Assignment – Linear Time Series

Galeran Subileau
Rémi Fouchérand

Academic Year 2024–2025

ENSAE Paris

1 The Data

1.1 Question 1

We chose to model the series (X_t) which is The CVS-SJO index of industrial production of agricultural and forestry machinery (available here). The data comprises monthly observations spanning from 1990 to February 2025. It is base 100 in 2021, meaning it represents the monthly variation of the production of agricultural and forestry machinery compared to that of 2021. By performing simple edits to the series, and adding a time variable T to the observations, we are able to plot our series (Figure 1). Note that the two last values of our series were intentionally removed in order to test our future predictions.

1.2 Question 2

Before computing ARMA models on our series, we have to check that it is stationary. As the chosen series is already corrected for seasonality (CVS-SJO), we do not have to worry about seasonal trends, and only need to check for stationarity.

Before performing any Unit Root test, we first need to check for the presence of a trend. We do this by regressing X_t on T and a constant. Table 1 shows us that there is no significant trend, but that the series exhibits a clear constant. We thus decide to run an ADF test on the series with only a constant, also making sure we take into account autocorrelation of the residuals (See Table 2). The test rejects the null hypothesis of the presence of a Unit Root, thus indicating the series is likely to be stationary. However, we have several reasons to doubt about the validity of this result: first by looking at the plotted series, which does not really seems stationary, second because our ADF test needed to consider autocorrelation of the residuals up to 11 lags, which is a lot and weakens our results. In order to further investigate, we also perform a Phillips-Perron test and a KPSS test on the same series; also registered in Table 2. The PP-test also yields good results and strongly rejects the null hypothesis of the presence of a Unit Root. Considering both previous tests only check for UR and not direct stationarity, we also take a look at the KPSS tests which has the advantage to test directly for stationarity. Here we get a p-value of 0.0775, which does not reject the null hypothesis of stationarity at the 5% level. However, this result is not really strong either and as the p-value is still very close to the rejection threshold, we cannot say with utter certainty that the series is in fact stationary. As stationarity does not really appear graphically, and to be more cautious as manipulating ARMA models on non-stationary series is impossible, we still decide to differentiate our series in order to ensure that it will be stationary.

We thus create the differentiated series $\Delta X_t = X_t - X_{t-1}$ and perform the same regression as before (Table 3). The differentiated series exhibits no trend and neither constant, we thus perform the same tests as before but this time with no constant and no trend. The results, shown in Table 4, are unambiguous and clearly show that our differentiated series is stationary.

1.3 Question 3

The comparison between both the raw and differentiated series is shown in Figure 2

2 ARMA Models

2.1 Question 4

As we differentiated our series 1 time in order to make sure it was stationary, we thus need to compute $ARIMA(p, 1, q)$ models if we want to predict future values of the raw series.
In order to determine the order p and q needed, we plot the ACF and PACF on 24 lags (i.e two full periodicities) for our differentiated series, which are shown in Figures 3 and 4. By reading the graphs, we choose accordingly orders $p^* = 5$ and $q^* = 2$. We decide not to take into account further bins outside the confidence interval, as this range has proven sufficient to find a good model.

We then need to check which models within orders p^* and q^* are valid and well adjusted. A model is considered valid if its residuals are not serially autocorrelated. To test this, we perform a Ljung-Box test with H_0 being the joint nullity of the serial correlation of residuals until an order k (we usually choose $k = 24$). Then, we check if the model is well-adjusted by looking at all its coefficients and checking if they are statistically significant. In order to perform those tests we use a set of automated functions fitting every possible model and checking for both conditions (see the code for more details). In the end, it appears that there are only three models within p^* and q^* that are valid and well-adjusted: $ARIMA(5, 1, 0)$, $ARIMA(4, 1, 1)$ and $ARIMA(4, 1, 2)$.

In order to choose between these three models, we compute each combination's Bayesian Information Criterion and Akaike Information Criterion (Tables 5 and 6). We can see that the AIC is minimized by $ARIMA(5, 1, 0)$ and that the BIC is minimized by $ARIMA(1, 1, 1)$.

As the latter has not been selected as a valid or well-adjusted model (See Table 9), we then choose $ARIMA(5, 1, 0)$, which is valid and well adjusted (See Tables 7 and 8).

2.2 Question 5

The differentiated series follows an ARMA(5,0), which is equivalent to an AR(5). So (X_t) follows an ARIMA(5,1,0). To lighten the notation, we denote $Y_t = \Delta X_t$, the model can be rewritten as:

$$Y_t = \sum_{i=1}^{5} \phi_i Y_{t-i},$$

where the ϕ_i are the AR coefficients estimated from the data. The explicit model is :

$$Y_t = -0.321Y_{t-1} - 0.228Y_{t-2} - 0.023Y_{t-3} - 0.171Y_{t-4} - 0.1415Y_{t-5},$$

Substituting back the differentiated form $Y_t = X_t - X_{t-1}$, we obtain the form :

$$X_t = 0.679X_{t-1} + 0,0927X_{t-2} + 0.205X_{t-3} - 0.149X_{t-4} + 0.0303X_{t-5} + 0.1415X_{t-6},$$

We are using an AR(5) on our differentiated series, so it is invertible (no moving average). So, in order to make predictions, we only have to check if there exists a causal solution for it to be a canonical ARMA. Hence, a causal stationary solution only exists if :

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3 - \phi_4 z^4 - \phi_5 z^5 \neq 0 \text{ for all z such that } |z| \leq 1$$

Figure 5 represents the inverse of the roots. As they are inside the unit circle, it means that the condition stated above is verified. We conclude that there exists a causal solution to our ARMA model.

