Time Series - Final Project

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística
Màster en Estadística i Investigació Operativa

Author : Maxime Jurado and Mathieu Marauri

Barcelona, Spain

Contents

Introduction

For this project we decided to work on the IPI serie. It is a monthly index of industrial production in Spain. The aim of this study is to fit a model to the serie in order to make previsions. During this analysis the outliers will be studied, if some exist, in order to adjust the model fitted.

Figure 1 shows the IPI serie.

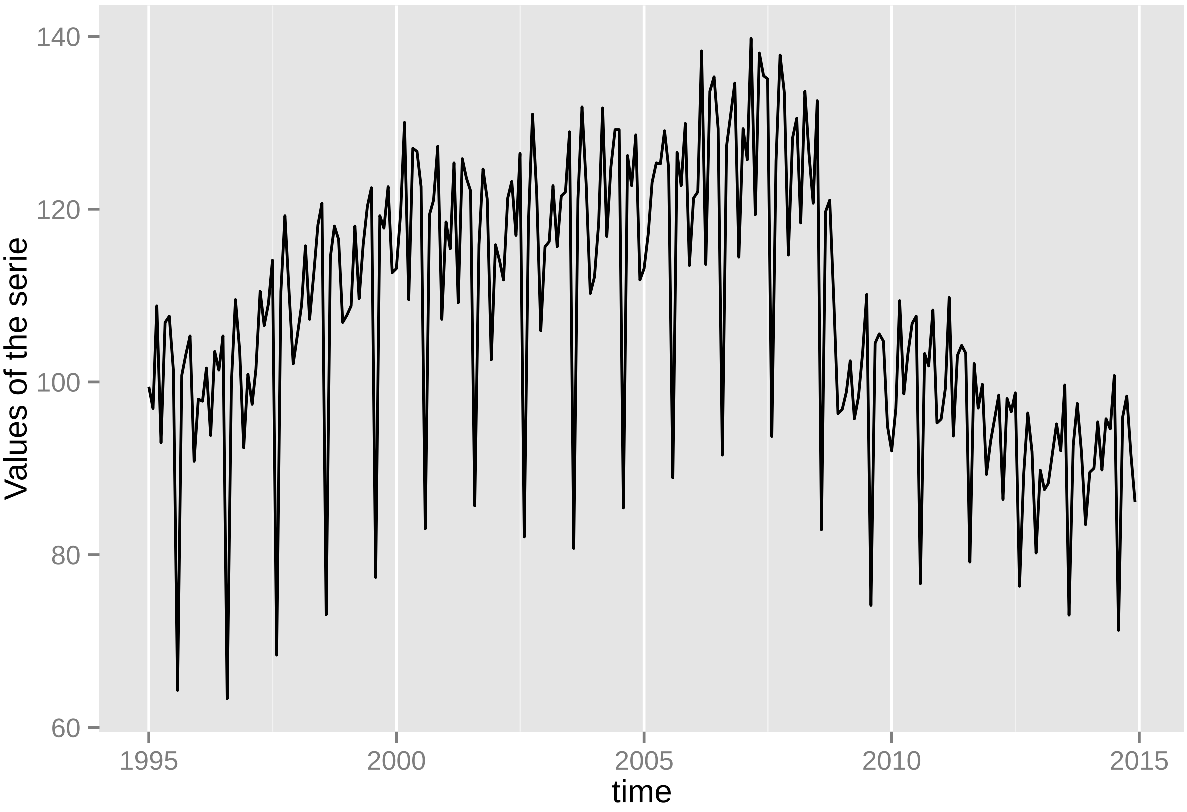


Figure 1: IPI serie.

Identification

Question a

The first thing that needs to be done is to check whether the variance is constant or not. Several graphs help answer that.

Figure 2 and 3 show the evolution of the variance depending on the mean.

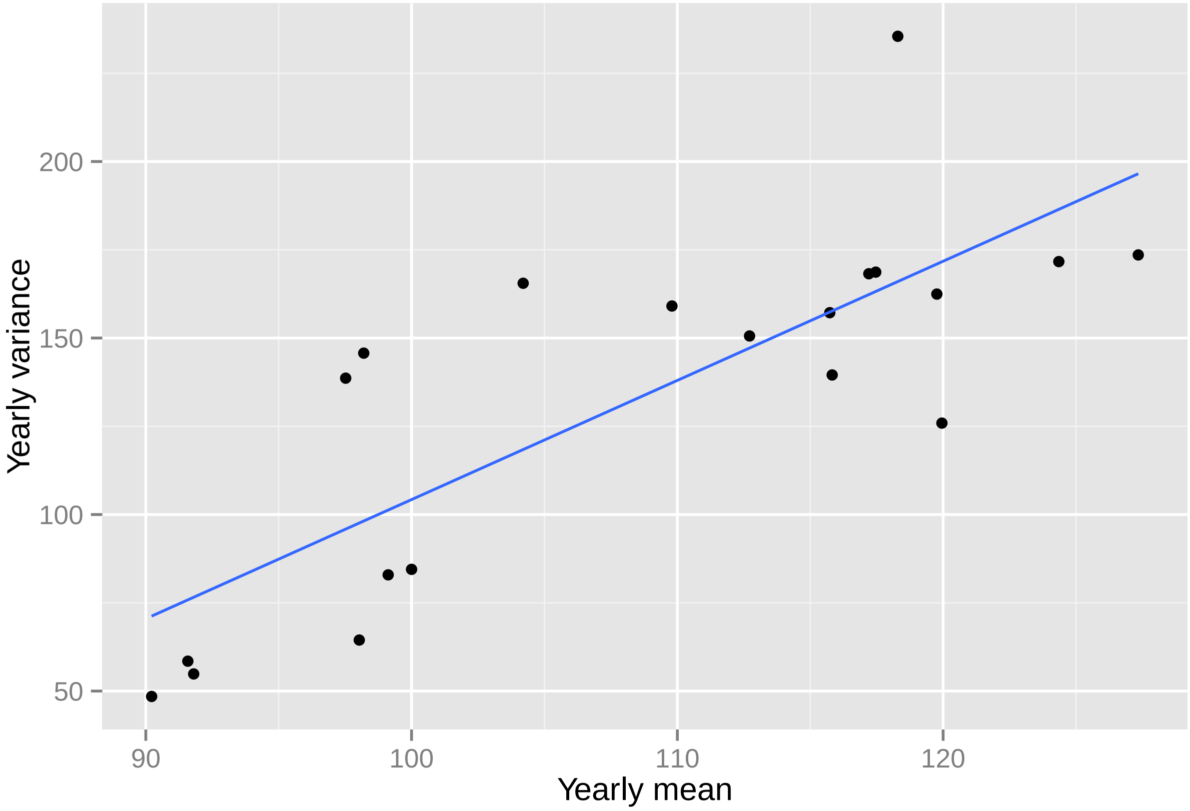


Figure 2: Evolution of the variance relatively to the mean for the original serie.

One can observe here that the linear relation between the variance and the mean does not have a null slope. It means that the variance is higher whenever the mean is higher.

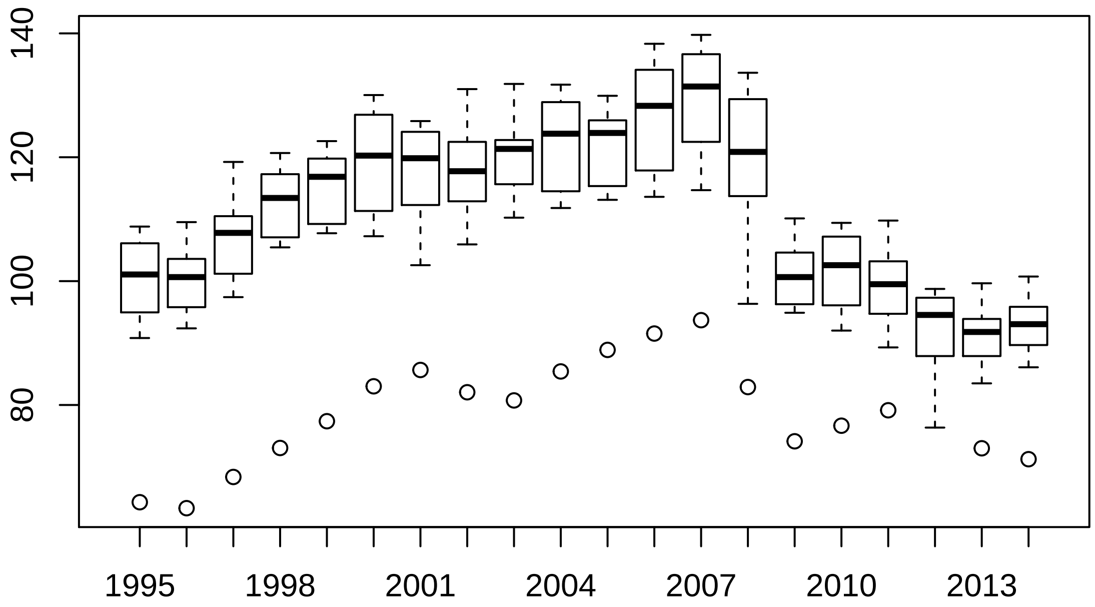


Figure 3: Boxplot of the ipi for each year.

One could observe a low variance in 2013 meanwhile in 2006 one could observe a high one. The variance of the original serie does not seem to be constant.

To correct this, a log-transformation is applied to the serie.

The previous graphs are done again in Figure 4 and 5. The transformation enhances the plots in a way that now one can consider that the variance is constant.

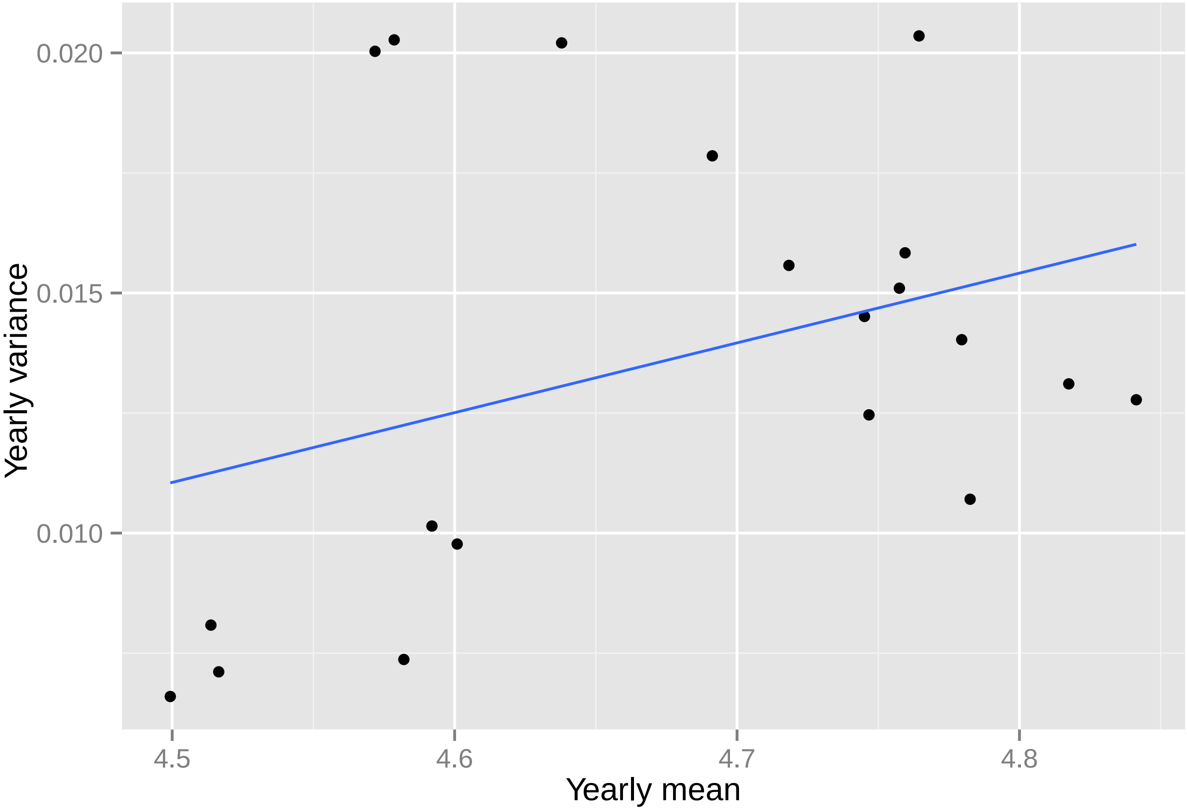


Figure 4: Evolution of the variance relatively to the mean for the log of the serie.

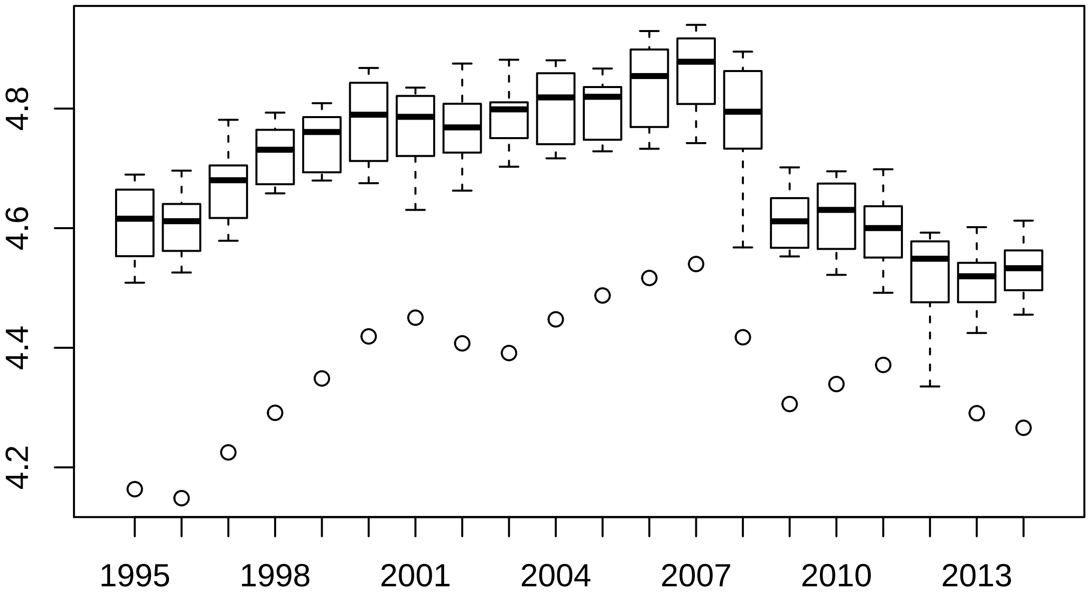


Figure 5: Boxplot of log(ipi) for each year.

Now that the serie has been transformed so that it has a constant mean, one has to check if there is seasonal pattern. The following plot (Figure 6) helps find a potential one.

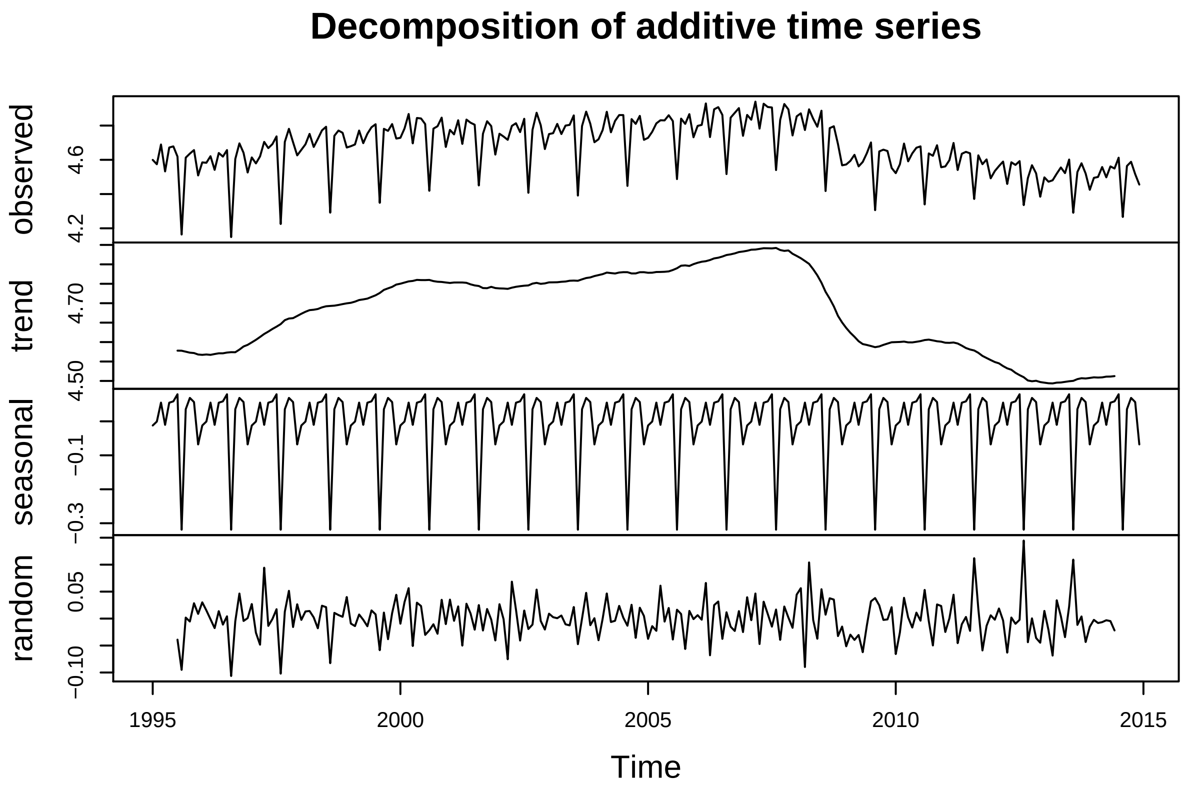


Figure 6: Decomposition of log(ipi).

It appears that the serie has a annual pattern. Hence a differentiation is needed.

Then the serie is to have a constant mean. By plotting the transformed serie one can have an idea whether the mean is constant or not. If not the serie needs to be differentiate again. In Figure 7 the logarithm of the serie, after differentiation is plotted.

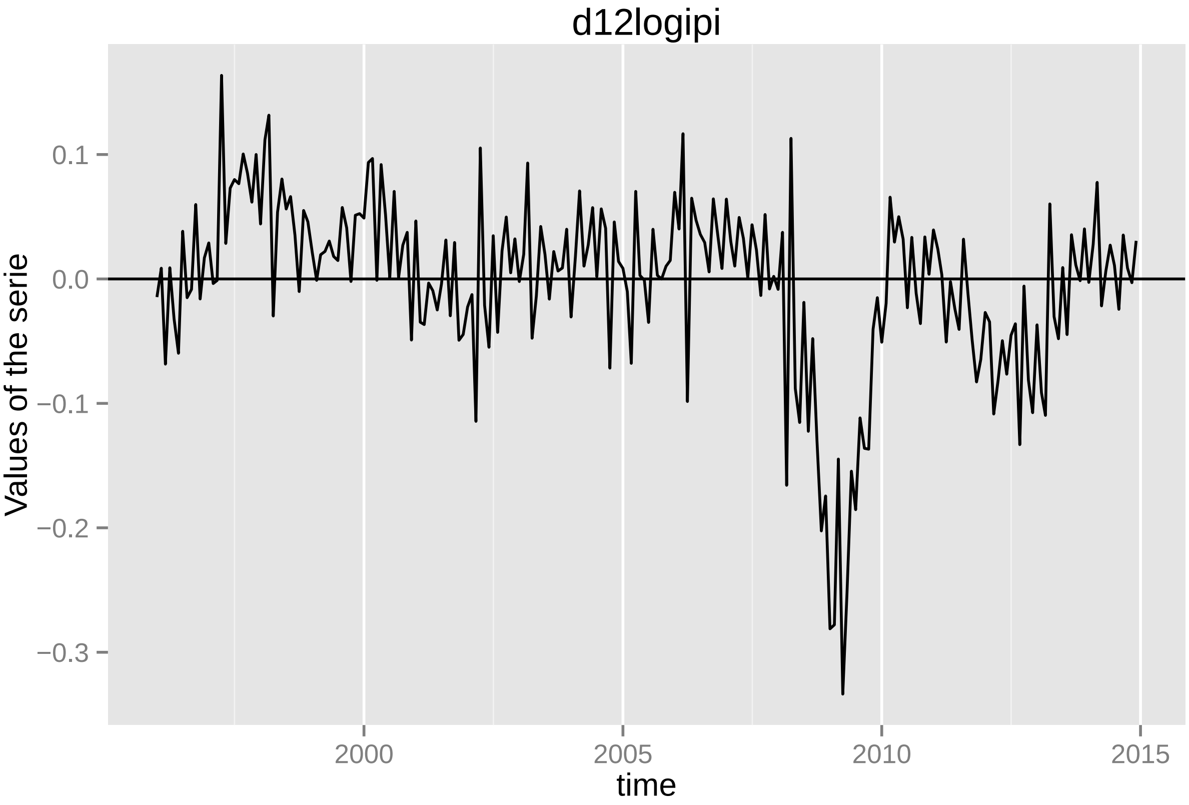


Figure 7: log(ipi) differentiate 12 times (seasonnality).

One could say that the mean is not constant so the serie is differentiate another time. The serie is plotted again (Figure 8).

