Time Series Analysis with SAS Evaluation

Sophie Gallet

DSTI - January 2020

EXERCISES

E1

Before looking at the data, I think about my goal: is it to analyze the data or to provide reliable predictions? As forecasting is not mentioned, and in the absence of any other context, I'll assume I'm asked by my colleague to perform an analysis of the time series data.

Looking at the raw data, I see two columns Date and Y, with 185 observations, ranging from September 1997 to January 2013. Data in the Date column are ordered and equally spaced in time (monthly intervals): we can plot it without any modification. The initial plot shows data that appears stationary (value of the time series is not dependent on time), with approximate mean of 62.6 and approximate standard deviation of 0.11.

I then use proc ARIMA, looking at the diagnostic plots as well as SCAN, ESACF and MINIC to **identify** candidate models: the PACF and even more clearly the IACF plots show an exponentially decreasing trend, while the ACF plot indicates a moving-average process of order 2. The ESACF and the MINIC's results are too in favor of a MA(2) model, while SCAN recommends first an ARMA(1,1) and second an ARMA(0,2) model. I note that they are no missing values.

Using proc ARIMA we now **estimate** the parameters for the 3 candidate models to see whether they're actually a good fit. We use the maximum likelihood method, recommended in class as the most accurate one. While the 3 models pass the Ljung-Box white noise test (high p-values for all lags, indicating we accept the alternative hypothesis that model residuals are white noise) and are close to be normally distributed, the parameters for MA(2) are all significant at the 0.05 cutoff (not the case for MA(3)). It also has the lowest standard error estimate and the lowest Akaike Information Criterion, which indicates the goodness of fit. The MA(2) model gives the following equation:

$Y_t = 62.60008 - 0.56908\varepsilon_{t-1} - 0.30819\varepsilon_{t-2} + u_t$, where u_t is white noise, and ε_{t-i} is the past error for data point *t-i*.

We finish this exercise by looking at the **forecast**. The model fit on past data seem correct, while the forecasted values end up quickly in a straight line centered at 62.60008 as we can't compute errors without the actual value.

In brief, we've been able to model our time series using an order 2 moving average model, which gave us the most satisfying fit among the various candidates model.

E2

For this exercise, I also assume that my goal is to analyze the time series.

Looking at the raw data, I see two columns again Date and Y, with 200 observations, ranging from June 1999 to September 2007. Data in the Date column are ordered, equally spaced in time (monthly intervals): it is a time series. The initial plot shows data that appears stationary, with approximate mean of 52.9 and approximate standard deviation of 0.11.

I'll proceed with the PROC ARIMA again, with the maximum likelihood option. There are no missing values. Looking at the plots, I'd probably choose a MA(1) as the PACF, IACF appear to be exponentially decreasing, while lag 1 is above the significant region, but as it's hard to say, I'll defer to identification methods. They suggest MA(3), ARMA(2,1) or ARMA(2,2).

I estimate the parameters for the 3 candidate models, and only ARMA(2,1) has parameters that are all significant. This model also passes the White Noise test.

The ARMA(2,1) model gives the following equation:

$Y_t = 52.90186 - 0.41283\ Y_{t-1} + 0.44116\ Y_{t-2} - 0.59390\varepsilon_{t-1} + \mu_t$, where u_t is white noise, and ε_{t-i} is the past error for data point *t-i*.

Looking finally at the fit and forecasts graph, our model seems like a good fit, and provides a 95% confidence interval centered at 52.9 of approximate width 0.4.

E3

Looking at the raw data, I see two columns Date and PercentUnemployed, with 61 observations, ranging from 1947 to 2007. Data in the Date column are ordered and equally spaced in time (yearly intervals): we can use it as is.

I perform the Ljung-Box White Noise probability test on the time series via PROC ARIMA. My hypothesis are:

- H0: The time series analyzed by the test is a white noise.
- Ha: The time series analyzed by the test is not a white noise.

