Sales forecasting model. Analysis of covariables and inclusion in the model

IX MODELLING WEEK

Manuel F. Avilés Lucas, Ubay Casanova Blancas, Imanol Gago Carro, Lidia Gómez-Tejedor Fernández

Contenido

INTRODUCTION

Sales forecasts are becoming more important to businesses.

Forecasting is the basis for developing plans to run the business and decide future strategy. If sales can be reasonably anticipated, costs and inventory can be controlled, and customer service enhanced.

If we can accurately predict our sales, we can generate an appropriate timetable to optimize the working hours of employees. This also produces a significant saving for the company.

Also, the expansion of the company depends strongly of the forecast sales, so it is important to have good mathematical models that help us in this goal.

In the restaurant industry, demand is highly patterned: sales figures represent a year periodicity, but also monthly, weekly and even hourly patterns within the day arise. Although each restaurant has its own characteristics, it is expected that a global pattern can be obtained for all as a whole.

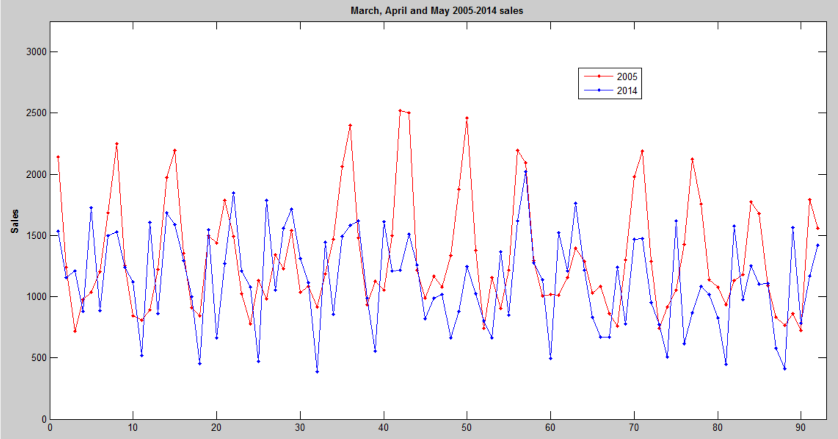


Figure 1. Sales pattern in different years.

OBJECTIVES

The company that challenges us to perform this work, Mapal Software, already has a forecasting system that uses an expert system analysis. Our goal is to improve this model. Also, we will try to create an automatic model that doesn’t need any human expert to predict future sales.

To do this, Mapal Software provide us data from 4 restaurants. All of them belong to the same restaurant chain. Obviously, each restaurant is located in a different place in the Madrid center. So we have to fit a non overfitted model which could be used in all the centers.

SELECTION OF DATA

In this kind of projects, the first decision to be made is about the data.

Mapal Software provides with daily data for each center:

- Sales.
- Location.
- Calendar special days.
- Current and past promotions.

Furthermore we could use data from seven different weather stations.

- Maximum and minimum temperature.
- Location.
- Precipitations.

We also had the chance to use temperature data for our forecasting model.

Splitting sales data

To fit the model, we split the data into two tables. First one was used in order to train the model (training data). Second one to validate it (validation data) with real data.

Finally, we have decided to split the training data from 1st of January of 2008 until 30th of June of 2013 and the validation data from 1st of July of 2013 until 21st of December of 2014. So, we had five years and a half to fit the model and one year and a half to check our model.

There are two reasons for this:

1- Promotional events of these centers start from 2008, and we want to capture all the possible information and the recent trend from the last years.