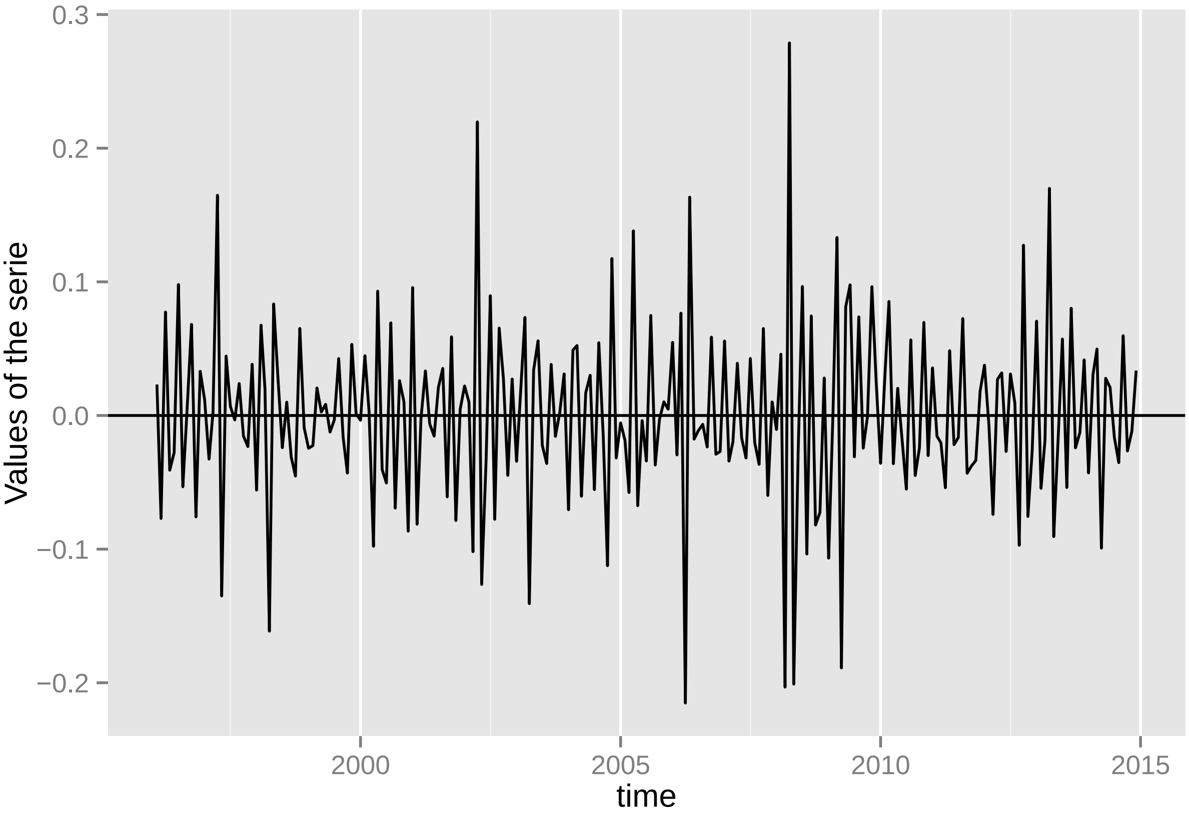


Figure 8: log(ipi) differentiate 12 times the 1 time.

Now the mean seems to be constant.

In order to select which transformation is actually the best the variance of the transformed series are compared. The smallest variance is preferred. Table 1 shows the results.

serie	variance
IPI	255.5872
log(IPI)	0.0240
$(1-B^12)log(IPI)$	0.0050
$(1-B)(1-B^12)log(IPI)$	0.0045

Table 1: Variances of each transformed serie.

The last serie is selected. It is the serie with a log transformation and two differentiation. One for the seasonality and one to have a constant mean. The serie can be written $(1-B)(1-B^12)log(IPI)$.

For now on the transformed serie will be referred simply as the serie.

Question b

In order to identify some possible models, the Auto-Correlation Function and the Partial ACF of the serie are plotted. Figure 9.

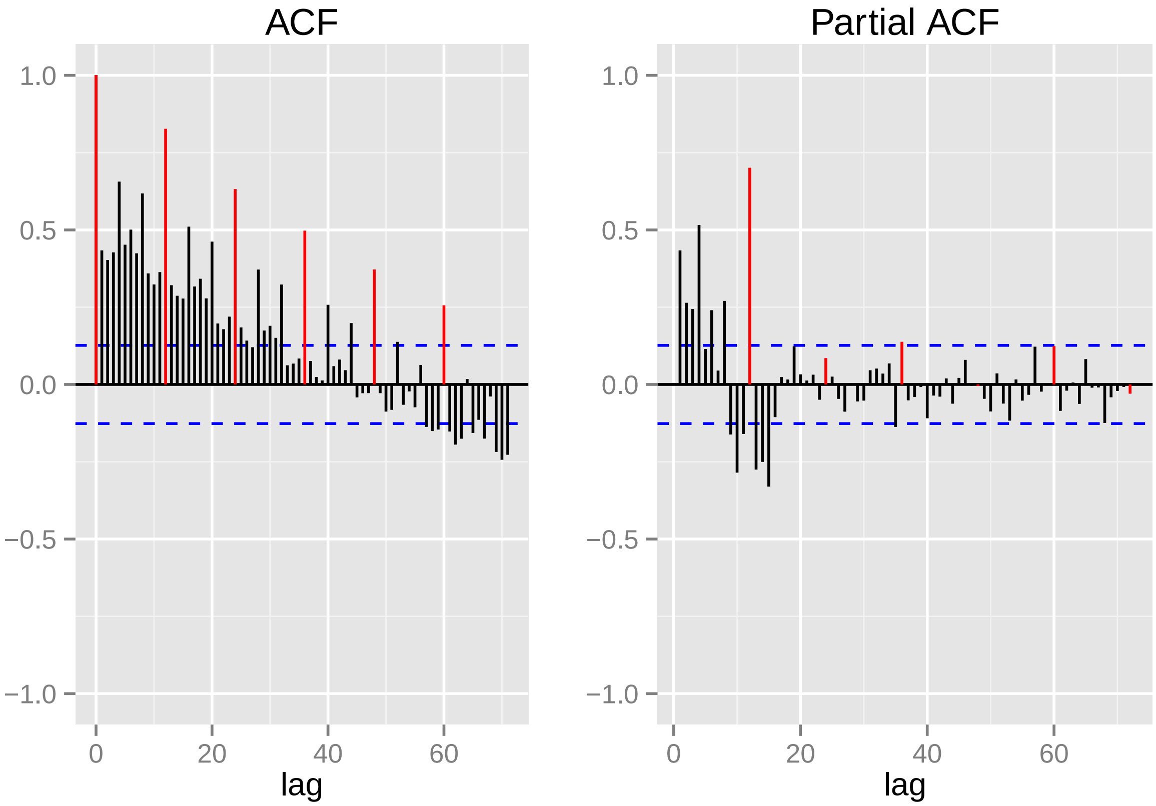


Figure 9: ACF and PACF of the serie.

The red lags are used to identify the seasonal part. One can consider an ARMA(3,2) or an ARMA(3,5) looking for the last significant lag. The ACF gives the MA part and the PACF the AR part. For the regular part, an AR(2) or an AR(6) can be choose.

Finally 4 models are possible:

- Model 1: $ARIMA(2,0,0)(3,1,2)_12$
- Model 2: $ARIMA(2,0,0)(3,1,5)_12$
- Model 3: $ARIMA(6,0,0)(3,1,2)_12$
- Model 4: $ARIMA(6,0,0)(3,1,5)_12$

Estimation

Before estimating the models one can try to see if the intercept is needed or not in the models. To do so the ratio $\frac{estimate}{standarderror}$ is computed. If it is lower than 2 then the coefficient can be set to zero. In every models the intercept was not significant so it was set to zero. The resulting models have a lower AIC. This confirms that the models are better without the intercept.

Now the ACF and PACF of each model are plotted and a comparison is made with the ACF and PACF of the serie. This makes possible eliminate 2 models. A residuals analysis will be performed later.

As an example the comparison between model 3 and the serie is presented in Figure 11. The other graphs are in the section Appendix.

Model 3: $ARIMA(6,0,0)(3,1,2)_12$.

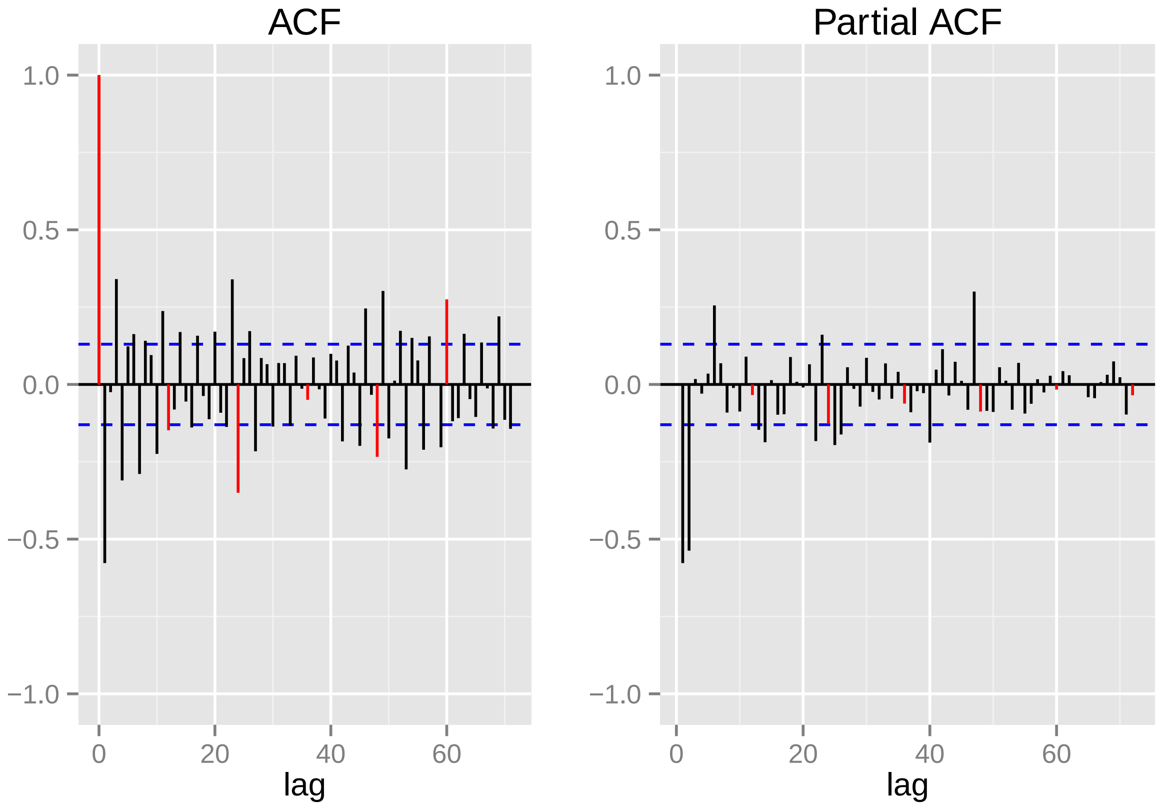


Figure 10: ACF and PACF of the serie and the model 3.

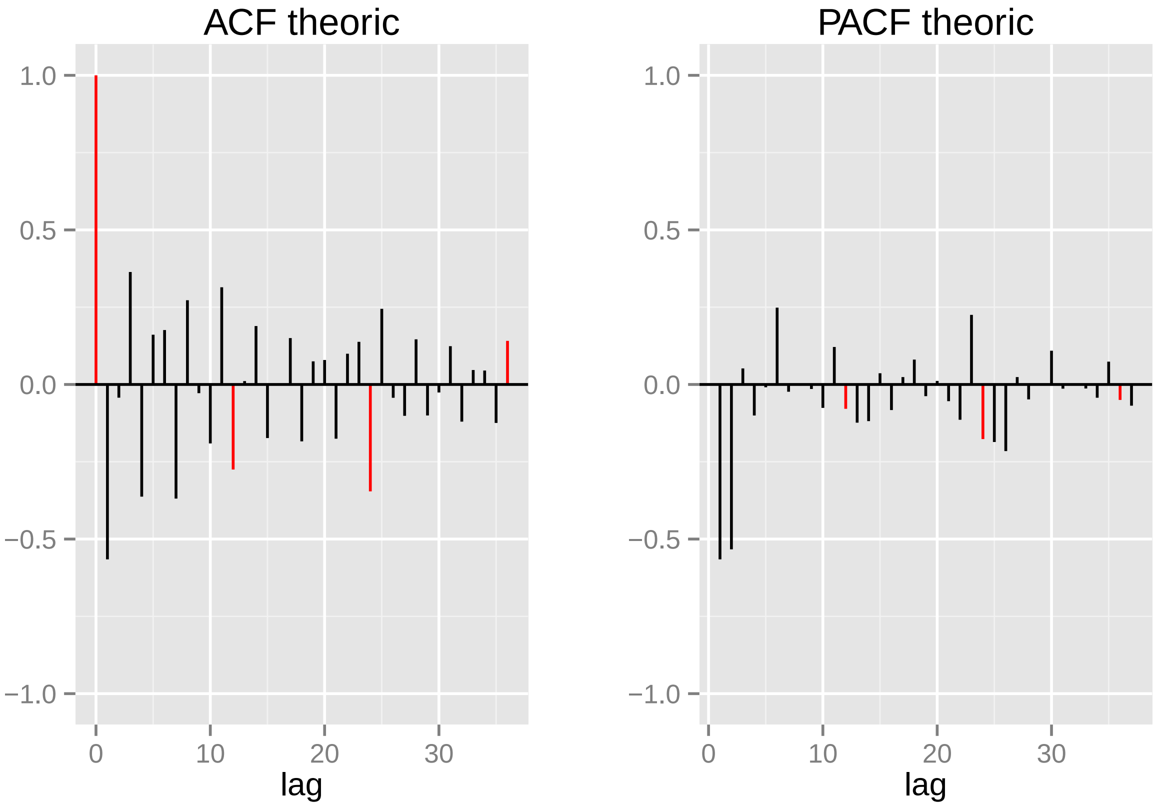


Figure 11: ACF and PACF of the serie and the model 3.

By comparing the red lags, the one for the seasonality, it appears that the model and the serie have the same significant lag with the same sign. The same is true for the first 12 lags that correspond to the regular part. One can conclude that model 3 is a good representation of the serie.

Such similitude were also observed for the model 4. Not so much for the models 1 and 2.

Therefore the complete analysis of residuals will be performed only on model 3 and 4.

After performing this analysis the models 3 and 4 were adjusted based on the significance of each coefficient. The non-significant coefficients were removed and the models were compared using the AIC. The model 4 did not change but the model 3 did. The coefficients for the ar3, ar4 and sma1 were fixed to 0 and the new model was an $ARIMA(6,1,0)(1,1,2)_{12}$ for the logarithm of the serie.

Validation

Question a

The analysis of the residuals is a way to validate the model. The residuals must verify the following hypotheses:

- They have the same variance. Homoscedasticity.
- They must follow a normal distribution.
- They have to be independent.

These hypotheses will be assessed graphically.

Model 3: $ARIMA(6,1,0)(1,1,2)_12$

We start by checking the homoscedasticity. To do so the residuals are plotted along with a scatter plot with a smooth. Figure 12.

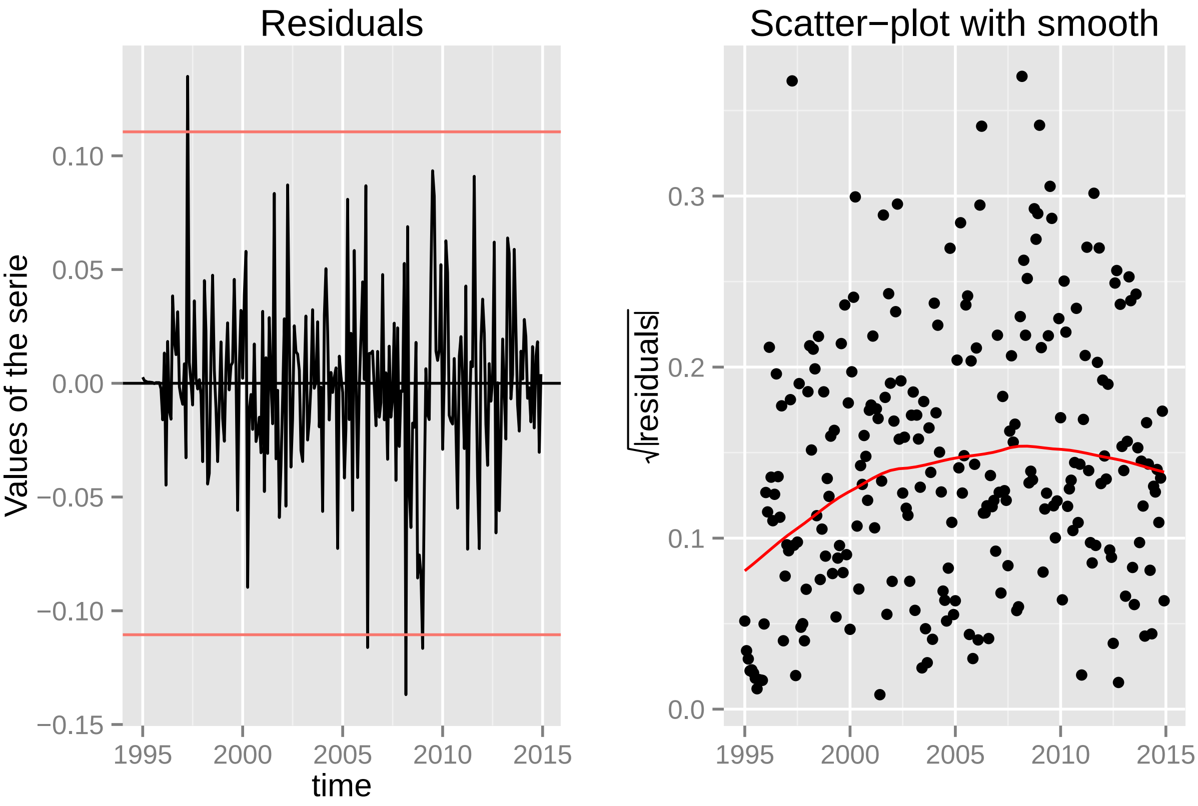


Figure 12: Residuals and scatter plot of the residuals.

Residuals do not go far outside the confidence bounds except for some values (outliers?). The scatter plot does not show any tendency even if the smooth is not completely straight. The homoscedasticity is verified.

Now the normality of the residuals is to be checked. To do so a qqplot and the histogram of the residuals along with the density of a normal are plotted. Figure 14

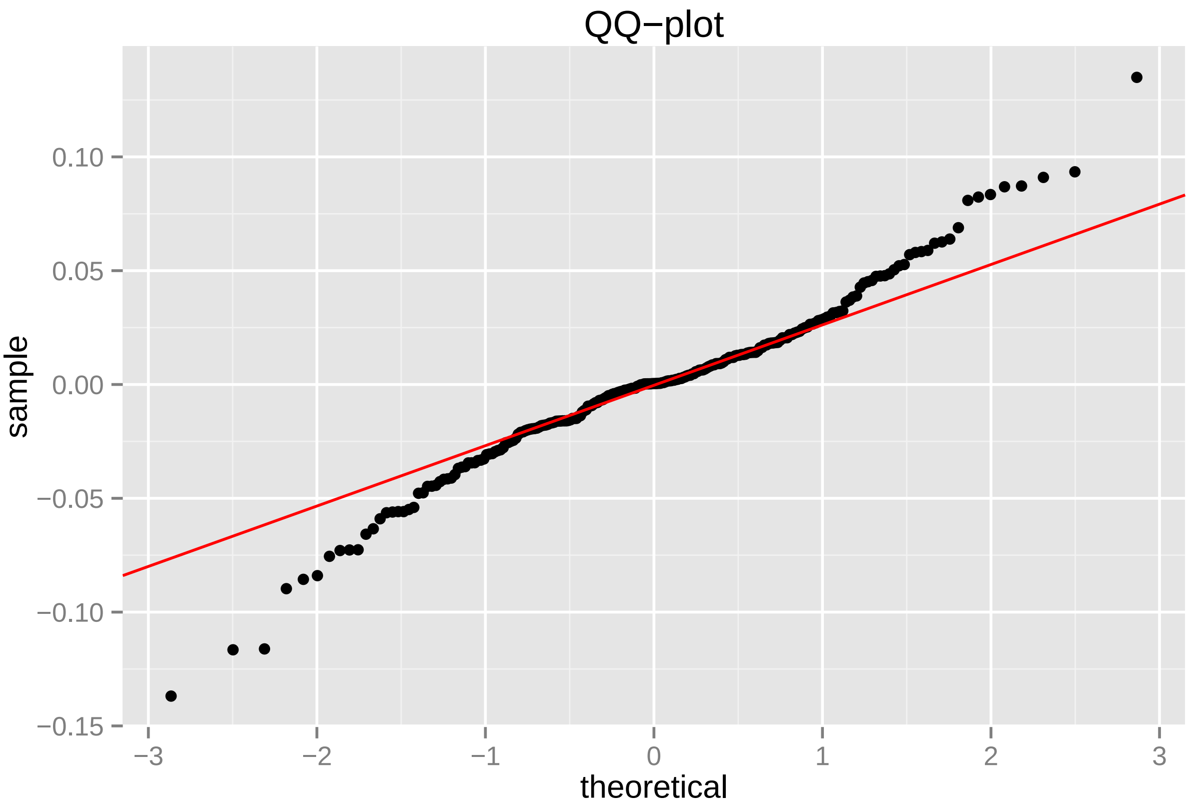


Figure 13: QQ-plot and histogram of the residuals.

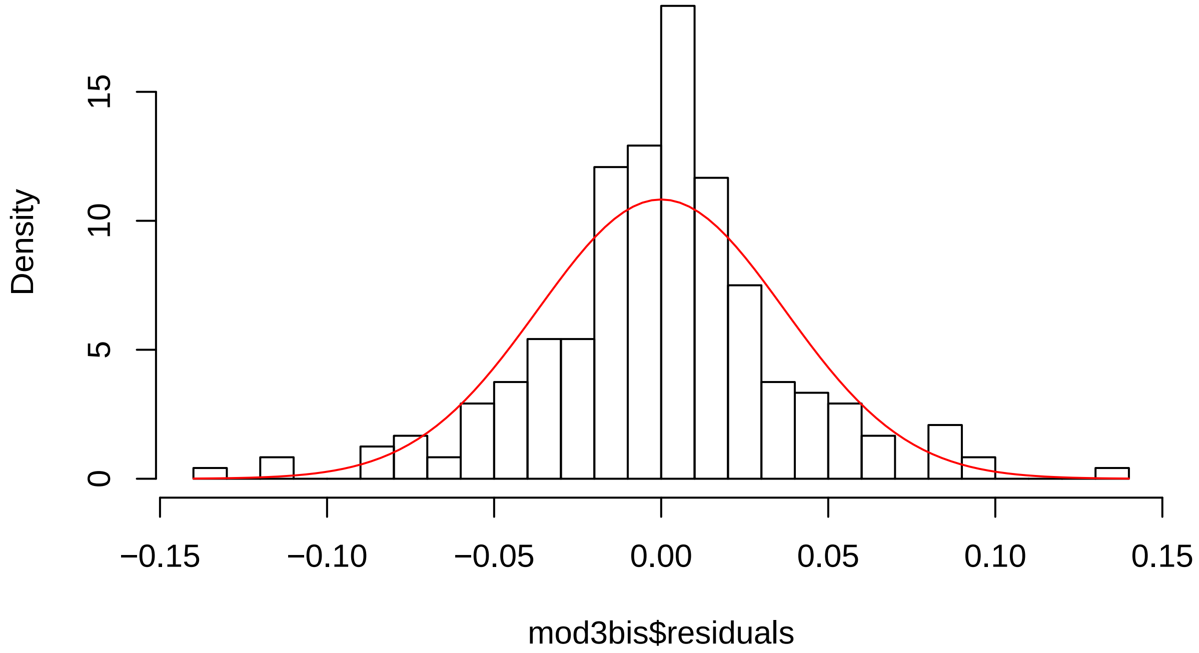


Figure 14: QQ-plot and histogram of the residuals.

