the next table, we show the model output explained. As it can be seen, the dynamics from input to output is very simple.

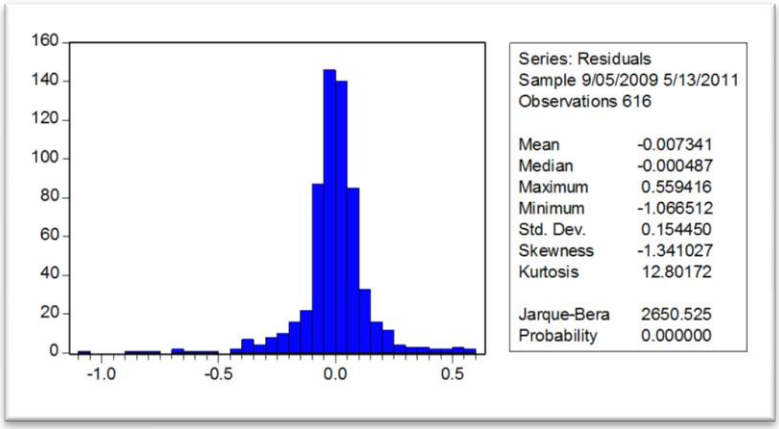
Variable	Coefficient(lag)	Explanation
$\Delta\log\left(Calls_{\ t}\right)$		Our target variable
$\Delta\log\left(NetError_{\ t}\right)$	0.20 (0)	An increase of 1% in billtype 3 or 4 **increases** 0.20% calls
$\Delta\left(NetBill2_{\ t}\right)$	-0.285(0) 0.141 (1)	An increase of 1000 billtype 2 over billtype1, **decreases** 1.4% calls
Long term seasonality	$-0.0021\sin\left(\frac{2\pi t}{365}\right)$ $+0.0010\cos\left(\frac{2\pi t}{365}\right)$	Is a parsimonious way to take long run cyclical movements
Daily pattern	$0.47\Delta Mon + 0.39\Delta Tue$ $+ 0.35\Delta wed$ $+ 0.12\Delta fri$ $+ 0.0\Delta sat$ $+ 0.0\Delta sun$	Dummy variables to catch deterministic daily fluctuations
Error term	$\frac{(1-0.36L-0.54L^2)(1-0.20L^7)}{(1+0.06L-0.23L^2)}$	Capturates autocorrelation pattern

Respect to model diagnosis, we can ensure a white noise structure in the residual, as it can be seen in the correlogram. Llung-Box Q stat allow us to say we could reject white noise hypothesis in the first 15 lags as of 13% significance, so, using 1%,5% and 10% as reference values, we can say this pattern is white noise.



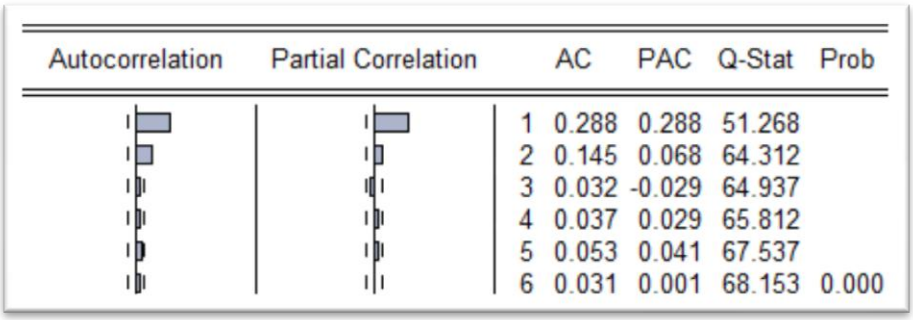
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.010	-0.010	0.0661	
		2	-0.014	-0.014	0.1801	
		3	0.051	0.050	1.7663	
		4	0.024	0.025	2.1354	
		5	-0.043	-0.041	3.2928	
		6	0.077	0.075	7.0023	0.008
		7	0.002	-0.000	7.0053	0.030
		8	0.034	0.040	7.7393	0.052
		9	-0.063	-0.068	10.218	0.037
		10	-0.026	-0.032	10.644	0.059
		11	-0.049	-0.050	12.182	0.058
		12	0.053	0.052	13.950	0.052
		13	-0.040	-0.032	14.944	0.060
		14	0.019	0.016	15.175	0.086
		15	0.008	0.011	15.215	0.124

But, analyzing error distribution trough residuals, we can see the non-normality apparience and the existence of fat tails:



Jarque-Bera test rejects the null hypothesis of normality, and as we can see in Skewness and Kurtosis, they are far to be the adequate values.

So, the next question could be: Are we dealing with a homocedastic variance in error term? The answer could be “No”. If we obtain the squared residual correlogram:



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.288	0.288	51.268	
		2	0.145	0.068	64.312	
		3	0.032	-0.029	64.937	
		4	0.037	0.029	65.812	
		5	0.053	0.041	67.537	
		6	0.031	0.001	68.153	0.000

We can see a possible AR structure for residual variance. So, it could be any GARCH-ARCH model type such as:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \hat{e}_{t-i}^2$$

Where we have written a GARCH(p). As we can see is a conditional model for residual variance. Is a conditional autoregressive model. The advantages of include this model are in efficience terms (so forecast accuracy could improve) and allow us to improve confidence bands.

