Time Series Project

Industry Sales for Printing and Writing Paper 1963-1972

Analysis

Xiaodan Zhang

STAT 429

Yinxiao Huang

University of Illinois at Urbana-Champaign

Dec.10, 2012

1. Introduction

The time series data we analyze is the industry sales for printing and writing paper (in Thousand of French francs) from January 1963 to December 1972. There are totally 120 data corresponding to each month. The data source is: Makridakis, Wheelwright and Hyndman (1998). Printing and writing papers are defined as paper grades that are used for newspapers, magazines, catalogs, books, commercial printing business forms, stationeries, copying and digital printing. In this project, we expect to fit a forecasting model for the monthly industry sales of printing and writing papers (in Thousand of French francs).

2. Data

In this section, we inspect for the overall trend and seasonality of the data.

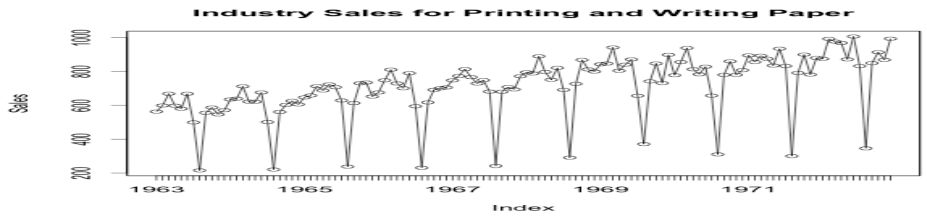


Figure 1.

First, we plot the original data and get Figure 1. From it we can see that in general, the paper sales increase over time and there seems a trend within each year: increasing in the beginning of the year, decreasing to the low ebb in the middle and later increasing again towards the end of the year. This indicates an overall increasing trend and seasonal effect in the paper sales.

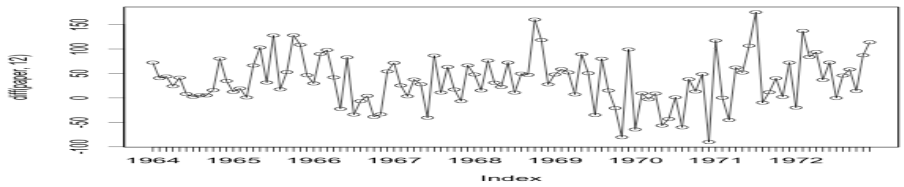


Figure 2.

Then, we get Figure 2 by taking the seasonal difference of the original data. We see that it is non-stationary. So, we additionally take a 1st difference operation (diff(diff(paper, 12))).

Figure 3.

In Figure 3, we see that the resulting series looks mean-stable and white. So, this is our transformed series after performing the seasonal difference and first difference operations over the original data.

3. Models

We divide the original data into two sets with 95% the data (114 data) in the training set and the rest (6 data) in the test set. We use the training set to build a good model and use the test set to test the goodness of fit of the fitted model.

First, we plot the ACF and PACF plots in Figure 4 of the training set part of the transformed series we get in last section.

Figure 4.

Preliminary Model Identification:

We fit an ARIMA(p,1,q)*(P,1,Q)$_{12}$ model for the time series data due to the seasonal difference and first difference operations to get a stationary time series. According to Figure 4, we notice that an MA(1) model fits here due to a clear cut off after lag 1 in the ACF plot and tail off in the PACF plot. That is, we could fit a seasonal ARIMA model with order (p,1,q)*(0,1,1)$_{12}$.

Model Estimation, Model Selection & Model Diagnostics:

Next, we use a global model search algorithm based on AIC, AICC, BIC criteria.

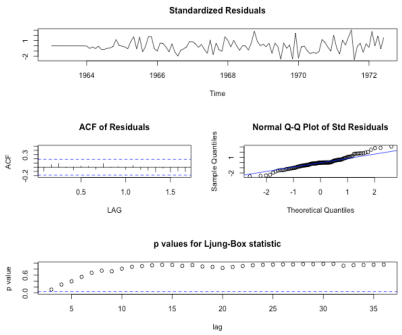
1) Model selection with BIC criterion

p/q	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	977.8228	908.6588	912.1078	916.4116	918.6319
[2,]	936.6782	911.9377	916.7357	921.6699	922.4532
[3,]	921.5839	916.7424	921.4379	919.3234	925.4904
[4,]	922.5623	916.9501	920.1845	925.0283	932.9344
[5,]	926.7444	919.5347	924.0779	930.5005	931.0230

Table 1. BIC values.

