ENSAE
ParisTech

École nationale
de la statistique
et de l'administration
économique

université
PARIS-SACLAY

ENSAE ParisTech

Modeling and Forecasting of the production of distilled alcoholic beverages in France

Time Series Project

Engineering curriculum
2nd Year

Johan Macq
Pierre Delanoue

Mai 2019

Preamble

For reasons of space, we decide to include all our graphs and tables of statistical test results in the appendix. When we comment on a graph or result, we are referring specifically to a figure in Section A of the Appendix.

1 Data

1.1 Question 1

The series chosen represents the evolution of the index of industrial production of distilled alcoholic beverages from January 1990 to January 2019, on a base 100 in 2015. The given index is raw and therefore does not include any transformation (logarithmic or other). These data are available on the INSEE website. In the rest of the project, we will note this series $(X_t)_{t \in T}$ where $T =$[1990-01,1990-02,...,2018-12,2019-01].

The figure 1 is the curve that represents the evolution of our data. It should be noted that the series does not seem to show any particular trend. On the other hand, several indices for annual seasonality appear. The objective of the next question is therefore to find a transformation that would make this series stationary.

1.2 Question 2

We want to justify the existence of a 12-month seasonality. To do this, we look at the autocorrelogram of our series which displays the evolution of the empirical autocorrelation function $\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}$ where $\widehat{\gamma}(.)$ is the empirical autocovariance function such that for $|h| < n$

$$\widehat{\gamma}(h) = \frac{1}{n} \sum_{t=|h|+1}^{n} (X_t - \overline{X}_n)(X_{t-|h|} - \overline{X}_n)$$

with $\overline{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t$.

When we analyze the figure 2, which represents the autocorrelogram of the series, we confirm our intuition of a seasonality of 12 months with peaks present for each multiple delay of 12.
The first step to make the series stationary is therefore to differentiate according to seasonality. So now we're going to focus on the series $(Z_t)_{t \in T} = (X_t - X_{t-12})_{t \in T}$.

To check if the series $(Z_t)_{t \in T}$ is stationary, we decide to perform an augmented Dickey-Fuller test, a Phillips-Perron test and then a KPSS test. We briefly recall the principle of these tests in the appendix and report our results in the table A.2.
The augmented Dickey-Fuller test leads us here to reject the null hypothesis of non-stationarity at the 1 percent level for $(Z_t)_{t \in T}$. The Phillips-Perron test also leads us to reject the null hypothesis of non-stationarity with a p-value of less than 1 percent. However, the two p-values of the KPSS tests are equal to 4%, which leads us to reject the null hypothesis of stationarity at the 5 percent level.
We therefore propose to add an additional differentiation to order 1. $(Y_t)_{t \in T} = (Z_t - Z_{t-1})_{t \in T}$ is then defined. Similarly, the 4 previous tests on the series $(Y_t)_{t \in T}$ are conducted, the results of which are recorded in the table A.2.
In the same way as for $(Z_t)_{t \in T}$, we reject the null hypotheses of non-stationarity of the augmented Dickey-Fuller and Phillips-Perron tests at the 1 percent level for $(Y_t)_{t \in T}$. However, this time we do not reject the zero hypotheses of stationarity of the KPSS tests because their p-values are higher than 10 percent. We can still see that the values of the respective statistics have improved significantly in favour of a stationary series.
Consequently, we will be able to continue our study with the series $(Y_t)_{t \in T}$ which is stationary.

1.3 Question 3

We represent $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ in the figure 6 in the appendix.

2 ARMA Models

2.1 Question 4

We will use a model of type SARIMA$_s$[(p,d,q),(P,D,Q)] for $(X_t)_{t\in T}$ (raw series), which seems adapted to the seasonal nature of our series. In addition, we decide to build our models over the period T = [1990-01,1990-02,...,2017-12,2018-01] in order to predict the last year of the production index and test our performance.

To do this, we will follow the Box-Jenkins procedure, which consists of four steps:

- Step 1 : Search d and D such that Y_t = (I-B)^d(I-B^s)$^D X_t$ is apparently stationary with B the delay operator. The previous study determined that s = 12, d = 1 and D = 1.
- Step 2: Examine autocorrelations and empirical partial autocorrelations from (Y_t) to multiple delays of s to identify P and Q.
- Step 3: Choose p and q to adjust the first order autocorrelations up to s-1 to an ARMA(p,q)
- Step 4: Estimate the model parameters by OLS. Many of them can be null.

Identification of P, Q, p and q:
By analyzing the autocorrelogram of (Y_t) in the appended figure 4, we quickly identify $Q = 1$ and $q = 1$. Indeed, we have significant autocorrelations at lags $12 = 1 \times 12$ and 1. Similarly, the partial autocorrelogram of (Y_t) in figure 5 leads to the conclusion that $P = 4$ and $p = 8$. Here, we have significant autocorrelations at lags $48 = 4 \times 12$ and 8.

Model validation:
We will now determine which models are valid among those possible. Two main conditions will be tested for this:

- Ljung-Box Test (coat rack) for delays of up to 36 : if a residue autocorrelation is different from 0 to 5%, then the model is rejected.
- Test of coefficient significance: if one of the coefficients associated with the highest orders is non-significant, the model is rejected. We test the null ($\beta = 0$) against the alternative ($\beta \neq 0$). Consequently, a coefficient β is said to be non-significant when the p-value given by $2 \times (1 - F(|\frac{\hat{\beta}}{se_\beta}|))$ is greater than 5%, with F the cumulative distribution function of a $\mathcal{N}(0, 1)$.

These conditions will be applied for all models such as P $\leq$ 4, Q $\leq$ 1, p $\leq$ 8 and q $\leq$ 1.
We now go from $Card(\llbracket 0:4 \rrbracket \times \llbracket 0:1 \rrbracket \times \llbracket 0:8 \rrbracket \times \llbracket 0:1 \rrbracket) = 180$ possible models to 6 models described by the orders in the table above.

Table 1: Valid models remaining after application of Ljung-Box and significance tests.

	P	D	Q	p	d	q
Model 1	0	1	1	8	1	0
Model 2	0	1	1	8	1	1
Model 3	2	1	0	8	1	0
Model 4	3	1	0	8	1	0
Model 5	4	1	0	8	1	0
Model 6	4	1	0	8	1	1

To select which of these models is the best, we calculate for each one the associated AIC and BIC information criteria. It appears that model 6 (SARIMA$_{12}$[(8,1,1),(4,1,0)]) minimizes the first criterion with an AIC = 5.64 and model 2 (SARIMA$_{12}$[(8,1,1),(0,1,1)]) minimizes the second criterion with a BIC = 4.76.
We therefore have two optimal models and need to choose one of them.