The normality can be considered verified even if we have quite heavy tails in the QQ-plot and some parts of the histogram are upper than the curve.

Here we check the independence of residuals. Figure 15.

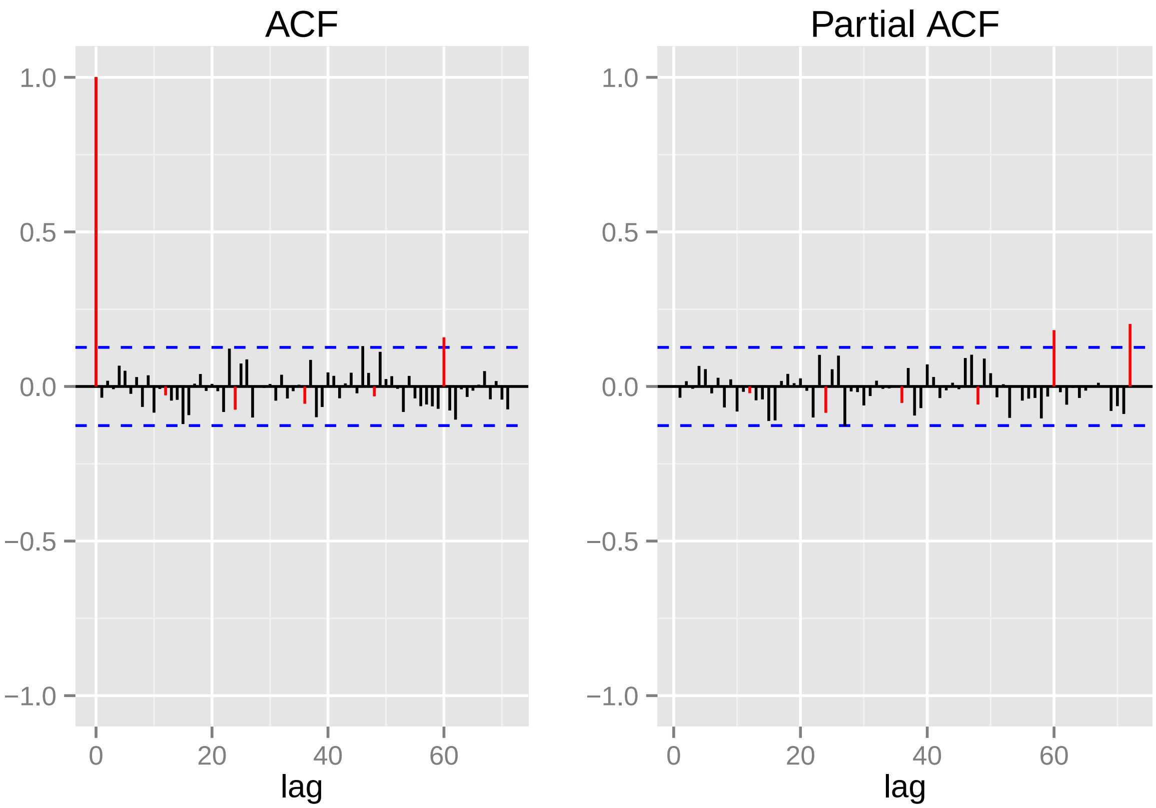


Figure 15: ACF and PACF of the residuals.

The residuals are independent. Indeed significant lags are far away from the origin.

The next step is to check for volatility. Figure 16.

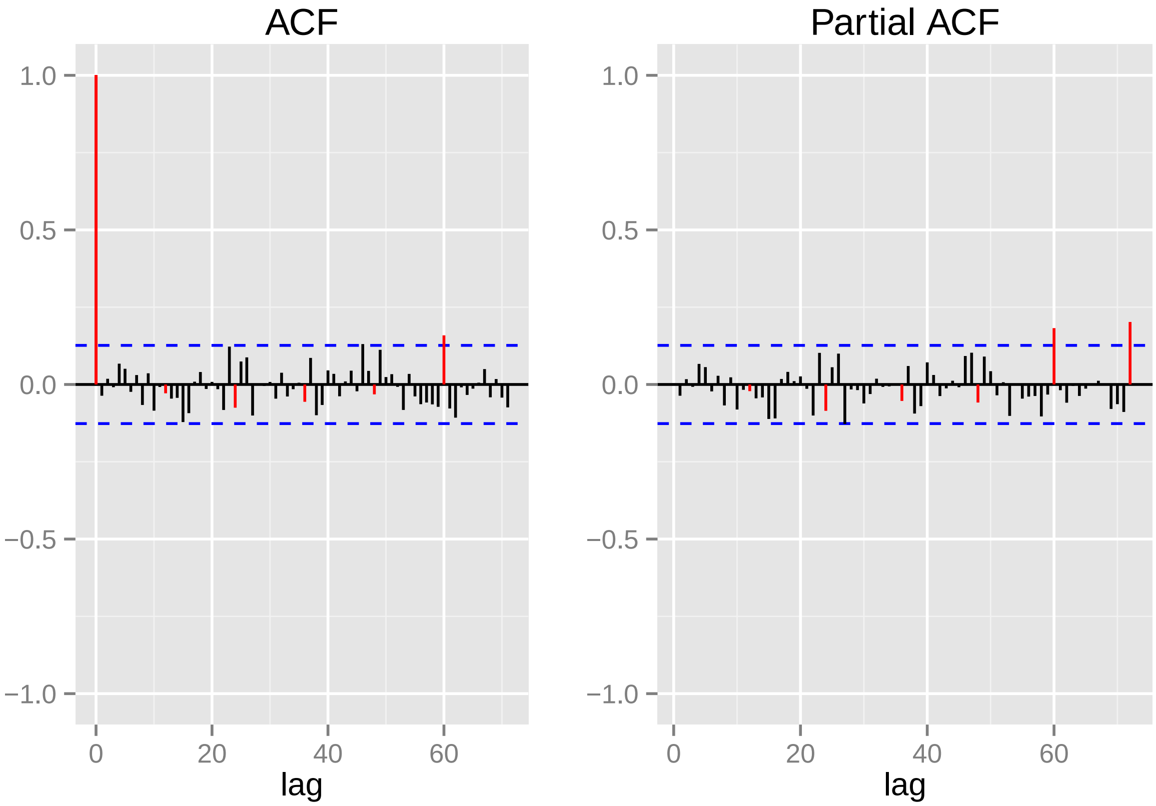


Figure 16: ACF and PACF of the residuals².

There is no volatility because the lags from the ACF and PACF of the squared residuals are not outside the confidence bounds, or far away from origin.

The last step is to perform a white noise test. To do so a Ljung-Box is performed. The resulting p-values are presented in Figure 17.

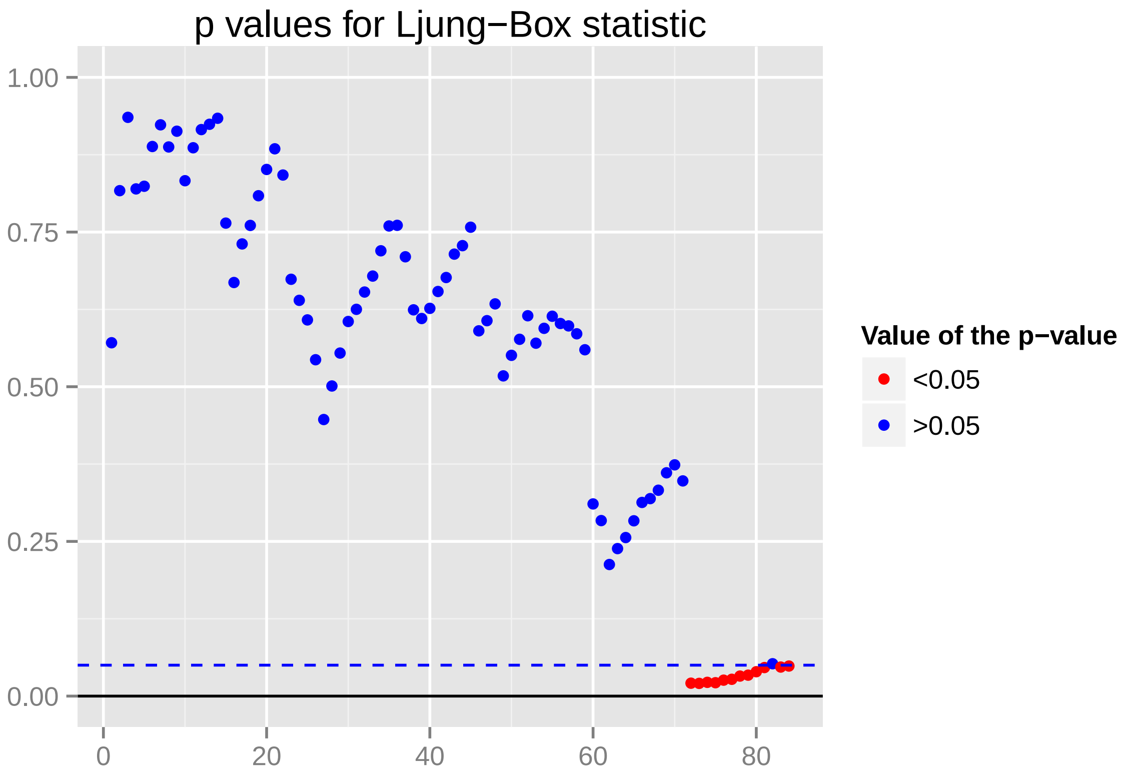


Figure 17: P-values of the Ljung-Box tests.

The fact that some p-values are below 0.05 and so that the null hypothesis of residuals being a white noise is rejected may be due to the fact that the residuals were maybe not really normally distributed.

Model 4 $ARIMA(6,0,0)(3,1,5)_12$

The same analysis was performed on the model 4. Figures are presented in section **??**. Mainly the same results are obtained except for the white noise tests. As one can see in Figure 30 no p-values are below 0.05.

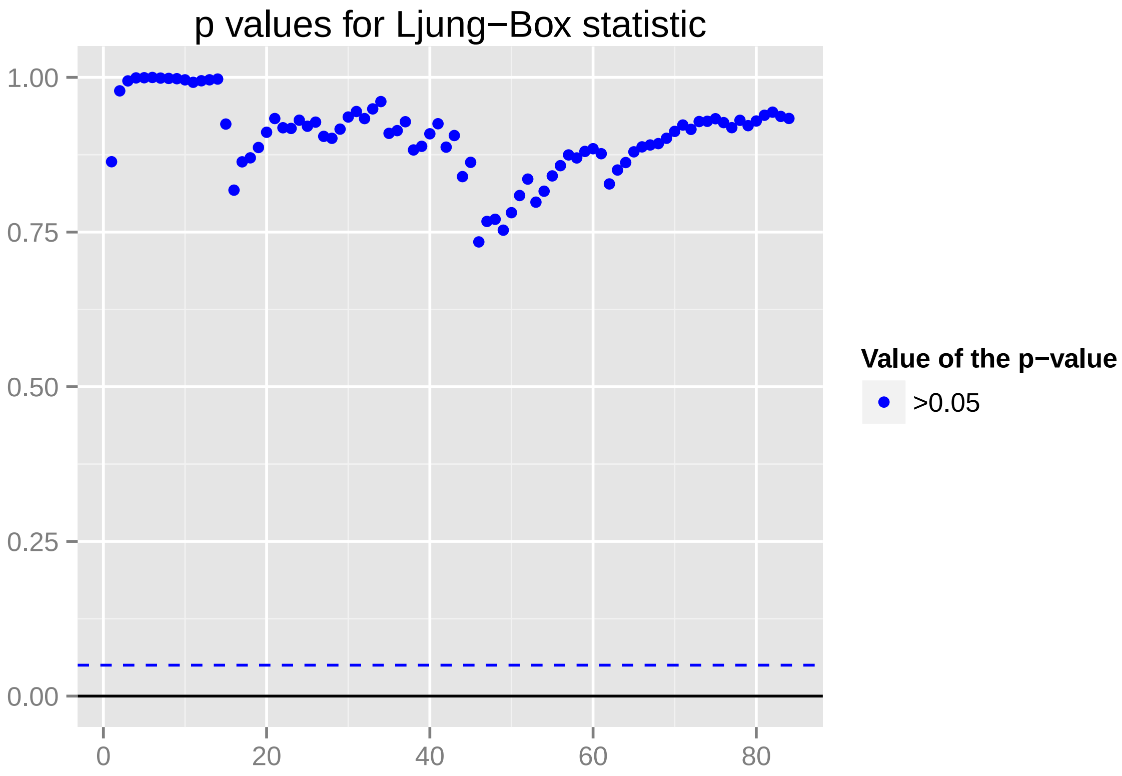


Figure 18: P-values of the Ljung-Box tests.

As a consequence model 3 seems to be a bit better than model 3. Besides it has a lower AIC (-857 vs -819).

Question b

A model is stationary and invertible if the AR characteristic polynomial roots and the MA characteristic polynomial roots respectively are greater than 1.

These roots were computed for the models 3 and 4 and they all were greater than 1 so the two models are invertible and stationary.

Question c

To check if the model is stable, the coefficients of the serie and those from the serie without the 12 last observation are compared. If they are mostly the same then the model is stable.

Table 2 presents the results for the model 3.

	Complete model	Truncate model
ar1	-0.8134	-0.8122
ar2	-0.4797	-0.4793
ar3	0.0000	0.0000
ar4	0.0000	0.0000
ar5	0.1236	0.1248
ar6	0.1988	0.2011
sar1	-0.4419	-0.4369
sma1	0.0000	0.0000
sma2	-0.5200	-0.5175

Table 2: Comparison of the coefficients for model 3

Since the coefficients are the same (mainly) the model 3 can be considered stable.

Table 4 presents the results for the model 4.

	Complete model	Truncate model
ar1	-0.7114	-0.7059
ar2	-0.3261	-0.3218
ar3	0.0603	0.0609
ar4	0.1108	0.1124
ar5	0.2150	0.2266
ar6	0.2045	0.2053
sar1	0.5295	0.5423
sar2	-0.7152	-0.7291
sar3	-0.3285	-0.3138
sma1	-1.2082	-1.2190
sma2	1.0480	1.0854
sma3	0.2160	0.2011
sma4	-0.5996	-0.5843
sma5	0.4025	0.4225

Table 3: Comparison of the coefficients for model 4

As before the coefficients are the same so the model 4 is stable.

Based on the truncate models forecasts were performed for the last year. It allows to check if the predictions are good or not since the values for the last year are available. Figure 19 and 20 shows the predictions made respectively with model 3 and 4 along with the confidence interval.

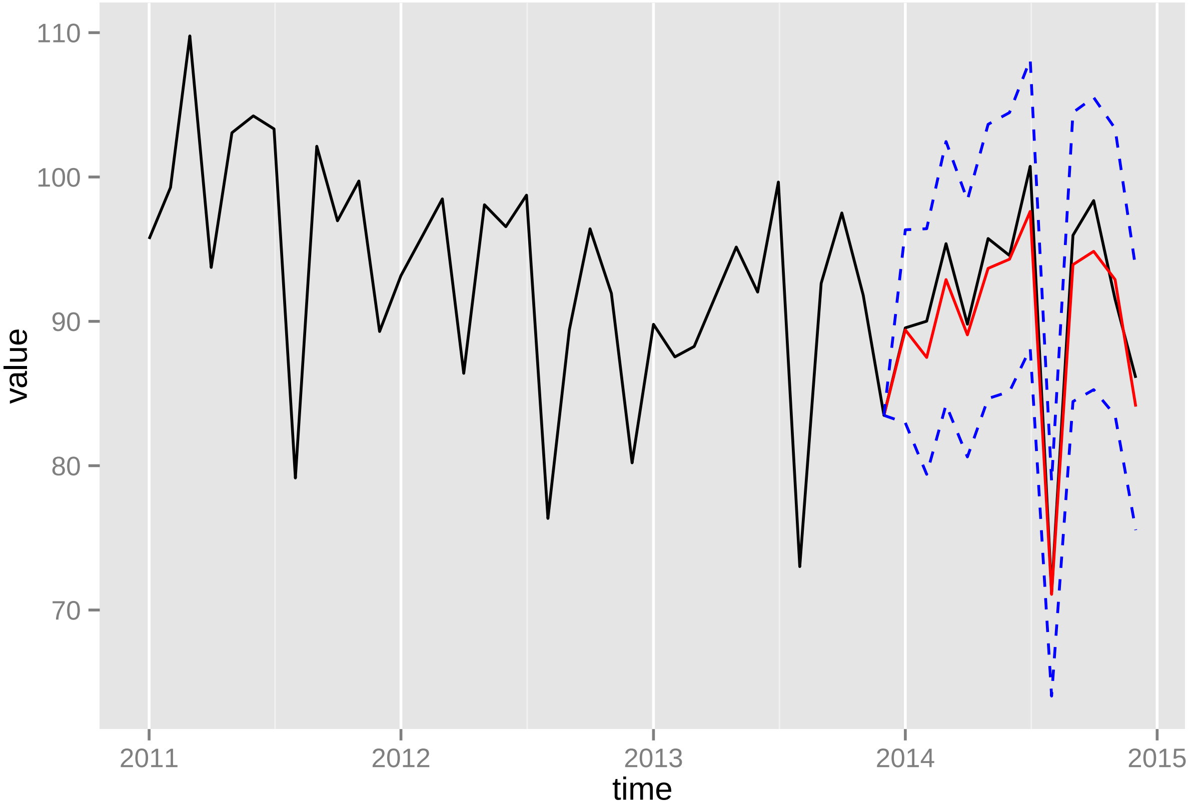


Figure 19: Predictions and confidence interval for the truncate model 3.

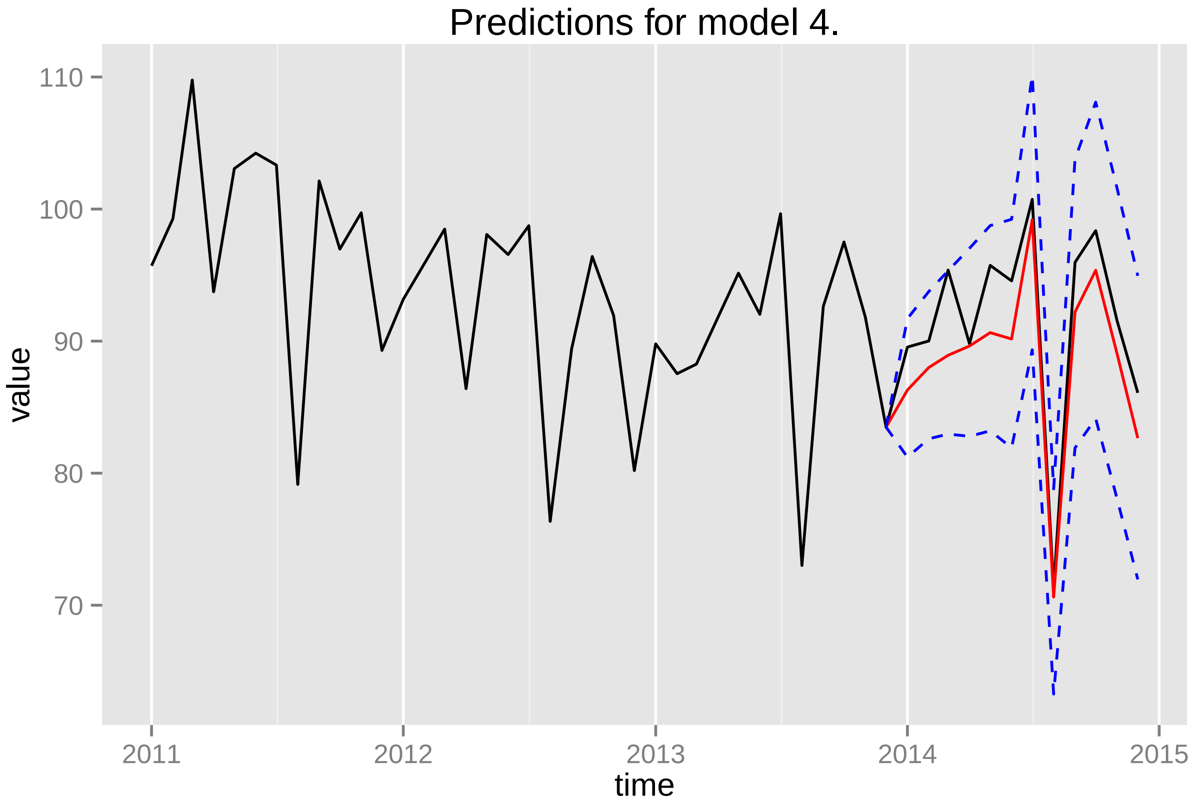


Figure 20: Predictions and confidence interval for the truncate model 4.

Question d

In order to compare the model for the prediction the Mean Square Prediction Error and the Mean Square Absolute Prediction Error are computed for the two models. The lower the values are, the better the model is.

- $MSPE = \sqrt{\frac{\sum(\frac{(obs-pred)}{obs})^2}{12}}$.
- $MSAPE = \sqrt{\frac{\sum\frac{|obs-pred|}{|obs|}}{12}}$

The model 3 has lower values than model 4. Therefore even if model 4 is better to describe the serie, model 3 is better for the forecast.

Predictions

Since model 3 is better for predictions (see section Question d) the long term forecasts will be done with this model.

Figure **??** displays the predicted values for the serie for the year 2015 along with the confidence interval.