Finally, we would like to check 2 things on our residuals : we would like them to be white noise and to be Gaussian (for the prediction). In Figure 6, we plot the residuals, their autocorrelation, and distribution. We see that they are centered around 0 and that their variance seems to be constant, and that their autocorrelation is null. This tends to indicate that they are weak white noise. Nevertheless, we can't be sure that they are Gaussian, because the distribution does not really match the normal law : they are too much concentrated around 0. We assume nonetheless that it is the case in the next questions.

3 Prediction

3.1 Question 6

We aim to predict the next two values of the process, X_{T+1} and X_{T+2}, based on the ARIMA(5,1,0) model selected in the previous steps. Let $\hat{X}_{T+1|T}$ denote the optimal linear forecast of X_{T+1} given the information up to time T.

Using this expression and the results from the previous question, the one-step and two-step ahead forecasts are computed as:

$$\hat{X}_{T+1|T} = 0.679X_T + 0,0927X_{T-1} + 0.205X_{T-2} - 0.149X_{T-3} + 0.0303X_{T-4} + 0.1415X_{T-5},$$

$$\hat{X}_{T+2|T} = 0.679\hat{X}_{T+1|T} + 0,0927X_T + 0.205X_{T-1} - 0.149X_{T-2} + 0.0303X_{T-3} + 0.1415X_{T-4},$$

Let us denote the forecast error vector as:

$$\tilde{X}_{T+1} = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \varepsilon_{T+1} \\ \varepsilon_{T+2} + (1+\phi_1)\varepsilon_{T+1} \end{pmatrix}.$$

Assuming the error terms ε_t are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$, the covariance matrix of the forecast error vector is:

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2(1+\phi_1) \\ \\ \sigma^2(1+\phi_1) & \sigma^2\left(1+(1+\phi_1)^2\right) \end{bmatrix}.$$

Thus, the vector $\tilde{X}_{T+1}$ follows a bivariate normal distribution with mean zero and covariance Σ. As a result, we can construct a $(1-\alpha)$ confidence region as the ellipsoid:

$$\left\{x \in R^2 : (x-\hat{x})^\top \Sigma^{-1}(x-\hat{x}) \le q_{\chi^2(2)}(1-\alpha)\right\},$$

where $\hat{x} = (\hat{X}_{T+1|T}, \hat{X}_{T+2|T})^\top$, and $q_{\chi^2(2)}(1-\alpha)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with 2 degrees of freedom.

On the marginal (univariate) level, the 95% confidence intervals for each forecast are:

$$X_{T+1} \in \left[\hat{X}_{T+1|T} \pm 1.96\hat{\sigma}\right],$$

$$X_{T+2} \in \left[\hat{X}_{T+2|T} \pm 1.96\hat{\sigma}\sqrt{1+(1+\phi_1)^2}\right].$$

Substituting the estimated values of ϕ_1 and σ^2 yields the final numerical intervals :

$$[57.10, 97.33]$$

$$[55.21, 103.83]$$

3.2 Question 7

The construction of the confidence region for the forecasted values X_{T+1}, X_{T+2}, etc., relies on the following key assumptions:

1. **Gaussian residuals**: The forecast errors are assumed to be independent and identically distributed (i.i.d.) and follow a normal distribution. This justifies the use of normal quantiles (e.g., 1.96 for 95% confidence) to build symmetric confidence intervals.
2. **Correct model specification**: The ARIMA model must accurately represent the dynamics of the time series. In particular, residuals should exhibit no significant autocorrelation (white noise) and no important structural components should be omitted.
3. **Known or precisely estimated parameters**: The model parameters are assumed to be known when computing the forecast intervals. In practice, they are estimated, and the additional uncertainty is typically neglected when the sample size is sufficiently large.
4. **Stationarity of the differentiated series**: The differentiated series is assumed to be stationary.

Violations of these assumptions may lead to incorrect or misleading confidence intervals. These assumptions are standard in the Box–Jenkins methodology and are implicitly used when computing forecast intervals with the `forecast` package in R.

3.3 Question 8

The Figure 8 represents our series and the estimations $\hat{X}_{T+1|T}$ and $\hat{X}_{T+2|T}$ (in red) and their true values (in black) as well as their univariate confidence intervals at level 95%. We can see that the true values X_{T+1} and X_{T+2} lie within the confidence intervals. Hence, our estimations are within expectations.