Having previously removed the last year of production to build the models, we will decide between the two finalists by comparing their forecast performance over the period 02/2018 to 11/2018 (the first ten months of the last year).The prediction of the last two months will be studied in Part 3 by the final model, for this reason, we do not take them into account here.

To do this, we will use the *sarima.for* function in R. Under the assumption of normal residuals, we can give the confidence interval to 95% of the forecasts. Here, the forecast error is directly calculated by the *sarima.for* function. The results are shown in the figures 7 and 8 in the appendix. The reader will note that the details of these calculations are not explained here because we will do so in Part 3.

The first model gives an RMSE of 7.81, while the second model gives an RMSE of 8.29. The model used is therefore SARIMA$_{12}$[(8,1,1,1),(4,1,0)] which minimizes AIC.

We know that for all (s, p, d, q, P, D, Q) such as $\tilde{X}_t$ is a SARIMA$_s$[(p,d,q),(P,D,Q)] with $\tilde{Y}_t = (I-B)^d(I-B^s)^D X_t$, we have the series $\tilde{Y}_t$ which is an ARMA(p+sP,q+sQ).

The deseasonalized and differentiated series $(Y_t)_{t\in T}$ is therefore an ARMA(56,1).

Limits of the model:
We have an error that is on average between 5 and 10%, this can be explained by the peak production present in October 2018. Indeed, this peak is the highest of all the years in the period considered and it leaves confidence bands at 95 percent, which misleads the model. Without the latter, we would certainly have had a better average forecast.
In addition, forecasts are made under the assumption of normal residuals. As can be seen in Figure 9 and in the Jarque-Bera test in question 6, the residuals are not normal, which is another limitation of the model.

We are aware that the choice of an ARMA(56,1) is less elegant than that of an ARMA(8,13) (the model that minimized the BIC). We would probably have had another result if we had decided between these models according to other criteria (Schwarz, Hannan-Quin,...). But since the goal here will be to have the best possible prediction (Part 3) and as we have enough data to use 56 months of data in the past, we are satisfied with the ARMA model (56,1).

3 Forecasts

3.1 Question 5

Let us return to the general case of ARMA prediction. Let be $\tilde{X}_t$ a causal and invertible ARMA(p,q) such that

$$\tilde{X}_t - \sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{t-i} = \tilde{\epsilon}_t - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{t-i}$$

First, we place ourselves in the unrealistic situation where we would have access to the infinite past of the process $\tilde{X}_t$ from the time T. In other words, we have access to $\tilde{X}_T, \tilde{X}_{T-1}, ..., \tilde{X}_1, \tilde{X}_0, \tilde{X}_{-1}$, etc. The optimal prediction of future values is therefore the conditional expectation in relation to the infinite past.
Therefore, the optimal prediction for $\tilde{X}_{T+1}$ would be:

$$\begin{aligned}
\mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...] &= \mathbb{E}[\sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{T+1-i} + \tilde{\epsilon}_{T+1} - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+1-i}|\tilde{X}_T, ..., \tilde{X}_1, ...] \\
&= \sum_{i=1}^{p} \tilde{\phi}_i \mathbb{E}[\tilde{X}_{T+1-i}|\tilde{X}_T, ..., \tilde{X}_1, ...] + \mathbb{E}[\tilde{\epsilon}_{T+1}] - \sum_{i=1}^{q} \tilde{\psi}_i \mathbb{E}[\tilde{\epsilon}_{T+1-i}|\tilde{X}_T, ..., \tilde{X}_1, ...] \\
&= \sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{T+1-i} - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+1-i}
\end{aligned}$$

Because $Vect(\tilde{X}_T, ..., \tilde{X}_1, ...) = Vect(\tilde{\epsilon}_T, ..., \tilde{\epsilon}_1, ...)$ and $\forall i > T,\ \epsilon_i \perp Vect(\tilde{X}_T, ..., \tilde{X}_1, ...)$

Similarly for X_{T+2} we have:

$$\begin{aligned}
\mathbb{E}[\tilde{X}_{T+2}|\tilde{X}_T, ..., \tilde{X}_1, ...] &= \mathbb{E}[\sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{T+2-i} + \tilde{\epsilon}_{T+2} - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+2-i}|\tilde{X}_T, ..., \tilde{X}_1, ...] \\
&= \tilde{\phi}_1 \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...] + \sum_{i=2}^{p} \tilde{\phi}_i \tilde{X}_{T+2-i} + \mathbb{E}[\tilde{\epsilon}_{T+2} - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+2-i}|\tilde{X}_T, ..., \tilde{X}_1, ...] \\
&= \tilde{\phi}_1 \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...] + \sum_{i=2}^{p} \tilde{\phi}_i \tilde{X}_{T+2-i} + \mathbb{E}[\tilde{\epsilon}_{T+2} - \tilde{\psi}_1 \tilde{\epsilon}_{T+1}] - \sum_{i=2}^{q} \tilde{\psi}_i \mathbb{E}[\tilde{\epsilon}_{T+2-i}|\tilde{X}_T, ..., \tilde{X}_1, ...] \\
&= \tilde{\phi}_1 \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...] + \sum_{i=2}^{p} \tilde{\phi}_i \tilde{X}_{T+2-i} - \sum_{i=2}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+2-i}
\end{aligned}$$

By noting $\tilde{X}_{T+1}^{|T} = \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...]$ and $\tilde{X}_{T+2}^{|T} = \mathbb{E}[\tilde{X}_{T+2}|\tilde{X}_T, ..., \tilde{X}_1, ...]$ we have the following equation system:

$$\begin{cases}
\tilde{X}_{T+1}^{|T} = \sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{T+1-i} - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+1-i} \\
\tilde{X}_{T+2}^{|T} = \tilde{\phi}_1 \tilde{X}_{T+1}^{|T} + \sum_{i=2}^{p} \tilde{\phi}_i \tilde{X}_{T+2-i} - \sum_{i=2}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+2-i}
\end{cases} \tag{1}$$

Which is equivalent to :

$$\begin{cases} \tilde{X}_{T+1}^{|T} = \tilde{X}_{T+1} - \tilde{\epsilon}_{T+1} \\ \tilde{X}_{T+2}^{|T} = \tilde{\phi}_1 \tilde{X}_{T+1}^{|T} + \tilde{X}_{T+2} - \tilde{\phi}_1 \tilde{X}_{T+1} - \tilde{\epsilon}_{T+2} + \tilde{\psi}_1 \tilde{\epsilon}_{T+1} \quad = \tilde{X}_{T+2} - \tilde{\epsilon}_{T+2} + \tilde{\epsilon}_{T+1}(\tilde{\psi}_1 - \tilde{\phi}_1) \end{cases} \tag{2}$$

Here we assume that the residues $\tilde{\epsilon}_t$ are Gaussian white noises (iid and centered) with a variance of σ^2.