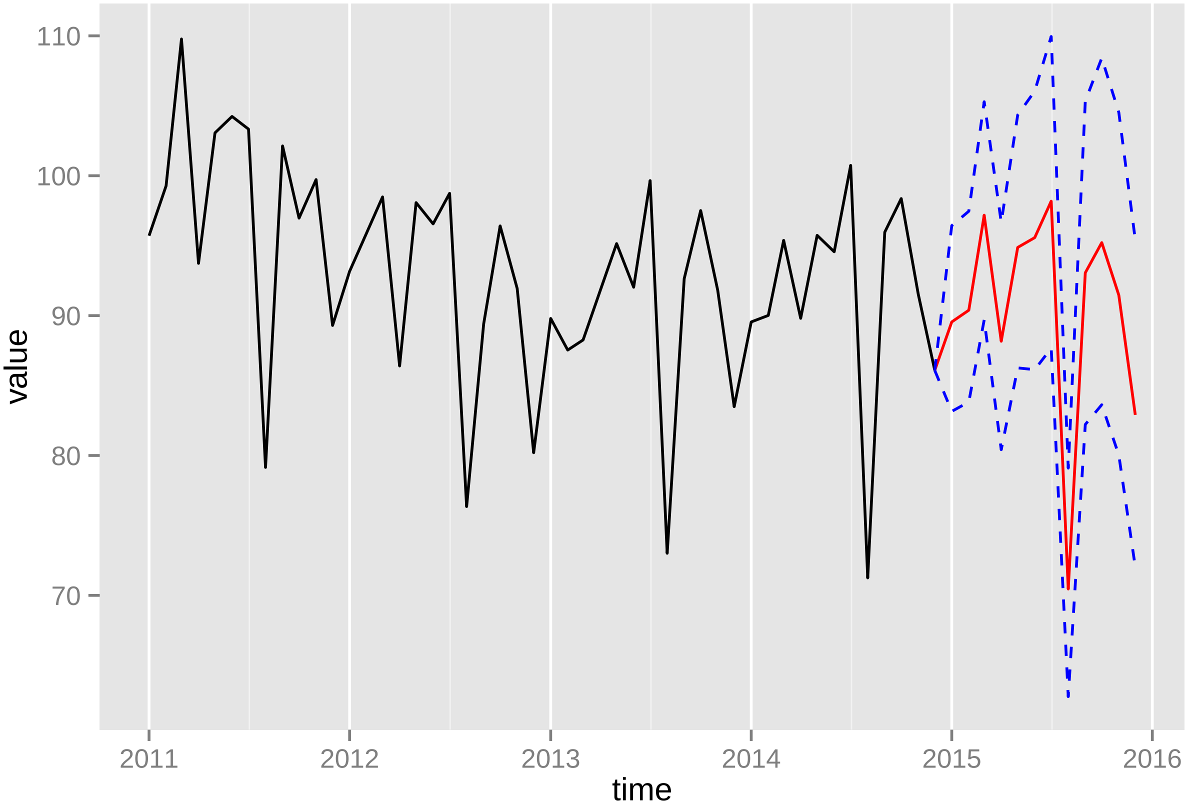


Figure 21: Predictions and confidence interval for the model 3.

Outliers treatment

Question a

In this section outliers are analyzed. For that an automatic detection of the outliers is applied. Table 3 presents the results.

Observations	Kind	Date	Evolution
28	AO	Apr 1997	109.08
62	TC	Feb 2000	105.33
88	AO	Apr 2002	107.43
124	AO	Apr 2005	106.99
159	LS	Mar 2008	94.52
160	AO	Apr 2008	108.35
166	LS	Oct 2008	95.32
168	LS	Dec 2008	91.41
175	LS	Jul 2009	107.38
181	AO	Jan 2010	94.41
220	LS	Apr 2013	104.96

Table 4: Summary of the outliers.

Note that *AO* stands for additive outlier and *LS* for level shift. Additive outliers means that it only affects one point. Instead of observing the predicted point on the linearized serie we observe another point. In level shift there is a break in the serie that is not corrected. The effect is permanent.

We can see that most of the additive outliers are in April, it is maybe due to the years with Easter in March instead of April.

2 level shift can be observed in 2008. They are probably due to the financial crisis since the shift is negative.

The last level shift is in July 2009 and has a positive effect. It may be explain by the recovery of the Spanish production.

Question b

Before comparing the predictions of the linearized serie versus the serie the whole serie can be compared and a model must be fitted on the linear serie.

Figure 22 shows the two series plotted on the same graph.

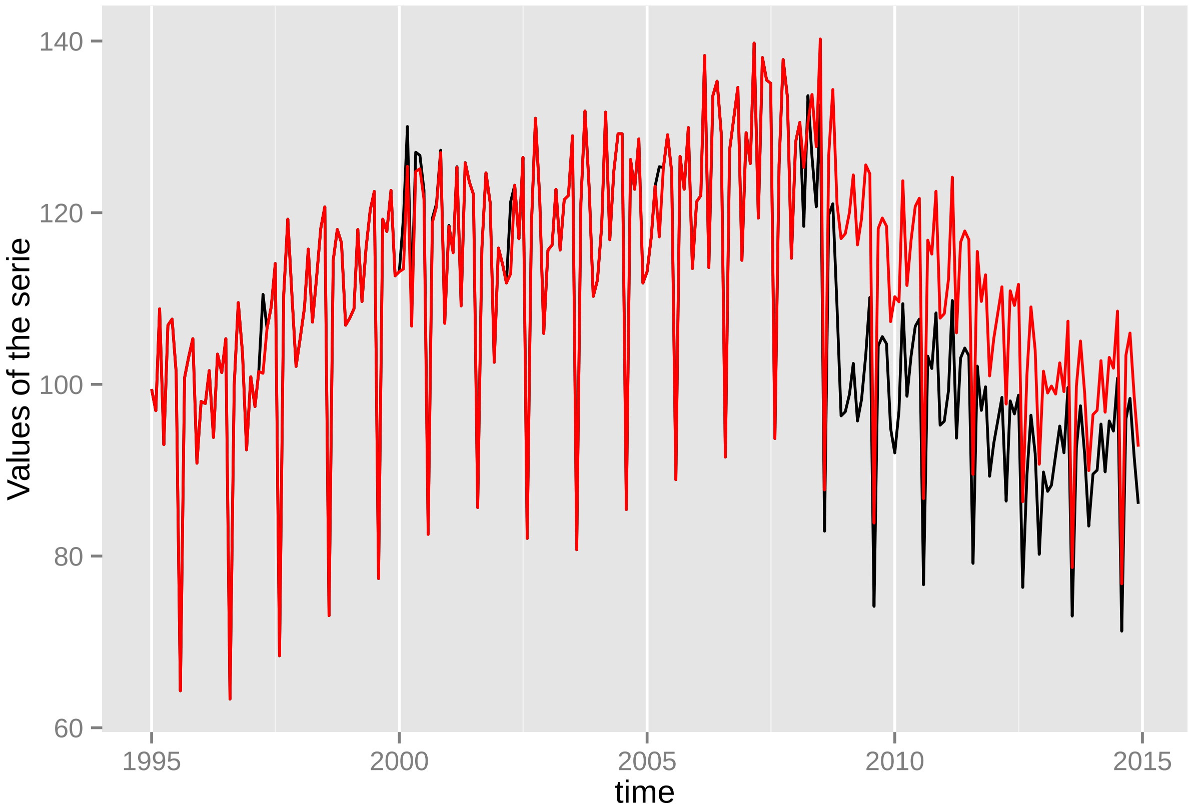


Figure 22: Observed vs linearized serie. The linearized serie is in red.

One can see that the linearized serie is different at the end of the period, after 2008, and that it is above the observed serie.

In the following graph the effects of the outliers on the serie of the logarithm are plotted. Figure 23

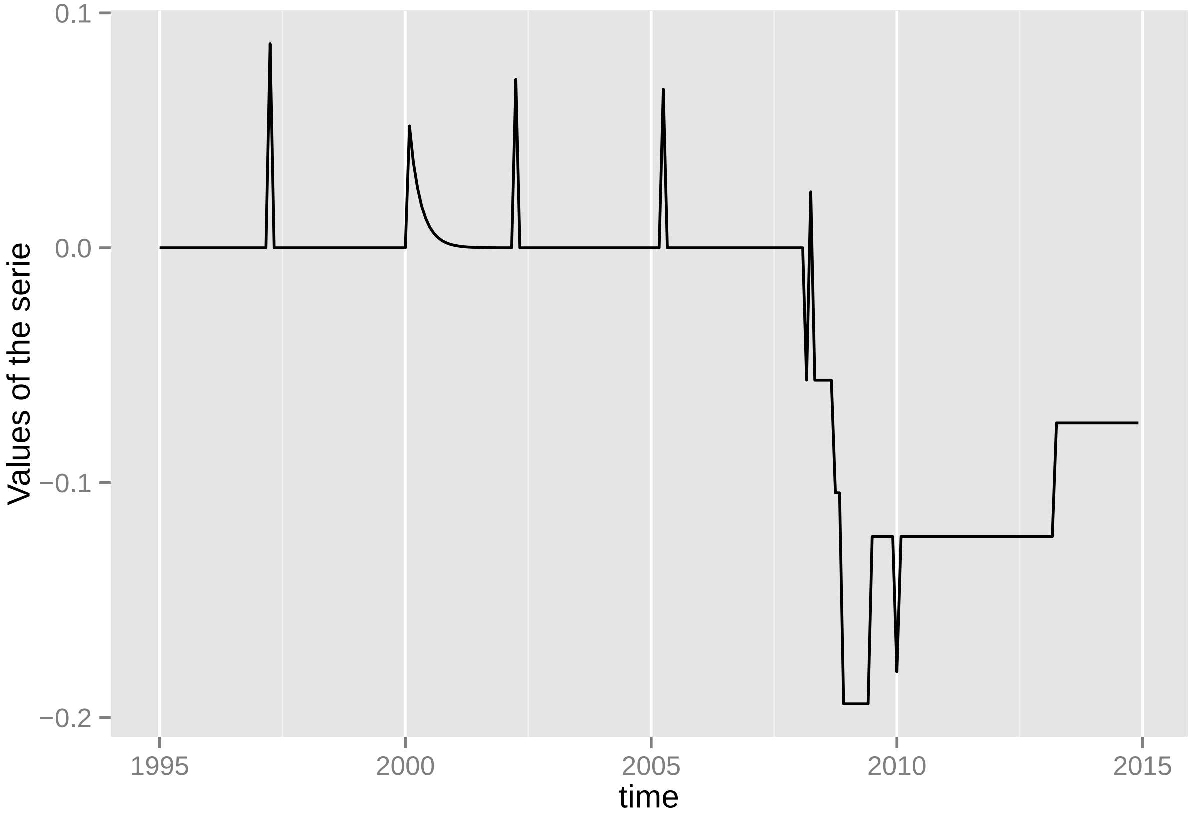


Figure 23: Effects of the outliers.

One can clearly see the different Additive Outliers (only a pick), the Transitive Change for the year 2000 and the several Level Shifts after 2008.

In order to fit a model on the linearized serie the ACF and PACF are plotted in Figure 24. They are done on the two times differentiated serie of the logarithm.

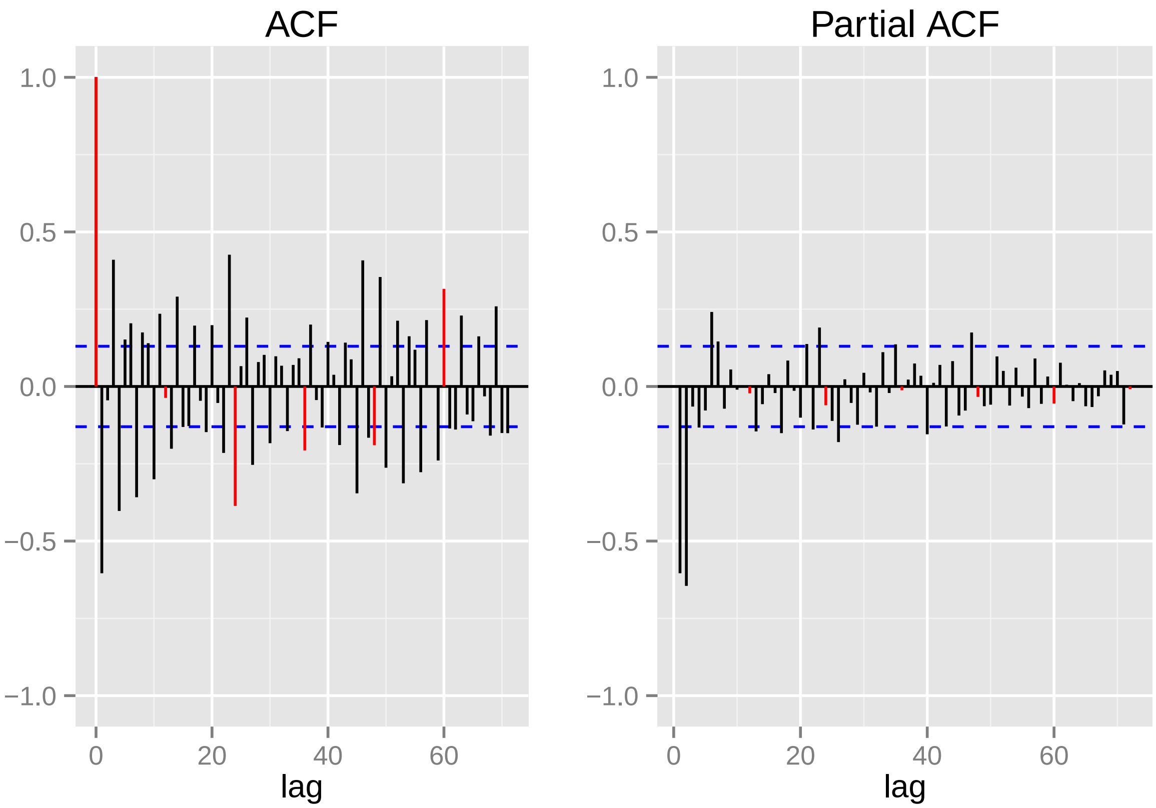


Figure 24: ACF and PACF of the linearized serie.

After validation, we decided to choose an $ARIMA(6,1,0)(1,1,2)_12$. The residuals analysis of this model can be seen in the section Residuals analysis for the linearized model.

As it was done before the truncated serie is used to get prediction. This way a comparison with the predictions of the model 3 for the all serie is possible.

Figure 25 shows the predictions and the confidence interval obtained with the linearized model.

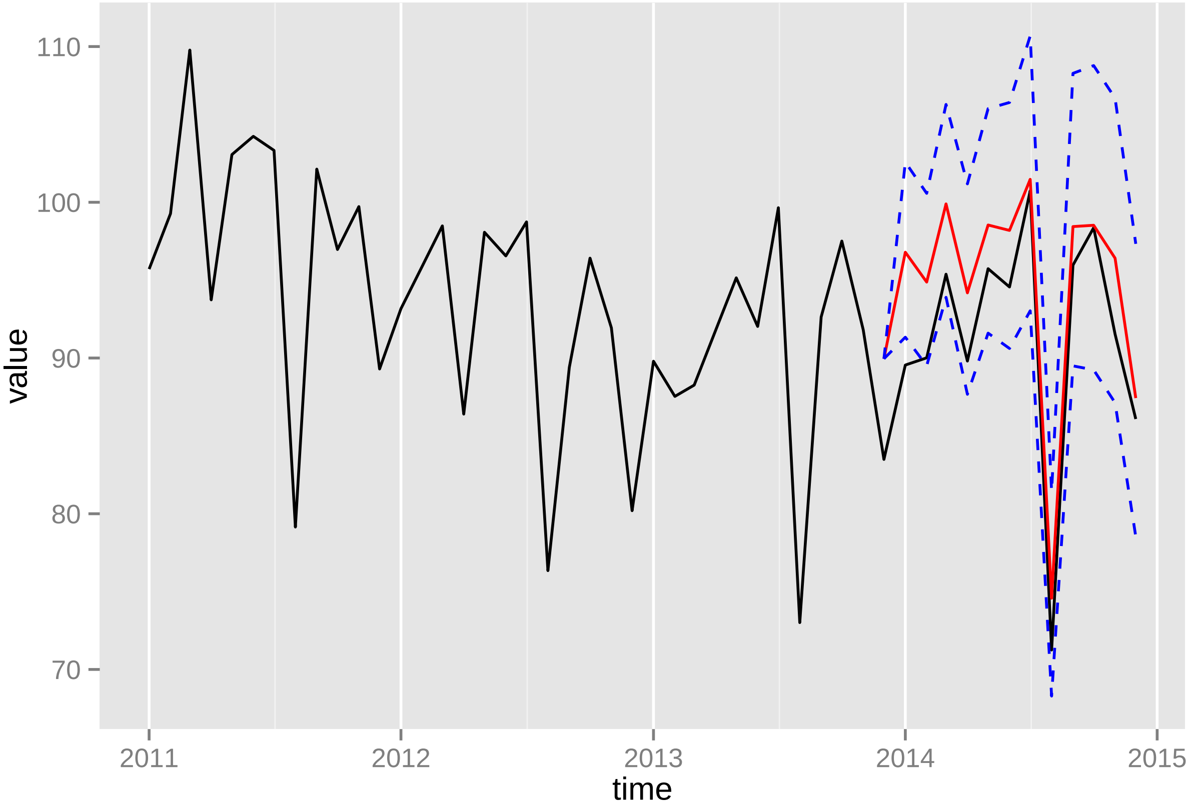


Figure 25: Predictions and confidence interval obtained with the linearized truncated model.

One can notice that the observed values are always in the confidence interval (except at the beginning).

To have a clear comparison, one can compute the Mean Square Prediction Error and the Mean Square Absolute Prediction Error. To compare the predictions of each model the values of these indicators are compared. As it was done previously small values are preferred.

It appears that the model obtained with the whole serie is better for prediction. It can be explained by the fact that there are a lot of level shift at the end and the linearized serie is above the true one.

Figure 26 shows the long term predictions made by the model obtained with the linearized serie.

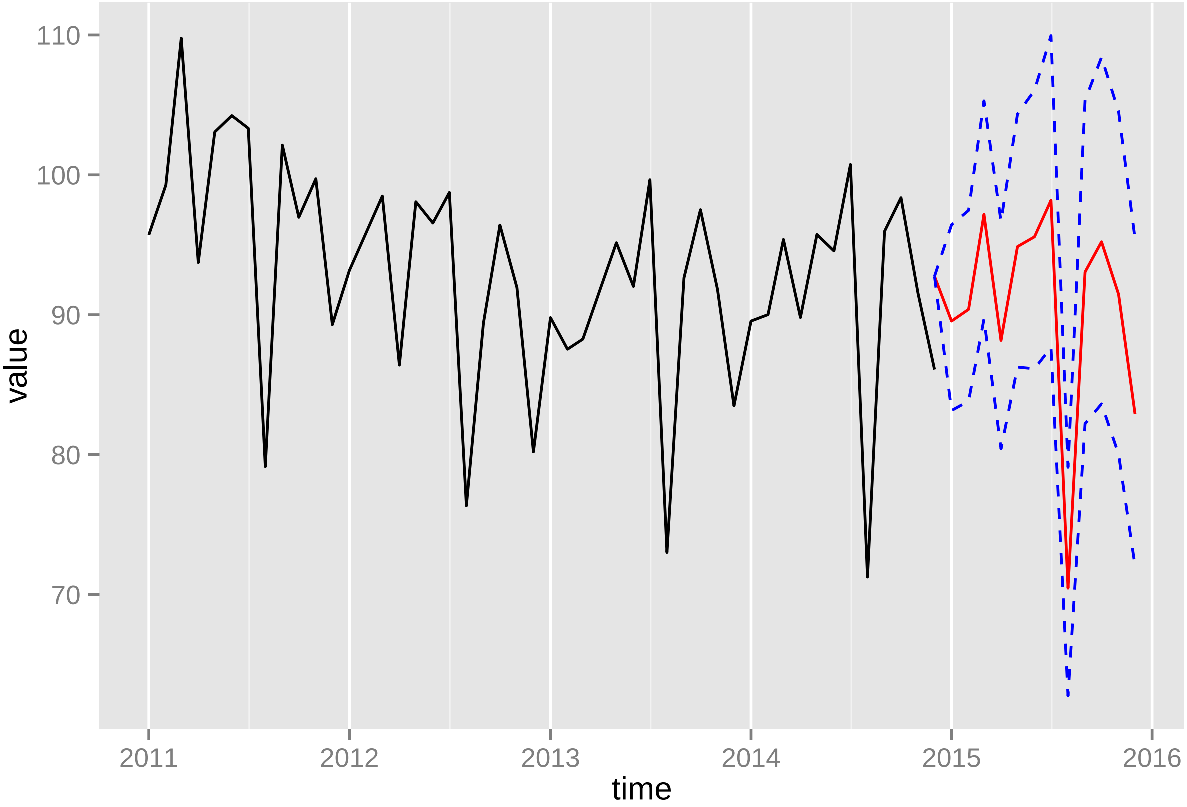
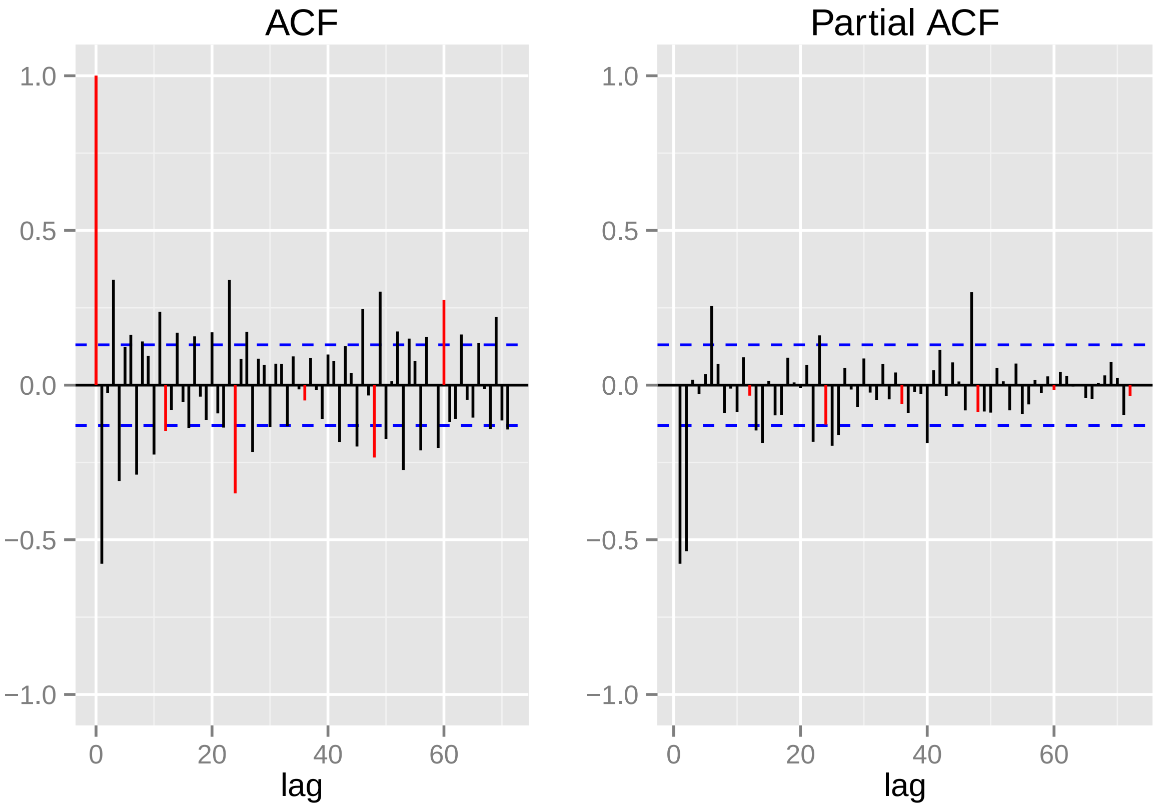


