Sales Forecasting Project

Planning the analysis

The company's supply chain management needs monthly sales forecast (four months) to plan the production of video games according to customer demand. For this purpose, a data set with the company's sales from January 2008 to September 2013 has been available.

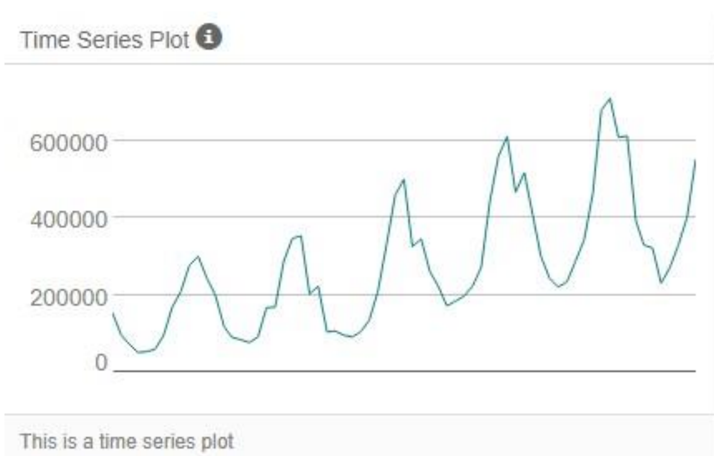
The problem is to predict future values based on historical results (predictive model rich in data). On the other hand, the target variable (monthly sales), besides being numeric, is based on time. Therefore, the statistical model indicated is the time series forecast.

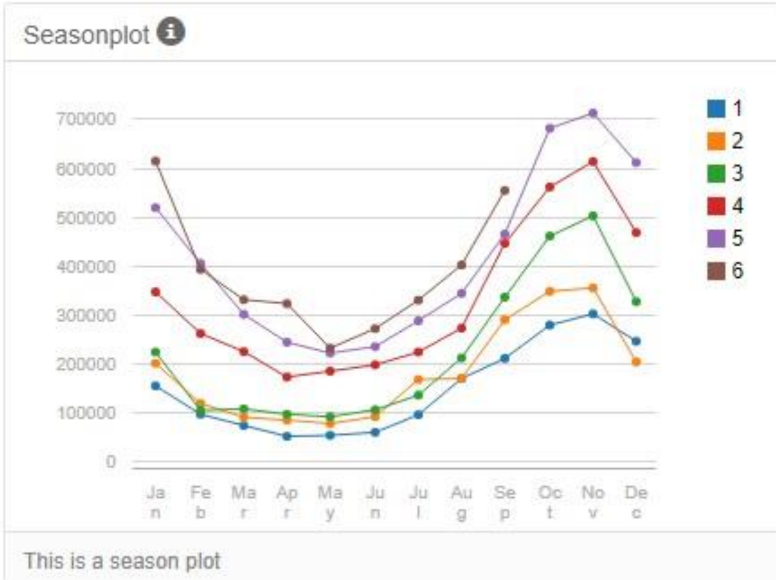
The dataset made available – sales from January 2008 to September 2013 – meets the requirements of the model, since the interval is continuous (there are no interruptions between the initial and the final period), the measures are sequential (consecutive periods) and with equal spacing (monthly), and each unit of time has only one corresponding data.

Since the model should predict four months, it is necessary to use a four-period validation sample; therefore, the months from June to September 2013 will be separated for later validation of the model. For the final forecast, however, all available data will be used: 69 months from 2008 to 2013.