Table 1 gives the minimum 908.6588 AIC value, suggesting an ARIMA(0,1,1)* (0,1,1)$_{12}$ model. It seems to be a reasonable model according to the diagnostics in Figure 5:

no outliers in standardized residuals plots since all the residuals are less than 2 standard deviations; the residuals do not have any trend; the Q-Q plot shows normality; the ACF of the residuals is not statistically different than zero for the lags, greater than zero; there is no correlation between residuals based on Ljung-box test.



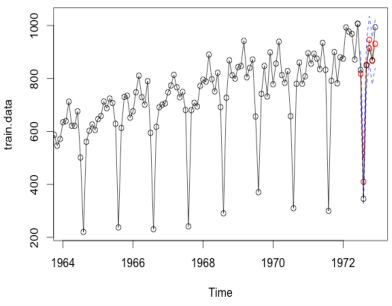


Figure 5. BIC Diagnostics. **Figure 6. Forecasting six data (index 115-120)**

According to Figure 6, we forecast six data ahead of the training data. The result looks good and all real test data are within the blue lines. The predicted values and corresponding forecast errors are as the following table:

1972	Jul	Aug	Sep	Oct	Nov	Dec
Predicted value	818.1773	409.9093	851.8402	946.0481	867.2199	930.9946
Standard eror	42.37122	42.96550	43.55167	44.13005	44.70095	45.26465
Test data	832.037	345.587	849.528	913.871	868.746	993.733
Forecast error	192.0899	4137.359038	5.346262	1035.363944	2.329062	3936.101181

Based on this model, we have a sum of residuals (forecast errors) of 9308.589.

2) Model selection with AIC criterion

p/q [,1] [,2] [,3] [,4] [,5]

[1,] 975.0353 903.0838 903.7453 905.2616 904.6945

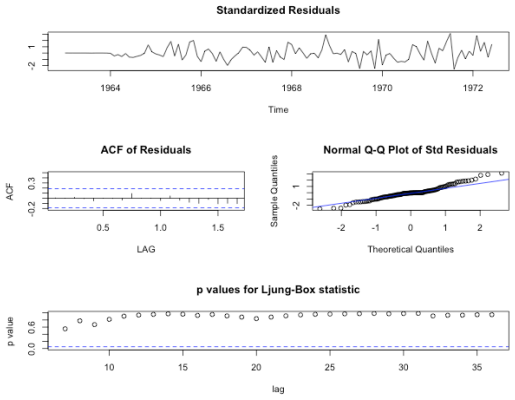
[2,] 931.1032 903.5752 905.5857 907.7324 905.7283

[3,] 913.2214 905.5924 907.5005 902.5984 905.9780

[4,] 911.4124 903.0127 903.4595 905.5159 910.6345

[5,] 912.8070 902.8098 904.5654 908.2006 905.9356 **Table 2. AIC values.**

Table 2 gives the minimum 902.5984 AIC value, suggesting an ARIMA(2,1,3)* $(0,1,1)_{12}$ model. It seems to be a reasonable model according to the diagnostics in Figure 7 (Similar diagnostics as the Figure 5 above).



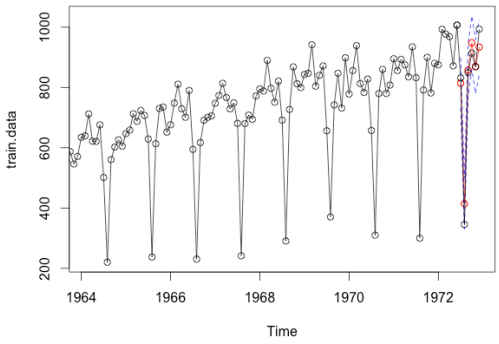


Figure 7. AIC Diagnostics. **Figure 8. Forecasting six data (index 115-120)**

According to Figure 8, we forecast six data ahead of the training data. The result looks good and all real test data are within the blue lines. The predicted values and corresponding forecast errors are as the following table:

1972	Jul	Aug	Sep	Oct	Nov	Dec
Predicted value	814.1930	414.3649	857.4423	948.2208	869.6031	933.7171
Standard eror	41.84164	41.93887	43.05790	44.22922	44.79957	45.34177
Test data	832.037	345.587	849.528	913.871	868.746	993.733
Forecast error	318.40978	4730.39629	62.63540	1179.91138	0.734692	3601.90477

Based on this model, we have a sum of residuals (forecast errors) of 9893.992.