Looking at the resulting table, first there are no missing values, and the time series' average is around 5.56, its standard deviation is around 1.46. The p-values for both lags are very small. We can thus reject the null hypothesis and accept the alternative hypothesis: the time series is not a white noise, and we can then proceed to modeling it.

E4

The goal of this exercise is forecasting.

Looking at the raw data, there are 4 columns including Date and Biscuits, with 52 observations, ranging from Jan 6 2008 to Dec 28 2008. Data in the Date column are ordered, equally spaced in time (weekly intervals): it is a time series with data ranging over a year. The initial plot shows data that appears stationary, with approximate mean of 376 and approximate standard deviation of 13.5. It almost looks like white noise, but the White Noise test is negative (at the 0.05 cutoff for lag 12).

As instructed, I use PROC ESM, along with PROC TIMESERIES to identify the best model. There doesn't appear to be seasonality, nor trend. I will stick to a simple model, and test the various candidates among simple exponential, double exponential, holt or damptrend. Damptrend appears a little bit better with the lowest RMSE (and null adjusted R squared..) so I will use this model for my predictions, which turn out to be on an almost straight line centered at 375. As I didn't find a trend or seasonality, my time series tools are not very helpful.

Note that I initially fitted an additional Winters & a seasonal model that both got perfect scores, but realized they were simply replicating data from a year ago identically for the forecasts, and were not performing well when using holdout. I wasn't able to find/tweak seasonality to a satisfying level.

CASE STUDY

My brief is to

- provide a 16 month forecast for 3 products
- detail the steps followed to choose each model
- write a quick report for the Sales Department

Forecasts with steps

Combined overview

I start looking at the combined raw data: the data is ordered by date, from September 15 to August 18, equally spaced (monthly intervals) and consists of 3 columns: product reference {FR001, ESA15, WW01A}, date and sales quantity. Next, I look at a combined graph that shows strikingly the dominance of the FR001 product sales over the next two product sales, as well as the big changes month-to-month for FR001 (143,000 sales in Oct 15 vs. 26,500 in Feb 16). While it may be tempting to focus mostly on the FR001 for these reasons, as I do not know the price of each product, and without further instructions, I will invest as much time in each of the 3 product forecasts.

Because it is simple, a good forecasting tool, and that I don't have flags or inputs to take into consideration, I'll use the Proc ESM method to build the forecasts of the 3 products [NOTE].

FR001

Using a simple plot as well as the Proc Timeseries for diagnostic, I can see seasonality and some small trend, although it's hard to say whether it's linear, increasing, or else. There doesn't appear to be any final effect.

I then use the PROC ESM to test several candidate models, take a look at their model fit and forecasts using the graphs, as well as their statistics. The Additional Winters model, which is meant to model seasonality with trend but without final effect, ends up being the best fit: it has the lowest RMSE, MAPE and AIC and the highest adjusted R squared, and the graphs look correct.

ESA15

Using the same steps as with FR001, I see in the time series seasonality, a linear increasing trend and perhaps a final effect. Testing several candidate models yield the same best fit as for FR001: the Additional Winters model. However I notice that the R squared and adjusted R squared are around 0.57, when they were around 0.86 for FR001: this second ESA15 model doesn't explain as much variation as the FR001 model is.

