GAUTHIER Jeanne
MASSIN Keryann

ENSAE 2nd year
Linear Time Series Project

Analysis of the French Industrial Production Index

May 15, 2023

Contents

1 Part I: the data

1.1 Question 1

What does the chosen series represent ? (sector, potential data processing, logarithmic transformation, etc.)

In this project, we will take a look at the French Industrial Production Index (IPI), available on INSEE website at this link: `https://www.insee.fr/fr/statistiques/serie/010537206`. This series is corrected from seasonal variations and working days, on a monthly frequency and makes it possible to follow the monthly evolution of industrial activity in France and in construction.
The initial series records the IPI from January 1990 to to March 2023 (base 100 in 2015). It is plotted on figure 1 below:

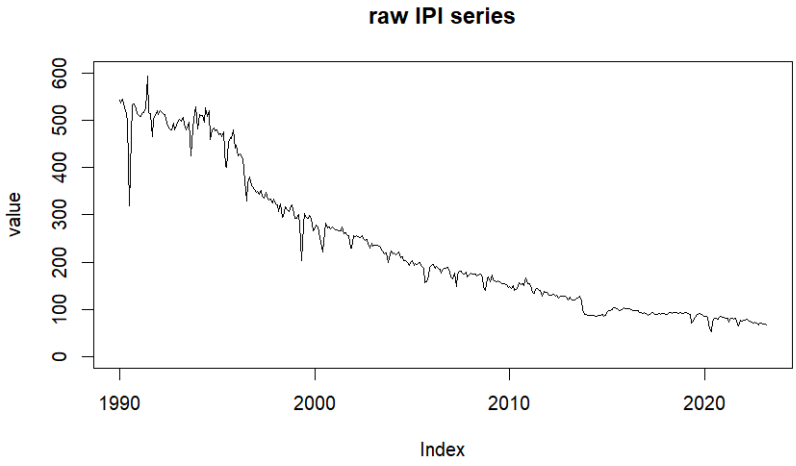

Figure 1: Raw initial series of French Industrial Production Index

As the series is already corrected from seasonal variations and working days, these data processing are unnecessary. However, the series' empirical variance seems proportional to its values. We correct this effect (thus diminishing heteroskedasticity) by applying a Box Cox transformation of parameter $\lambda = 0$. Namely, we take the logarithm, to obtain figure 2.

1.2 Question 2

Transform the series to make it stationary if necessary (differentiate it, correct the deterministic trend, etc.). Thoroughly justify your choices.

Clearly, figure 2 shows that log(IPI) is not stationary. To statistically back this affirmation, we conducted Augmented Dickey-Fuller and KPSS tests both on log(IPI) and on the first difference of log(IPI). Table 1 shows the results.

The Augmented Dickey-Fuller test rejects the null hypothesis of non-stationarity for the first difference of log(IPI) at the 1% level and does not reject it for log(IPI) at the 5% level.

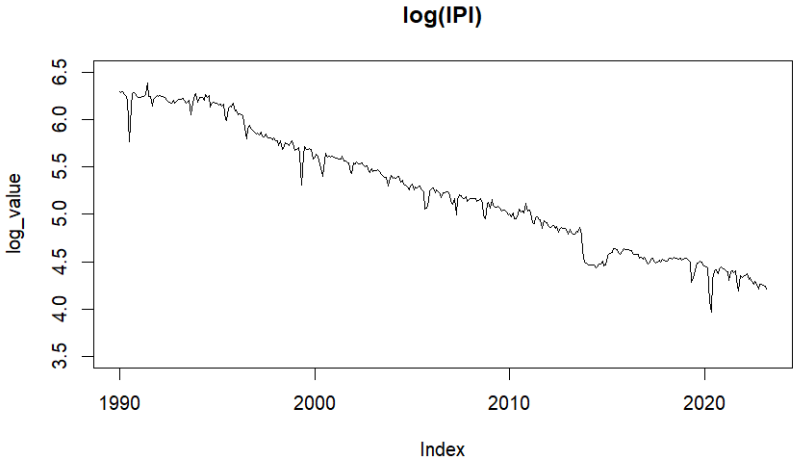

Figure 2: logarithm of IPI

The KPSS test rejects the null hypothesis of stationarity for log(IPI) at the 1% level but does not reject it for the first difference of log(IPI) at the 10% level.

As a result of these two tests, we can confidently argue that our corrected time series 'First difference of log(IPI)' (see figure 4) is stationary. Therefore, log(IPI) was indeed a I(1).

	ADF p-value	KPSS p-value
log(IPI)	0.06	0.01
first difference of log(IPI)	0.01	0.1

Table 1: ADF and KPSS tests for stationarity

1.3 Question 3

Graphically represent the chosen series before and after transforming it.