So we have that the forecast errors are Gaussian :

$$\tilde{X}_{T+1} - \tilde{X}_{T+1}^{|T} \sim \mathcal{N}(0, \sigma^2) \;\; and \;\; \tilde{X}_{T+2} - \tilde{X}_{T+2}^{|T} \sim \mathcal{N}(0, \sigma^2(1 + (\tilde{\psi}_1 - \tilde{\phi}_1)^2)\ )$$

By noting $q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}$ the $1 - \frac{\alpha}{2}$ level quantile of a $\mathcal{N}(0,1)$, we have the following α level confidence intervals:

$$\tilde{X}_{T+1} \in \left[\tilde{X}_{T+1}^{|T} - q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sigma, \tilde{X}_{T+1}^{|T} + q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sigma\right]$$

and

$$\tilde{X}_{T+2} \in \left[\tilde{X}_{T+2}^{|T} - q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sigma \sqrt{1 + (\tilde{\psi}_1 - \tilde{\phi}_1)^2}, \tilde{X}_{T+2}^{|T} + q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sigma \sqrt{1 + (\tilde{\psi}_1 - \tilde{\phi}_1)^2}\right]$$

Since we don't have access to the infinite past but only to $X_T, ..., X_1$, we'll have to estimate $\tilde{X}_{T+1}^{|T}$ and $\tilde{X}_{T+2}^{|T}$.

For that we will put $\tilde{X}_0 = \tilde{X}_{-1} = ... = 0$.
Since our ARMA model is invertible, we have the following:

$$\forall t, \tilde{\epsilon}_t = \tilde{X}_t + \sum_{j=1}^{\infty} b_j \tilde{X}_{t-j} \quad o\grave{u} \quad \frac{\tilde{\Phi}(z)}{\tilde{\Psi}(z)} = \sum_{j=0}^{\infty} b_j z^j$$

We will therefore estimate $\tilde{\epsilon}_t$ by $\widehat{\tilde{\epsilon}}_t = \tilde{X}_t + \sum_{j=1}^{t-1} b_j \tilde{X}_{t-j}$ for all $t > 0$.

Thus we will estimate $\tilde{X}_{T+1}^{|T}$ and $\tilde{X}_{T+2}^{|T}$ respectively by $\widehat{\tilde{X}}_{T+1}^{|T}$ and $\widehat{\tilde{X}}_{T+2}^{|T}$ such that:

$$\begin{cases} \widehat{\tilde{X}}_{T+1}^{|T} = \sum_{i=1}^{min(p,T)} \tilde{\phi}_i \tilde{X}_{T+1-i} - \sum_{i=1}^{min(q,T)} \tilde{\psi}_i \widehat{\tilde{\epsilon}}_{T+1-i} \\ \widehat{\tilde{X}}_{T+2}^{|T} = \tilde{\phi}_1 \widehat{\tilde{X}}_{T+1}^{|T} + \sum_{i=2}^{min(p,T)} \tilde{\phi}_i \tilde{X}_{T+2-i} - \sum_{i=2}^{min(q,T)} \tilde{\psi}_i \widehat{\tilde{\epsilon}}_{T+2-i} \end{cases}$$

In our study of the Y_t series, we do not have access to the coefficients $(\phi_i)_{i \in [\![1:p]\!]}$ and $(\psi_i)_{i \in [\![1:q]\!]}$. Our predictions of future values will therefore use the formula explained above in the general case but by estimating the coefficients $(\phi_i)_{i \in [\![1:p]\!]}$ and $(\psi_i)_{i \in [\![1:q]\!]}$ using the Durbin-Levinson algorithm.

3.2 Question 6

Our confidence interval is based on the assumption that the residues follow a Gaussian centred law with a variance of σ^2. To test this hypothesis, we perform a Jarque-Bera test on the residues $\widehat{\epsilon}_t$. The details of this test are explained in the appendix.

The test gives us a statistic $JB = 96,4$ for a p-value of less than $2,2e{-}16$. The null hypothesis of normality is therefore rejected. So we must consider the confidence interval with caution because it is not quite accurate.

Empirical residues are linear combinations of the treated variable Y_t, which is itself centered because it is differentiated and seasonally deseasonalized. The residues are centered.

3.3 Question 7

The forecasts and the 95% confidence interval for the last two months are shown in Figure 10. We note that our prediction and the actual value of the index are within the confidence interval. In addition, the performance for the month of January 2019 is very satisfactory. On the other hand, as we have specified within the limits of the model, the confidence interval should be taken with caution because the assumption of normal residuals is not respected.

3.4 Question 8

The necessary condition is to have an instantly Granger causality.

Let be $(\tilde{X}_t)$ and $(\tilde{Y}_t)$ two series.
It is said that $(\tilde{Y}_t)$ instantly Granger-causes $(\tilde{X}_t)$ if $(\tilde{Y}_{T+1})$ is useful to predict $(\tilde{X}_{T+1})$ at the date T.

With our notations, this gives

$$\mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ..., \tilde{Y}_{T+1}, ..., \tilde{Y}_1, ...] \neq \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...]$$

We note that, because $Vect(\tilde{X}_T, ..., \tilde{X}_1, ...) \subset Vect(\tilde{X}_T, ..., \tilde{X}_1, ..., \tilde{Y}_{T+1}, ..., \tilde{Y}_1, ...)$, the above inequality implies that the new estimator is "better" than the other because it projects over a larger space.

Taking the theoretical situation from question 5 as a starting point, the information provided would come from the fact that we would have the following estimate:

$$\begin{aligned}\mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ..., \tilde{Y}_{T+1}, ..., \tilde{Y}_1, ...] &= \sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{T+1-i} - \sum_{i=1}^{q} \tilde{\psi}_i \tilde{\epsilon}_{T+1-i} + \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}] \\ &= \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ...] + \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]\end{aligned}$$

because $\mathbb{E}[\tilde{\epsilon}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ..., \tilde{Y}_{T+1}, ..., \tilde{Y}_1, ...] = \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]$ which represents our gain of information in comparison to the situation in question 5 (if not null).