3.4 Question 9

Let us consider an additional stationary time series (Y_t) observed up to time T, with the particularity that Y_{T+1} becomes available before X_{T+1}. We wish to assess under which circumstances the observation of Y_{T+1} can enhance the prediction of X_{T+1}.

The key concept here is **Granger causality**. In a predictive framework, we say that Y_t Granger-causes X_t if the inclusion of the past values of Y_t leads to a statistically significant improvement in the prediction of X_{t+1}, compared to using only the past of X_t. Formally, this means that:

$$\mathbf{E}[X_{t+1} \mid X_t, X_{t-1}, \dots, Y_t, Y_{t-1}, \dots] \neq \mathbf{E}[X_{t+1} \mid X_t, X_{t-1}, \dots].$$

To evaluate this hypothesis, we can fit a **Vector AutoRegressive (VAR)** model of the form:

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = A_1 \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \cdots + A_p \begin{pmatrix} X_{t-p} \\ Y_{t-p} \end{pmatrix} + \varepsilon_t,$$

where each A_i is a coefficient matrix and ε_t is a zero-mean white noise vector. We then conduct a **Granger causality test** by checking whether the coefficients on the lagged values of Y_t in the equation for X_t are jointly zero.

In statistical terms, the null hypothesis is:

$$H_0 : \text{"}Y_t \text{ does not Granger-cause } X_t\text{"} \quad \Longleftrightarrow \quad A_{i,12} = 0 \text{ for all } i = 1, \dots, p.$$

If this hypothesis is rejected using an F-test or Wald test, we conclude that Y_t contains useful predictive information for X_t. Consequently, the knowledge of Y_{T+1} can be exploited to refine the prediction of X_{T+1}.

Appendix

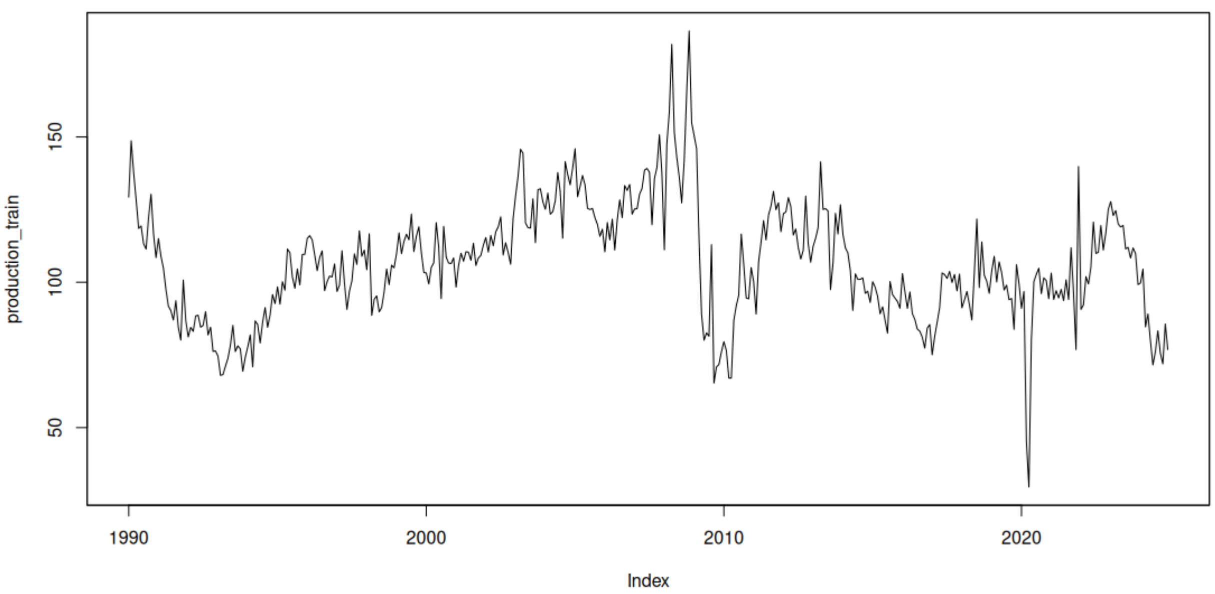

Figure 1: (X_t): The CVS-SJO production index of agricultural and forestry machinery (Base 100 in 2021)

