Project Report – STAT 429

Team: Siva kumar Gorantla (sgorant2), Hani Ramezani(hrameza2)

Introduction

This project aims to establish a forecasting model for the monthly industry sales of printing and writing papers (in Thousands of french francs). Data were collected from January 1963 to December 1972. (1) (Source: Makridakis, Wheelwright and Hyndman (1998)). First, in data description, we inspect for overall trend and seasonality of the data. Second, model selection and diagnostics are performed and based on the results, the best model is selected. Third, the spectral analyses are conducted for the data and finally discussion and concluding remarks are provided at the end.

Data Description

In this section, the stationarity of the data is verified. From Figure 1(a), the plot of the time series suggests an increasing trend and a seasonal behavior in the sales. The ACF values (Fig 1(b)) are all positive and decreasing at a slow rate. The auto-correlations also show a seasonal trend with peaks at 12, 24, and 36. This lets us take the difference to remove seasonal (12 month) component.

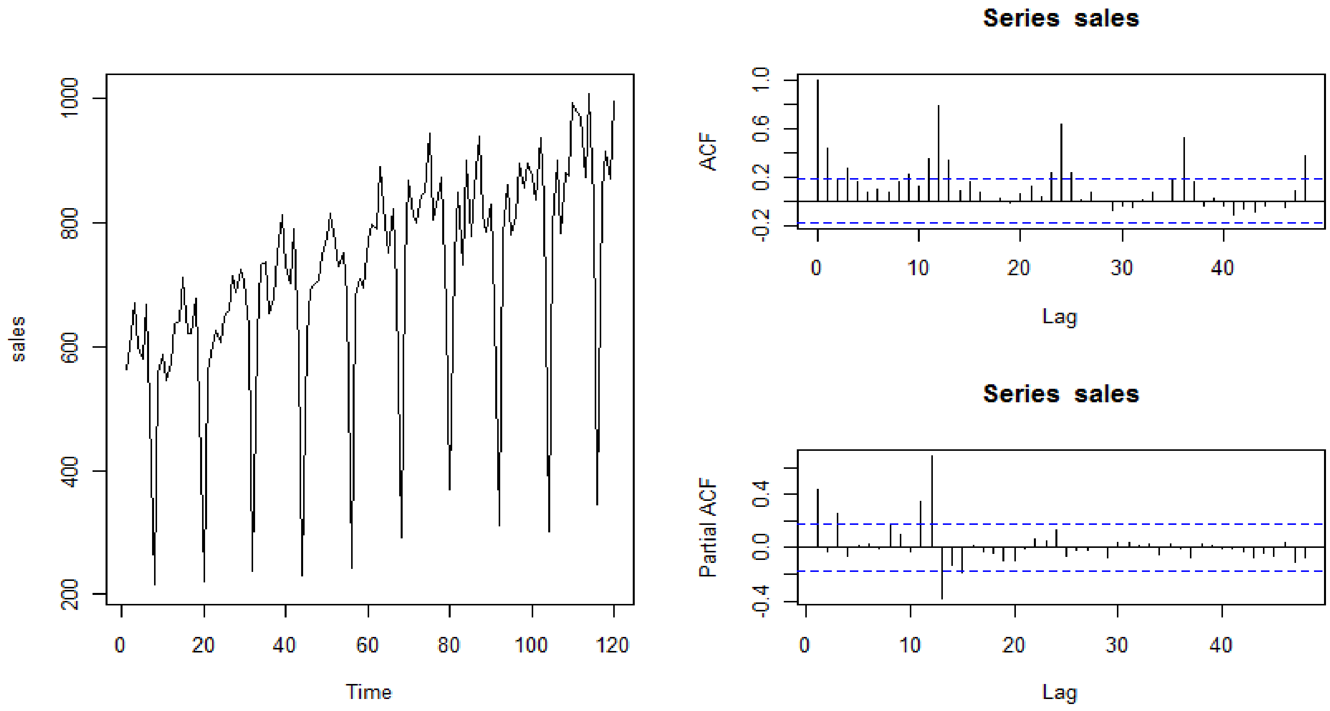


Fig 1: (a) Sales of printing and writing paper between 1963-1972, **(b)** ACF and PACF of sales data

The plot after performing seasonal difference (diff(sales,12)) shown in Fig 2(a) is clearly non-stationary. P-value from the Lobato test is less than 5% verifying non-stationarity. To make the data stationary, we additionally perform a first difference operation (diff(diff(sales,12))) . The results are shown in Fig 2(b).

