Linear Time Series Assignment

ARIMA modelling of a time series

Tony LAUZE - Tien-Thinh TRAN-THUONG

April 2024

1 Part I : The data

1.1 What does the chosen series represent ?

For this project, we chose to work on the time series of the mineral and other bottled waters production. The series is a French Industrial Production Index (IPI) series, taken from the INSEE's time series databank. More precisely, we will work on the series that is corrected from seasonal variations and working days (CVS-CJO), covering the 1990/01-2024/01 period with a monthly frequency, resulting in 410 observations. The index is with base 100 in 2021. Both the raw and the corrected series can be observed below (see **Figure 1**).

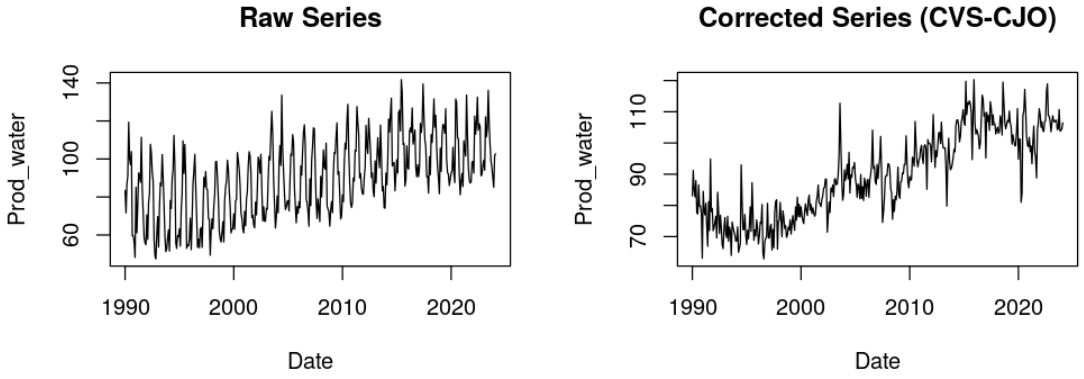


Figure 1: Comparison of the raw and corrected series of the mineral water production from 1990/01 to 2024/02

On the raw series, a clear seasonality can be seen, as well as a slight trend. On the corrected series, the seasonality seems to be gone (which is all the point of the CVS-CJO), and an increasing trend can be observed, which might be linear, from around 1995 to 2017. Apart from a positive peak in the summer of 2003 (which may be attributed to the notorious heatwave that took place in France at that time), and a negative peak for the months of March and April 2020 (which are the firsts months of the COVID Crisis), no outliers or particular periods can be found in the series.

1.2 Transform the series to make it stationary if necessary

Given no seasonality seems to arise, and no change in the variability is noticeable through time, we only need to deal with the trend. A possibility would be to estimate the trend that occurs from around 1995 to 2017 (by supposing it is linear) and then to remove it from the series. However, this trend does not seem to continue after 2017, and before 1995 the series is rather decreasing.

Therefore, we will instead try to differentiate the series, that is, to apply the operator $\Delta U_t = U_t - U_{t-1}$. The result can be seen in **Figure 2**.