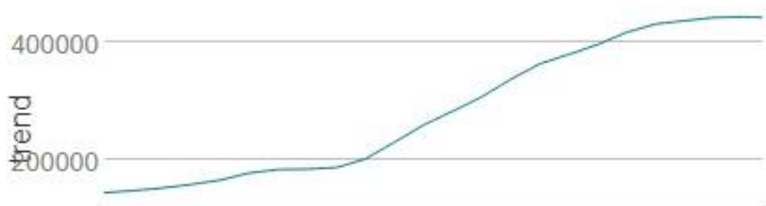
Determining the components trend, seasonality and error

The Alteryx software's "TS Plot" interactive report shows the following graphs for the time series under analysis:

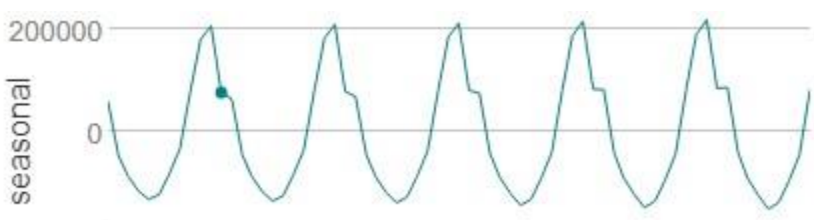




The “Decomposition Plot” of this graph shows a linear trend, so that this component must be applied in the model in the additive form:



The “Decomposition Plot” also evidences a growth of seasonality over time, although very subtle, which results in the need to apply this component in the model in the multiplicative form:



As for the error component, the “Decomposition Plot” does not present a constant variance in time (peaks and valleys are not uniform over time), which is why it must be applied in the model in the multiplicative form:

Thus, the model to be constructed will be ETS (M, A, M).

Building the models

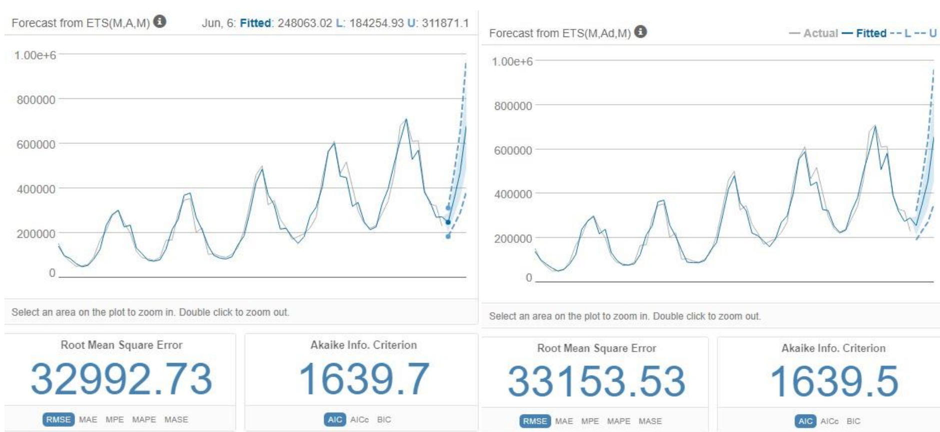
As the components error, trend and seasonality should be applied in the ETS model in a multiplicative, additive and multiplicative form, respectively, the model constructed was ETS (M, A, M).

The comparison between the ETS models with and without the dampening component reveals a better ETS_damp performance, since the forecast values are closer to the real values:

Actual and Forecast Values:

Actual	ETS_MAM	ETS_MAM_damp
271000	248063.01908	255966.17855
329000	351306.93837	350001.90227
401000	471888.58168	456886.11249
553000	679154.7895	656414.09775

On the other hand, as the AIC (Akaike Info. Criterion) of both ETS models – without and with dampening – are very similiar (1,639.7 e 1,639.5) – an analysis of the error measures is necessary to define the best option. It should be noted that the AIC measures the relative quality of a statistical model, as it balances the quality of fit of the model and its complexity.