3) Model selection with AICc criterion

It seems that the selected model under AICc criteria agrees with BIC criteria, we omitted the table and plots here. The AICc value is 903.1864.

In conclusion, the sum of square errors (forecast errors) for the AIC, BIC, AICc models are 9893.992, 9308.589 and 9308.589 respectively. Since the best forecast has the minimal forecast error, we choose between AICc and BIC models. According to their model diagnostics, they both perform well. In addition, because AICc and AIC are asymptotically equivalent, but AICc outperforms AIC in small samples; BIC is consistent, we choose BIC model as the best here.
The corresponding coefficient of non-seasonal MA (1) model was -0.6815, and the corresponding coefficient of seasonal MA (1) model was -0.8319.

Spectral density analysis:

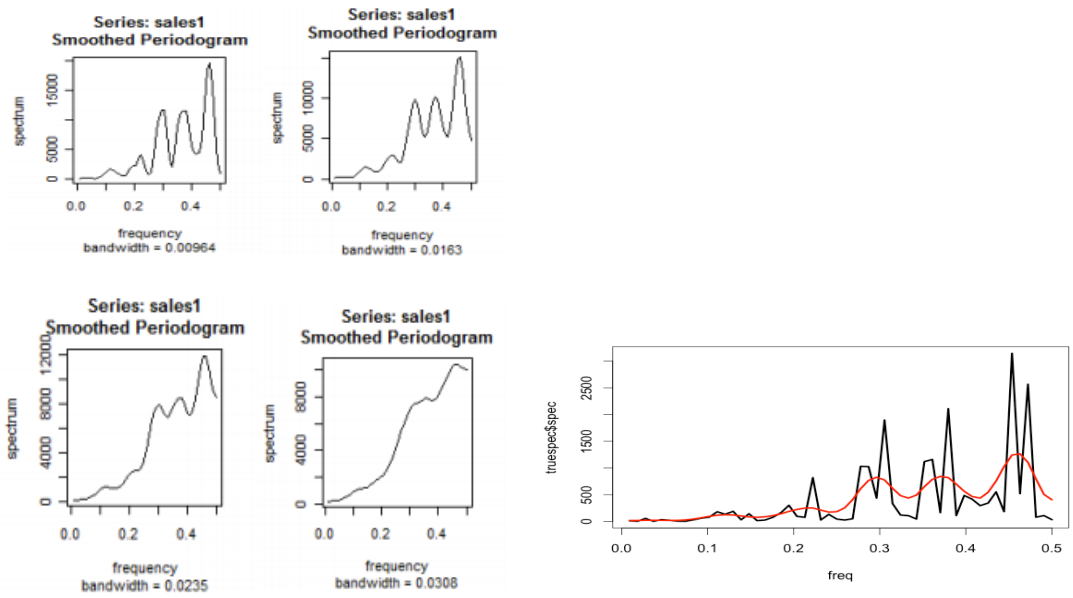


Figure 9. Spectral estimates using "modified.daniell" kernel with k=c(1,1),c(2,2),c(3,3) and c(4,4) spans & Figure 10. Spectral Density with span(2,2)
From Figure 9, we could see that smoothing with larger bandwidth is bad because it will result in combining peaks to a single flat peak when there are multiple peaks in the actual spectrum. We find that K=c(2,2) span on the top right gives the best spectral estimate, neither over smoothing nor missing peaks.

4. Conclusions

In this project, we considered the monthly industry sales for printing and writing papers time series from 1963 to 1972. After removing seasonality and trend from the original data, we fit ARIMA models based on the minimum AIC, BIC, and AICc selection criteria. The ARIMA(0,1,1)*$(0,1,1)_{12}$ model chosen by BIC criterion was good at diagnostics and was good to do the predictions. Spectral density function was estimated by smoothing the periodograms. Various bandwidths were tried, and the best smoothing window was span(2,2). The spectrum matched with the ARIMA(0,1,1)*$(0,1,1)_{12}$ chosen before with the 12-month seasonal effect in the series.