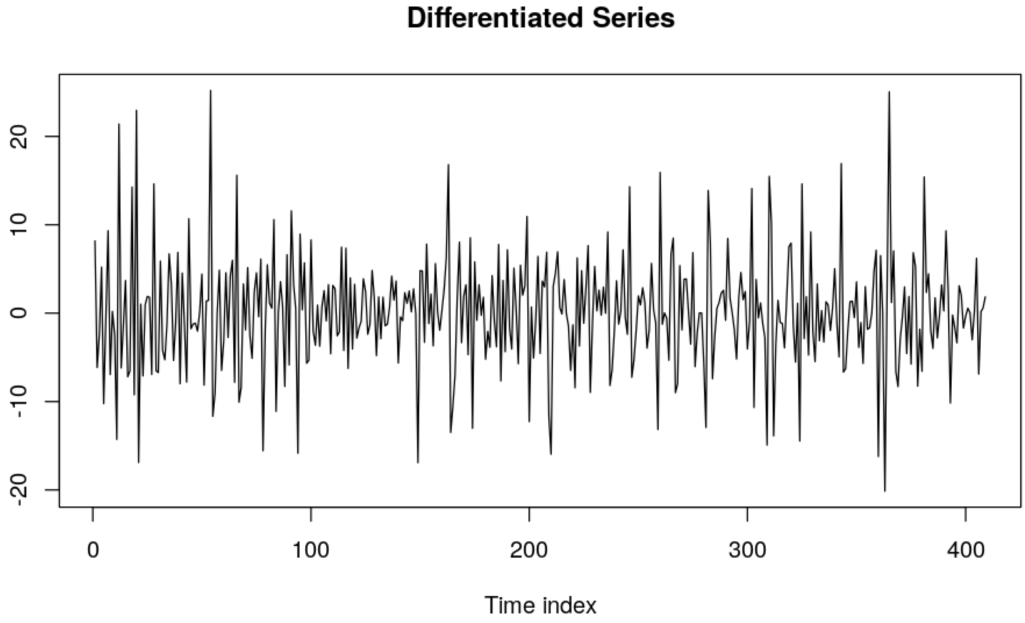


Figure 2: Observation of the differentiated series

The one-time differentiated series seems to be centered around zero, no clear change in variability can be seen: this series looks stationary. To check if it is, we will proceed to some tests.

First, we can check whether a unit root is present in our series, using the augmented Dickey-Fuller (ADF) test. Regarding the specification of the test, the series being centered around zero and presenting no trend, we use the regression with no intercept (constant) nor time trend. To check the hypothesis of absence of trend, we can make a simple regression of the series on time index (see the regression table in **Table 1**): both the intercept and coefficient of the time index are almost equal to zero and are definitely non significant. We can thus be confident about the specification of the ADF test. Regarding the number of lags, 7 are required so that the Ljung-Box test cannot reject the null-hypothesis of absence of auto-correlation among the residuals of the model until order 24. Notice that this number of lags is in accordance with the value of p that will be chosen later. The results of the performed ADF test is presented in **Table 2**, and it leads to reject the null hypothesis of the presence of a unit root; which is in favor of stationarity.

Table 1: Regression of the series on time

	Dependent variable:
	diff_prod_water
index	0.0004 (0.003)
Constant	−0.017 (0.641)
Observations	409
R^2	0.00004
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Results of the Augmented Dickey-Fuller Test

PARAMETER:	
Lag Order	7
Speficication	No constant, no time-trend
STATISTIC:	
Dickey-Fuller	-12.2605
P VALUE:	
	< 0.01

Then, we can perform a KPSS test in which the null hypothesis is the level-stationarity of the series. The p-value being greater than 0.1 (see **Table 3**), we cannot reject the null-hypothesis; which is once again in favor of stationarity.

Table 3: Results of the KPSS Test

Truncation lag parameter:	
Lag Order	5
KPSS Level:	
Value	0.023193
P VALUE:	
	> 0.1

1.3 Graphically represent the chosen series before and after transforming it

Our transformation is thus limited to a simple differentiation of the series. The transformed series can be seen below next to the original corrected series.