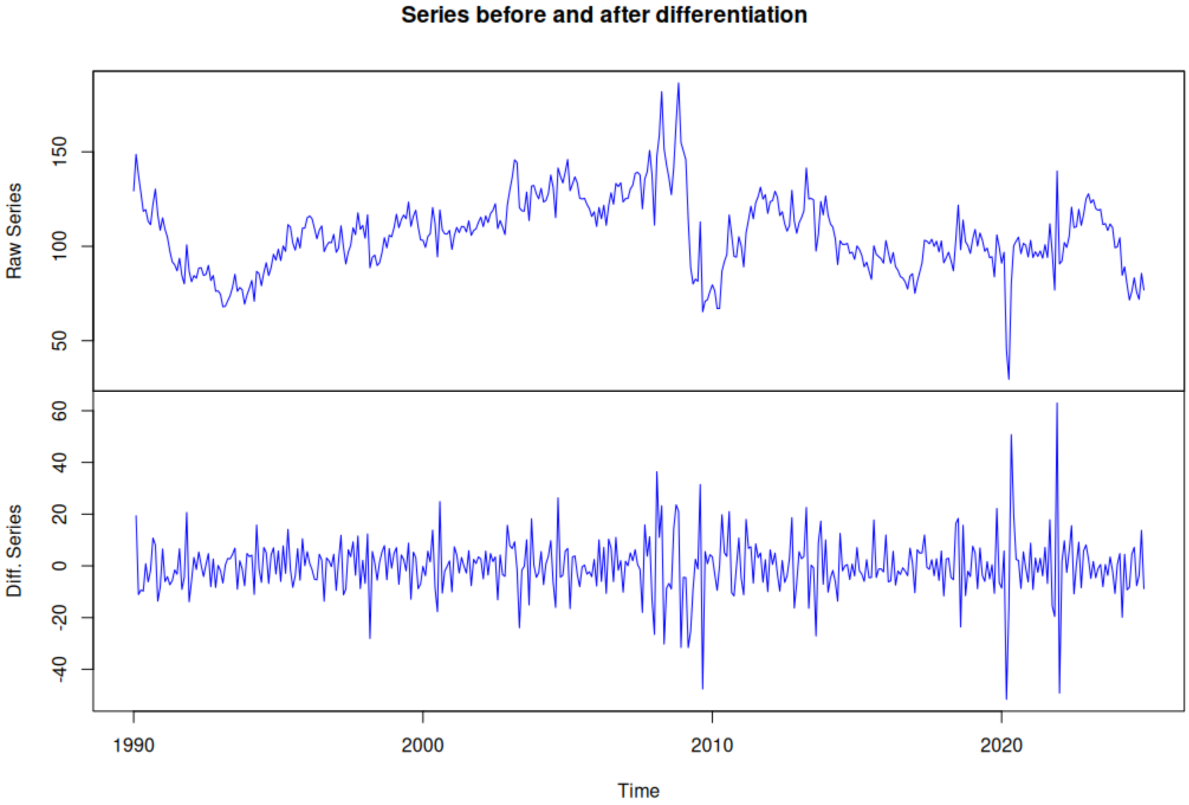

Figure 2: Comparison between (X_t) and ΔX_t

Table 1: Regression of (X_t) on time T and a constant

	Dependent variable:
	(X_t)
Time	-0.008
	(0.008)
Constant	107.917***
	(1.947)
Observations	422
R^2	0.002
Adjusted R^2	-0.0002
Residual Std. Error	19.962 (df = 420)
F Statistic	0.928 (df = 1; 420)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Unit Root Test Results on the raw series

Component	**ADF Test**	**PP Test**	**KPSS Test**
Null Hypothesis	Unit root	Unit root	Stationarity
Specification	Constant only	Drift, no trend	Drift, no trend
Test Statistic	–3.4913	–49.8	0.399
p-value	< 0.01	≤ 0.01	0.0775
Lag used	11	5	4

Table 3: Regression of (ΔX_t) on time T and a constant

	Dependent variable:
	(ΔX_t)
Time	-0.0003
	(0.004)
Constant	-0.060
	(1.078)
Observations	421
R^2	0.00001
Adjusted R^2	-0.002
Residual Std. Error	11.005 (df = 419)
F Statistic	0.006 (df = 1; 419)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Unit Root Test Results on the differenciated Series

Component	**ADF Test**	**PP Test**	**KPSS Test**
Null Hypothesis	Unit root	Unit root	Stationarity
Specification	No constant	No drift, no trend	No drift, no trend
Test Statistic	–12.8521	–442	0.0904
p-value	< 0.01	≤ 0.01	≥ 0.10
Lag used	4	5	4