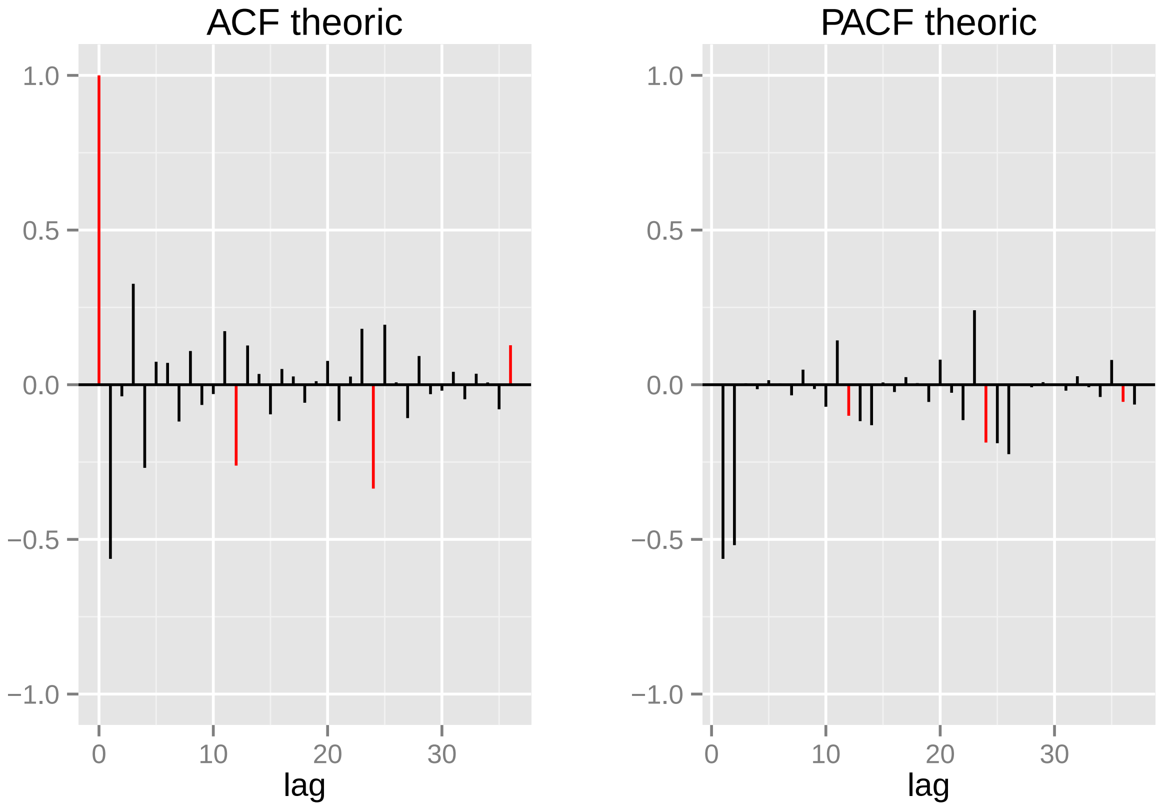
Figure 26: Long term predictions and confidence interval obtained with the linearized truncated model.

Appendix

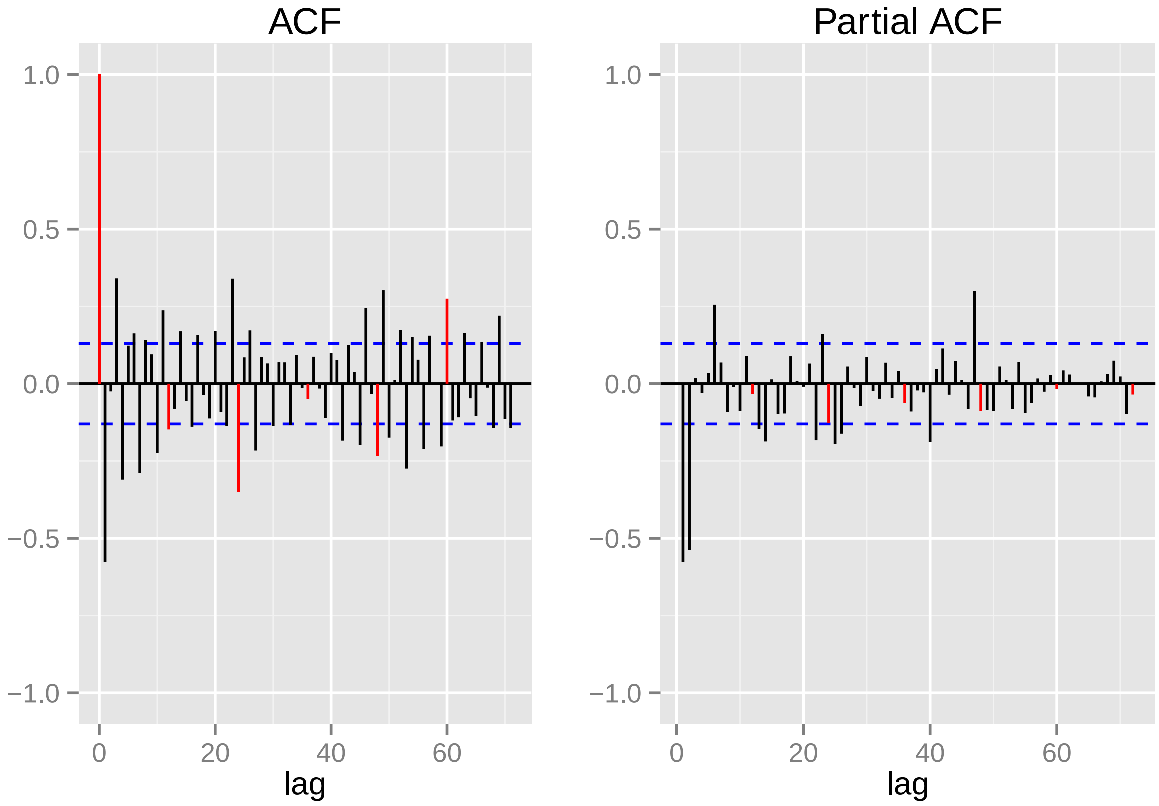
Comparison of ACF and PACF between models and the serie.

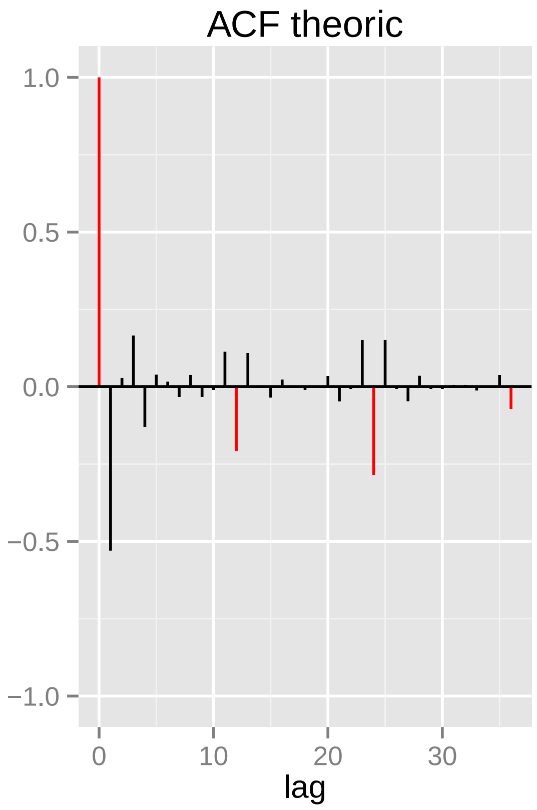
Model 1: $ARIMA(2,0,0)(3,1,2)_12$

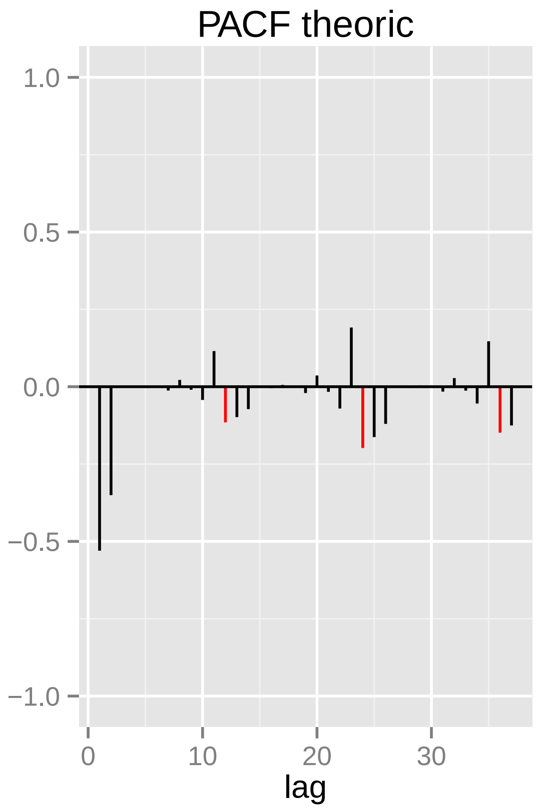



ACF theoric
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
lag
PACF theoric
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
lag