We can finally compare our two series: the initial raw series and the stationarized one. We plotted them in figures 3 and 4.

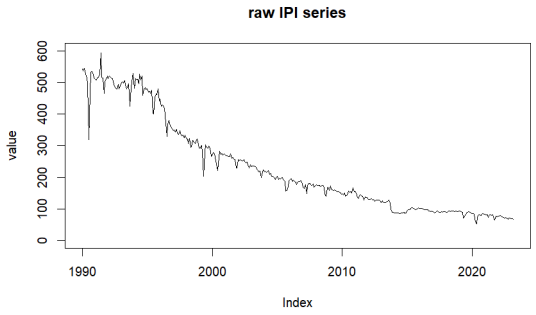

Figure 3: Raw initial series of French Industrial Production Index

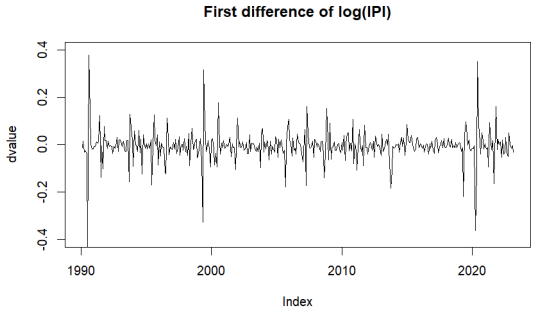

Figure 4: Corrected time series

2 Part II: ARMA models

2.1 Question 4

Pick (and justify your choice) an ARMA(p, q) model for your corrected time series X_t. Estimate the model parameters and check its validity.

In this part, we will attempt to find the ARMA(p, q) model that best fits our corrected time series X_t (see figure 4). We assume stationarity of X_t. To select the right model, we will first take a look at autocorrelation function (ACF) and partial autocorrelation function (PACF). This will allow us to find maximal orders p_{max} and q_{max} of the ARMA(p, q) model. To evaluate p_{max} (resp. q_{max}), we analyze PACF (resp. ACF). On figures 5 and 6, we see that

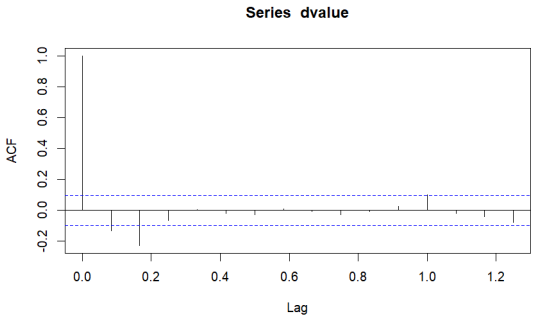

Figure 5: Autocorrelation function

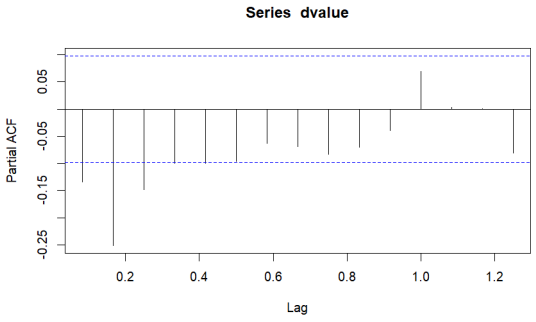

Figure 6: Partial autocorrelation function

empirical autocorrelations and partial autocorrelations decrease rapidly. Hence, an ARMA(p, q) can rightfully be fitted on the series.
From figure 5, we see that autocorrelation is statistically different from 0 for lags $0, 1, 2$, so we pick $q_{max} = 2$. From figure 6, we see that partial autocorrelation is statistically different from 0 for lags $1, 2, 3$ so we pick $p_{max} = 3$.

We are therefore looking for an ARMA(p, q) model that satisfies three conditions:

- $p \in \{0, 1, 2, 3 = p_{max}\}$ and $q \in \{0, 1, 2 = q_{max}\}$
- The model is valid, which means that its residuals are not autocorrelated.
  We can check it using the Ljung-Box test with null hypothesis being the joint nullity of autocorrelations until order k. We chosed $k = 24$.
- The model is well-adjusted, which means that its coefficients are statistically significant.
  We can check it using the usual approach to test the significance of a linear regression's coefficients, the t-Student test.

From the twelve possible models, only three are valid and well-adjusted: ARMA(2, 1), MA(2) and ARMA(1, 2). To chose the best fitted, we compare their AIC and BIC in table 2.