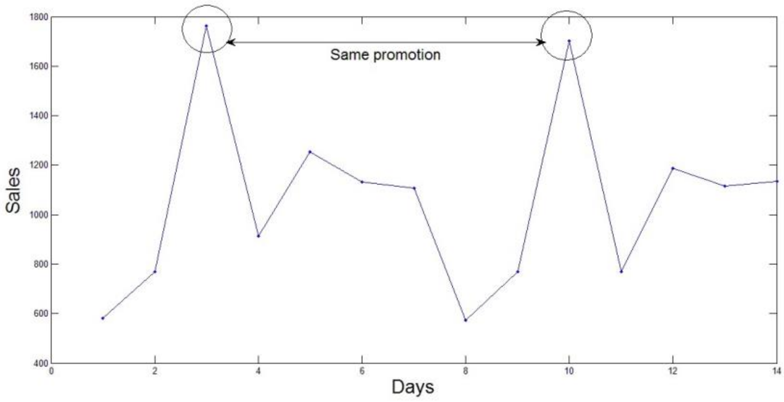


Figure 2. Promotional events

In the above chart, we can check how the promotions strongly affect to the sales.

2- Besides that, the behavior of the people from 2008 has changed due to the global economical crisis. We can check it in the next charts.

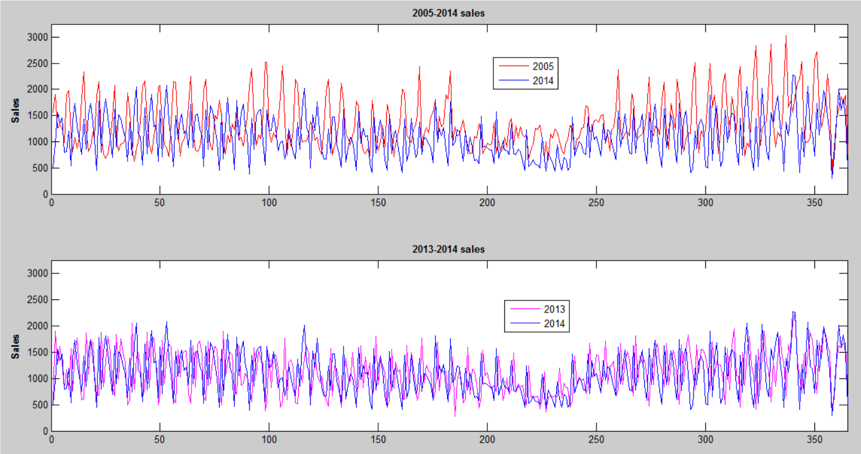


Figure 3. Comparison of patterns in different years

We can see how the sales patterns are similar in 2013-2014 and different in 2005-2014.

Weather data

The first decision we had to take is about the useful stations for our centers. Stations like Navacerrada are far away from the center of Madrid, so we decided not to include them in our model. Also, Colmenar Viejo Station has been removed due to its remoteness.

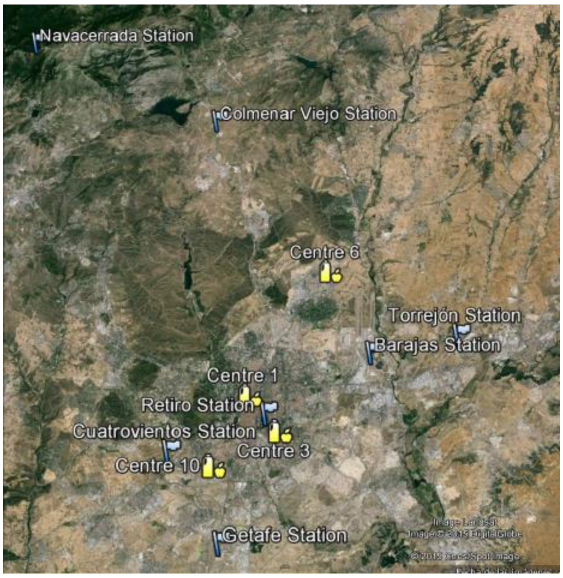


Figure 4. Location of centers and weather stations.

To take into account every station for each center, we have calculated a weighted average. This weighted average is calculated for minimum and maximum temperatures and for precipitation data. Thus, we can add information of all the stations in each center.

IMPLEMENTED MODEL

First of all if we take a quick look at figure 3, we can see very easily how there're consistent periodic fluctuations over time also known as seasonality. Besides all these ups and downs and this not linear general behavior is quite related to ARIMA analysis in time series. As we should expect when other approaches to this matter like multilinear regression were tried errors were much higher.