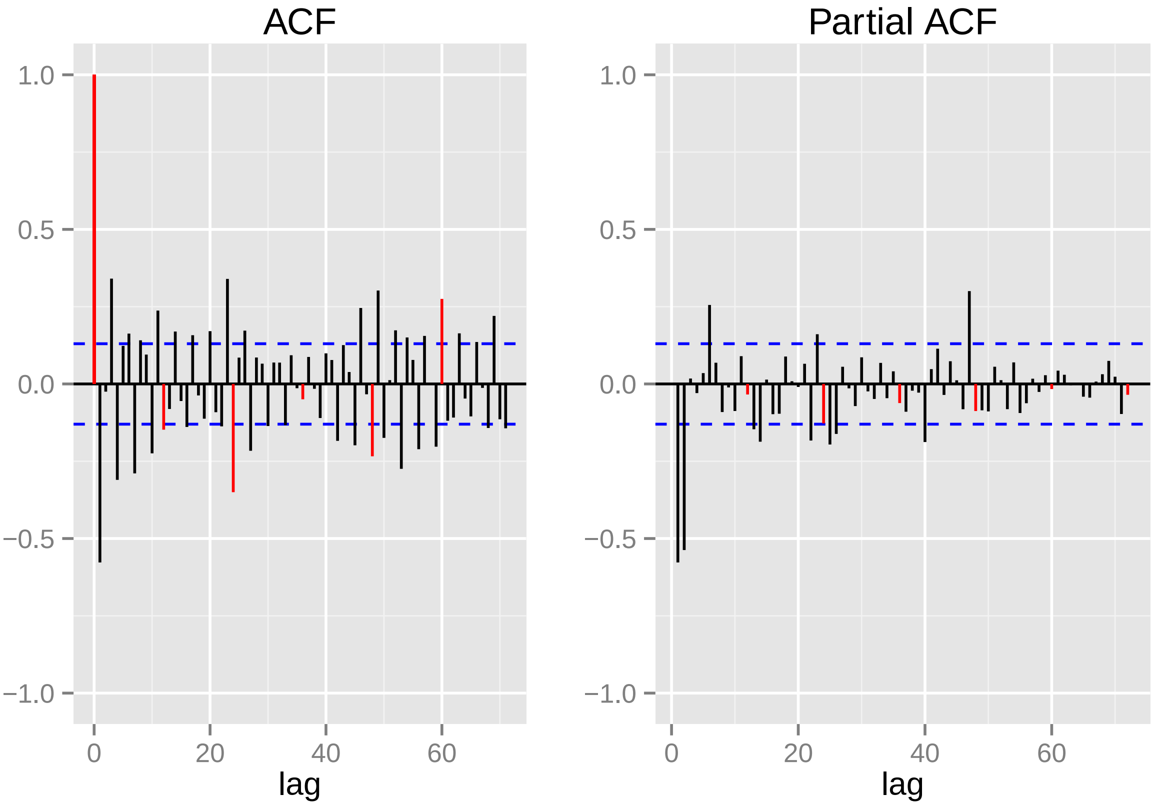
Model 2: $ARIMA(2,0,0)(3,1,5)_12$



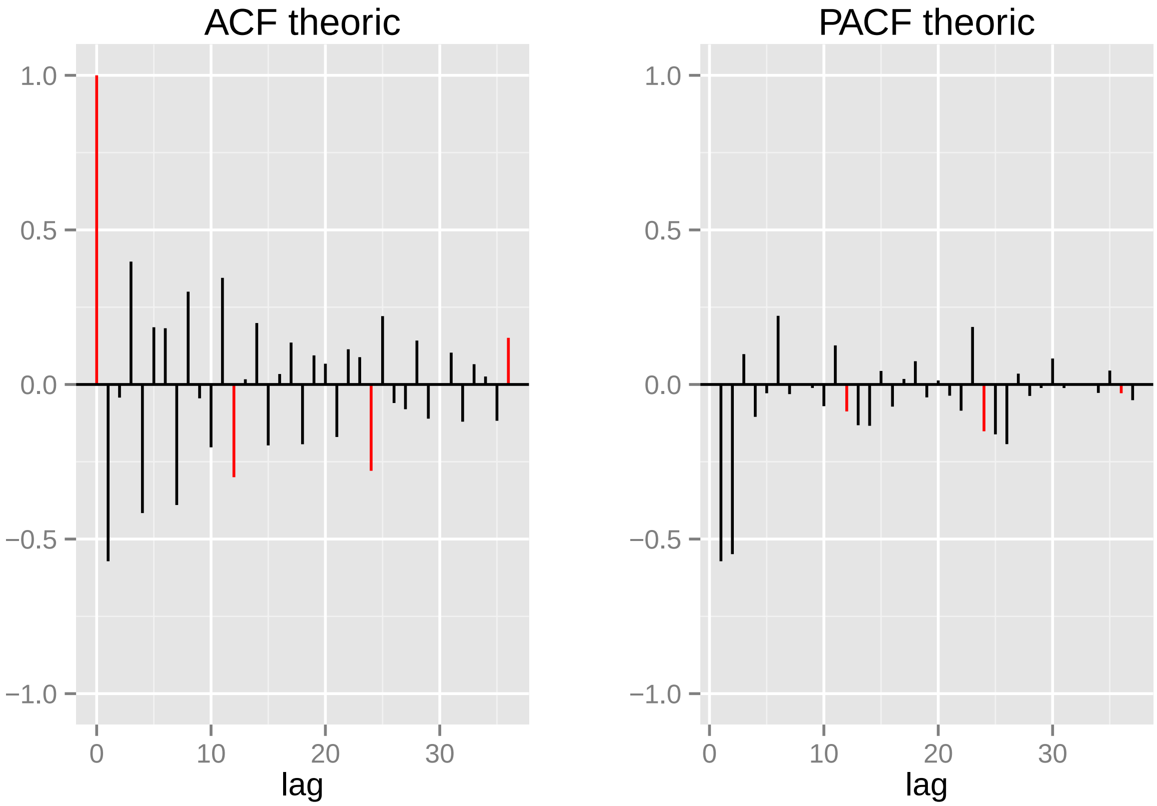

ACF theoric
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
lag


PACF theoric
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
lag

Model 3: $ARIMA(6,0,0)(3,1,2)_12$

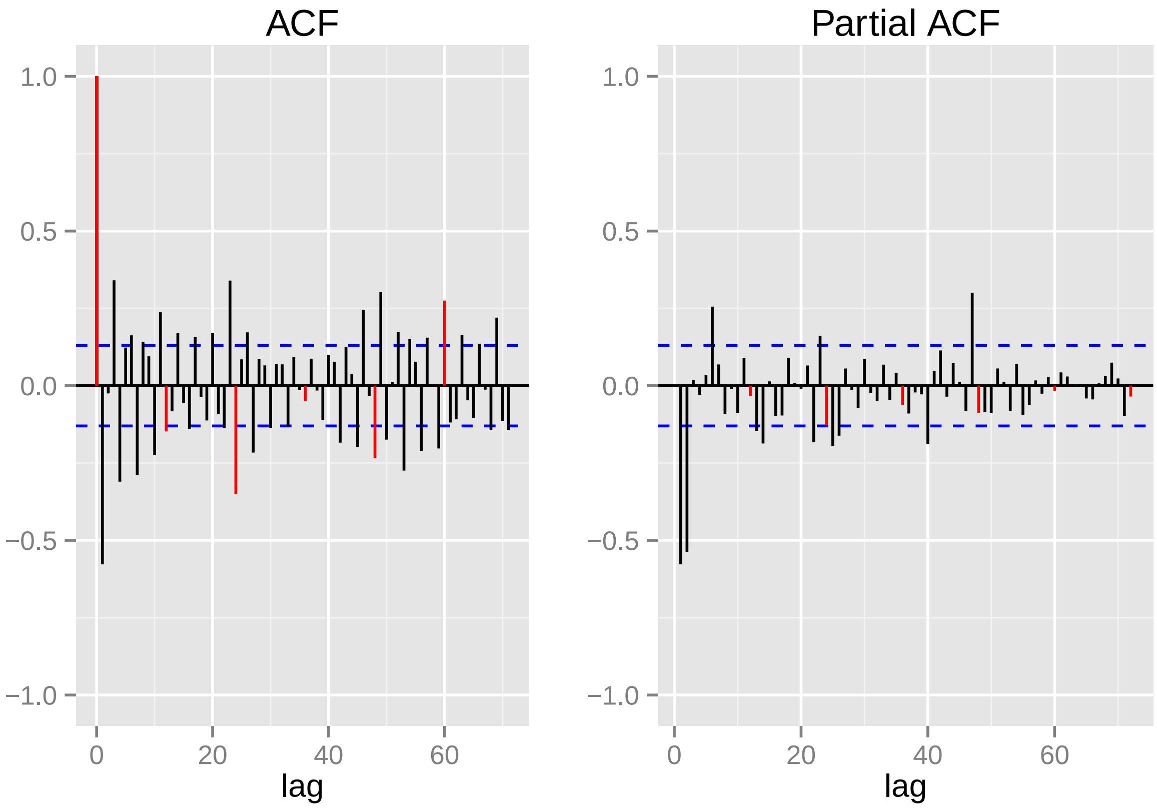


apendix-1.pdf

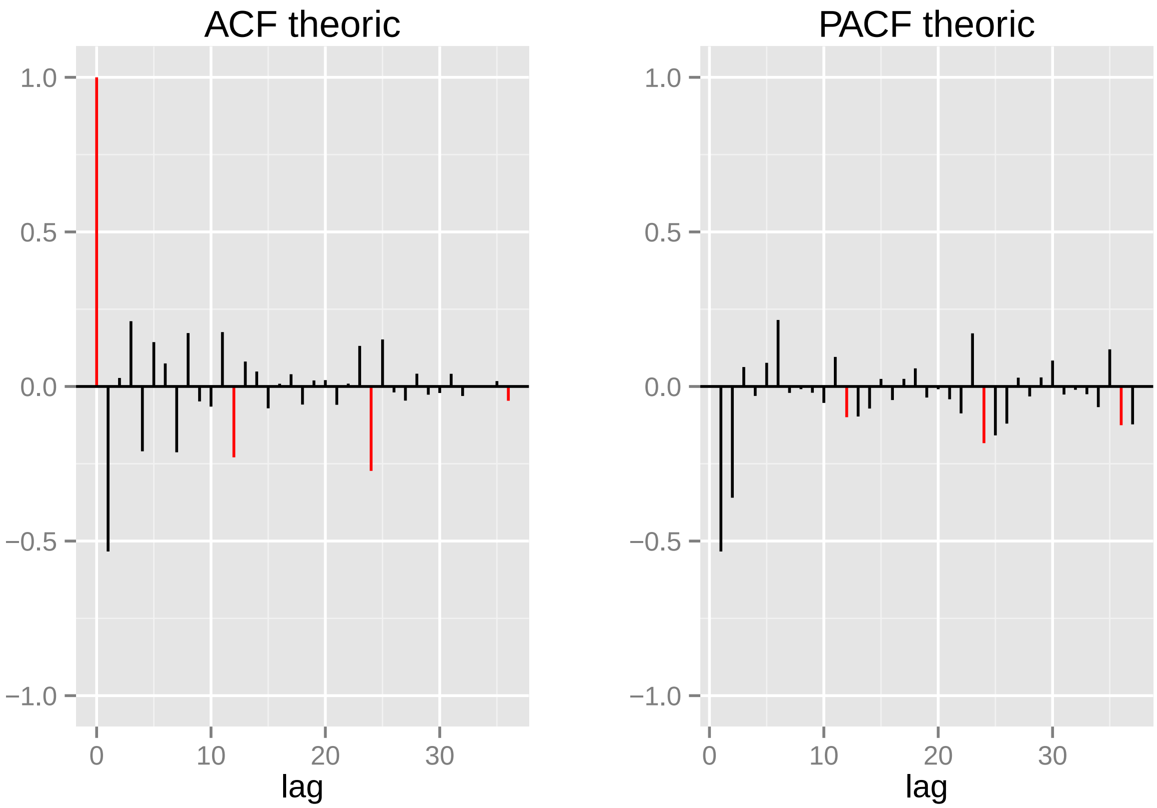


apendix-2.pdf

Model 4: $ARIMA(6,0,0)(3,1,5)_12$



appendix-1.pdf



appendix-2.pdf

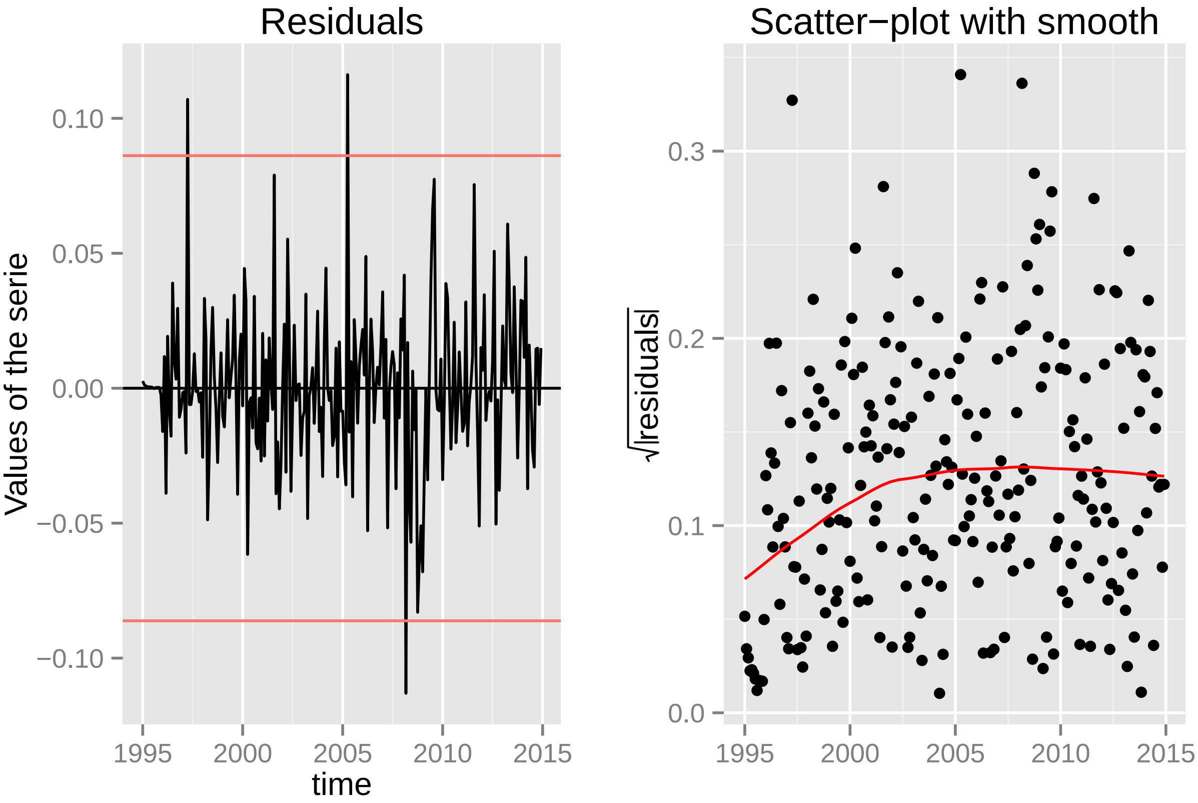
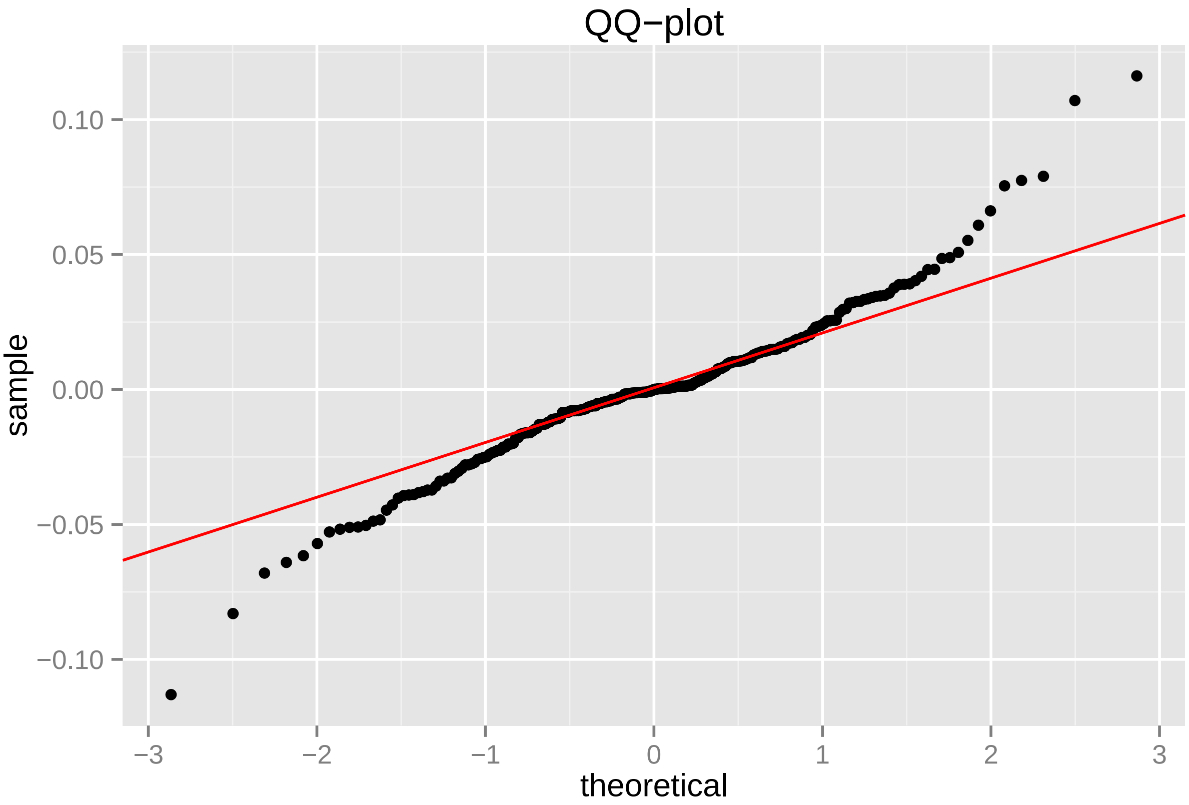
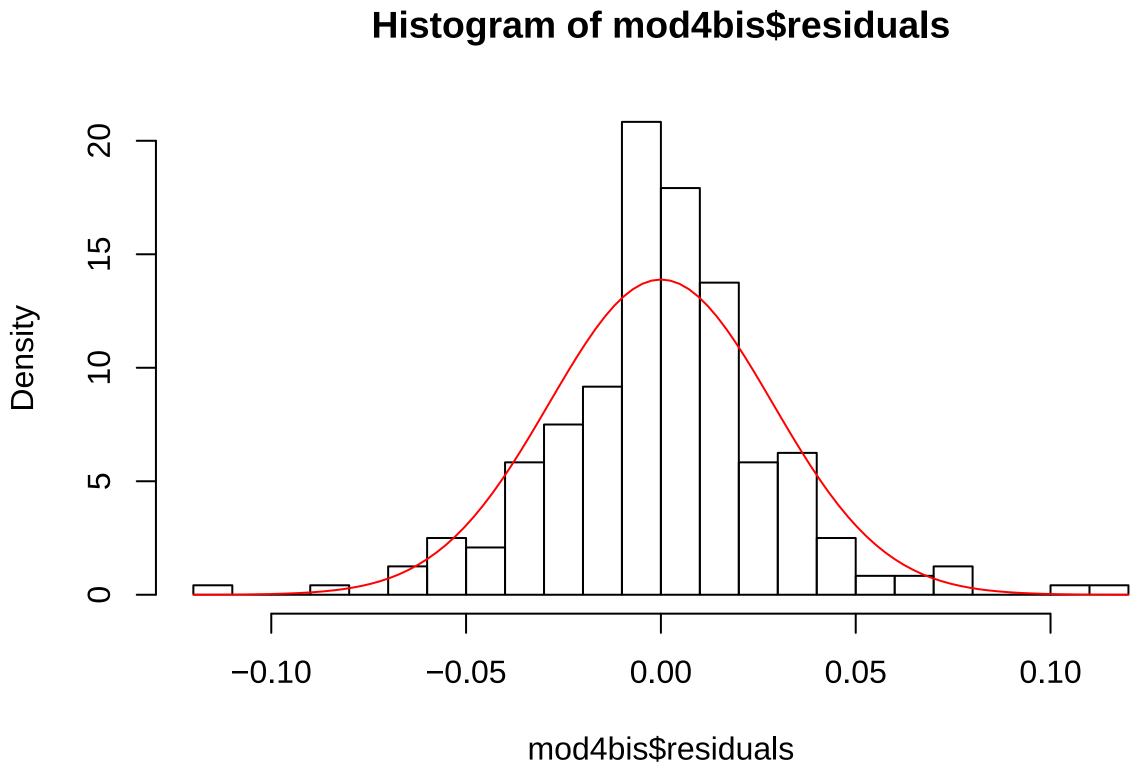


Figure 27: Residuals and scatter plot of the residuals.

Residuals analysis of model 4.




Histogram of mod4bis$residuals
Density
0
5
10
15
20
−0.10
−0.05
0.00
0.05
0.10
mod4bis$residuals

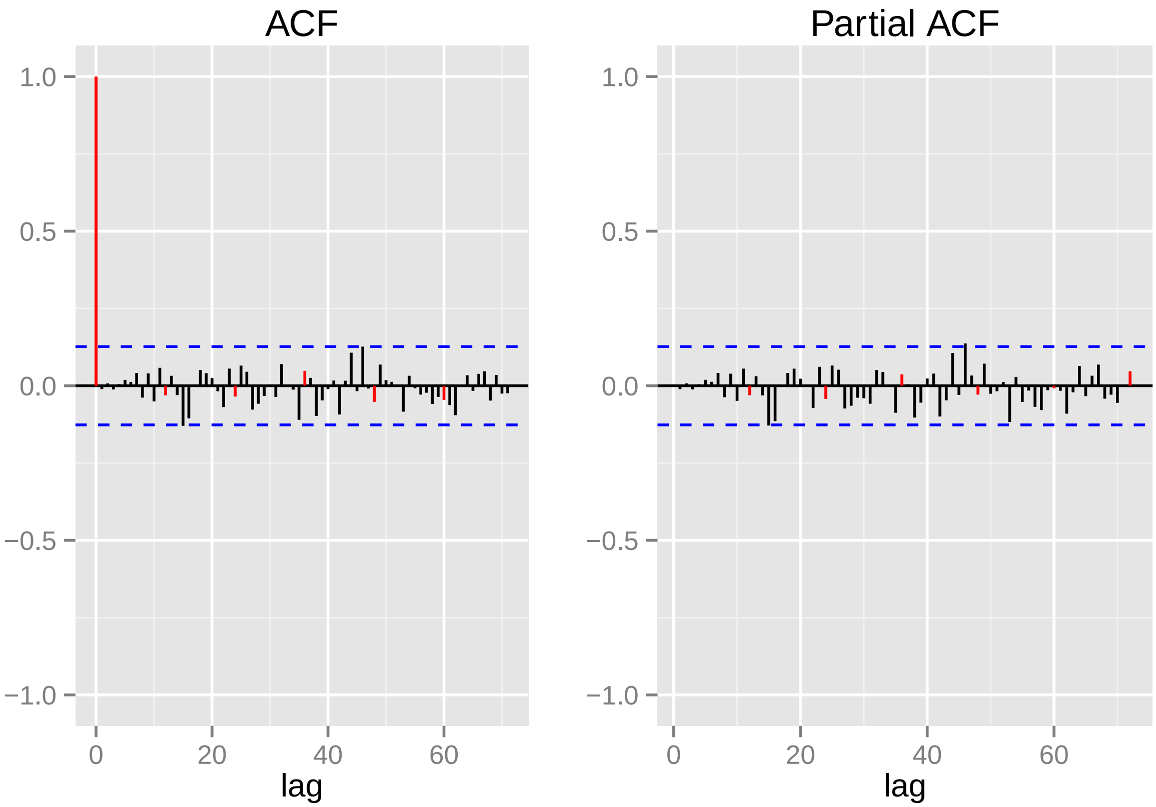


Figure 28: ACF and PACF of the residuals.

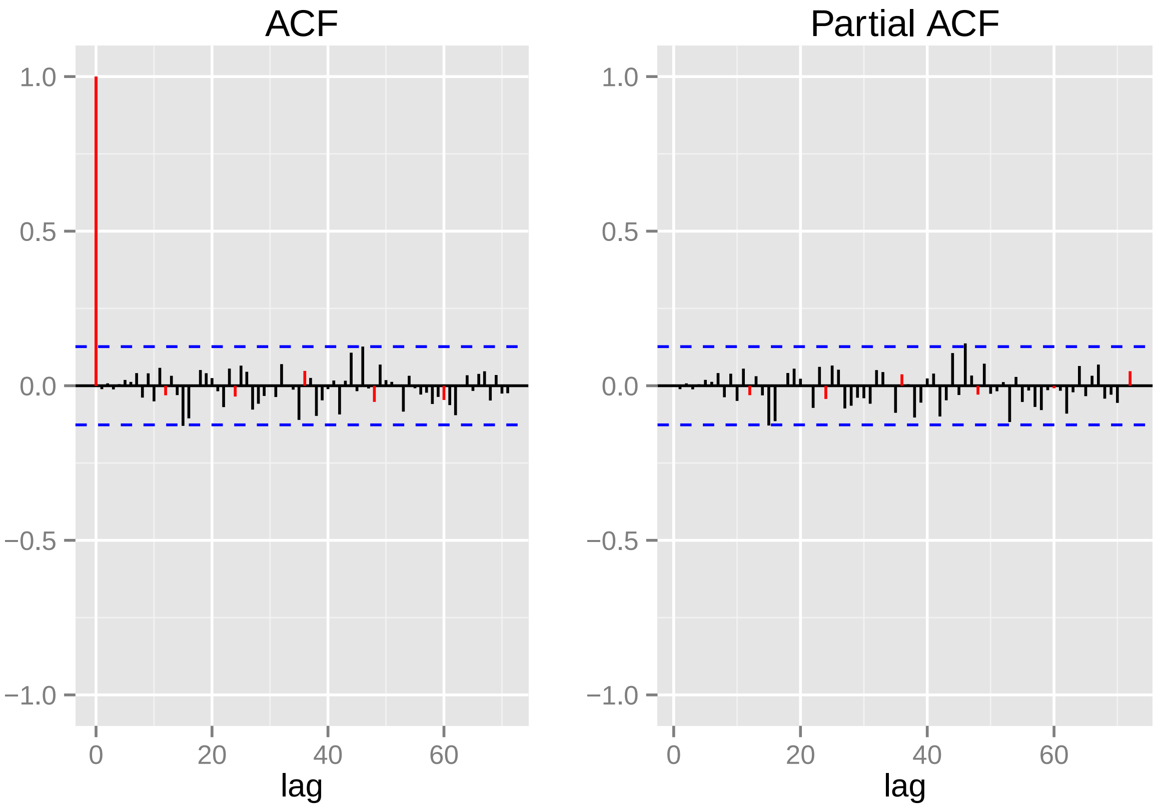


Figure 29: ACF and PACF of the residuals².

appendix-1.pdf

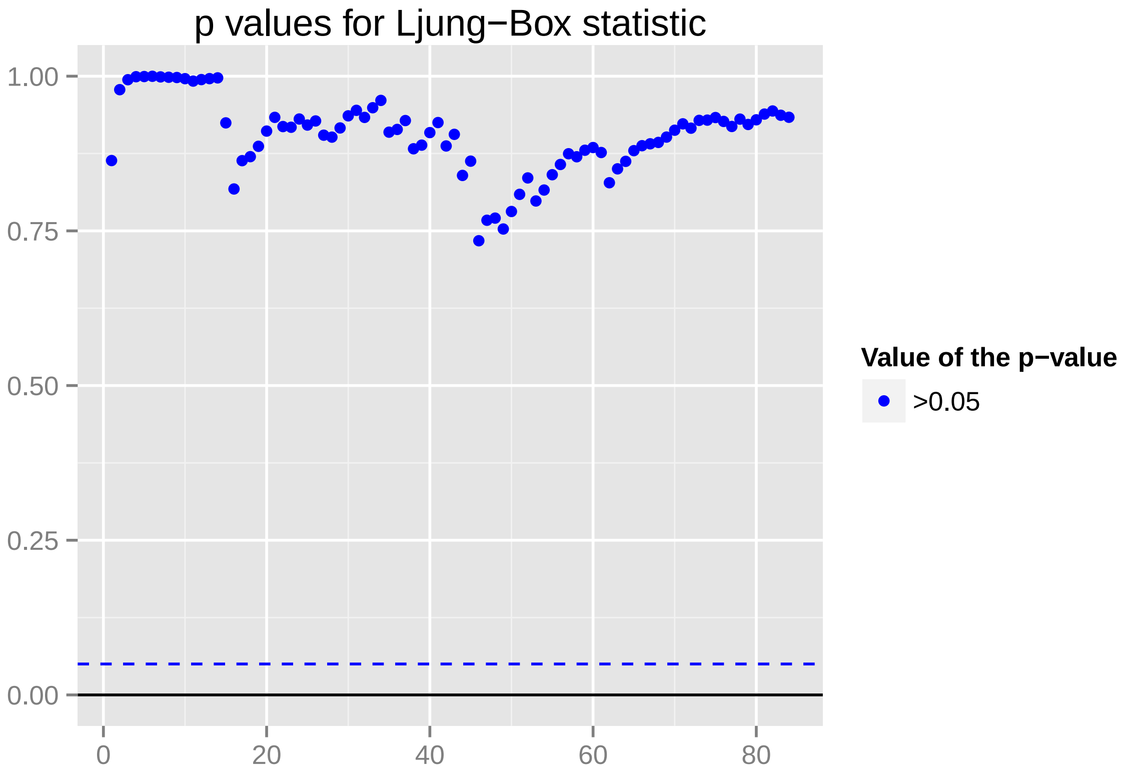


Figure 30: P-values of the Ljung-Box tests.

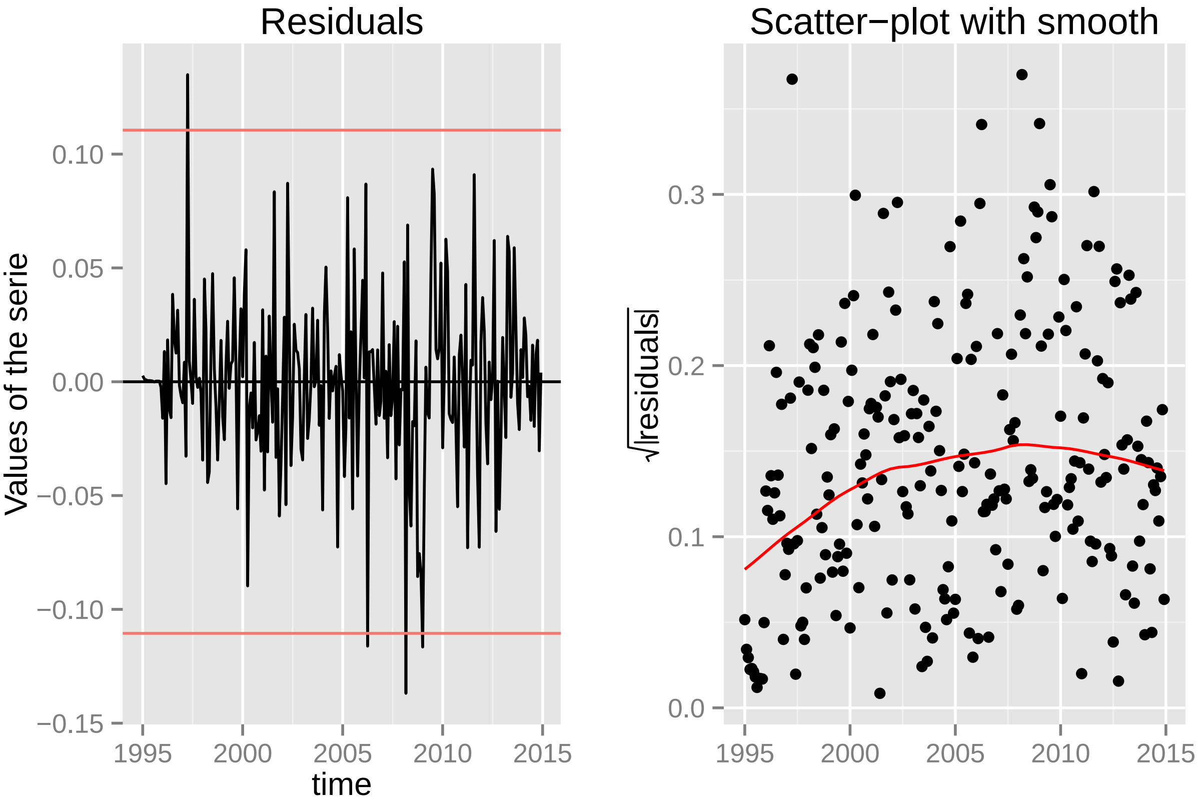
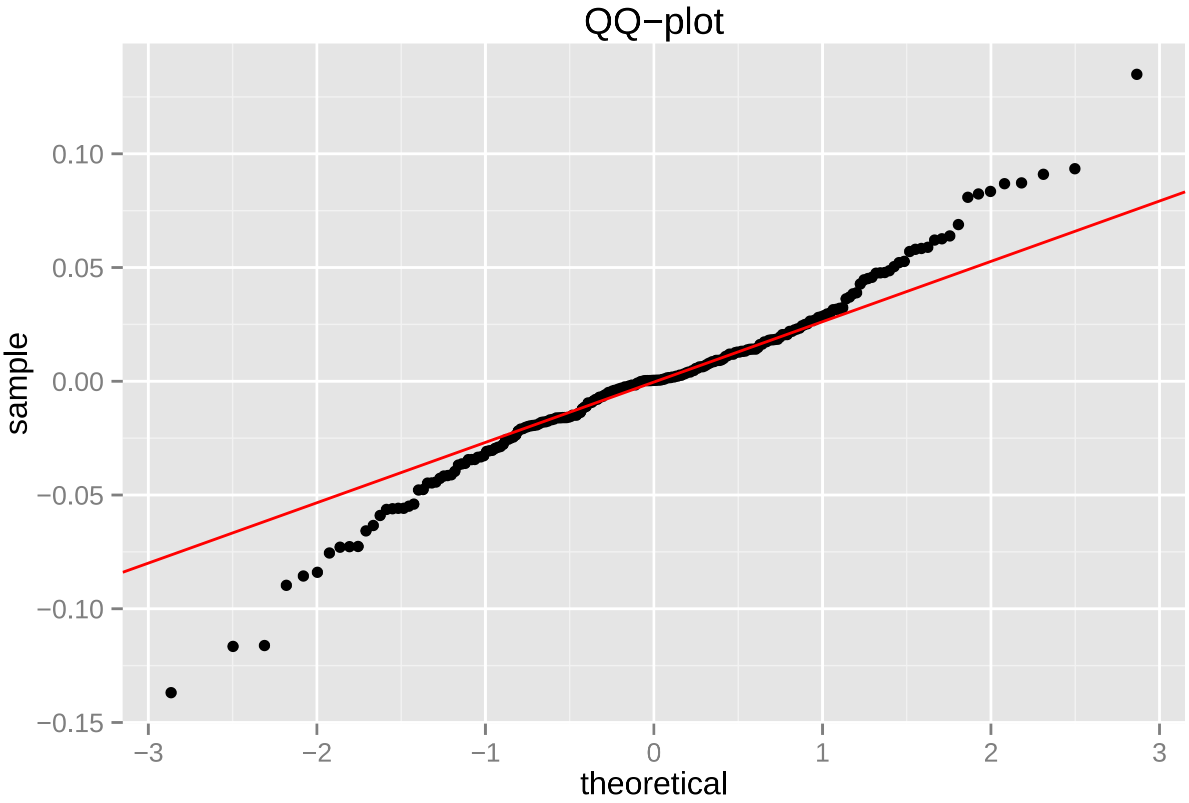
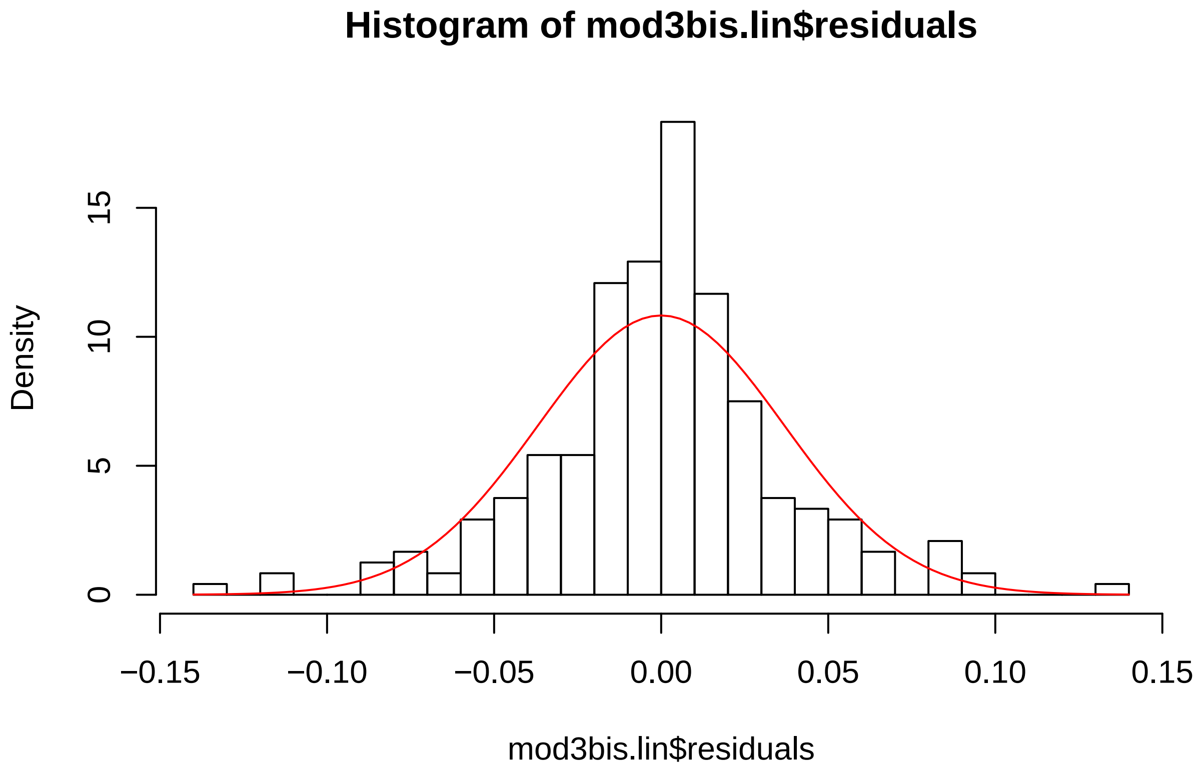