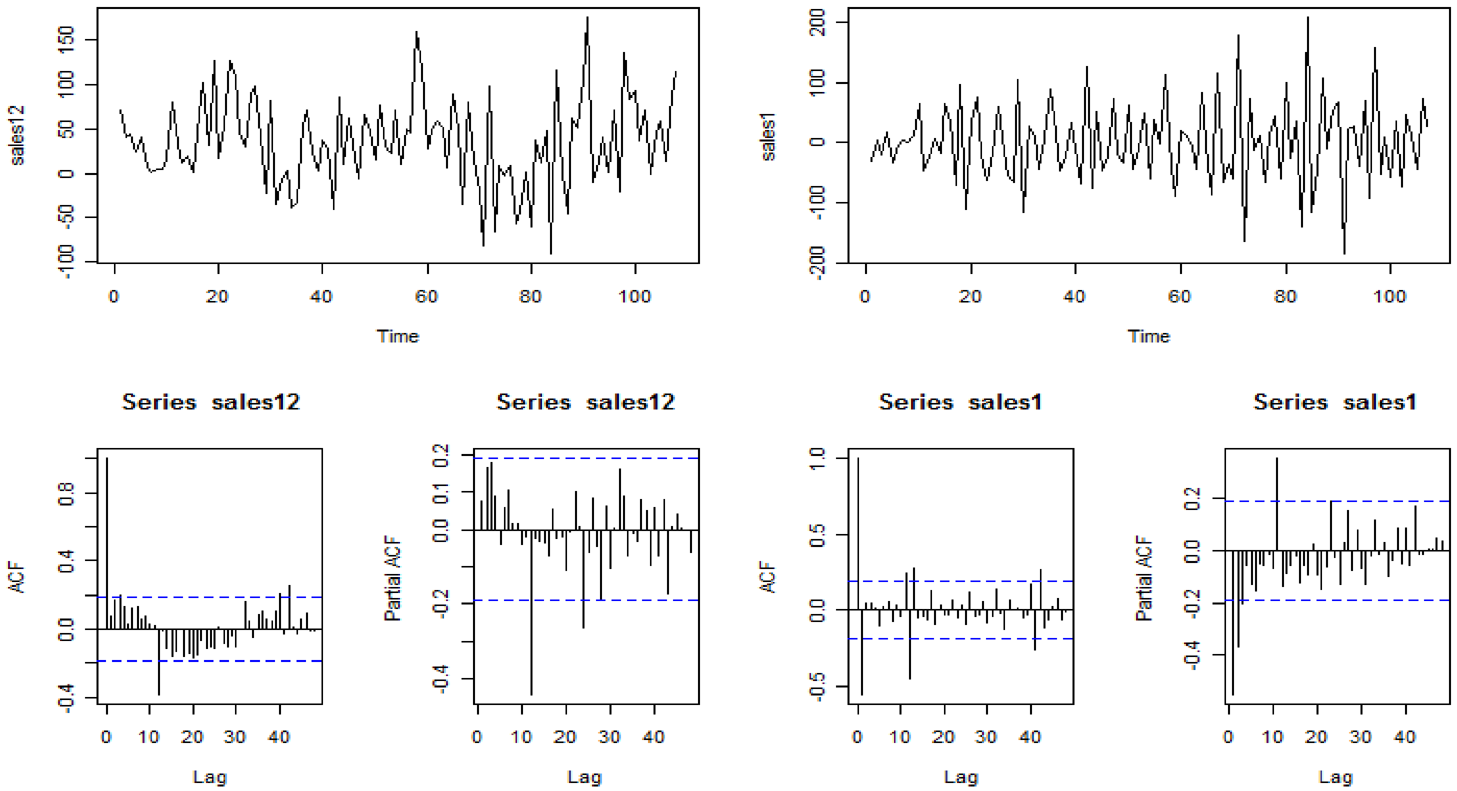


Fig 2: (a) ACF and PACF of diff(sales,12) data , **(b)** ACF and PACF of diff(diff(sales,12)) data

P-value for the Lobato test is more than 10% and also the pattern of the data in Figure 2(b) as well as its ACF and PACF imply that we can reasonably assume that the data is stationary. We use this stationary data for further analysis.

Thus by performing the seasonal difference and first difference operations over the sales data, we obtained a stationary time-series. At this point we divide the data into training and test (95%-5%) data.

Model Selection and Diagnostics

Preliminary Model Identification:

We fit an ARIMA(p,d,q) X (P,D,Q)s model for the sales data, where s = 12. The values of d=1 and D=1 because of the seasonal difference and first difference operations to get stationary time series.