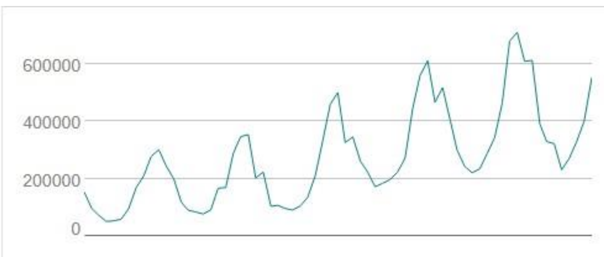


Accuracy Measures:

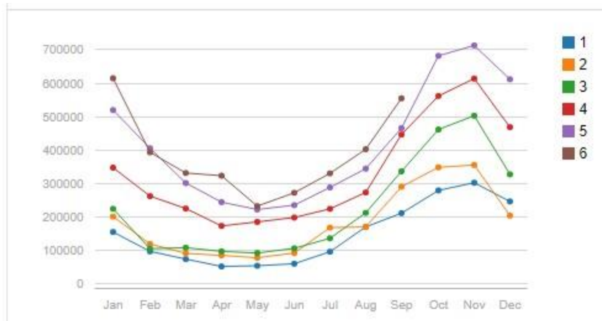
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MAM	-49103.33	74101.16	60571.82	-9.7018	13.9337	1.0066
ETS_MAM_damp	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116

The MASE measurement, understood as the absolute mean error of the model divided by the absolute mean value of the first difference in the series, measures the relative reduction in error in relation to a naïve model and must be less than 1. This measure indicates the best model as the ETS_damp. The RMSE measure, in turn, represents the standard deviation of the differences between predicted and realized values. Considering that the measure for the ETS model with damping is the smallest, this is the best model to continue the analysis.

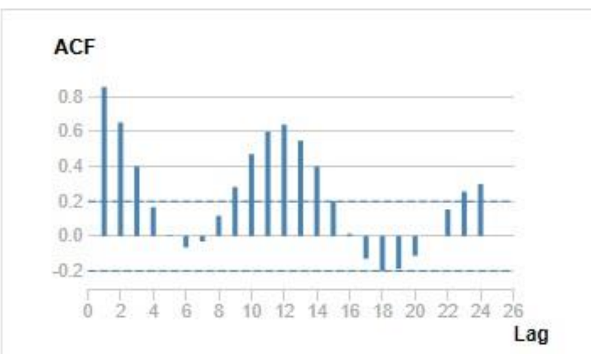
For the construction of the ARIMA model, it is necessary to verify initially if the time series is stationary, with mean and variance constant over time. In case, as the series has tendency and seasonality, it must be transformed into stationary first. The report generated by “TS Plot” tool from the sales data confirms that the series does not have this characteristic:



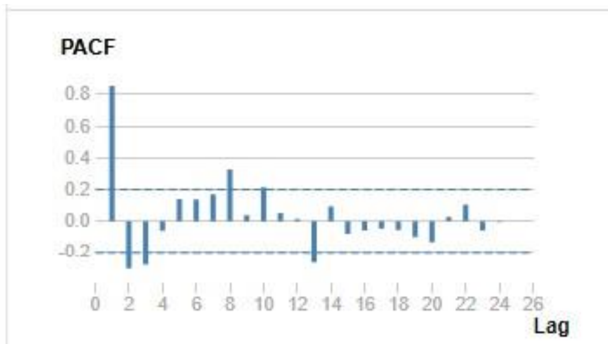
This is a time series plot



This is a season plot

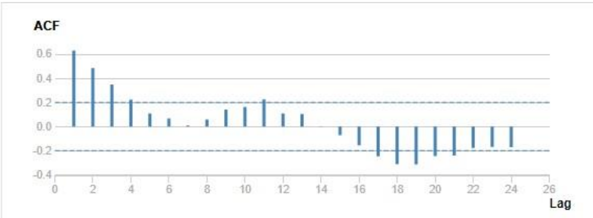


This is an autocorrelation plot

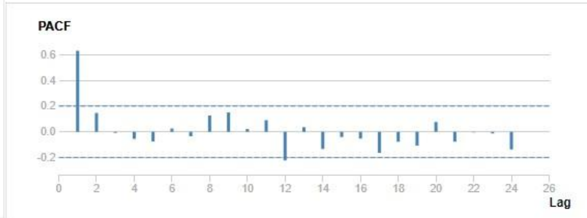


This is an partial autocorrelation plot

The result of a first differentiation – seasonal – can be observed in the following graphs:



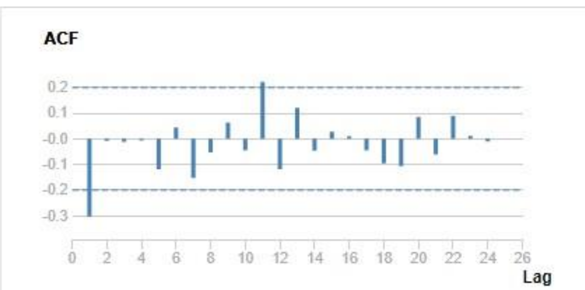
This is an autocorrelation plot



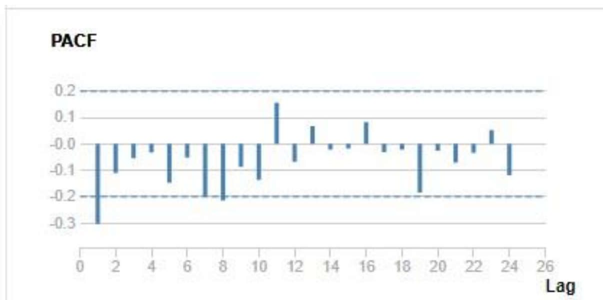
This is an partial autocorrelation plot

The ACF autocorrelation graph, however, shows that the series is not yet stationary and that it must be differentiated again, since the variance is not yet constant. In addition, the PACF graph of this seasonal part of the model shows a Lag 0 positive, which denotes the non-existence of the moving average component. Thus, the ARIMA model, in the seasonal part, should have the terms “p”, “d”, “q” equal to 0,1,0.

After a new differentiation, the series can already be considered stationary, so the term I or “d” of the ARIMA model must be 1 for both the seasonal and non-seasonal parts:



This is an autocorrelation plot



This is an partial autocorrelation plot

On the other hand, the graphs above show a negative correlation in Lag 1; in addition, the ACF truncates abruptly after the first lag and the PACF falls more gradually. These are indicators that the term to be used in the model is MA or moving average component. The autoregressive component (AR), in this case, is 0. Thus, the ARIMA model, in the non-seasonal part, should have the terms “p”, “d”, “q” equal to 0,1,1.

The ARIMA model to be used, therefore, is ARIMA (0,1,1) (0,1,0) [12]. Through Alteryx software, it can be observed that its main measures are: AIC de 1.256,5967, RMSE de 36.761,5281 e MASE de 0,3646:

Method: ARIMA(0,1,1)(0,1,0)[12]

Call:
Arima(Monthly.Sales, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

	ma1
Value	-0.378032
Std Err	0.146228

sigma^2 estimated as 1722385234.94439: log likelihood = -626.29834

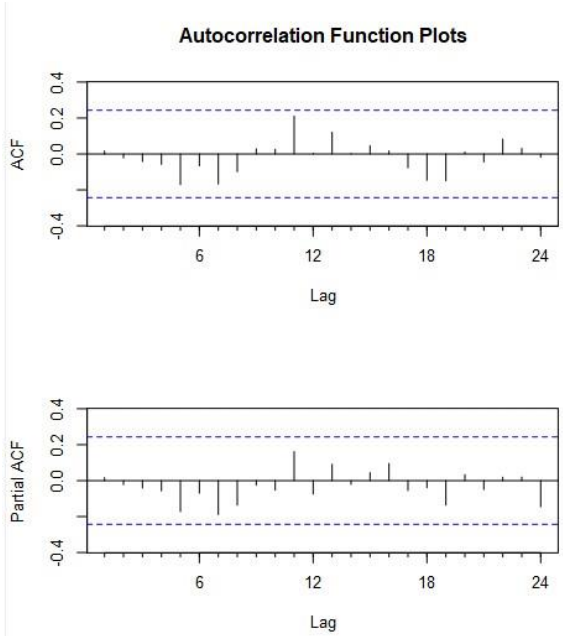
Information Criteria:

AIC	AICc	BIC
1256.5967	1256.8416	1260.4992

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145

Ljung-Box test of the model residuals:
Chi-squared = 16.4458, df = 23, p-value = 0.83553



Predicting Sales

The ETS_damp model (M,A,M) presented AIC of 1,639.5 and the ARIMA model presented AIC of 1,256.5967. Since ARIMA has a smaller AIC, it should be considered the best time series prediction model for the specific case.

Furthermore, after the union of both models and their connection with the Alteryx "TS Compare" tool, the respective comparison report reveals smaller measeures for the ARIMA model, and it should be noted that MASE measure of this model (0,45) is lesser than the ETS (0,81):

Accuracy Measures:

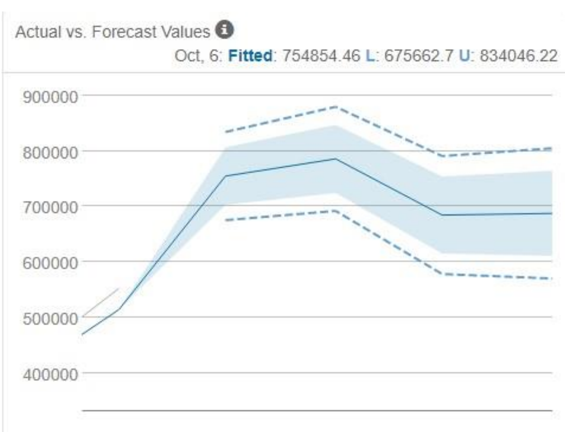
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MAM_damp	-41317.07	60176.47	48833.98	-8.3683	11.1421	0.8116
ARIMA_011_010	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532

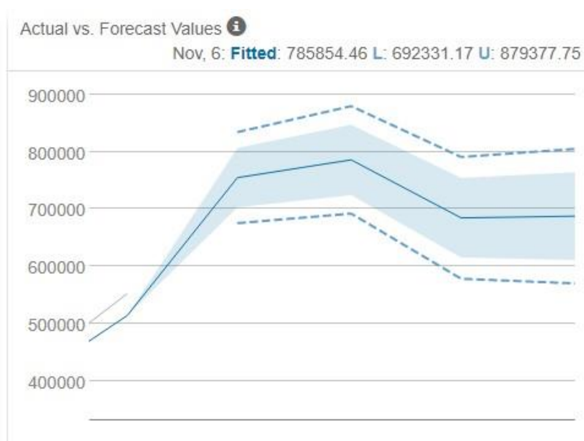
With respect to the errors found in the sample, the following graph, which compares the actual data with the data predicted by the models, shows that ARIMA prediction values are closer to those observed:

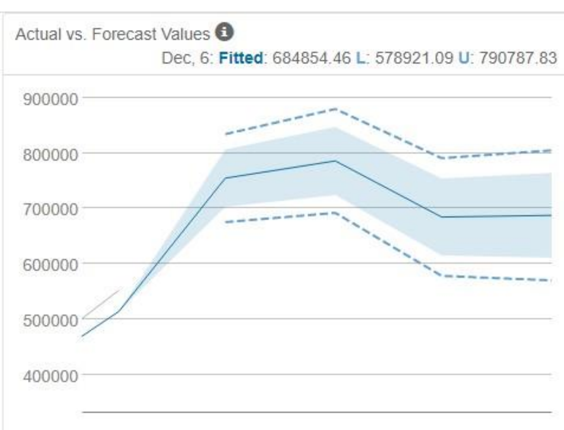
Actual and Forecast Values:

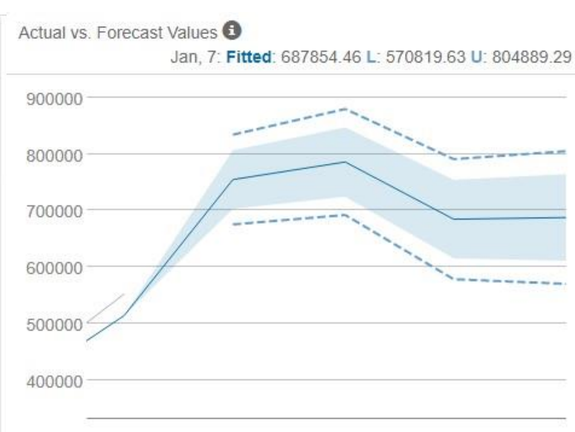
Actual	ETS_MAM_damp	ARIMA_011_010
271000	255966.17855	263228.48013
329000	350001.90227	316228.48013
401000	456886.11249	372228.48013
553000	656414.09775	493228.48013

From the ARIMA model (0,1,1)(0,1,0)[12] and considering a confidence interval between 80% and 95%, the forecast of demand for the months of October, November and December of 2013 and January of 2014 is 754,854.46; 785,854.46, 684,854.46 e 687,854.46, respectively:

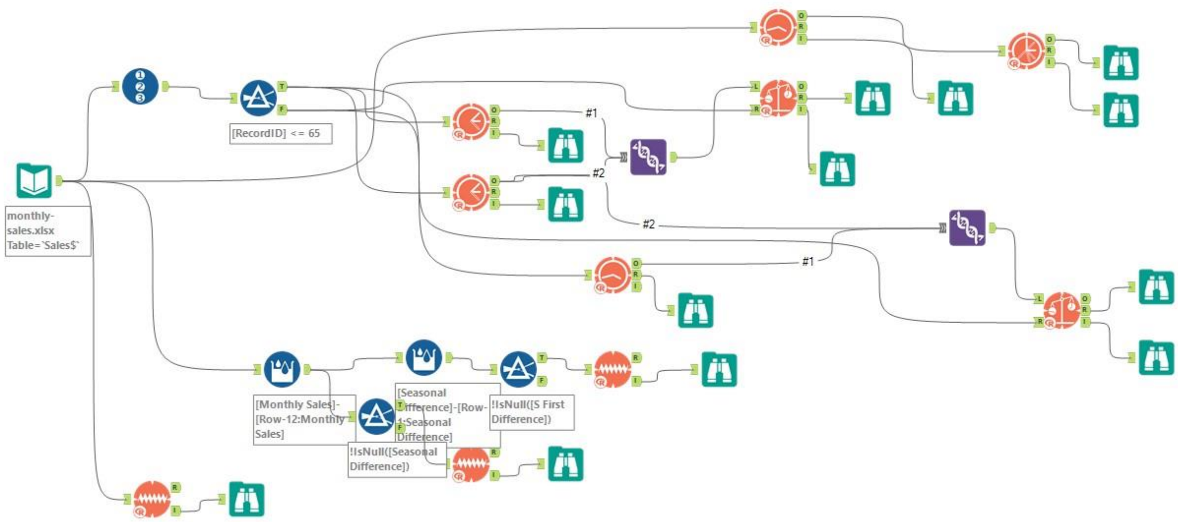








Flow in Alteryx:



<u>Credits:</u>

Curso Nanodegree Análise de Dados da Udacity
Slack Udacity
https://www.midomenech.com.br/lean-seis-sigma/downloads/artigos/157identifica%C3%A7%C3%A3o-modelo-arima-p,d,q.html
https://pt.wikipedia.org/wiki/S%C3%A9rie_temporal
http://www.portalaction.com.br/series-temporais/35-medidas-de-acuracia
http://ucanalytics.com/blogs/arima-models-manufacturing-case-study-example-part-3/
https://www.youtube.com/watch?v=abBzvDijEnM