In order to build the model up we used data from 'Center 1' because it was considered like a good case to represent all the other restaurants. But we should notice that we recalculate all coefficient estimates using data from each center. So at the end we got fitted models for every center that's reason we obtained similar errors between centers.

SARIMA $(\mathbf{p}, \mathbf{d}, \mathbf{q}) \times (\mathbf{P}, \mathbf{D}, \mathbf{Q})_s$

Box-Jenkins methodology was followed. So identification, estimation and diagnosis were taken steps during all modelization process.

In one hand, identified orders from our SARIMA model (p, d, q) represent stationarity structure of our time series. In the other hand (P, D, Q) specify seasonality.

Afterwards doing all B-J modelization we got this model: SARIMA $(1, 0, 0) \times (0,1,1)_7$

This means that sales value at time t is affected by:

$$\underbrace{(1-\phi_1 B^1)}_{AR}\underbrace{(1-B^7)}_{I} X_t = \underbrace{(1-\theta_7 B^7)}_{MA} a_t$$

1-. Which sales value we got at time t-1. Autorregresive part. AR

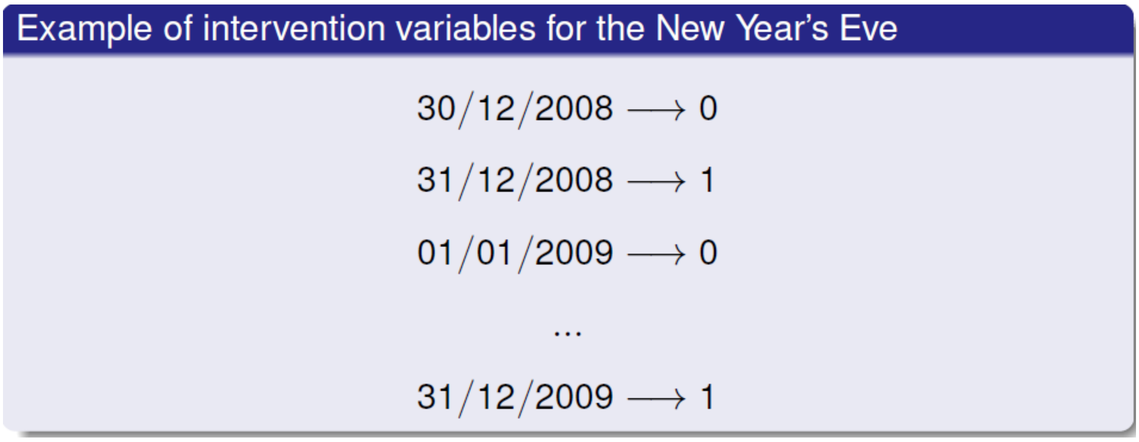
2-. Which error we did estimating same value the week before. Seasonal moving averages. MA(7).

Interventions + Transfer function

In order to include in our model some very important calendar effects we coded many binary variables which were calculated to improve the fitting at certain dates.

$$S_{t^*}(t) = \begin{cases} 1, & \text{if } t = t^* \\ 0, & \text{if } t \neq t^* \end{cases}$$

For example, It’s quite logic to think that when it’s a special day people will tend to visit more a restaurant that’s why we introduced a variable like this one for New Year’s Eve.



Example of intervention variables for the New Year’s Eve

$$30/12/2008 \longrightarrow 0$$

$$31/12/2008 \longrightarrow 1$$

$$01/01/2009 \longrightarrow 0$$

$$\ldots$$

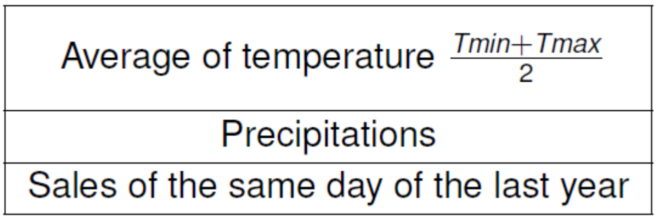
$$31/12/2009 \longrightarrow 1$$

Many interventions were tried and not all turned to be significant but the ones included really improved the general error with a relevant decline.