Figure 31: Residuals and scatter plot of the residuals.

Residuals analysis for the linearized model




Histogram of mod3bis.lin$residuals
Density
0
5
10
15
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
mod3bis.lin$residuals

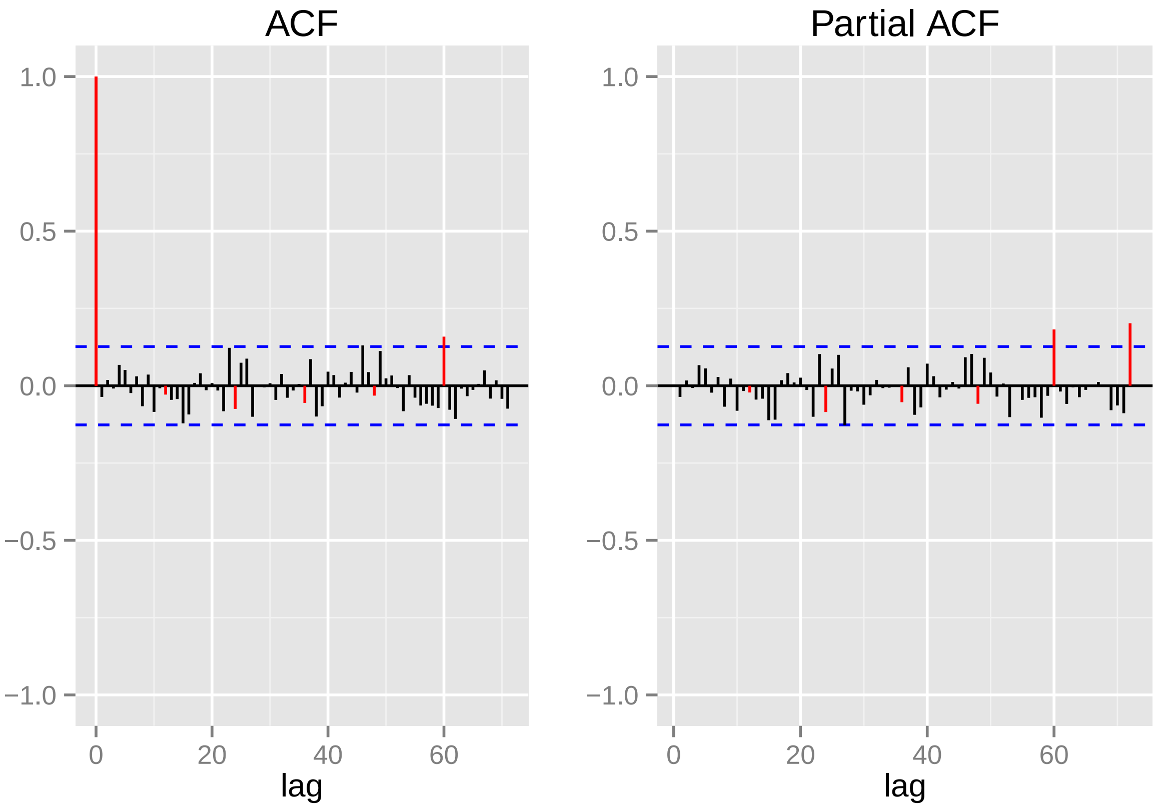


Figure 32: ACF and PACF of the residuals.

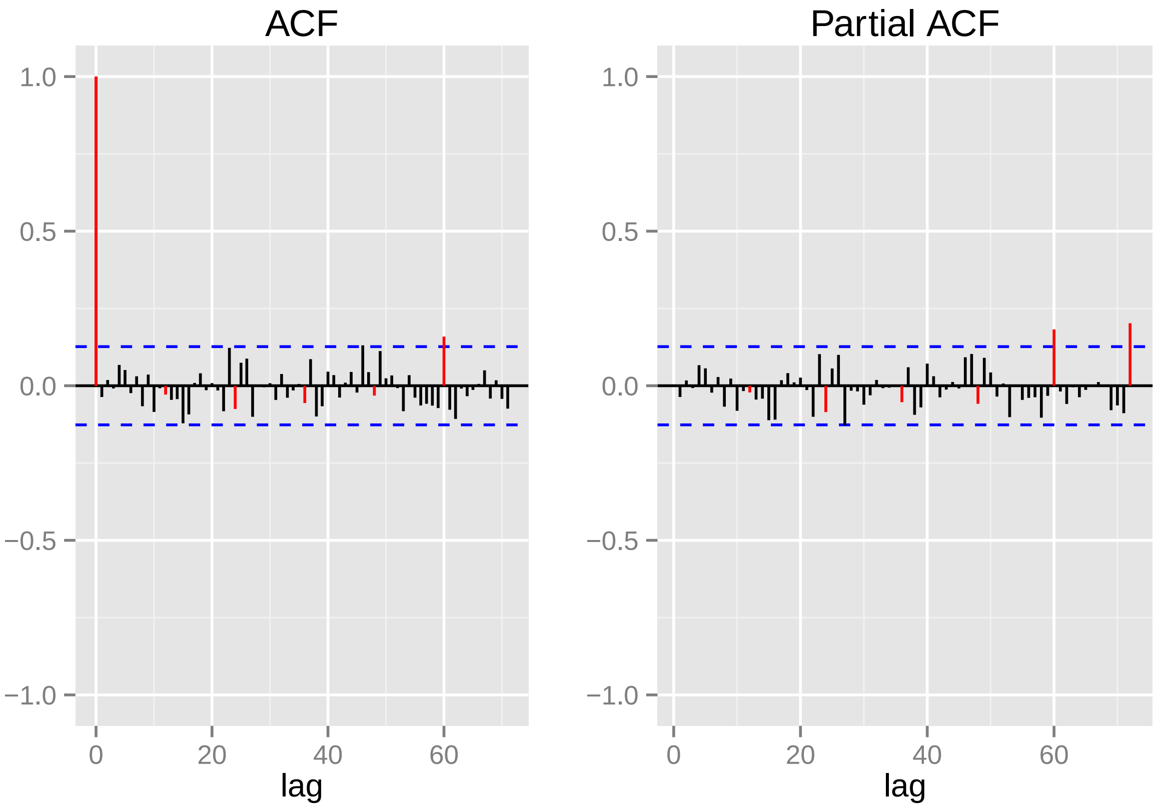


Figure 33: ACF and PACF of the residuals².

appendix-1.pdf

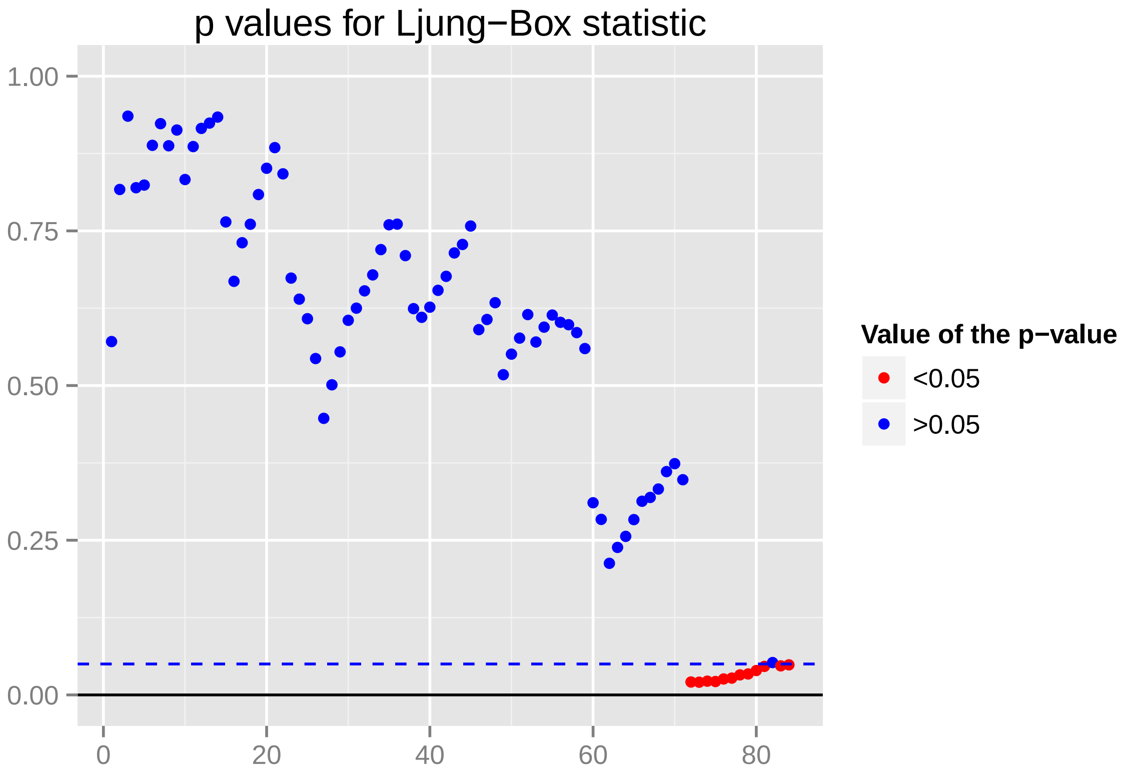


Figure 34: P-values of the Ljung-Box tests.

R code

```
# Packages ----------------------------------------------------------------

library("xtable")
library("ggplot2")
library("gridExtra")
library("zoo")
source('PlotTimeSeriesFunctions.R')
source("outlierTreatment.r")


# Data --------------------------------------------------------------------

ipi <- window(ts(read.table("Data/IPI.dat"),start=1990,freq=12),start=1995)

tsggplot(ipi,title="IPI")


# Identification ----------------------------------------------------------

## Question a

m <- apply(matrix(ipi,nr=12),2,mean)
v <- apply(matrix(ipi,nr=12),2,var)
qplot(m,v,xlab="Yearly mean",ylab="Yearly variance ",main="IPI") +
  stat_smooth(method="lm", se=FALSE)

boxplot(ipi~floor(time(ipi)))
# Variance does not seem to be constant.

logipi <- log(ipi)

m <- apply(matrix(logipi,nr=12),2,mean)
v <- apply(matrix(logipi,nr=12),2,var)
qplot(m,v,xlab="Yearly mean",ylab="Yearly variance ",main="logIPI") +
  stat_smooth(method="lm", se=FALSE)

boxplot(logipi~floor(time(logipi)))
# It looks better for the boxplots.

plot(decompose(logipi))
monthplot(logipi)
# There is a clear seasonal pattern

d12logipi <- diff(logipi,12)

tsggplot(d12logipi,"d12logipi") + geom_hline(y=0)

d1d12logipi <- diff(d12logipi,1)

tsggplot(d1d12logipi) + geom_hline(y=0)
```

```r
var <- as.data.frame(c(var(ipi),var(logipi),var(d12logipi),var(d1d12logipi)))
colnames(var) <- c("variance")
print(xtable(var,digits=4,caption = "Variances of each transformed serie."))

# We select d1d12logipi

ipi.t <- d1d12logipi


## Question b

acfts(ipi.t)
# ARMA(3,2) or ARMA(3,5) for the seasonal part.
# AR(6) or AR(2) for the regular part.

# The two possible models are: ARIMA(6,0,0)(3,1,2)12 or
# ARIMA(6,0,0)(3,1,5)12 or ARIMA(2,0,0)(3,1,2)12 or ARIMA(2,0,0)(3,1,5)12


# Estimation -------------------------------------------------------------

mod1 <- arima(ipi.t,order=c(2,0,0),seasonal=list(order=c(3,0,2),period=12))
mod2 <- arima(ipi.t,order=c(2,0,0),seasonal=list(order=c(3,0,5),period=12))
mod3 <- arima(ipi.t,order=c(6,0,0),seasonal=list(order=c(3,0,2),period=12))
mod4 <- arima(ipi.t,order=c(6,0,0),seasonal=list(order=c(3,0,5),period=12))

mod1bis <- arima(logipi,order=c(2,1,0),seasonal=list(order=c(3,1,2),period=12))
mod2bis <- arima(logipi,order=c(2,1,0),seasonal=list(order=c(3,1,5),period=12))
mod3bis <- arima(logipi,order=c(6,1,0),seasonal=list(order=c(3,1,2),period=12))
mod4bis <- arima(logipi,order=c(6,1,0),seasonal=list(order=c(3,1,5),period=12))
# We first check if the constant is needed. Since it is not we work with
# the models without the intercept. (estimate/se<2)

## Model 1

acfmodel(mod1)
acfts(ipi.t)
acfmodel(mod1bis)
# Really good for ACF, quite good for PACF.

## Model 2

acfmodel(mod2)
acfts(ipi.t)
acfmodel(mod2bis)
# Quite good for ACF and PACF.

## Model 3

acfmodel(mod3)
acfts(ipi.t)
acfmodel(mod3bis)
# Really good for ACF and PACF.
```

```
## Model 4

acfmodel(mod4)
acfts(ipi.t)
acfmodel(mod4bis)
# Really good for ACF and PACF, better than the previous one.

# We can to take out the non-significant parameters to see if it improves
# the model. (estimates/se<2). We will do such a analysis on model 3 and 4
# since they seem to be the best.

mod3bis.adjusted <- arima(logipi,order=c(6,1,0),seasonal=list(order=c(1,1,2),
                          period=12),fixed=c(NA,NA,0,0,NA,NA,NA,0,NA))
# After several test this model is better (aic smaller). Is it still
# a good fit?

acfts(ipi.t)
acfmodel(mod3bis.adjusted)
# We keep this model as the new mod3bis.

mod3bis <- mod3bis.adjusted

mod4bis.adjusted <- arima(logipi,order=c(6,1,0),seasonal=list(order=c(3,1,5),
                          period=12),fixed=c(NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
                                             NA,0,NA,NA))
# The aic is better with all the parameters.

# We will perform the residuals analysis on the models 3 and 4.


# Validation ---------------------------------------------------------------

## Question a

# model 3

# Constant variance (homoscedasticity)
resid <- residplot(mod3bis)
scatter <- scatterggplot(mod3bis)
grid.arrange(resid,scatter,ncol=2)
# Pretty good

# Normal residuals
qqggplot(mod3bis)
# Almost ok

hist(mod3bis$residuals,breaks=20,freq=F)
curve(dnorm(x,mean=0,sd=sd(mod3bis$residuals)),col=2,add=T)
# Almost ok

# Independance of the residuals
acfts(mod3bis$residuals,"Residuals")
# Independant (significant lag are far away)
# Volatility
acfts(mod3bis$residuals,"Residuals²")
# No volatility

# White noise test (above 0.05 => wn)
ljunggplot(mod3bis)
# Someissues at the end

# model 4

# Constant variance (homoscedasticity)
resid <- residplot(mod4bis)
scatter <- scatterggplot(mod4bis)
grid.arrange(resid,scatter,ncol=2)
# Some outliers but seem constant

# Normal residuals
qqggplot(mod4bis)
# Almost ok

hist(mod4bis$residuals,breaks=20,freq=F)
curve(dnorm(x,mean=0,sd=sd(mod4bis$residuals)),col=2,add=T)
# Seem normal

# Independance of the residuals
acfts(mod4bis$residuals,"Residuals")
# Independant

# Volatility
acfts(mod4bis$residuals,"Residuals²")
# No volatility

# White noise test (above 0.05 => wn)
ljunggplot(mod4bis)
# OK


## Question b

cat("\nModul of AR Characteristic polynomial Roots: ",
    Mod(polyroot(c(1,-mod3$model$phi))),"\n")
cat("\nModul of MA Characteristic polynomial Roots: ",
    Mod(polyroot(c(1,mod3$model$theta))),"\n")

length(Mod(polyroot(c(1,-mod3$model$phi)))[Mod(polyroot(c(1,-mod3$model$phi)))<1])
length(Mod(polyroot(c(1,mod3$model$theta)))[Mod(polyroot(c(1,mod3$model$theta)))<1])
# Stationary and invertible

cat("\nModul of AR Characteristic polynomial Roots: ",
    Mod(polyroot(c(1,-mod4$model$phi))),"\n")
cat("\nModul of MA Characteristic polynomial Roots: ",
    Mod(polyroot(c(1,mod4$model$theta))),"\n")
```

```
length(Mod(polyroot(c(1,-mod4$model$phi)))[Mod(polyroot(c(1,-mod4$model$phi)))<1])
length(Mod(polyroot(c(1,mod4$model$theta)))[Mod(polyroot(c(1,mod4$model$theta)))<1])
# Stationary and invertible


## Question c

# Stability

# We look at the same serie without the last 12 observations.
ultim <- c(2013,12)
pdq <- c(6,1,0)
PDQ <- c(1,1,2)

ipi2 <- window(ipi,end=ultim)
logipi2 <- log(ipi2)

# Model 3 and 4 for the truncated serie.
mod3bis2 <- arima(logipi2,order=pdq,seasonal=list(order=PDQ,period=12),
                  fixed=c(NA,NA,0,0,NA,NA,NA,0,NA))
mod4bis2 <- arima(logipi2,order=c(6,1,0),seasonal=list(order=c(3,1,5),period=12))

coef3 <- data.frame(mod3bis$coef,mod3bis2$coef)
colnames(coef3) <- c("Complete model","Truncate model")
print(xtable(coef3,align=c("c","c","c"),digits=4,
             caption="Comparison of the coefficients for model 3"))
# The model 3 is stable.

mod4bis$coef
mod4bis2$coef
# The model 4 is stable

# Prediction for model 3.

pred3 <- predict(mod3bis2,n.ahead=12)
pr3 <- ts(c(tail(logipi2,1),pred3$pred),start=ultim,freq=12)
se3 <- ts(c(0,pred$se),start=ultim,freq=12)

tl3 <- ts(exp(pr3-1.96*se3),start=ultim,freq=12)
tu3 <- ts(exp(pr3+1.96*se3),start=ultim,freq=12)
pr3 <- ts(exp(pr3),start=ultim,freq=12)

tspredggplot(ipi,pred=pr3,upperb=tu3,lowerb=tl3,
             title="Predictions for model 3.")

# Prediction for model 4.

pred4 <- predict(mod4bis2,n.ahead=12)
pr4 <- ts(c(tail(logipi2,1),pred4$pred),start=ultim,freq=12)
se4 <- ts(c(0,pred4$se),start=ultim,freq=12)

tl4 <- ts(exp(pr4-1.96*se4),start=ultim,freq=12)
tu4 <- ts(exp(pr4+1.96*se4),start=ultim,freq=12)
```

```
pr4 <- ts(exp(pr4),start=ultim,freq=12)

tspredggplot(ipi,pred=pr4,upperb=tu4,lowerb=tl4,
             title="Predictions for model 4.")

## Question d

obs <- window(ipi,start=ultim)

mod3bis2.EQM <- sqrt(sum(((obs-pr3)/obs)^2)/12)
mod3bis2.EAM <- sum(abs(obs-pr3)/obs)/12

mod4bis2.EQM <- sqrt(sum(((obs-pr4)/obs)^2)/12)
mod4bis2.EAM <- sum(abs(obs-pr4)/obs)/12

# We check if the first model is better than the second one.
mod3bis2.EQM-mod4bis2.EQM<0
mod3bis2.EAM-mod4bis2.EAM<0

# Model 3 is better based on these two indicators.


# Predictions ------------------------------------------------------------

# We know work with the entire serie and only the model 3.

ipi1 <- window(ipi,end=ultim+c(1,0))
logipi1 <- log(ipi1)

pred <- predict(mod3bis,n.ahead=12)
pr <- ts(c(tail(logipi1,1),pred$pred),start=ultim+c(1,0),freq=12)
se <- ts(c(0,pred$se),start=ultim+c(1,0),freq=12)

tl1<-ts(exp(pr-1.96*se),start=ultim+c(1,0),freq=12)
tu1<-ts(exp(pr+1.96*se),start=ultim+c(1,0),freq=12)
pr1<-ts(exp(pr),start=ultim+c(1,0),freq=12)

tspredggplot(ipi1,pred=pr1,upperb=tu1,lowerb=tl1,
             title="Predictions for model 3.")


# Outliers treatment -----------------------------------------------------

## Question a

mod.atip3 <- outdetec(mod3bis,dif=c(1,12),crit=2.6,LS=T)

atipics3 <- mod.atip3$atip[order(mod.atip3$atip[,1]),]
meses <- c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct",
           "Nov","Dec")

atipics <- data.frame(atipics3,Fecha=paste(meses[(atipics3[,1]-1)%%12+1],
                      start(logipi)[1]+((atipics3[,1]-1)%/%12)),
                      perc.Obs=exp(atipics3[,3])*100)
atipics <- data.frame(atipics[,1],atipics[,2],atipics[,5],atipics[,6])
colnames(atipics) <- c("Observations","Kind","Date","Evolution")
print(xtable(atipics,align=c("c","c","c","c","c"),digits=2,
             caption="Summary of the outliers."))

# Linearized serie vs serie
logipi.lin <- lineal(logipi,mod.atip3$atip)
ipi.lin <- exp(logipi.lin)
ts.data.frame <- data.frame(date=as.Date(as.yearmon(time(ipi))),
                            as.matrix(ipi))
colnames(ts.data.frame) <- c("time","value")
tsl.data.frame <- data.frame(date=as.Date(as.yearmon(time(ipi.lin))),
                             as.matrix(ipi.lin))
colnames(tsl.data.frame) <- c("time","valuel")
ggplot(data=ts.data.frame, mapping=aes(x=time, y=value)) + geom_line() +
  geom_line(data=tsl.data.frame,aes(x=time,y=valuel),colour="red") +
  theme(panel.grid.major.y=element_blank(),
                       panel.grid.minor.y=element_blank()) +
  ylab("Values of the serie")

# Effect of the outliers
tsggplot(logipi-logipi.lin)

## Question b

d12logipi.lin <- diff(logipi.lin,12)
d1d12logipi.lin <- diff(d12logipi.lin,1)
acfts(d1d12logipi.lin)

mod3bis.lin <- arima(logipi,order=c(6,1,0),seasonal=list(order=c(1,1,2),
                     period=12),fixed=c(NA,NA,0,0,NA,NA,NA,0,NA))

# Validation of the model
resid <- residplot(mod3bis.lin)
scatter <- scatterggplot(mod3bis.lin)
grid.arrange(resid,scatter,ncol=2)

qqggplot(mod3bis.lin)

hist(mod3bis.lin$residuals,breaks=20,freq=F)
curve(dnorm(x,mean=0,sd=sd(mod3bis.lin$residuals)),col=2,add=T)

acfts(mod3bis.lin$residuals,"Residuals")

acfts(mod3bis.lin$residuals,"Residuals²")

ljungggplot(mod3bis.lin)


## truncated lineal serie

ultim <- c(2013,12)
```

```r
pdq <- c(6,1,0)
PDQ <- c(1,1,2)

ipi2.lin <- window(ipi.lin,end=ultim)
logipi2.lin <- log(ipi2.lin)

# Model lineal for the truncated serie.
mod3bis2.lin <- arima(logipi2.lin,order=pdq,seasonal=list(order=PDQ,
                      period=12),fixed=c(NA,NA,0,0,NA,NA,NA,0,NA))

pred <- predict(mod3bis2.lin,n.ahead=12)
pr <- ts(c(tail(logipi2.lin,1),pred$pred),start=ultim,freq=12)
se <- ts(c(0,pred$se),start=ultim,freq=12)

tl1<-ts(exp(pr-1.96*se),start=ultim,freq=12)
tu1<-ts(exp(pr+1.96*se),start=ultim,freq=12)
pr1<-ts(exp(pr),start=ultim,freq=12)

tspredggplot(ipi1,pred=pr1,upperb=tu1,lowerb=tl1,
             title="Predictions for model lineal.")
tspredggplot(ipi1,pred=pr1,upperb=tu1,lowerb=tl1,
             title="Predictions for model 3.")

# Comparison of the predictions
obs <- window(ipi,start=ultim)

mod3bis2.EQM <- sqrt(sum(((obs-pr3)/obs)^2)/12)
mod3bis2.EAM <- sum(abs(obs-pr3)/obs)/12

mod3bis2.lin.EQM <- sqrt(sum(((obs-pr1)/obs)^2)/12)
mod3bis2.lin.EAM <- sum(abs(obs-pr1)/obs)/12

mod3bis2.EQM-mod3bis2.lin.EQM<0
mod3bis2.EAM-mod3bis2.lin.EAM<0

# Long term predictions

ipi1.lin <- window(ipi.lin,end=ultim+c(1,0))
logipi1.lin <- log(ipi1.lin)

pred.lin <- predict(mod3bis.lin,n.ahead=12)
pr.lin <- ts(c(tail(logipi1.lin,1),pred.lin$pred),start=ultim+c(1,0),freq=12)
se.lin <- ts(c(0,pred.lin$se),start=ultim+c(1,0),freq=12)

tl1.lin<-ts(exp(pr.lin-1.96*se.lin),start=ultim+c(1,0),freq=12)
tu1.lin<-ts(exp(pr.lin+1.96*se.lin),start=ultim+c(1,0),freq=12)
pr1.lin<-ts(exp(pr.lin),start=ultim+c(1,0),freq=12)

tspredggplot(ipi1.lin,pred=pr1.lin,upperb=tu1.lin,lowerb=tl1.lin,
             title="Predictions for model lineal.")
```