In addition, the prediction error with this estimator would become :

$$\tilde{X}_{T+1} - \mathbb{E}[\tilde{X}_{T+1}|\tilde{X}_T, ..., \tilde{X}_1, ..., \tilde{Y}_{T+1}, ..., \tilde{Y}_1, ...] = \tilde{\epsilon}_{T+1} - \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]$$

However we have :

$$\begin{aligned}Var(\tilde{\epsilon}_{T+1} - \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]) &= \mathbb{E}[(\tilde{\epsilon}_{T+1} - \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}])^2] \\ &= \mathbb{E}[\tilde{\epsilon}_{T+1}^2 - \mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]^2] \\ &= Var(\tilde{\epsilon}_{T+1}) - \mathbb{E}[\mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]^2] \\ &\leq Var(\tilde{\epsilon}_{T+1})\end{aligned}$$

Therefore, the variance of the error of the new estimator is lower.

Thus, it would be useful to find a good form of estimator for $\mathbb{E}[\tilde{\epsilon}_{T+1}|Y_{T+1}]$ and test the dependency hypotheses to have a more efficient estimator.

A Figures and Tests Tables

A.1 Figures

Figure 1: Production Index of Distilled Alcoholic Beverages in Metropolitan France from January 1990 to January 2019 (Monthly)

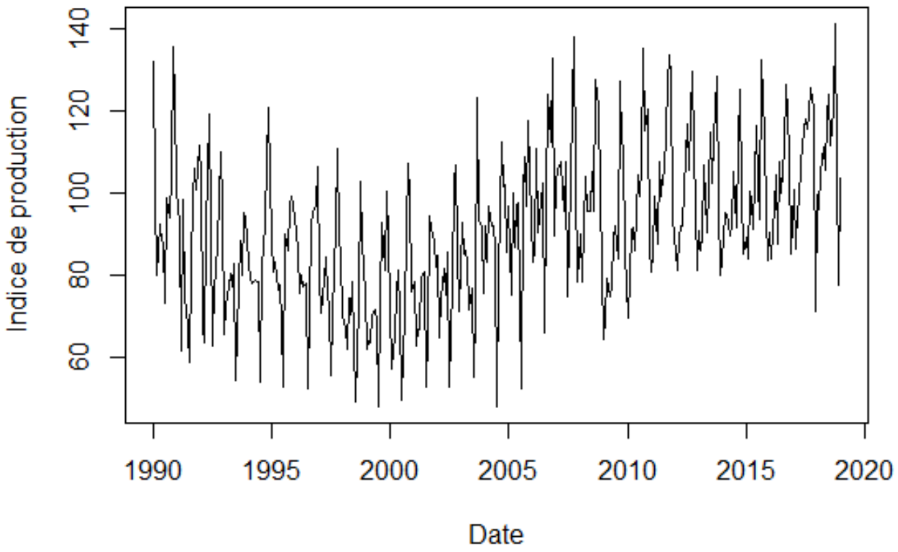


Figure 2: Autocorrelogram of the undifferentiated series ($(X_t)_{t\in T}$)

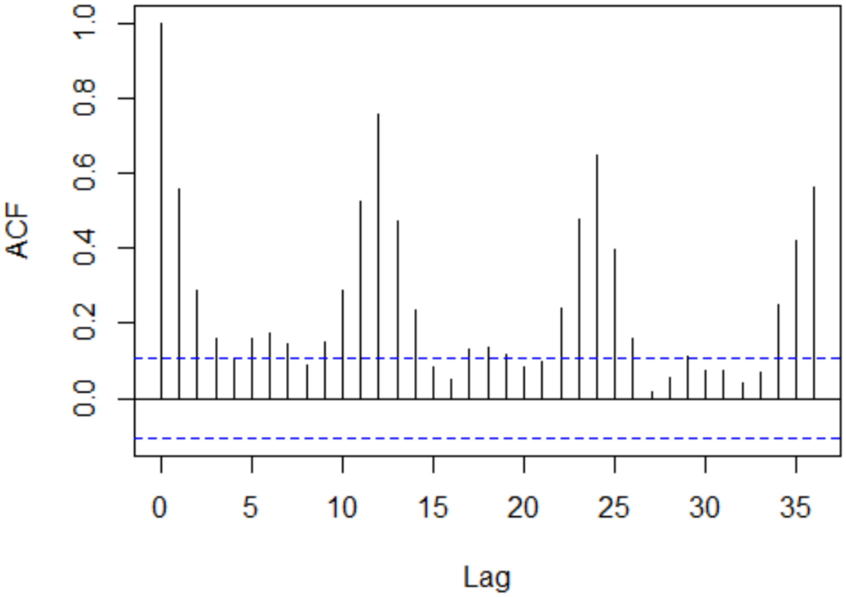


Figure 3: Autocorrelogram of the differentiated series at 12 ($(Z_t)_{t\in T}$)

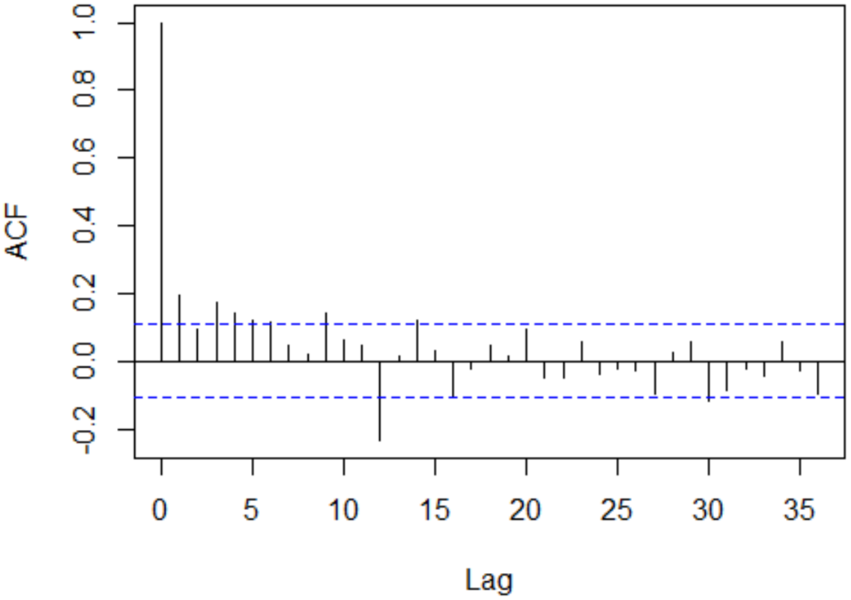


Figure 4: Autocorrelogram of the differentiated series at 1 of the differentiated series at 12 $((Y_t)_{t\in T})$