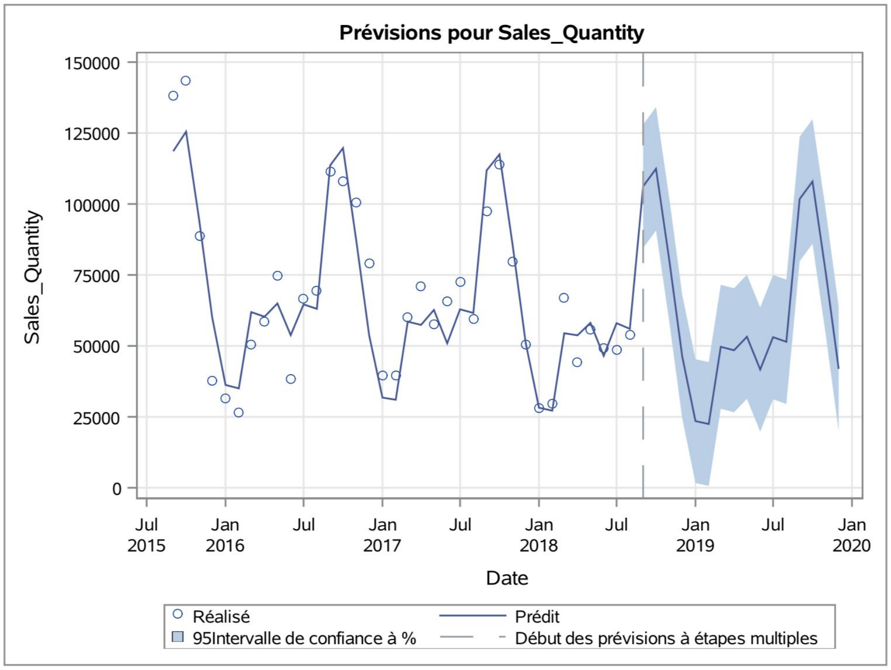
WW01A

For the third product, I use the same steps again, but I quickly notice there are 3 missing data for Jan 16, Jul 16 and Sep 16, bringing the total count of observations to 33. Looking at the diagnostic plots I see seasonality, a linear increasing trend, possibly a final effect (decreasing). Without context, it is hard to categorize the first spike (Oct 15, ~6000) as an outlier, a terrific month, or the result of a promotion. When I move forward I find that the additional Winters is again the best model (smallest RMSE, MAPE, highest adjusted R squared), but results in a disappointing adjusted R squared of 0.26.

Sales Report

This report and the analysis behind it were made using the provided past data for the 3 products: FR001, ESA15 and WW01A. I analyzed the trends in the sales evolution for the products so far, and provided 16-month forecasts, along with 95% confidence intervals. The tables with the numeric forecasts can be sent if needed.