Observing the ACF in Fig 2(b), the coefficient at lag 1 is very significant showing that q=1 is possible. Also, from PACF, the peak at 12-th coefficient signifies Q=1 possibility. Preliminary analysis suggests that ARIMA(0,1,1)X(0,1,1) $_{12}$ is a good fit. Estimating the parameters, the preliminary model in eq(1) has seasonal and first difference on the LHS, seasonal and non-seasonal moving-average components on the RHS.

$$\nabla_{12}\nabla\hat{x}_t = (1-0.8319B)(1-0.6816B^{12})\widehat{w}_t \qquad (1)$$

Model Selection & Estimation:

The order of the models selected based on the criteria of minimum AIC, BIC, and AICC are shown in Table1. Table 2 shows the coefficients and the corresponding standard deviation of the models, obtained by the maximum likelihood method.

Table1: Model selection, based on the Minimum Information Criteria for diff(diff(sales,12))

Criteria	Minimum Information Criteria	Selected Model	
		(p,0,q)	Seasonal (P,0,Q)
AIC	1060.173	(0,0,1)	(0,0,1)
BIC	760.6487	(0,0,1)	(1,0,2)
AICC	755.8005	(0,0,2)	(2,0,2)

Table2: Models coefficients and corresponding standard deviation

Criteria	Model, Fitted
AIC	$\nabla_{12}\nabla\hat{x}_t = (1-0.8319B)(1-0.6816B^{12})\widehat{w}_t$ *Se: (0.0661) (0.1060)*
BIC	$(1-\ 0.8880B^{12})\nabla_{12}\nabla\hat{x}_t = (1-0.9064B)(1-1.6534B^{12}+0.7334\ B^{24})\widehat{w}_t$ *Se: (0.6576) (0.0679) (0.9424) (0.6801)*
AICC	$(1+0.4541B^{12}+0.3600B^{24})\nabla_{12}\nabla\hat{x}_t = (1-0.9978B-0.0021B^2)(1-0.2700B^{12}-0.0044B^{24})\widehat{w}_t$ *Se: (0.3560) (0.1749) (0.1057) (0.0897) (0.3732) (0.2523)*

Model Diagnostics:

Diagnostic tests are made to detect any possible problem in the models. The results of the diagnostic test are shown in Figures 5, to 7 for the AIC, BIC and the AICC models repectively.

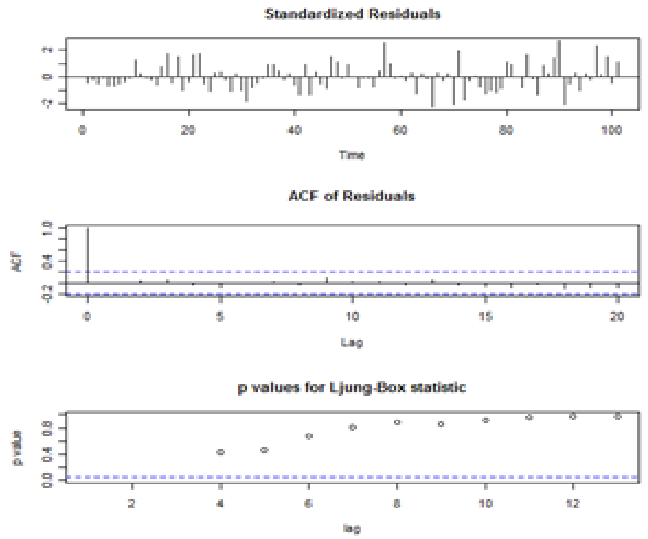


Fig5: Diagnostics for the AIC model

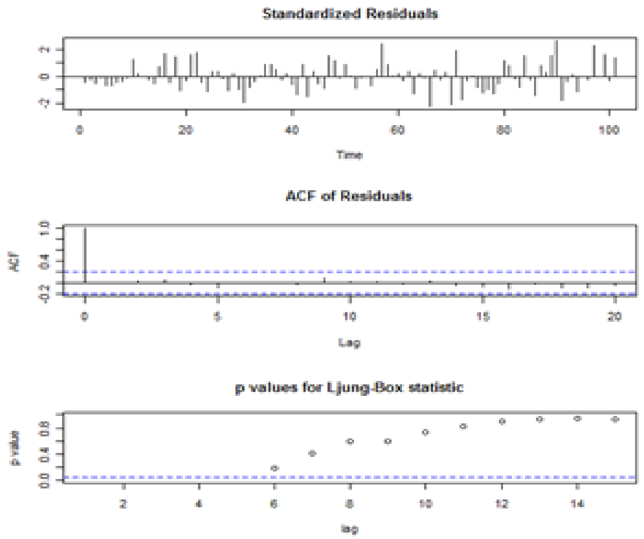


Fig6: Diagnostics for the BIC model

Comments: The standardized residuals plots do not show any outlier, since all the residuals are less than 2 standard deviations. Also the residuals do not have any trend. The ACF of the residuals is not statistically different than zero for the lags, greater than zero. The p-value for the Ljung-Box test is low at lag 7 of the AICC model, other than this, there is no problem with the Ljung-Box statistics. However, the Q-Q plots, in Figure8, show, a little deviation from the normality of the residuals.