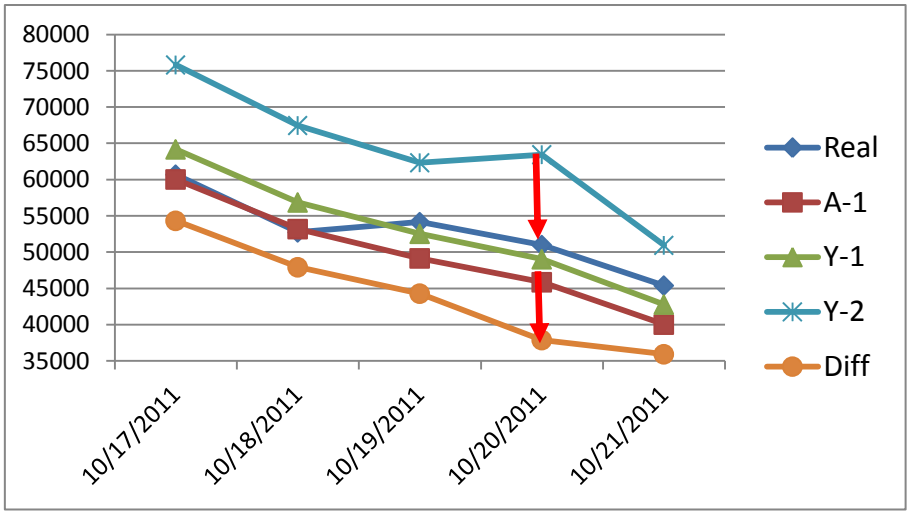
But at this point we decided not to modelize error variance due to its sensibility to atypical data. Until we don´t have knowledge about our atypical data (we need company feedback) we don´t think is correct to go the next step.

Some naives models

We will consider different naives models:

- Y-1: consider the number of calls of previous year.
- A-1: adjust the **Y-1** model with a real coefficient proportional to the difference with real data.
- Diff: apply the same variation (%) of the number of calls from Y-1 to Y-0 than from Y-1 and Y-2.

The next graph shows the evolution of this models in a week of our study.



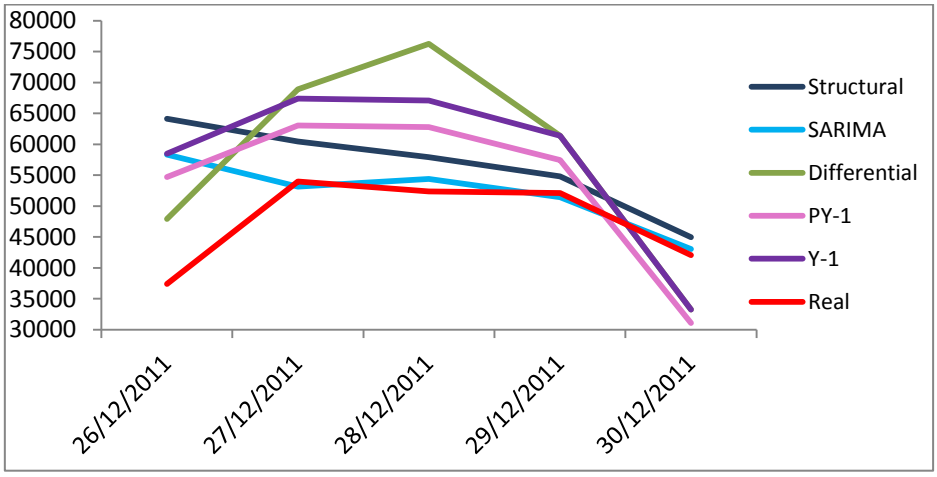
Evolution of naives models in a week

Validation and prediction

In this section we will analyze the mean error that our models done. In this table we can see the mean relative error in the validation model. It shows that *structural model* have the smallest error.

Structural	Differential	A-1	Y-1
9,31	12,40	10,32	11,68

Figure shows the prediction of each model in a week and the real value. We can see that the model who have smaller error is also the model who makes the best adjust to the real data.



Comparison of the different models

Following *Clements and Hendry* we have considerate a model based on the weigthened of the solutions generated by other models (Structural, ARIMA, Differential, A-1). The idea is make a linear combination of all models and study what combination makes the relative error minimum. We will name it the *weighted model*.

We define $J_{err}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ the relative error function who are computed considering the weights $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \Theta=[0,1]^4$. We minimize J_{err} in Θ by considering a Newton method and we obtain that the values of alpha variables are the following:

Estructural	SARIMA	Differential	PT-1
α_1 =0.6	α_2 =0	α_3 =0	α_4 =0.4

So the best combination includes structural method and A-1 method. For instance a week from inputs generated by Iberdrola.

Figure shows a weekly prediction using weighted model and the maximum and minimum value for each day of the simple models.

Conclusions and perspectives

Conclusions

- We have generated models with low error values (8%-15%).
- We have predicted the future number of calls from an unknown time interval.

Perspectives

- Comparer our solution when the real data are available.
- We would like to have additional variables such as campaigns, discretional events, etc.