References

Time Series Data Library and DataMarket (2012).
http://datamarket.com/data/set/22l3/industry-sales-for-printing-and-writing-paper-in-thousands-of-french-francs-january-1963-december-1972#!display=line&ds=22l3

Appendix

```
library(rdatamarket)
paper <- dmseries("http://data.is/Ur24KI")
#-------Data Section begins--------
plot(paper, type="o", ylab="Sales",main="Industry Sales for Printing and Writing Paper")
#plot(paper,type="o", ylim=c(0,1200), xlab=" ",ylab=" ")
#par(new=TRUE)
#plot(SMA(paper), col=2,ylim=c(0,1200), xlab="Index",
#         ylab="Car Sales", main="Moving Average Plot for Car Sales (in red)")
plot(diff(paper,12), type="o")
plot(diff(diff(paper,12)), type="o", ylab="Paper Sales",main="Monthly Industry Sales for Printing and
Writing Paper")
#-------Model Section begins--------
n = length(paper)
n # 120
train.size = round(n*0.95)
train.size # 114
train.data = paper[1:114]
test.data = paper[115:120]
paper.transformed = diff(diff(train.data, 12)) # on train.data
acf2(paper.transformed,60)
#---
P=4
Q=4
crit<-matrix(0,P+1,Q+1)
for (j in 0:P)
{
   for (k in 0:Q)
   {
      dataML<-arima(train.data,order=c(j,1,k),seasonal=list(order=c(0,1,1),period=12),method="ML")
      #AICC
      crit[j+1,k+1]<-n*log(dataML$sigma)+2*(j+k+1)*n/(n-j-k-2)
      #BIC
      #crit[j+1,k+1]<-n*log(dataML$sigma)+(j+k+1)*log(n)
      #AIC
      #crit[j+1,k+1]<-n*log(dataML$sigma)+2*(j+k+1)
   }
}
#locate the minimum information criteria
crit
min(crit)
# BIC&AICc
sarima(train.data, 0,1,1,0,1,1,12)
```

```
bic.pr=sarima.for(train.data,6,0,1,1,0,1,1,12)
bic.pr
lines(paper[114:120],type="o")
# AIC
sarima(train.data, 2,1,3,0,1,1,12)
aic.pr=sarima.for(train.data,6,2,1,3,0,1,1,12)
aic.pr
lines(paper[114:120],type="o")
# forcast errors
#pred=bic.pr$pred
pred = aic.pr$pred
fe=rep(0,6)
for(i in 1:6)
{
   fe[i]<-(pred[i]-test.data[i])^2
}
fe
sum(fe)
#-----spectral density analysis begins--------
data=diff(diff(paper,12))
T=108
freq<-(1:(T/2))/T
per<-(abs(fft(data)))^2/T
per<-per[2:(T/2+1)]
truespec<-spec.pgram(data, taper=0, log="no")
Umax<-max(max(truespec$spec),max(per))
plot(freq,per,type="l",ylab="spectral density",ylim=c(0,Umax))
lines(freq,truespec$spec,ylim=c(0,Umax),col=2,lwd=3)
#par(mfrow=c(2,2))
k1 = kernel("modified.daniell", c(1,1))
truespec1<-spec.pgram(data, kernel =k1,taper=0, log="no")
per.ave1 = spec.pgram(data, kernel =k1, taper=0, log ="no",plot=F)$spec
plot(freq,per,type="o",ylab="spectral density",ylim=c(0,Umax))
lines(freq,truespec1$spec,ylim=c(0,Umax),col=2,lwd=3)

k2 = kernel("modified.daniell", c(2,2))
truespec2<-spec.pgram(data, kernel =k2,taper=0, log="no")
per.ave2 = spec.pgram(data, kernel =k2, taper=0, log ="no",plot=F)$spec
plot(freq,per.ave2,type="l",ylab="spectral density",ylim=c(0,Umax))
lines(freq,truespec2$spec,ylim=c(0,Umax),col=2,lwd=3)

k2 = kernel("modified.daniell", c(2,2))
truespec2<-spec.pgram(data, kernel =k2,taper=0, log="no")
k3 = kernel("modified.daniell", c(3,3))
```

```
truespec3<-spec.pgram(data, kernel =k3,taper=0, log="no")
k4 = kernel("modified.daniell", c(4,4))
truespec4<-spec.pgram(data, kernel =k4,taper=0, log="no")
Umax<-max(max(truespec$spec),max(per))

#abline(v=1/12, lty="dotted")
k2 = kernel("modified.daniell", c(2,2))
per.ave2 = spec.pgram(data, kernel =k2, taper=0, log ="no",plot=F)$spec
plot(freq,truespec$spec,type="l",col=1,lwd=3)
lines(freq,per.ave2,col=2,lwd=3,type="l")
```