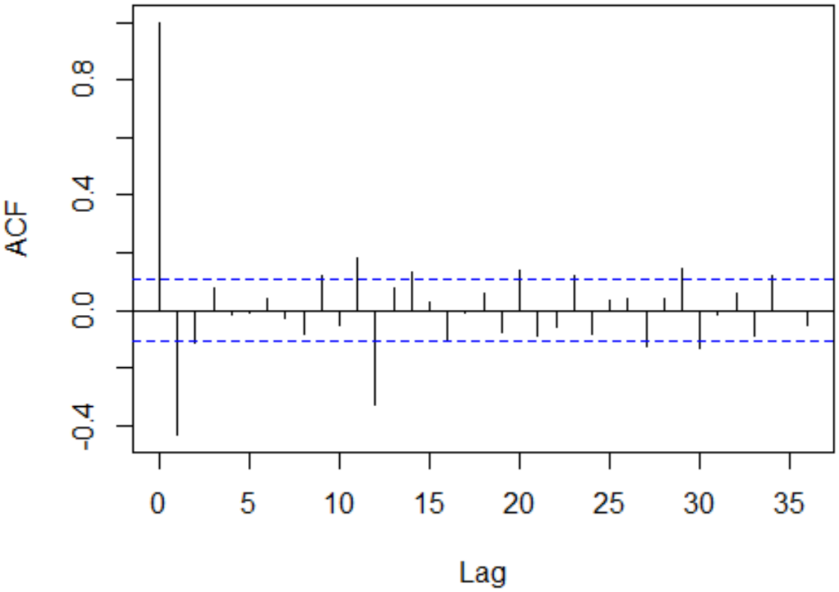


Figure 5: Partial autocorrelogram of the differential series at 1 of the differential series at 12 $((Y_t)_{t\in T})$

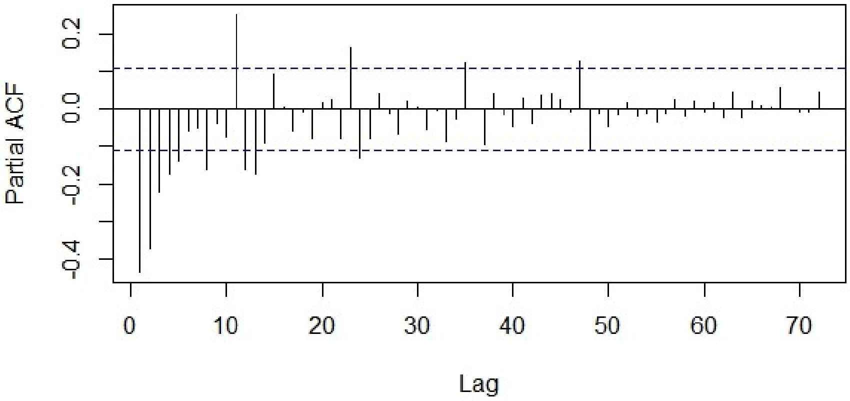


Figure 6: Time representation of the gross series $(X_t)_{t\in T}$ (top) then its twice differentiated series $(Y_t)_{t\in T}$ (down)

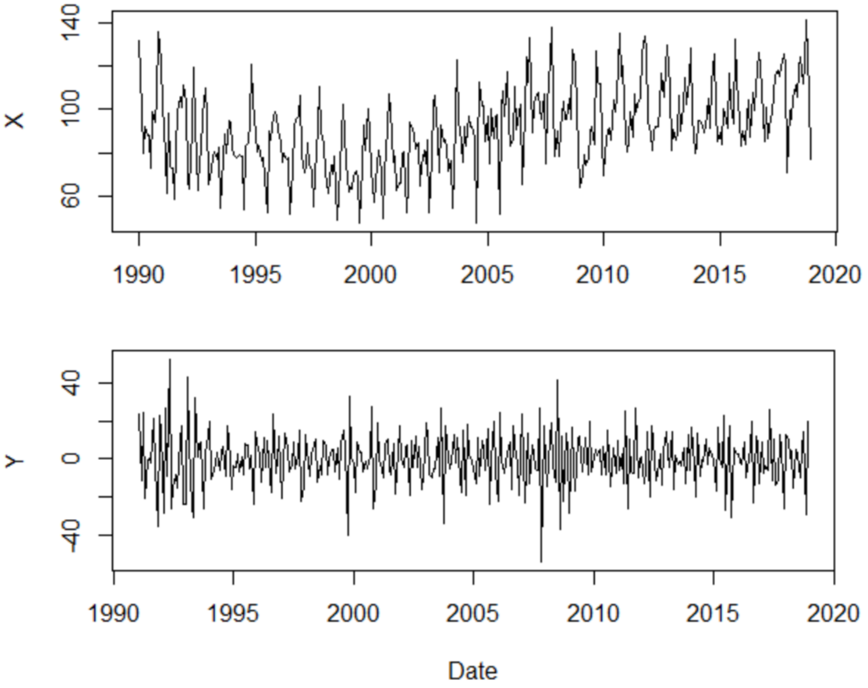


Figure 7: Prediction of the index with the SARIMA$_{12}$[(8,1,1,1),(4,1,0)] model that minimizes AIC, for the period 02/2018 - 11/2018

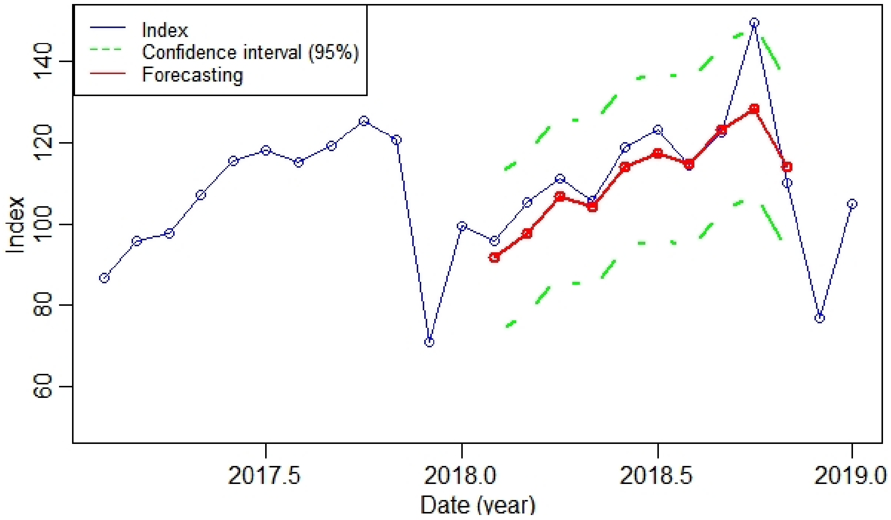


Figure 8: Prediction of the index with the SARIMA$_{12}$[(8,1,1,1),(0,1,1)] model that minimizes BIC, for the period 02/2018 - 11/2018