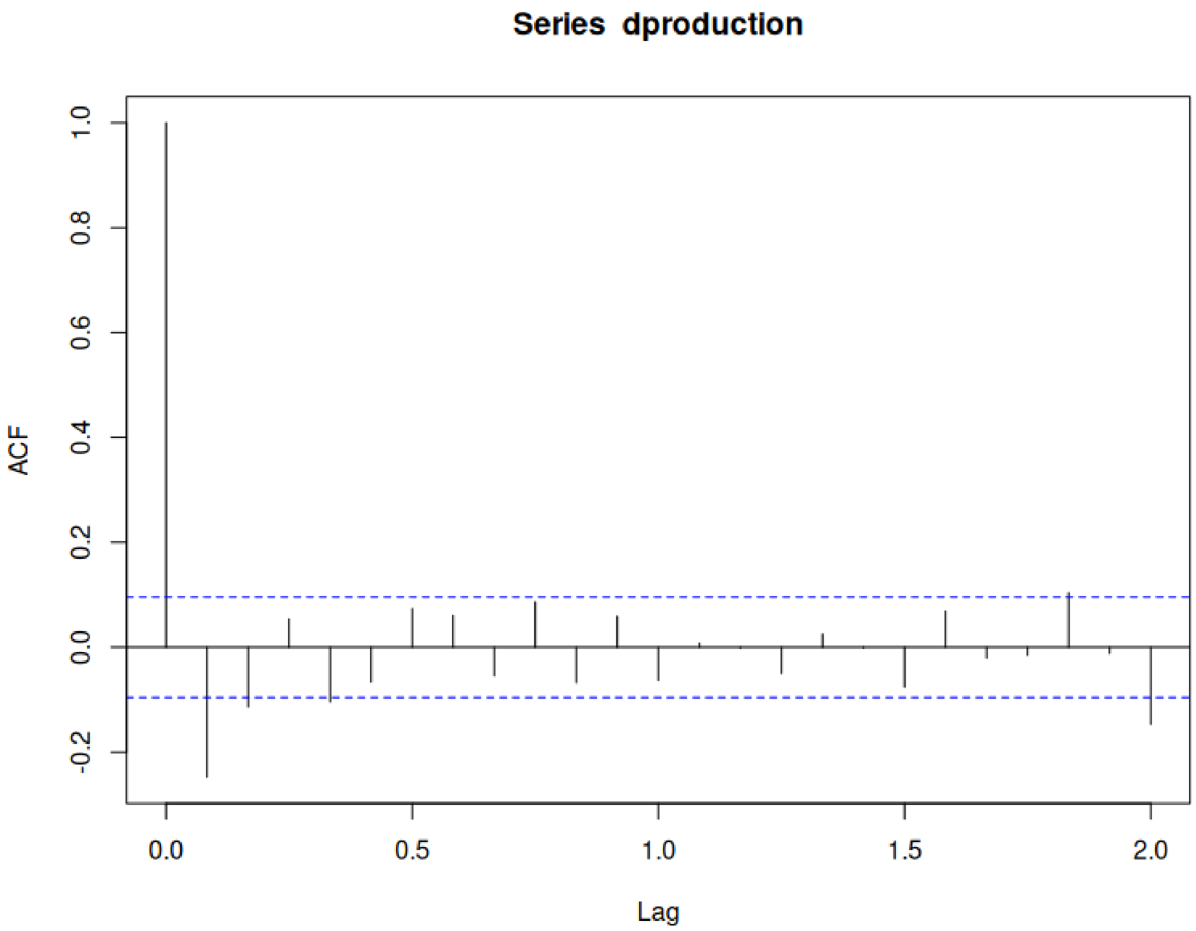

Figure 3: ACF for the differentiated series

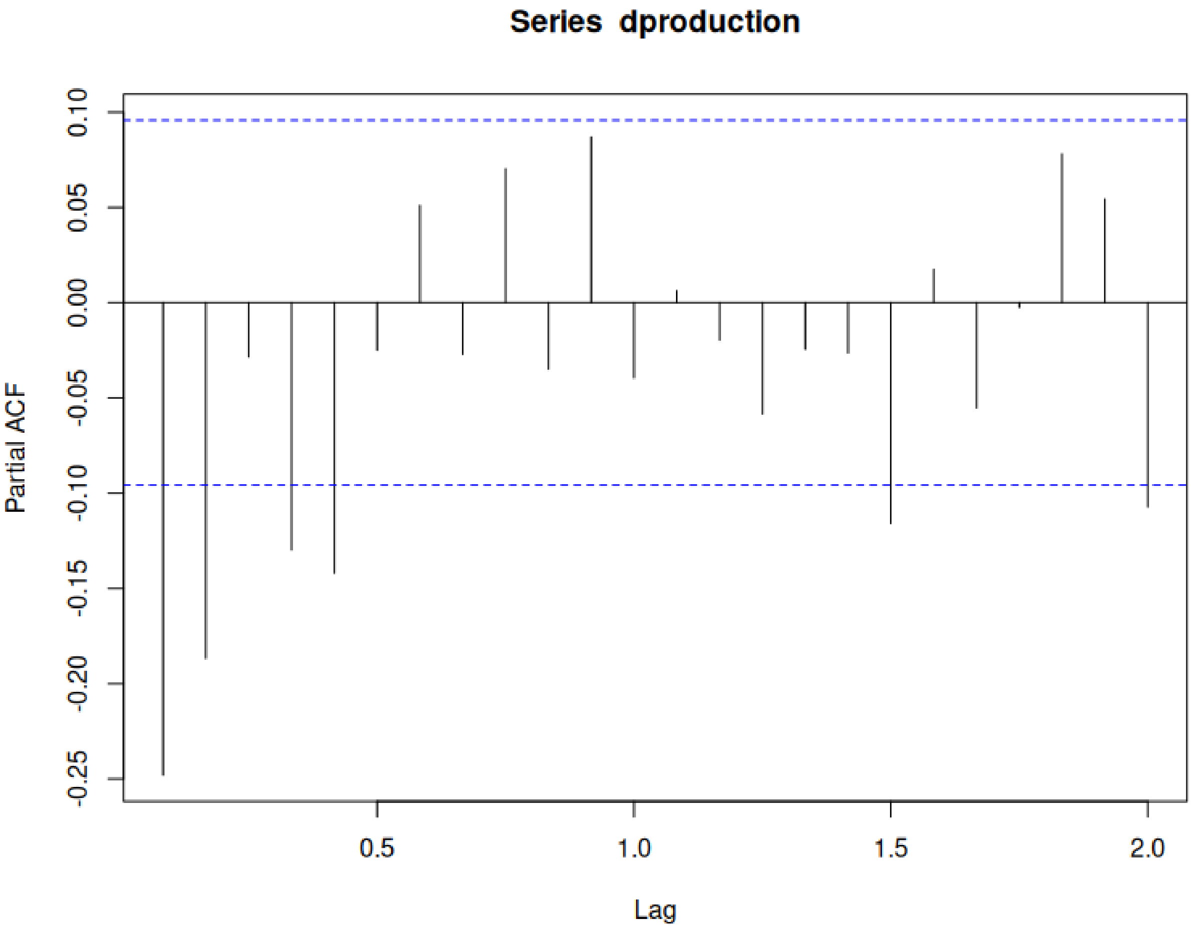

Figure 4: PACF for the differentiated series

Table 5: Akaike Information Criterion

	q=0	q=1	q=2
p=0	3200.06	3163.03	3156.45
p=1	3175.31	3156.29	3158.24
p=2	3162.55	3158.26	3159.52
p=3	3164.25	3155.96	3157.75
p=4	3159.29	3155.89	3155.00
p=5	**3152.89**	3154.76	3156.20

Table 6: Bayesian Information Criterion

	q=0	q=1	q=2
p=0	3204.10	3171.10	3168.57
p=1	3183.38	**3168.40**	3174.39
p=2	3174.66	3174.41	3179.71
p=3	3180.40	3176.14	3181.98
p=4	3179.48	3180.12	3183.26
p=5	3177.11	3183.02	3188.50

Table 7: Estimated Coefficients of ARIMA(5,1,0)

	AR(1)	**AR(2)**	**AR(3)**	**AR(4)**	**AR(5)**
Estimate	-0.3210	-0.2283	-0.0961	-0.1718	-0.1415
Std. Error	0.0484	0.0503	0.0513	0.0504	0.0486