Before we talked a about data from weather stations. In SARIMA models the way we can implement these covariables is through transfer functions.

These transfer functions are a simplified version of an ARIMA model for every covariable. So what we’re doing is fitting an SARIMA for every covariable in order to be able to use it within our prior SARIMA.

In the next table we refer all covariables considered and included after their significance was tested.



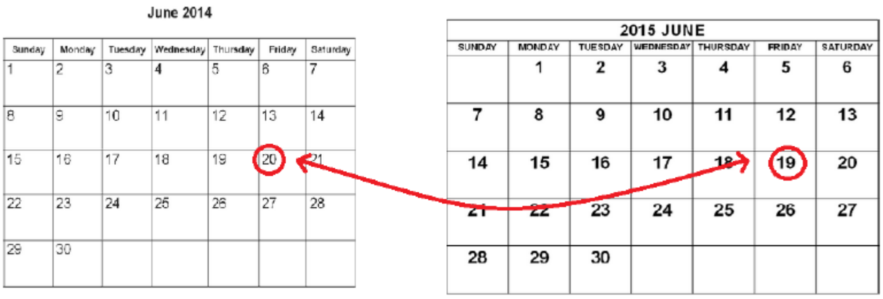
Average of temperature $\frac{Tmin+Tmax}{2}$
Precipitations
Sales of the same day of the last year

Let’s take a moment to explain covariable ‘Sales of the same day of the last year’. This covariable calculates for each date equivalent day from the year before and same weekday. We can understand this more easily looking at this next figure.

The comparable day is defined as the same day of the week in the calendar for the year of comparison.

Example

Friday 19^{th} June 2015 $\Longleftrightarrow$ Friday 20^{th} June 2014.



June 2014

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

2015 JUNE

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

Figure 5 Example of equivalent day.

Low frequency curve

Finally we implemented a low frequency curve in order to obtain better predictions in medium and long term.

This idea came from the fact that with SARIMA model as soon as we leave behind the highest order of our model (max(P,D,p,q)) we cannot provide accurate predictions only intercept will be given. Using smooth polynomials we can avoid this issue and get much better predictions.

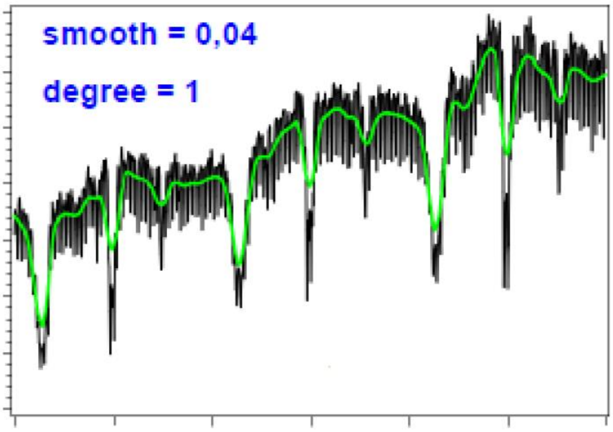


Figure 6. Adjusted low frequency curve to our time series.

RESULTS

Improvement of the error

The model made for forecasting the sales of each center has changed a lot from its first version to the final one. We fitted the model according to restaurant nº 1. The first model we implemented was a SARIMA $(1,0,0) \times (0,1,1)_7$ and the error committed by this primitive model was almost 18%. This error was high, so we tried to decrease it including holidays and other special dates in the model. The improvement was significant and we achieved an error of 16.32%, more or less.

Then, we decided to fit our model introducing a low frequency curve. At this point, the error was 15.67%.

We thought promotions and weather conditions could be significant and their inclusion would improve the error considerably. However, the difference between previous errors and the current one was actually only of 0.13%.