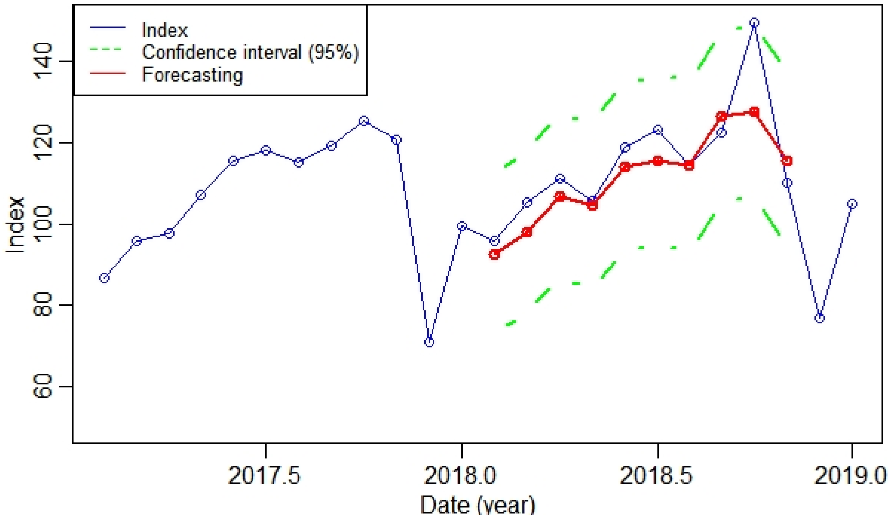


Figure 9: Q-Q Plot to visualize the assumption of normality of the SARIMA model residues$_{12}$[(8,1,1,1),(4,1,0)]

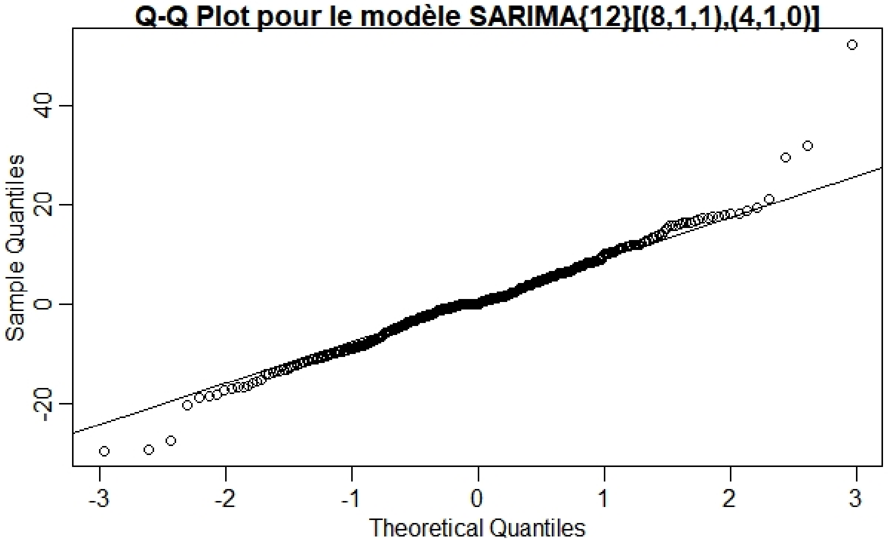
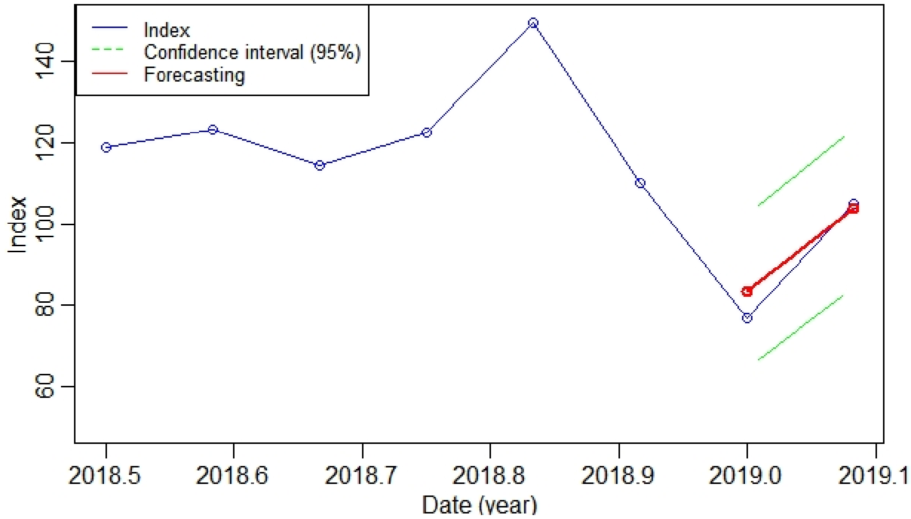


Figure 10: Prediction of the index with the SARIMA$_{12}$[(8,1,1,1),(4,1,0)] model that minimizes AIC, for the period 12/2018 - 01/2019 (last 2 months)



A.2 Test Tables

Table 2: Tests performed in question 2

Test	Result	Lag order	p-value
Augmented Dickey-Fuller on Z	-5.5196	6	<0.01
Phillips-Perron on Z	-312.34	5	<0.01
KPSS Trend on Z	0.14877	5	0.0477
KPSS Level on Z	0.55027	5	0.03034
Augmented Dickey-Fuller on Y	-10,26	6	<0,01
Phillips-Perron on Y	-385.07	5	<0.01
KPSS Trend on Y	0.017399	5	>0.1
KPSS Level on Y	0.028383	5	>0.1

A.3 Analysis of an autocorrelogram

A.4 Dickey-Fuller test

The Dickey-Fuller test is used to test the presence of a unit root in an autoregressive model. More formally, if we consider an AR(1) model: $y_t = \rho y_{t-1} + u_t$, we calculate the test statistic t_n under the null hypothesis H0 : $\rho = 1$ (ie the series is not stationary). By comparing t_n to the values referenced in the Dickey-Fuller table, we conclude on the rejection or non-rejection of H0. The principle of the augmented Dickey-Fuller test is the same and applies to more general series.