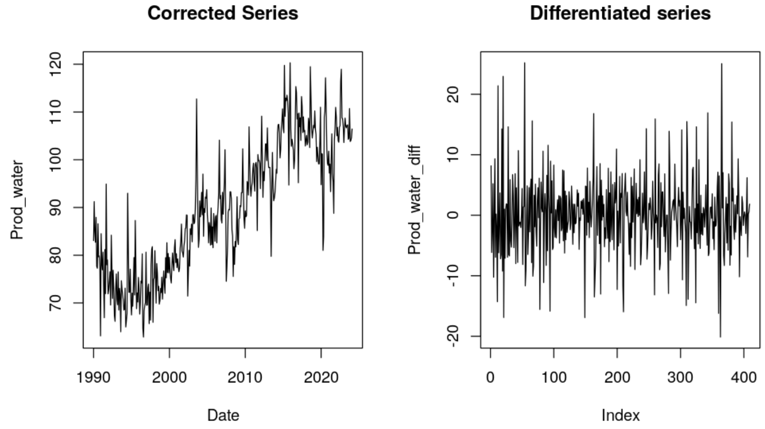
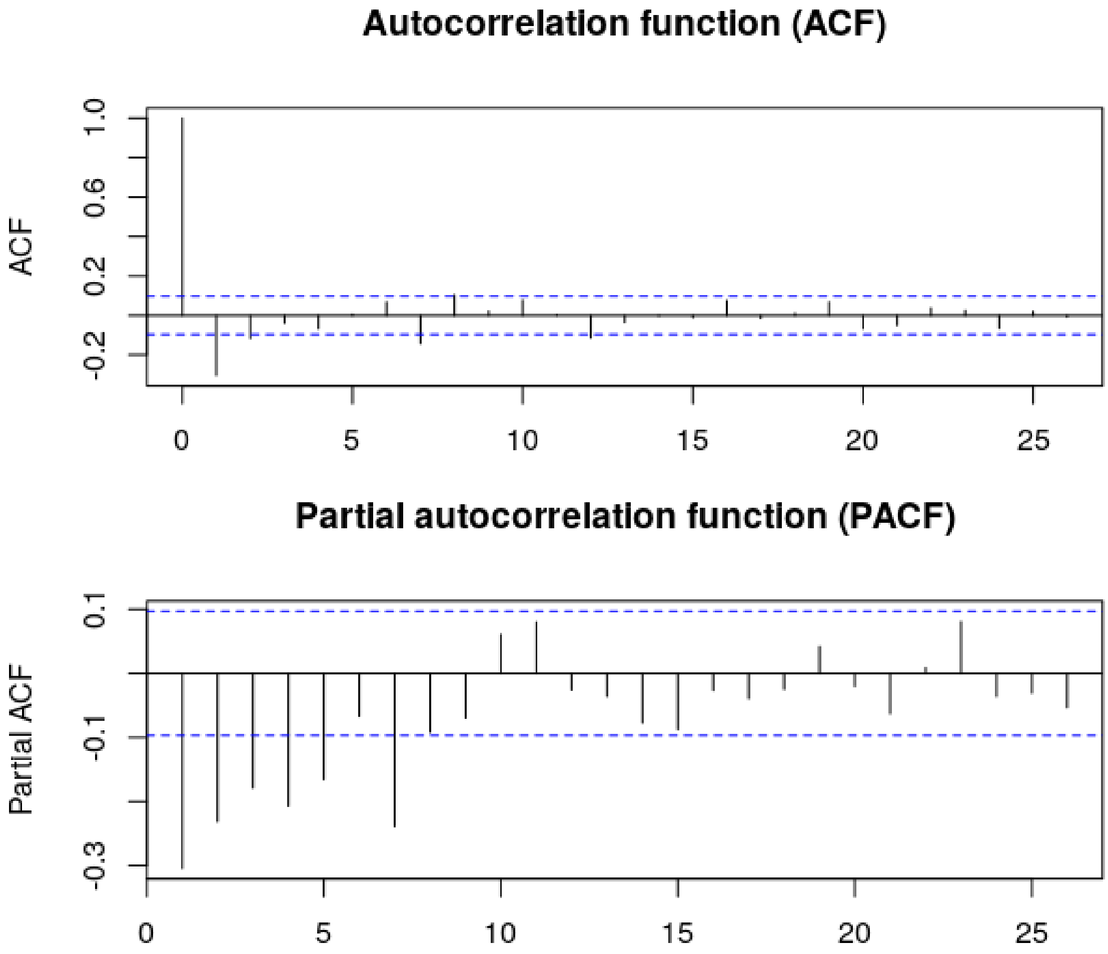


Figure 3: Observation of the corrected series and the differentiated series

2 Part II : ARMA models

2.1 Pick an ARMA(p,q) model for your corrected time series

To choose the parameters for our ARMA model, we will rely on the ACF and PACF plots of our transformed series, which can be see in **Figure 4**. The values of the auto-correlations fall under the confidence intervals right after the first lag, which suggests the choose q=1. The values of the partial auto-correlations decrease strongly after lag 7, which would lead us to choose p=7.


Autocorrelation function (ACF)

Partial autocorrelation function (PACF)

Figure 4: Plot of the ACF and PACF, until lag order 24

However, choosing p=7 would lead to a rather complex model. To be sure that such a model is necessary, will will try all values for p between 1 and 7, and use information criteria (AIC/BIC) to pick to most parsimonious model. The results of the models can be seen in **Table 4**. Both AIC and BIC are minimized by the choice p=7; we therefore choose the final model ARMA(7,1). The coefficients obtained for this model can be found in **Table 5**, confirming the significance of the coefficients.