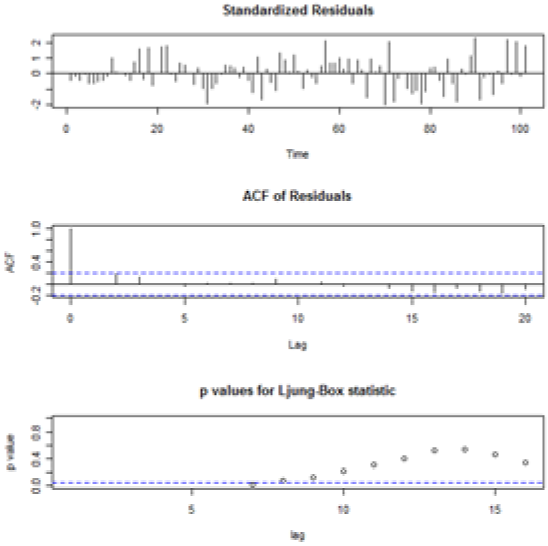


Fig7: Diagnostics for the AICC model

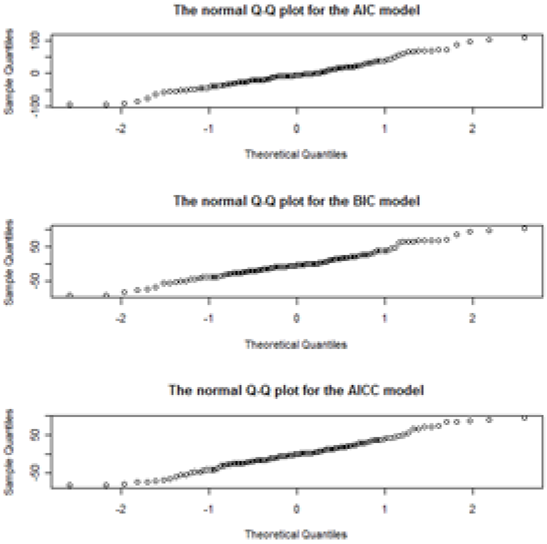


Fig8: Normal Q-Q plot for the three models

Prediction: The models were used to predict the next six steps (the last 5% of the data). The actual observations were shown versus the predicted values in Figure 9. The sum of square error (SSE) for the AIC, BIC, and the AICC models are 9308.59, 7092.821, and 9230.34. Therefore the model, based on the BIC criterion has the lowest SSE and is selected as the best model.

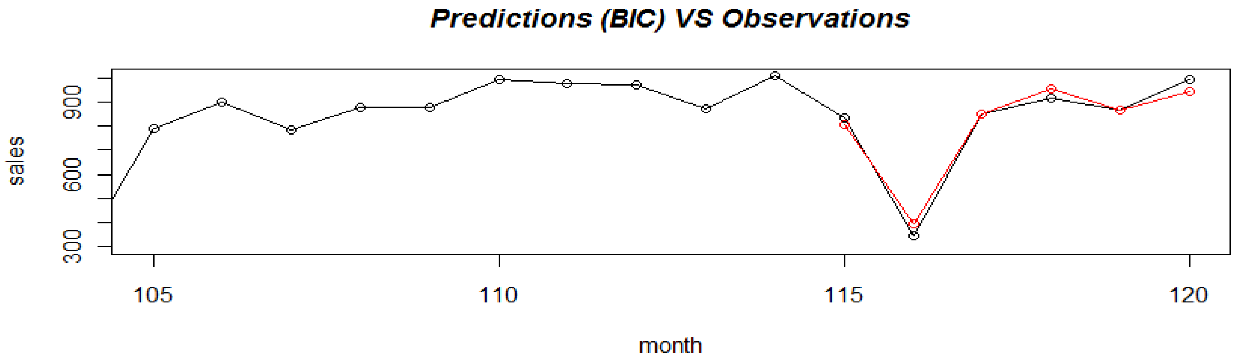


Fig9: The actual observations versus the predicted values

Spectogram:

K=c(2,2) gives a good spectral estimate. Smoothing with longer window size is not good because of there are multiple peaks in the actual spectrum and wide-window will result in combining these peaks as a single wide peak as seen in Fig 10(d).