A.5 Phillips-Perron test

The Phillips-Perron test allows the stationarity of the model to be tested in the same way as the Dickey-Fuller test. We test the null hypothesis H0 : $\rho = 1$ (i.e. the series is not stationary) in semi-parametric models of the form : $X_t = c + bt + \rho X_{t-1} + u_t$ where u_t is a very general error term (not necessarily a white noise).

A.6 KPSS test

The Kwiatkowski-Phillips-Schmidt-Shin test is used to test the null hypothesis of a stationary model (trend). Their model is $y_t = \xi t + r_t + \epsilon_t$ where $r_t = r_{t-1} + u_t$ and r_0 serves as constant and (u_t) is iid. $xi \ = 0$ if there is no deterministic trend. The test statistic does not depend on the law of ϵ_t, so it is robust to the non-normality of residues.

A.7 Ljung-Box test

The Ljung-Box test, also known as the coat rack test, is used to test the correlation of a time series. It is defined by its null hypothesis H0: the data are independently distributed (i.e. the correlations are null) and by its test statistic $Q = n(n+2)\sum_{k=1}^{h} \frac{\hat{\rho_k}^2}{n-k}$ where n is the sample size, $\hat{\rho_k}^2$ is the empirical autocorrelation of delay k and h is the number of delays tested. Under H0, Q follows a $\chi^2_{(h)}$ and therefore for a confidence level α, the rejection region is $[Q > \chi^2_{1-\alpha,h}]$ where $\chi^2_{1-\alpha,h}$ is the $1-\alpha$ quantile of a χ^2 distribution at h degrees of freedom.

A.8 Jarque-Bera test

The Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The null is : the data do have a normal distribution ie a skewness of 0 and an expected excess kurtosis of 0 (which is the same as a kurtosis of 3).

The test statistic JB is defined as : $JB = \frac{n-k+1}{6}(S^2 + \frac{(C-3)^2}{4})$ where n is the sample size, S the sample skewness

and C the sample kurtosis.
The JB statistic asymptotically has a chi-squared distribution with two degrees of freedom, so the statistic can be used to test the null.

B Code

B.1 Question 1

```
#install.packages("lubridate")
library(lubridate)

#Donnes
data_0 = read.csv('valeurs_mensuelles.csv',sep=';')

#Dimensions
dim(data_0) # (352,3)
data = data_0[352:4,1:2] #Remplacer 4 par 3 si on veut Fevrier 2019

#Noms Colonnes
names(data) <- c("date","indice")

#On transforme les dates en format timestamp
date_format <- function(x) paste(x,"-01 00:00:00",sep="")
data$date<-ymd_hms(sapply(data$date,date_format))

#Graphique
plot(x=data$date, y=data$indice, type='l', ylab='Indice de production',xlab='Date')
```

B.2 Question 2

```
#install.packages('tseries')
library(tseries)


#Donnes
data = data[-349,1:2] #Ici on ne veut pas Fevrier 2019
serie <- ts(data$indice)
serieD <- (serie - lag(serie,-12)) # On enleve la saisonnalite D=1
serieD2 = (serieD - lag(serieD,-1)) #on rajoute une diffrence : d=1, toujours D=1

#Autocorrlogrammes
acf_serie = acf(serie, lag.max = 36)
acf_serieD = acf(serieD, lag.max = 36)
acf_serieD2 = acf(serieD2, lag.max = 36)

#Tests
##Test de Dickey-Fuller
adf1<-adf.test(serieD) #L'hypothse de non-stat est rejete  1%, pas de racine unitaire
adf2<-adf.test(serieD2) #L'hypothse de non-stat est rejete  1%, pas de racine unitaire


##Test de Phillips-Perron
pp1 <- pp.test(serieD) #Mieux que DF car par d'hyp sur les rsidus , H0 rejete -> stationnaire
pp2 <- pp.test(serieD2) #Mieux que DF car par d'hyp sur les rsidus , H0 rejete -> stationnaire

##Test de KPSS
kpss1 <- kpss.test(serieD, null='Trend') # PAS STATIONNAIRE : H0 = stationnaire
kpss2 <- kpss.test(serieD, null='Level') # PAS STATIONNAIRE : H0 = stationnaire
kpss3 <- kpss.test(serieD2, null='Trend') # STATIONNAIRE : H0 = stationnaire
kpss4 <- kpss.test(serieD2, null='Level') # STATIONNAIRE : H0 = stationnaire

#adf1
```

```
#adf2
#pp1
#pp2
#kpss1
#kpss2
#kpss3
#kpss4
```

B.3 Question 3

```
# Serie X
plot(x=data$date, y=serie, type='l', ylab='X',xlab='Date')

# Serie Y
dates_seriesD2 <- data$date[(length(data$date)-length(serieD2)+1):length(data$date)]
plot(x=dates_seriesD2, y=serieD2, type='l', ylab='Y',xlab='Date')
```