Residual variance ($\hat{\sigma}^2$):	105.4
Log-likelihood:	-1570.44
AIC:	3152.89

Table 8: Autocorrelation of the residuals for ARIMA(5,1,0)

	lag	pval
1	1.00	
2	2.00	
3	3.00	
4	4.00	
5	5.00	0.85
6	6.00	0.98
7	7.00	0.63
8	8.00	0.70
9	9.00	0.52
10	10.00	0.52
11	11.00	0.50
12	12.00	0.40
13	13.00	0.46
14	14.00	0.50
15	15.00	0.41
16	16.00	0.49
17	17.00	0.55
18	18.00	0.51
19	19.00	0.56
20	20.00	0.62
21	21.00	0.68
22	22.00	0.59
23	23.00	0.63
24	24.00	0.17

Table 9: Autocorrelation of the residuals for ARIMA(1,1,1)

	lag	pval
1	1.00	
2	2.00	0.72
3	3.00	0.45
4	4.00	0.20
5	5.00	0.21
6	6.00	0.14
7	7.00	0.10
8	8.00	0.15
9	9.00	0.10
10	10.00	0.11
11	11.00	0.14
12	12.00	0.11
13	13.00	0.15
14	14.00	0.18
15	15.00	0.15
16	16.00	0.19
17	17.00	0.23
18	18.00	0.17
19	19.00	0.18
20	20.00	0.22
21	21.00	0.27
22	22.00	0.20
23	23.00	0.22
24	24.00	0.03