FR001

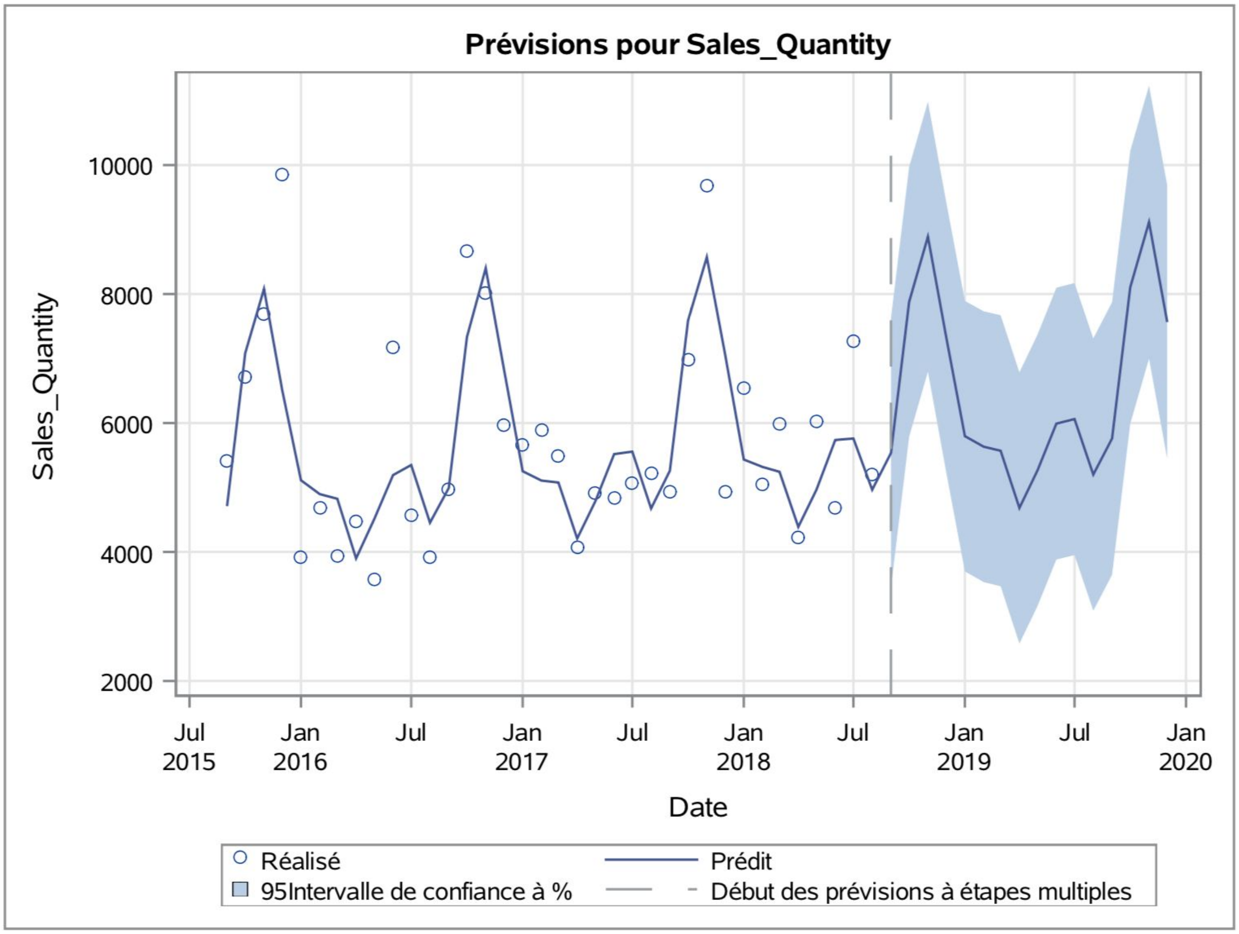


The sales of FR001 are seasonal: they tend to be the lowest in January and February (25-40,000), somewhat stable from March to August (50-70,000) and they peak in September, October (100 - 140,000), to then sharply decrease again to the year low. There is in addition a slightly decreasing trend.

The forecasts reflect these two elements: we're expecting the sales peaks for 2018 and 2019 to be in October and September, and the lows in January, February. Overall, if the current conditions remain, we're expecting slightly lower numbers than in the past.

Note that the evolution of sales for this product is somewhat easily modelized: we're confident in our analysis and in our forecasts, if conditions remain the same. This translates on the graph into very thin 95% confidence intervals.