B.4 Question 4

```
##### Question 4 ######

res_acf <- acf(serieD2,main='')
# --> Q*=1, q*=1
res_pacf <- pacf(serieD2, lag.max=72, main='')
# --> P*=4, p*=8

# La fonctino Qtest renvoie les p-val des Ljung-Box test (Portmanteau) pour un lag maximal k
Qtest <- function(series, k) {
  t(apply(matrix(1:k), 1, FUN=function(l) {
    pval <- Box.test(series, lag=l, type="Ljung-Box")$p.value
    return(c("lag"=l,"pval"=pval))
  }))
}

signif <- function(estim){ #fonction de test des significations individuelles des coefficients des ordres max
  coef <- estim$fit$coef
  se <- sqrt(diag(estim$fit$var.coef))
  t <- coef/se
  pval <- (1-pnorm(abs(t)))*2
  return(rbind(coef,se,pval))
}

# Les sous-modeles possibles sont tous les SARIMA{12}(P,1,Q)(p,1,q) tels que P<=4 et Q<=1,
# p<=8 et q<=1.
# On choisira un modele siginificatif (les coefficients associes aux ordres les
# plus eleves de retards sont individuellement siginificatives) et les residus
# non autocorreles

Pmax = 4
pmax = 8
Qmax = 1
qmax = 1
lag = 36
m2=c()
i=c() #seulement pour suivre l'avancement

for (P in seq(0,Pmax)){
  for (Q in seq(0,Qmax)){
    for (p in seq(0,pmax)) {
      for (q in seq(0,qmax)) {
        i=rbind(i,c(P,Q,p,q))
        modele=try(sarima(serie,p,1,q,P,1,Q,12,no.constant=FALSE, details = FALSE))
        if(class(modele) == 'try-error'){next}
```

```r
        if (length(which(Qtest(modele$fit$residuals,lag)[,2]>0.05)) == lag ){ # si le bruit est non correle on garde
          if(length( which( c(signif(modele)[3,p],signif(modele)[3,p+q],signif(modele)[3,p+q+P],signif(modele)[3,p+q+P+Q]) < 0.05) ) == 4){
            m2=rbind(m2,c(P,Q,p,q))
          }
        }
      }
    }
  }
}

stargazer(m2)
#m2
#      [,1] [,2] [,3] [,4]
#[1,]    0    1    8    0
#[2,]    0    1    8    1
#[3,]    2    0    8    0
#[4,]    3    0    8    0
#[5,]    4    0    8    0
#[6,]    4    0    8    1


##### Selectionner le meilleur avec min AIC, BIC

AIC2 = c()
BIC2 = c()

for ( i in seq(1,dim(m2)[1]) ){
  P = m2[i,1]
  Q = m2[i,2]
  p = m2[i,3]
  q = m2[i,4]
  #modele=try(arima(serie, c(p,0,q), seasonal = list(order=c(P,1,Q),period=12)))#, method='CSS')
  modele = try(sarima(serie,p,1,q,P,1,Q,12,no.constant=FALSE, details=FALSE))
  if ( class(modele)=="try-error"){next}#si converge pas
  else{
    #AIC = c(AIC,AIC(modele))
    AIC2 = c(AIC2,modele$AIC)
    BIC2 = c(BIC2,modele$BIC)#c(BIC,BIC(modele))
  }
}

which(AIC2==min(AIC2))
which(BIC2==min(BIC2))
m2[6,] # P=4, Q=0, p=8, q=1 #min AIC
m2[2,] # P=0, Q=1, p=8, q=1 #min BIC

est4 = sarima(serie,8,1,1,4,1,0,12,no.constant=FALSE, details = FALSE) #min AIC
est42 = sarima(serie,8,1,1,0,1,1,12,no.constant=FALSE, details = FALSE) #min BIC


##### Prediction de la derniere annee de production

#pred4
pred4 = sarima.for(serie,n.ahead=10,8,1,1,4,1,0,S=12) #Min AIC
ypred4 = pred4$pred

inf4 = ypred4 -1.96*pred4$se #int conf 95%
sup4 = ypred4 +1.96*pred4$se

plot(x=seq(2017+1/12,2019,1/12),data$indice[326:349],col='blue',type='o',ylim=c(50,150), ylab='Index',xlab='Date (year)') # bleu = vraie serie
lines(x=seq(2017+1/12,2019,1/12), c(rep('',12),ypred4,rep('',2)),type="o",col='red',lwd=2)
lines(x=seq(2017+1/12,2019,1/12),c(rep('',12),as.vector(inf4),rep('',2)),col='green', type='c',lwd
=2)
lines(x=seq(2017+1/12,2019,1/12),c(rep('',12),as.vector(sup4),rep('',2)),col='green', type='c',lwd=2)
legend('topleft', legend=c("Index","Confidence interval (95%)","Forecasting"),
       col=c("blue","green","red"),lty=1:2, cex=0.8)
sqrt(mean((data$indice[338:347]-ypred4)**2))


#pred42
pred42 = sarima.for(serie,n.ahead=10,8,1,1,0,1,1,S=12) #Min BIC
ypred42 = pred42$pred

inf42 = ypred42 -1.96*pred42$se #int conf 95%
sup42 = ypred42 +1.96*pred42$se

plot(x=seq(2017+1/12,2019,1/12),data$indice[326:349],col='blue',type='o', ylim=c(50,150), ylab='Index',xlab='Date (year)') # bleu = vraie serie
lines(x=seq(2017+1/12,2019,1/12), c(rep('',12),ypred42,rep('',2)),type="o",col='red',lwd=2)
lines(x=seq(2017+1/12,2019,1/12),c(rep('',12),as.vector(inf42),rep('',2)),col='green', type='c',lwd=2)
lines(x=seq(2017+1/12,2019,1/12),c(rep('',12),as.vector(sup42),rep('',2)),col='green', type='c',lwd=2)
legend('topleft', legend=c("Index","Confidence interval (95%)","Forecasting"),
       col=c("blue","green","red"),lty=1:2, cex=0.8)

sqrt(mean((data$indice[338:347]-ypred42)**2))
```

B.5 Question 5

B.6 Question 6

```
## Question 6
# limite : normalite residus

qqnorm(est4$fit$residuals, main='Q-Q Plot pour le modele SARIMA{12}[(8,1,1),(4,1,0)]')
qqline(est4$fit$residuals) #residus normaux

jarque.bera.test(est4$fit$residuals)
```

B.7 Question 7

```
## Question 7

serieQ7 <- ts(data$indice, start=0, end=346)
predQ7 = sarima.for(serieQ7,n.ahead=2,8,1,1,4,1,0,S=12)

ypred7 = predQ7$pred

inf7 = ypred7 -1.96*predQ7$se #int conf 95%
sup7 = ypred7 +1.96*predQ7$se

plot(x=seq(2018+6/12,2019+1/12,1/12),data$indice[342:349],col='blue',type='o', ylim=c(50,150), ylab='Index',xlab='Date (year)') # bleu = vraie serie
lines(x=seq(2018+6/12,2019+1/12,1/12), c(rep('',6),ypred7),type="o",col='red',lwd=2)
lines(x=seq(2018+6/12,2019+1/12,1/12),c(rep('',6),as.vector(inf7)),col='green', type='c')
lines(x=seq(2018+6/12,2019+1/12,1/12),c(rep('',6),as.vector(sup7)),col='green', type='c')
legend('topleft', legend=c("Index","Confidence interval (95%)","Forecasting"),
       col=c("blue","green","red"),lty=1:2, cex=0.8)

sqrt(mean((data$indice[348:349]-ypred42)**2))
```

B.8 Question 8