Finally, we got the best model until now implementing sales of the equivalent day of the previous year variable. It is important to recall the need of using this variable and not using the sales of the same day of the previous year. Including all those variables we obtained an error of 15.19%.

In the following table we can see a summary of the evolution of the model and which the error was in each step:

MODEL DEVELOPING	ERROR (%)
SARIMA $(\mathbf{1,0,0}) \times (\mathbf{0,1,1})_7$	17.98
+ Holidays	16.32
+ Loess	15.67
+ Promotions	15.57
+ Temperature and precipitations	15.54
+ Sales of equivalent day of the previous year	15.19

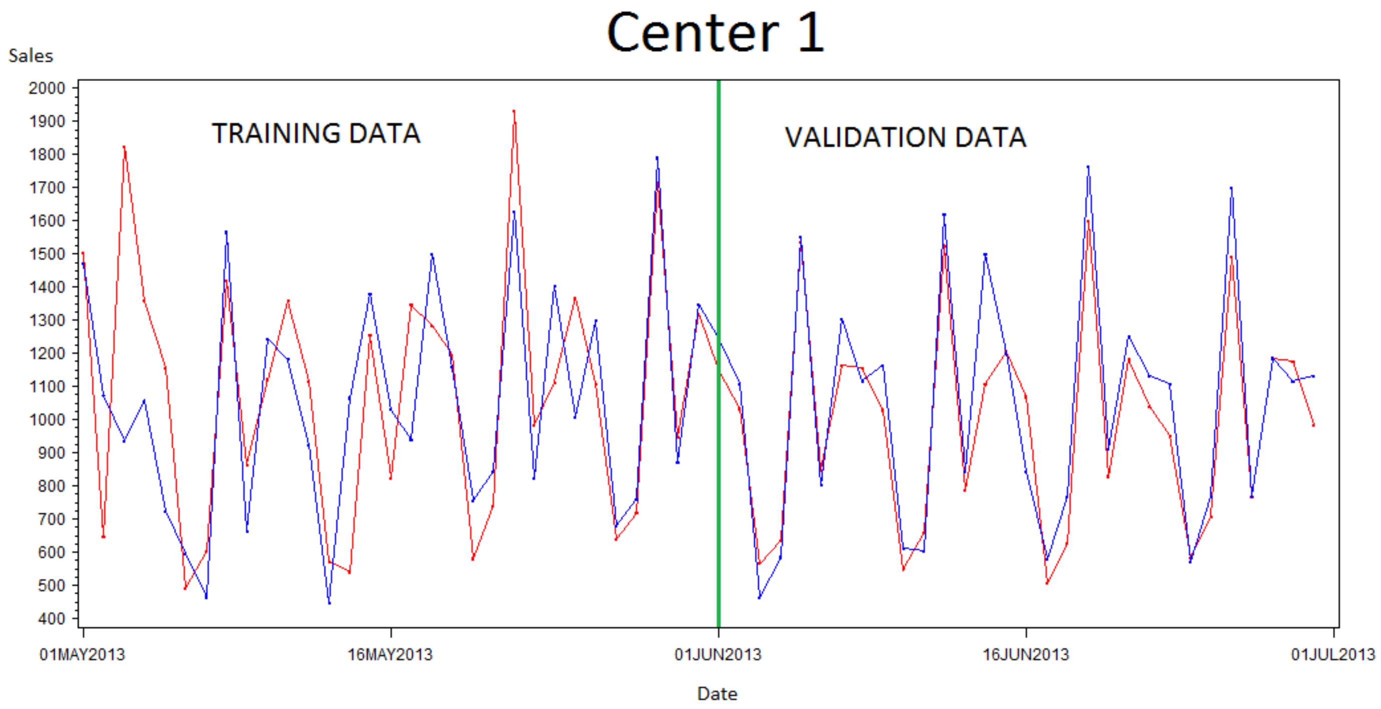
Other centers

Once we had a definitive model for Center 1, we recalculate all coefficients with the another three models. These are the results we obtained:

CENTER	TRAINING ERROR (%)	VALIDATION ERROR (%)
1	15.19	15.04
3	12.26	12.77
6	22	12.32
10	14.34	12.59

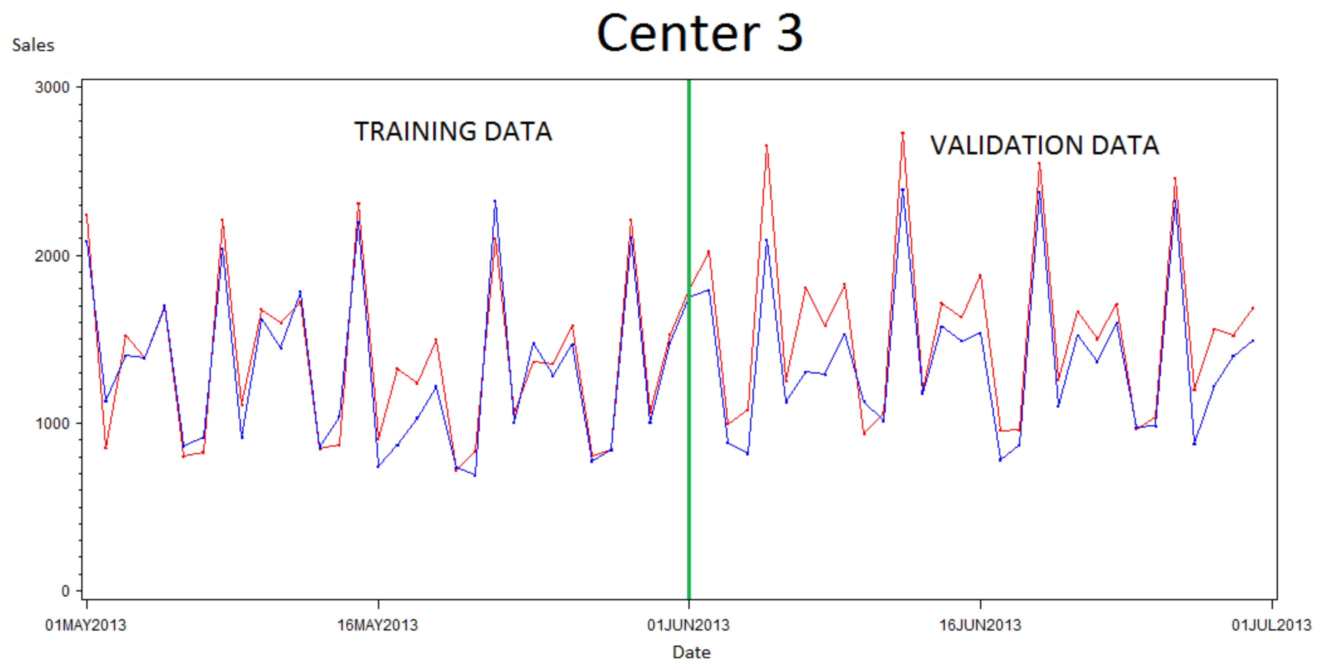
We realized that errors committed with training data for centers 1, 6 and 10 were larger than those that were committed in validation data. Moreover, the difference between them was small because the data from training and validation were similar. However, it is easy to see that in Center 6, there is a big difference between training and validation data. After studding this unique case, we reached to the conclusion that the "error" was because Center 6 hadn't the same properties as the others: it was in the outskirts of Madrid and it's located inside a mall.