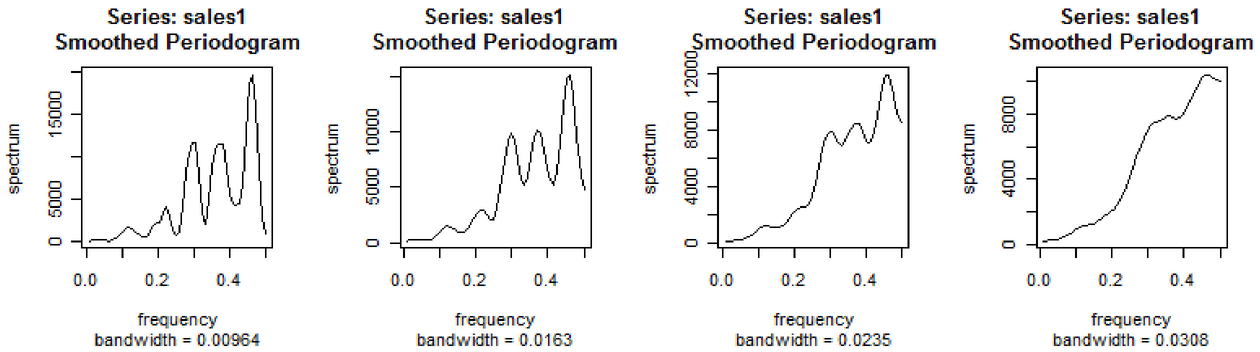
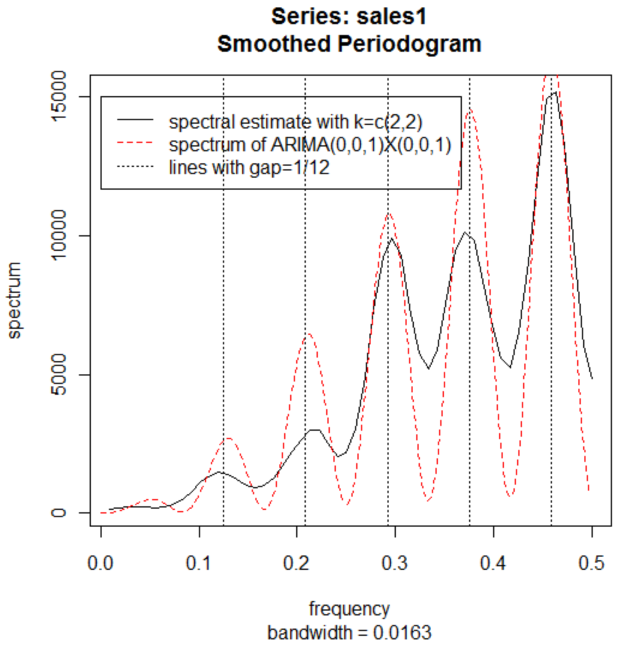


Fig 10: Spectral estimates using modified Daniell kernel with k=c(1,1),c(2,2),c(3,3) and c(4,4)

Analysis of spectrum:

- Presence of harmonics: Spectrum has peaks at intervals of 1/12 → periodicity of time series = 12 (months).
- High-pass filter structure: The spectrum (frequency gain) increases with frequency and suggests presence of more zeros than poles (high-pass filter) from a filtering perspective.

We also plot the ARIMA(0,0,1)X(0,0,1) (which has 2 zeros in frequency response, q=1 and Q=1) with parameters found in Table 1 to show the existence of zeros and harmonics.



Conclusion

In this project statistical models were developed for time series data, representing monthly industry sales for printing and writing papers. After removing, seasonality and trend from the data, ARIMA models were fit to the data based on the minimum AIC, BIC, and AICC selection criteria. The model chosen by BIC criterion gives least squared error over the test data. We conclude that ARIMA(0,1,1)X(1,1,2)$_{12}$ chosen by BIC criterion is a good fit for the data verified by the model diagnostics.

Spectral density function was estimated by smoothing the periodogram. Various bandwidths were tried and the best smoothing window is chosen to be c(2,2). Presence of harmonics in the spectrum suggest a 12-month seasonal component in the data. The spectrum matches with the ARIMA(0,1,1)X(0,1,1)$_{12}$ model.

For the practical purposes, more data collection is recommended to come up with a better model. Furthermore, investigation of innovative methods like GARCH is suggested to explore further improvements in the model development.

Reference

1)http://robjhyndman.com/tsdldata/data/writing.dat