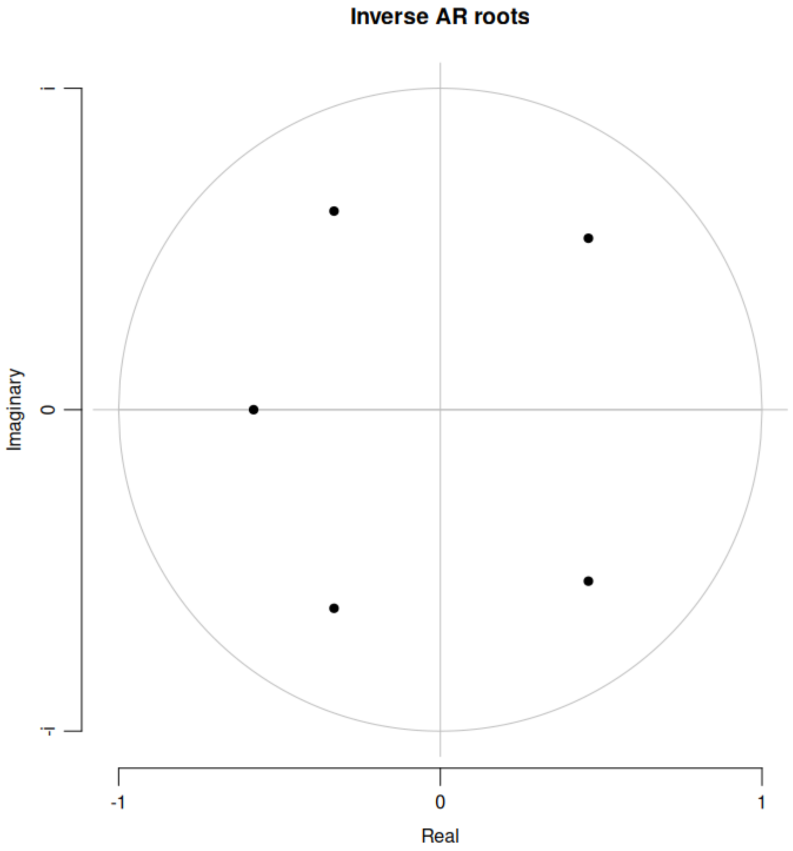

Figure 5: Inverse roots for the selected ARIMA model

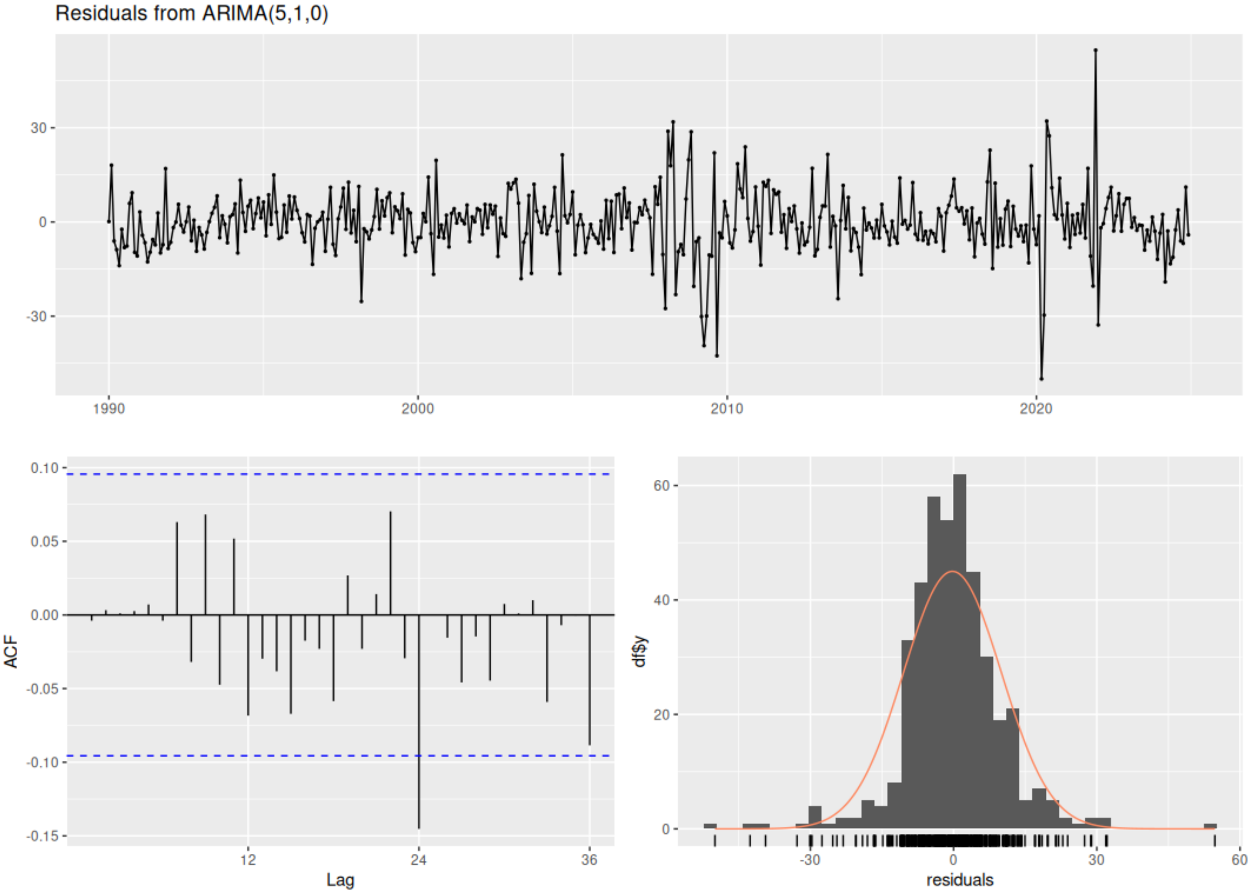

Figure 6: Normality of the residuals for the selected ARIMA model

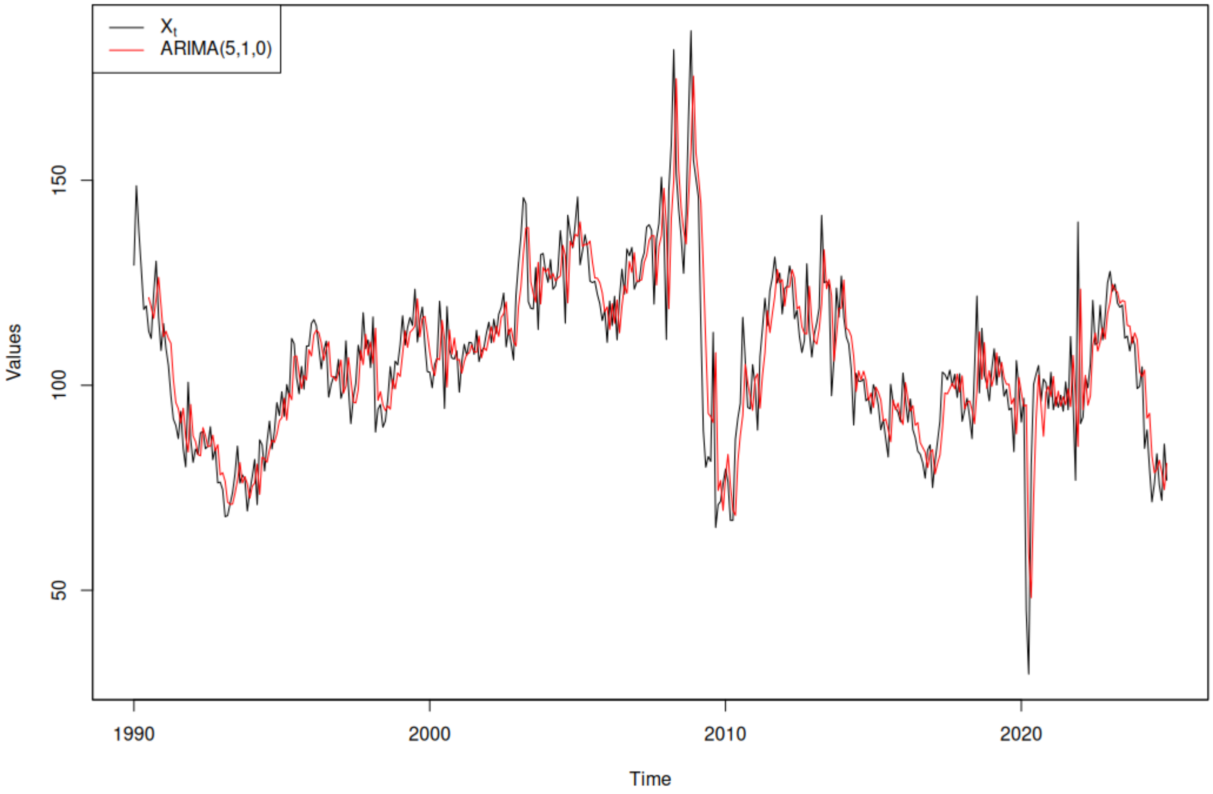