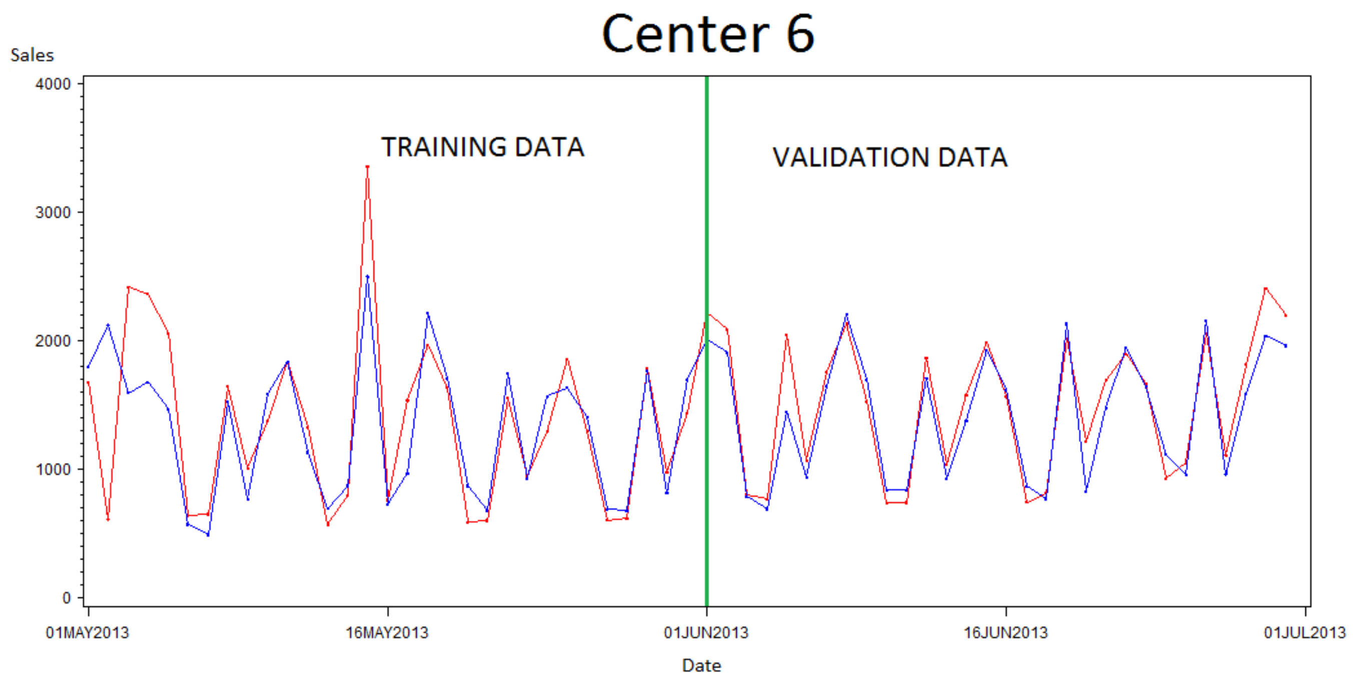
In the following charts, we can check easily the improvement of the error:

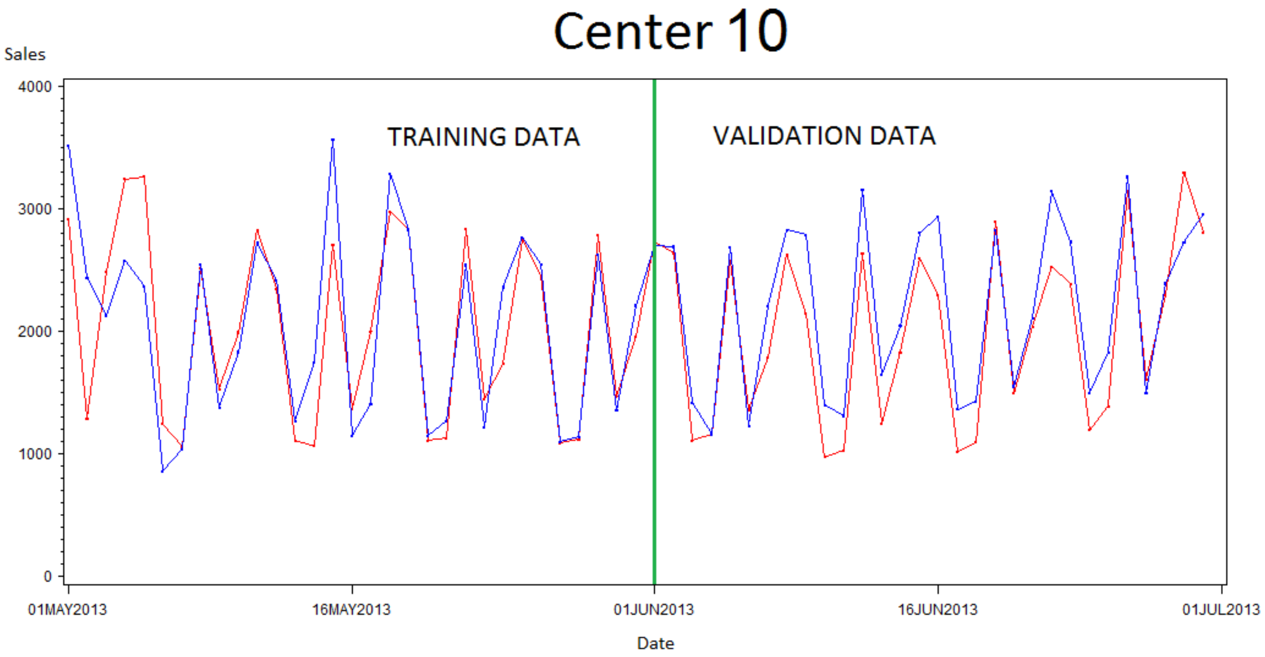


In the graph, appears in red colour our prediction and painted in blue we can see the real values of the sales. We can appreciate how, as time goes by, the forecasts are better because the time series become more and more stable. In the left part of the figure (training data), the forecasts we've done are not always accurate. Nevertheless, in the right side of the figure we are closer to real data.

For restaurant 3, 6 and 10 we obtained similar graphs. As we can see in the table of errors, Centers 6 and 10 have better predictions with validation data. On the contrary, Center 3 has a smaller error in training data:







CONCLUSIONS

1. SARIMA (1,0,0)x(0,1,1)$_7$ provides a good and not overfitted model for the restaurant data considered.
2. The inclusion in the model of variables that adjust the forecasting for days with promotions and holidays make an important decreasing of the error.
3. It has been proof how appropriate was to fit a loess smoothing polynomial, specially for medium term forecasting.
4. Incorporating the data for temperature and precipitations has not given much impact in the improvement of the forecast.
5. The sales of the equivalent day of the previous year has a significant performance in our model.
6. The training error obtained oscillates between 12% and 22%, that is, in a range about 10%. However, for three of the four centers, their training errors are quite close, they differ in less of 3%.
7. The error in the validation dataset goes from 12% to 15%, so it is quite similar for all the centers considered.
8. The errors in both, training and validation datasets are closed. That means that the model is not overfitted so is able to get the general performance (trend, cycle, …) from the past of the series.

9. We have obtained similar error rates in all the centers so, despite the model have been fit for one of them, the model is simple enough to make a good forecasting with different restaurants. So It is not a model for a particular restaurant but it is a model that explains well the behaviour of the sales in kind of centers considered.

FURTHER WORK

Some further tasks could be considered in order to improve our model:

1. Work with some new calendar variables. For example, adding intervention variables for each season or days when special events take place, such as important football matches.
2. Also we could include some social and economic (p.g. GDP) data in order to model other effects, external to the restaurant but that affects its sales. This would allow us to include some factors such as the influence of an economic crisis.
3. In addition, if data from many different restaurants was available, we could try to identify clusters within the centers with similar performance and then adjust the model for each cluster.
4. As new data is generated every day (not only for sales but weather data, promotions, etc.), it is recommended to implement an automatic process and schedule it to be executed periodically (every week or every month, for instance) so we could have constant updated coefficients in our model.
5. Finally, we could consider to implement a tool that detects if the sales suffer a significant change. For example, a threshold could be fixed so that, if the error rises above it, we could make some adjustments in the model: adding or removing variables, including the center into another cluster, and so on.