Table 4: Comparison of ARMA models with q=1 and different values for p

p	AIC	BIC
1	2574.057	2586.098
2	2575.519	2591.574
3	2576.731	2596.800
4	2576.680	2600.762
5	2578.598	2606.694
6	2580.566	2612.676
7	2570.502	2606.625

Table 5: Coefficients of the estimated ARMA model

	Coefficient	Standard Error	t-value	p-value
ar1	-0.1981	0.1465	-1.3526	0.1769
ar2	-0.3111	0.0859	-3.6225	0.0003
ar3	-0.2848	0.0794	-3.5842	0.0004
ar4	-0.2818	0.0720	-3.9147	0.0001
ar5	-0.1998	0.0688	-2.9044	0.0039
ar6	-0.1226	0.0589	-2.0797	0.0382
ar7	-0.2202	0.0528	-4.1687	0.0000
ma1	-0.3375	0.1471	-2.2950	0.0222

2.2 Write the ARIMA(p,d,q) model for the chosen series

The model for the initial series is thus an ARIMA(7,1,1). We can write the formula for this model as follows (where B is the simple lag operator):

$$\begin{aligned} X_t = {} & (1-B)\epsilon_t \\ & - 0.1981(1-B)X_{t-1} \\ & - 0.3111(1-B)X_{t-2} \\ & - 0.2848(1-B)X_{t-3} \\ & - 0.2818(1-B)X_{t-4} \\ & - 0.1998(1-B)X_{t-5} \\ & - 0.1226(1-B)X_{t-6} \\ & - 0.2202(1-B)X_{t-7} \\ & - 0.3375\epsilon_{t-1} \end{aligned} \tag{1}$$

3 Part III : Prediction

3.1 Write the equation verified by the confidence region of level α on the future values (X_{T+1}, X_{T+2}).

The equations of the confidence regions can be written as follows:

$$IC_{T+1} = \hat{X}_{T+1} \pm z_{\alpha/2} \times se_{\hat{X}_{T+1}}$$
$$IC_{T+2} = \hat{X}_{T+2} \pm z_{\alpha/2} \times se_{\hat{X}_{T+2}}$$

Where:
$z_{\alpha/2}$ is the quantile of the standard normal distribution for a confidence level of $1 - \alpha$,
$se_{\hat{X}_{T+1}}$ is the standard error of the prediction $\hat{X}_{T+1}$,
$se_{\hat{X}_{T+2}}$ is the standard error of the prediction $\hat{X}_{T+2}$.

3.2 Give the hypotheses used to get this region

We consider here that the series's residuals are Gaussian. According to the course, we also need to suppose that the values of our series are uncorrelated with those of the underlying ARMA process.

3.3 Graphically represent this region for $\alpha = 95\%$. Comment on it.

Below we plot the predictions of our ARIMA(7,1,1) model for the T+1 and T+2 periods, as well as the confidence intervals for $\alpha = 95\%$. We can notice that the confidence intervals are growing with time.

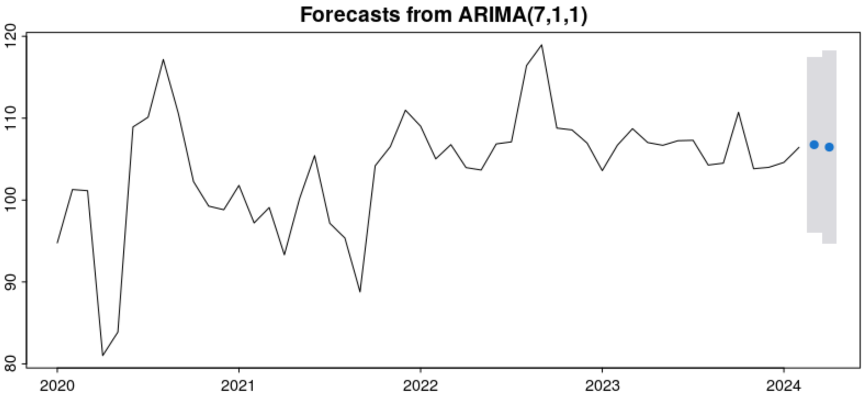


Figure 5: Plot of the predictions of the ARMA(7, 1, 1) model for the T+2 and T+1 periods, with the 95% confidence intervals

3.4 Open question