Figure 7: Plot of the selected ARIMA(5,1,0) against the actual series

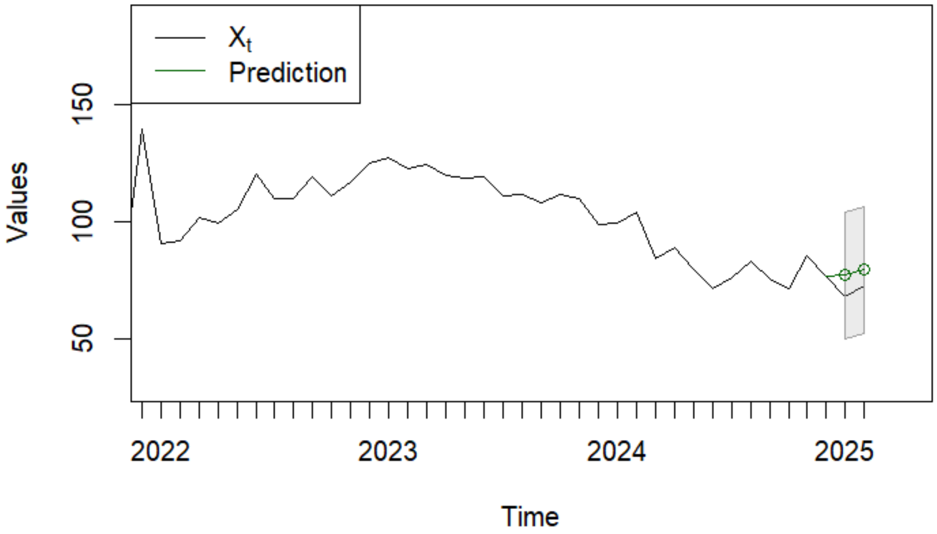

Figure 8: Forecasted values and their 95% confidence intervals, with actual values for comparison