ESA15

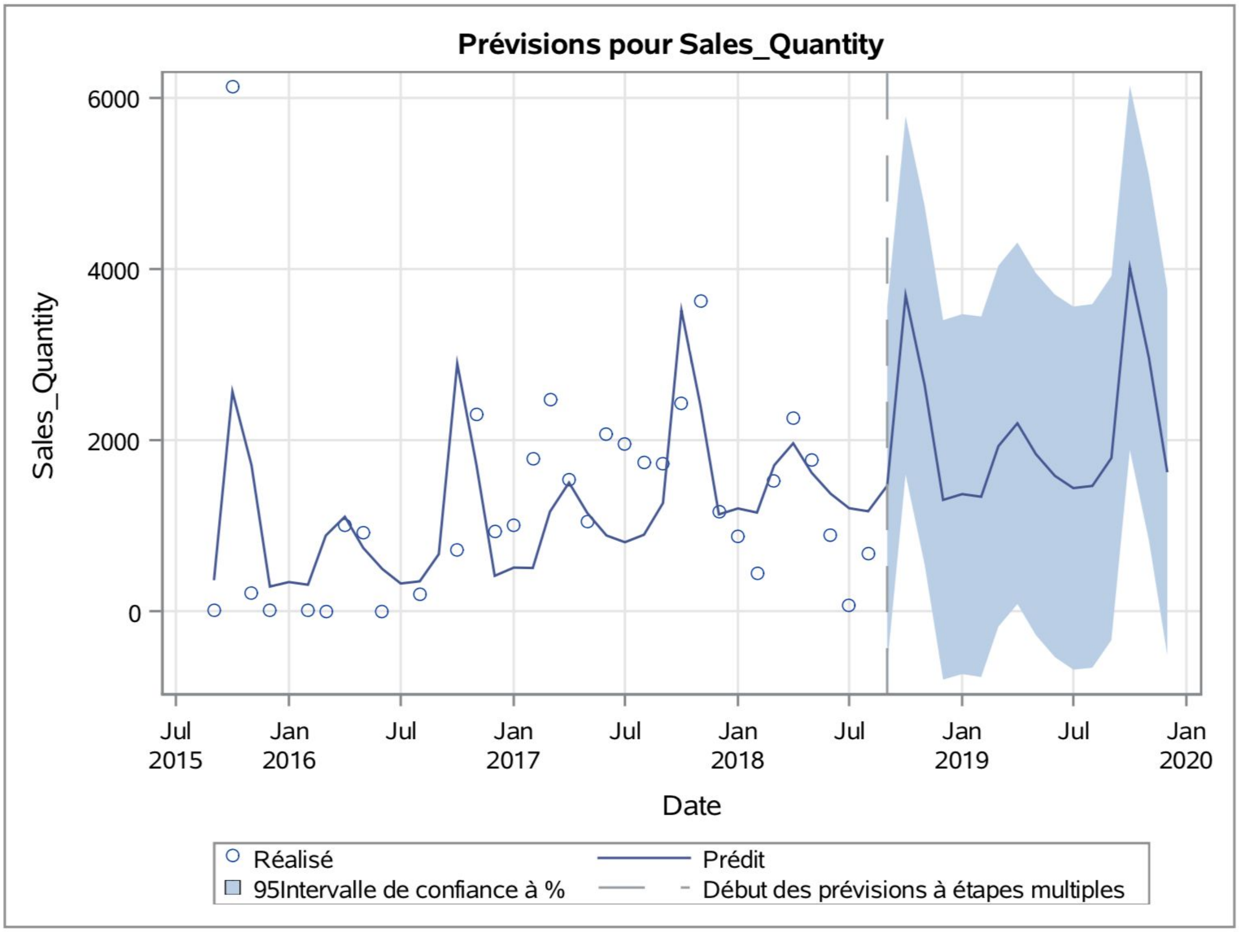


Sales of ESA15 are also seasonal but in a slightly different way: there's a peak in November, sometimes the month before or after in the 8,000 - 9,000 range, but outside of peak season, sales oscillate between 4-6,000. There is a slow increasing trend, which results in sales forecasts that are higher than past sales quantities.

On the graph, you can see that our 95% confidence intervals are wider than for FR001, as the sales evolution for this product is harder to model and forecast.

WW01A

Due among other things to missing values and the unexplained peak of October 2015, the accuracy of this model is not adequate. For information, we're still attaching the graph and a brief, but both should be taken with caution.



There is a seasonal trend, with two peaks, the highest one above 2,000 happens in October/November, and a second smaller one happens around March/April. There's an increasing trend: sales are overall growing. Note that the October 2015 value doesn't fit in the model.

Our forecasts show this increasing trend, and the wide 95% confidence interval is the graphical display of the uncertainty in this model.

NOTES

As is work in real life, this exercise came with time constraints. When reviewing my work, I noticed a few things that could be improved, or could be used as “learnings” for future similar works:

- Provide more context to the augmented Dickey Fuller test in E1, E2.
- Take a closer look at outliers
- In the case study, I first developed models using PROC ARIMA but wasn’t satisfied with the results. I realize now that it could be interesting to include as inputs the sales quantity of other products and/pr work more in depth with trends.

ANNEXE

The code for the exercises and the case study can be found in the zip file, along with the generated reports and graphs from the code for each section.