Functions ggplot

```r
# List of plot functions for object of class time series.

# Plot of the time serie --------------------------------------------------

tsggplot <- function(ts,title=NULL){ # plot the time serie.
  # ts must be a monthly times serie. A title can be add, must be character.
  ts.data.frame <- data.frame(date=as.Date(as.yearmon(time(ts))),as.matrix(ts))
  colnames(ts.data.frame) <- c("time","value")
  ggplot(data=ts.data.frame, mapping=aes(x=time, y=value))+geom_line() +
    ggtitle(title) + theme(panel.grid.major.y=element_blank(),
    panel.grid.minor.y=element_blank()) + ylab("Values of the serie")
}


# Plot of the ACF and PACF of a time serie --------------------------------

acfts<-function(ts,title=NULL){ # acf and pacf for a ts object.
  ts.data.frame <- data.frame(c(1:length(ts)),ts)
  colnames(ts.data.frame) <- c("time","value")
  ts.acf<-acf(ts, plot=FALSE,lag.max=71)
  ts.pacf<-pacf(ts, plot=FALSE,lag.max=72)
  ci <- 0.95
  clim0 <- qnorm((1 + ci)/2)/sqrt(ts.acf$n.used)
  clim <- c(-clim0,clim0)
  hline.data <- data.frame(z=c(0,clim),type=c("base","ci","ci"))
  acfPlot <- ggplot(data.frame(lag=c(0:71),acf=ts.acf$acf)) +
    geom_hline(aes(yintercept=z,colour=type,linetype=type),hline.data) +
    geom_linerange(aes(x=lag,ymin=0,ymax=acf),
                   colour=c(rep(c("red",rep("black",11)),6))) +
    scale_colour_manual(values = c("black","blue")) +
    scale_linetype_manual(values =c("solid","dashed")) + ylab("") +
    ggtitle("ACF") + scale_y_continuous(limits=c(-1, 1)) +
    geom_segment(aes(x = 0, y = 0, xend = 0, yend = 1),colour="red")
  pacfPlot <- ggplot(data.frame(lag=c(1:72),pacf=ts.pacf$acf)) +
    geom_hline(aes(yintercept=z,colour=type,linetype=type),hline.data) +
    geom_linerange(aes(x=lag,ymin=0,ymax=pacf),
                   colour=c(rep(c(rep("black",11),"red"),6))) +
    scale_colour_manual(values = c("black","blue")) +
    scale_linetype_manual(values =c("solid","dashed")) +
    ggtitle("Partial ACF") + scale_y_continuous(limits=c(-1, 1))
  grid.arrange(acfPlot,pacfPlot,ncol=2,main=title)
}


# Plot of the ACF and PACF of a ARMAacf -----------------------------------

acfmodel <- function(model){ # acf and pacf for an ARiMA model.
  modelacf <- as.data.frame(ARMAacf(model$model$phi,model$model$theta,
                                    lag.max=36))
  modelacf <- cbind(seq(0,length(modelacf[[1]])-1,by=1),modelacf)
  colnames(modelacf) <- c("lag","acf")
  acfPlot <- ggplot(modelacf,aes(lag,acf)) +
    geom_segment(aes(x=lag,y=0,xend=lag,yend=acf),
                 colour=c(rep(c("red",rep("black",11)),3),"red")) +
    geom_hline(y=0) + scale_y_continuous(limits=c(-1, 1)) +
    ggtitle("ACF theoric") + ylab("")
  modelpacf <- as.data.frame(ARMAacf(model$model$phi,model$model$theta,
                                     lag.max=37,pacf=TRUE))
  modelpacf <- cbind(seq(1,length(modelpacf[[1]]-1),by=1),modelpacf)
  colnames(modelpacf) <- c("lag","pacf")
  pacfPlot <- ggplot(modelpacf,aes(lag,pacf)) +
    geom_segment(aes(x=lag,y=0,xend=lag,yend=pacf),
                 colour=c(rep(c(rep("black",11),"red"),3),"black")) +
    geom_hline(y=0) + scale_y_continuous(limits=c(-1, 1)) +
    ggtitle("PACF theoric") + ylab("")
  grid.arrange(acfPlot,pacfPlot,ncol=2)
}


# Plot of the residuals ---------------------------------------------------

# Regular plot
residplot <- function(model){
  ci <- 3*sd(model$residuals)
  clim <- c(-ci,ci)
  hline.data <- data.frame(z=clim,type=c("ci","ci"))
  tsggplot(model$residuals) + geom_hline(aes(yintercept=z,linetype=type,
    colour=type),hline.data) + geom_hline(y=0) + ggtitle("Residuals")
}

# QQ-plot
qqggplot <- function(model){
  ts.data.frame <- data.frame(c(1:length(model$residuals)),model$residuals)
  colnames(ts.data.frame) <- c("time","residuals")
  qtype <- 7
  y <- quantile(model$residuals[!is.na(model$residuals)], c(0.25, 0.75))
  x <- qnorm(c(0.25, 0.75))
  slope <- diff(y)/diff(x)
  intercept <- y[1L] - slope * x[1L]
  ggplot(ts.data.frame, aes(sample = residuals)) + geom_point(stat = "qq") +
    geom_abline(intercept=intercept,slope=slope,color="red") +
    ggtitle("QQ-plot")
}

# Square root of the absolute residuals
scatterggplot <- function(model){
  resid <- model$residuals
  x <- sqrt(abs(resid))
  y <- NULL
  lpars=list(col=2)
  span = 2/3
  family = c("symmetric","gaussian")
  evaluation = 50
  degree = 1
  xlabel <- if (!missing(x)){
    deparse(substitute(x))
  }
  ylabel <- if (!missing(y)){
    deparse(substitute(y))
  }
  xy <- xy.coords(x, y, xlabel, ylabel)
  x <- xy$x
  y <- xy$y
  xlab <- xy$xlab
  ylab <- if (is.null(ylab)){
    xy$ylab
  }
  x1 <- data.frame(x=x,y=y)
  pred <- loess.smooth(x,y, span=span, degree=degree, family=family,
                       evalution=evaluation)
  ylim = range(y, pred$y,na.rm = TRUE)
  linh <- as.data.frame(c(list(pred),lpars))
  colnames(linh) <- c("pred","lpars","col")
  ggplot() + geom_point(data=x1, aes(x=x, y=y)) +
    geom_line(data=linh,aes(x=pred,y=lpars),colour="red") +
    ylab(expression(sqrt(abs(residuals)))) + xlab("") +
    ggtitle("Scatter-plot with smooth")
}

# Ljung-Box test plot
ljungggplot <- function(model){
  gof.lag <- 7*frequency(get(model$series))
  rs <- model$residuals
  nlag <- gof.lag
  pval <- numeric(nlag)
  for (i in 1L:nlag){
    pval[i] <- Box.test(rs, i, type = "Ljung-Box")$p.value
  }
  df <- data.frame(c(1:nlag),pval)
  test <- factor(df$pval<0.05)
  df <- cbind(df,test)
  colnames(df) <- c("lag","pval","test")
  ggplot(data=df,aes(x=lag, y=pval)) + geom_point(aes(colour=factor(test))) +
    geom_hline(y=0.05,linetype="dashed",colour="blue") + geom_hline(y=0) +
    ggtitle("p values for Ljung-Box statistic") + ylab("") + xlab("") +
    coord_cartesian(ylim=c(-0.05, 1.05)) +
    scale_colour_manual(values=c("TRUE"="red","FALSE"="blue"),
                        name="Value of the p-value",
                        breaks=c("TRUE","FALSE"),labels=c("<0.05",">0.05"))
}


# Plot of the arma roots in a unite circle --------------------------------

plotarmaroots <- function(object){
  if(class(object) != "Arima")
    stop("object must be of class Arima or ar")
```

```
if(class(object) == "Arima"){
  parvecphi <- object$model$phi
}
if(length(parvecphi) > 0){
  last.nonzero <- max(which(abs(parvecphi) > 1e-08))
  if (last.nonzero > 0){
    arroots <- structure(list(roots=polyroot(c(1,-parvecphi[1:last.nonzero])),
                              type="AR"), class='armaroots')
  }
} else{
  arroots <- structure(list(roots=numeric(0),type="AR"),class='armaroots')
}
parvectheta <- object$model$theta
if(length(parvectheta) > 0){
  last.nonzero <- max(which(abs(parvectheta) > 1e-08))
  if (last.nonzero > 0){
    maroots <- structure(list(roots=polyroot(c(1,parvectheta[1:last.nonzero])),
                              type="MA"), class='armaroots')
  }
} else{
  maroots <- structure(list(roots=numeric(0),type="MA"),class='armaroots')
}
rootsma <- as.data.frame(1/maroots$roots)
testma <- factor(Mod(rootsma[[1]])>1)
rootsma <- cbind(rootsma,testma)
colnames(rootsma) <- c("invmaroots","test")
rootsar <- as.data.frame(1/arroots$roots)
testar <- factor(Mod(rootsar[[1]])>1)
rootsar <- cbind(rootsar,testar)
colnames(rootsar) <- c("invarroots","test")
xc <- 0
yc <- 0
r <- 1
arplot <- ggplot(data=rootsar,aes(x=Re(invarroots), y=Im(invarroots))) +
  geom_point(aes(colour=factor(testar))) +
  ggtitle(paste("The",length(arroots$roots),"inverse",arroots$type," roots")) +
  xlab("Real") + ylab("Imaginary") + annotate("path",x=xc+r*cos(seq(0,2*pi,
  length.out=100)),y=yc+r*sin(seq(0,2*pi,length.out=100))) + coord_fixed() +
  scale_colour_manual(values=c("TRUE"="red","FALSE"="blue"),name="1/Roots",
                      breaks=c("TRUE","FALSE"),labels=c(">1","<1")) +
  geom_hline(y=0,linetype="dashed") + geom_vline(x=0,linetype="dashed")

maplot <- ggplot(data=rootsma,aes(x=Re(invmaroots), y=Im(invmaroots))) +
  geom_point(aes(colour=factor(testma))) +
  ggtitle(paste("The",length(maroots$roots),"inverse",maroots$type," roots")) +
  xlab("Real") + ylab("Imaginary") + annotate("path",x=xc+r*cos(seq(0,2*pi,
  length.out=100)),y=yc+r*sin(seq(0,2*pi,length.out=100))) + coord_fixed() +
  scale_colour_manual(values=c("TRUE"="red","FALSE"="blue"),name="1/Roots",
                      breaks=c("TRUE","FALSE"),labels=c(">1","<1")) +
  geom_hline(y=0,linetype="dashed") + geom_vline(x=0,linetype="dashed")

grid.arrange(arplot,maplot,ncol=2,
             main="Inverse roots of the charasteristic polynomial")
```

```
}


# Plot of the predictions with the confidence interval --------------------

tspredggplot <- function(ts,pred,upperb,lowerb,title=NULL){ # plot the time serie.
  # ts must be a monthly times serie. A title can be add, must be character.
  ts.data.frame <- data.frame(date=as.Date(as.yearmon(time(ipi))),as.matrix(ipi))
  colnames(ts.data.frame) <- c("time","value")
  pred.data.frame <- data.frame(date=as.Date(as.yearmon(time(pred))),
                                as.matrix(pred))
  colnames(pred.data.frame) <- c("time","pred")
  upperb.data.frame <- data.frame(date=as.Date(as.yearmon(time(upperb))),
                                  as.matrix(upperb))
  colnames(upperb.data.frame) <- c("time","upperb")
  lowerb.data.frame <- data.frame(date=as.Date(as.yearmon(time(lowerb))),
                                  as.matrix(lowerb))
  colnames(lowerb.data.frame) <- c("time","lowerb")
  ts.data.frame <- tail(ts.data.frame,48)
  pred.data.frame <- tail(pred.data.frame,48)
  upperb.data.frame <- tail(upperb.data.frame,48)
  lowerb.data.frame <- tail(lowerb.data.frame,48)
  ggplot(data=ts.data.frame, mapping=aes(x=time, y=value))+geom_line() +
    geom_line(data=pred.data.frame, mapping=aes(x=time, y=pred),colour="red") +
    geom_line(data=upperb.data.frame, mapping=aes(x=time, y=upperb),colour="blue",
              linetype="dashed") +
    geom_line(data=lowerb.data.frame, mapping=aes(x=time, y=lowerb),colour="blue",
              linetype="dashed") +
    ggtitle(title) + theme(panel.grid.major.y=element_blank(),
                           panel.grid.minor.y=element_blank())
}
```

Outliers detection functions

```
# Outliers detection function ------------------------------------------------------

outdetec<-function(object,dif=c(0,0),crit,LS=T)
{
residuals<-object$residuals
m<-length(residuals)

piweight<--ARMAtoMA(ar=-object$model$theta, ma=-object$model$phi, lag.max=m+sum(dif))

if (dif[1]!=0) for(i in 1:dif[1]) piweight<-c(piweight,0)-c(-1,piweight)
if (length(dif)>1){
for (i in 2:length(dif)){
if (dif[i]>1) piweight<-c(piweight,rep(0,dif[i]))-c(rep(0,dif[i]-1),-1,piweight)
}
}
piweight<-piweight[1:m]

atip<-NULL
```

```
num<-NULL
type<-NULL
wcoeff<-NULL
LCrit<-NULL
if (crit<=0) {cat("The Critical value may be positive") }

va<-mean(residuals^2)

c<-cumsum(piweight)-1

d<-rep(0,m)
delta<-0.7
d[1]<-piweight[1]-delta
for (i in 2:m) d[i]<-delta*d[i-1]+piweight[i]

sum1<-1+sum(piweight*piweight)
sum2<-1+sum(c*c)
sum3<-1+sum(d*d)

maxL<-crit+1

while (maxL>crit)
{   ka1<-sum1
    ks1<-sum2
    kt1<-sum3
      maxL<-0
    for (i in 1:m)
    {

        suma1<-sum(residuals[i:m]*c(1,-piweight[1:(m-i)]))
        suma2<-sum(residuals[i:m]*c(1,-c[1:(m-i)]))
        suma3<-sum(residuals[i:m]*c(1,-d[1:(m-i)]))

        ka1<-ka1 - piweight[m-i+1]*piweight[m-i+1]
        w_ao<-suma1/ka1
        v_ao<-va/ka1
        l_ao<-w_ao/sqrt(v_ao)

        ks1<-ks1 - c[m-i+1]*c[m-i+1]
        w_ls<-suma2/ks1
        v_ls<-va/ks1
        l_ls<-w_ls/sqrt(v_ls)

        kt1<-kt1 - d[m-i+1]*d[m-i+1]
        w_tc<-suma3/kt1
        v_tc<-va/kt1
        l_tc<-w_tc/sqrt(v_tc)

        if(abs(l_ao)>maxL)
        {   maxL<-abs(l_ao)
            t<-i
            w<-w_ao
            v<-v_ao
            ts<-"AO"
        }
        if(abs(l_ls)>maxL & LS==T & i!=m)
        {   maxL<-abs(l_ls)
            t<-i
            w<-w_ls
            v<-v_ls
            ts<-"LS"
        }
        if(abs(l_tc)>maxL & i!=m)
        {   maxL<-abs(l_tc)
            t<-i
            w<-w_tc
            v<-v_tc
            ts<-"TC"
        }
    }


    if(maxL > crit){
        if(ts=="AO") residuals[t:m]<-residuals[t:m]+w*c(-1,piweight[1:(m-t)])
        if(ts=="LS") residuals[t:m]<-residuals[t:m]+w*c(-1,c[1:(m-t)])
        if(ts=="TC") residuals[t:m]<-residuals[t:m]+w*c(-1,d[1:(m-t)])

        val<-mean(residuals^2)
        l<-w/sqrt(v*val/va)
        va<-val

        num<-c(num,t)
        type<-c(type,ts)
        wcoeff<-c(wcoeff,w)
        LCrit<-c(LCrit,abs(l))

        atip<-data.frame(Obs=num,type_detected=type,W_coeff=wcoeff,ABS_L_Ratio=LCrit)

    }
}
return(list(atip=atip,sigma2=va,resid=residuals))
}


# Linearization function --------------------------------------------------

lineal<-function(serie,atip)
{

m<-length(serie)
for(i in 1:nrow(atip))
{
    t<-atip[i,1]
    ts<-atip[i,2]
    w<-atip[i,3]
    if(ts=="TC")    serie[t:m]<-serie[t:m]-w*c(1,0.7^(1:(m-t)))
    if(ts=="LS")    serie[t:m]<-serie[t:m]-w
    if(ts=="AO")    serie[t]<-serie[t]-w
}

return(serie)

}
```