	ARMA(2, 1)	MA(2)	ARMA(1, 2)
AIC	-1078.649	-1075.884	-1079.288
BIC	-1058.717	-1059.938	-1059.356

Table 2: AIC and BIC for the different ARMA(p, q) models

According to table 2, AIC and BIC are respectively minimized for ARMA(1, 2) and MA(2) models. In addition, the ARMA(1, 2) model presents a BIC close to the MA(2) model's. We thus consider that ARMA(1, 2) is the best fit for X_t.

2.2 Question 5

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

For $(p, d, q) = (1, 1, 2)$, the model is well-adjusted (the coefficients are significant) and is valid (the residuals are not autocorrelated), so our series log(IPI) (see figure 2) fits with an ARIMA(1,1,2).
Let Y_t be our raw series in figure 1. With $X_t = \log(Y_t) - \log(Y_{t-1})$, ϵ_t the residuals at time t, and by evaluating the coefficients of the model, we finally have:

$$X_t = 0.208X_{t-1} + \epsilon_t - 0.426\epsilon_{t-1} - 0.214\epsilon_{t-2} \tag{1}$$

3 Part III: Prediction

Denote T the length of the series. Assume the series residuals are Gaussian.

In the following, we denote T the length of the series and we assume the series residuals are gaussian, that is $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$. We have a ARMA(1,2) which translates:

$$X_t = \phi_1 X_{t-1} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} \tag{2}$$

3.1 Question 6

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

Given $\mathbb{E}[\epsilon_{T+h} \mid X_T, X_{T-1}, \dots] = 0$, $\forall h > 0$, optimal predictions satisfy:

$$\begin{cases} \hat{X}_{T+1|T} = \phi_1 X_T + \theta_1\epsilon_T + \theta_2\epsilon_{T-1} \\ \hat{X}_{T+2|T} = \phi_1 \hat{X}_{T+1|T} + \theta_2\epsilon_T \end{cases} \tag{3}$$

Let us calculate prediction erros $X_{T+1} - \hat{X}_{T+1|T}$ and $X_{T+2} - \hat{X}_{T+2|T}$. It writes:

$$\hat{X} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix} \tag{4}$$

Thus, using (2):

$$X - \hat{X} = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + (\phi_1 + \theta_1)\epsilon_{T+1} \end{pmatrix} \tag{5}$$

We can now compute the variance of prediction errors:

$$\begin{cases} \mathbb{V}(X_{T+1} - \hat{X}_{T+1|T}) = \mathbb{V}(\epsilon_{T+1}) = \sigma^2 \\ \mathbb{V}(X_{T+2} - \hat{X}_{T+2|T}) = \mathbb{V}(\epsilon_{T+2} + (\phi_1 + \theta_1)\epsilon_{T+1}) = \sigma^2(1 + (\phi_1 + \theta_1)^2) \end{cases} \tag{6}$$

$X - \hat{X}$ thus follows a normal distribution of mean $\mu = 0$ and variance Σ, that is:

$$X - \hat{X} \sim \mathcal{N}(0, \sigma^2) \quad \text{where} \quad \Sigma = \sigma^2 \begin{pmatrix} 1 & \phi_1 + \theta_1 \\ \phi_1 + \theta_1 & 1 + (\phi_1 + \theta_1)^2 \end{pmatrix} \tag{7}$$

We see that $\det(\Sigma) = \sigma^2$, so Σ is invertible if and only if $\sigma^2 > 0$, which is true by assumption. According to the lectures, we finally have $(X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \sim \chi^2(2)$. It follows that the confidence region R_α of level α verifies, for all $\alpha \in [0, 1]$:

$$R_\alpha = \{X \in \mathbb{R}^2 \mid (X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \leq q^{1-\alpha}_{\chi^2(2)}\} \tag{8}$$

where $q^{1-\alpha}_{\chi^2(2)}$ is the $(1-\alpha)$-quantile of $\chi^2(2)$ distribution.

3.2 Question 7

Give the hypothesis used to get this region.

To get the previous results, we made some hypothesis:

- The model is perfectly known
- The coefficients obtained in part 2 are correct
- The white noise follows a normal distribution $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$
- $\sigma^2 > 0$

3.3 Question 8

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

3.4 Question 9 - Open question

Let Y_t a stationary time series available from $t = 1$ to T. We assume that Y_{T+1} is available faster than X_{T+1}. Under which condition(s) does this information allow you to improve the prediction of X_{T+1}? How would you test it/them?

Under the assumption that Y_{T+1} is available faster than X_{T+1}, we can use Y_{T+1} to predict X_{T+1} if and only if

$$\hat{X}_{T+1|\{X_t,Y_t|t<T\}\cup\{Y_{T+1}\}} \neq \hat{X}_{T+1|\{X_t,Y_t|t<T\}} \tag{9}$$

i.e if and only if Y_t instantaneously causes X_t in the Granger sense.