Assignment : Linear Times Series

Authors :
Maël TREMOUILLE | Charles Bouchet

1 The Data

Question 1

The chosen series (X_t) corresponds to Indice CVS-CJO de la production industrielle (base 100 en 2021) - Travail des grains (NAF rév. 2, niveau classe, poste 10.61) from January 1990 to March 2025. The data series corresponds to the seasonally and calendar-adjusted industrial production index (CVS-CJO), base 100 in 2021, for grain milling activities (NACE Rev. 2, class level, code 10.61). We drop the two last values to make predictions later.

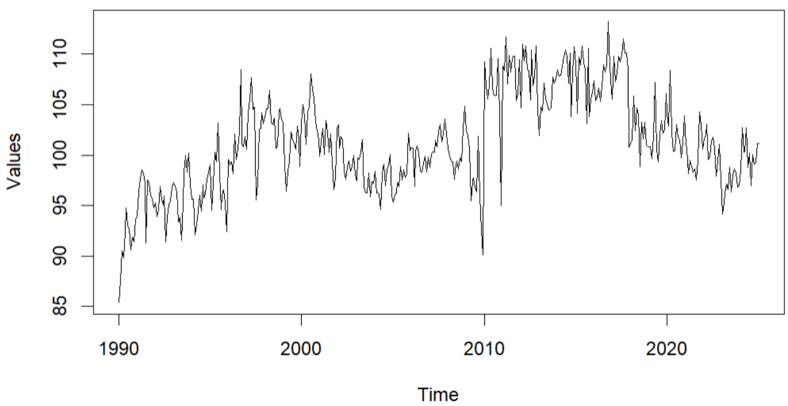


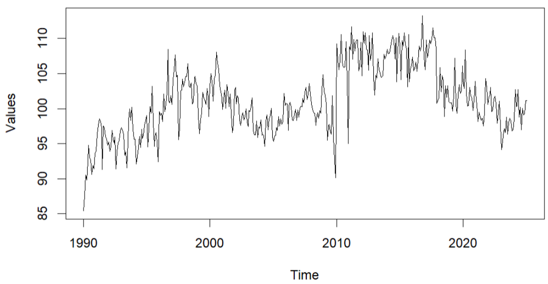
FIGURE 1 – Seasonally and calendar-adjusted industrial production index (base 100 in 2021) – Grain milling

Question 2

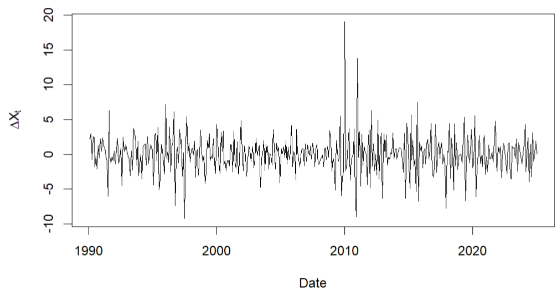
We begin by examining whether the series is stationary. To get a preliminary idea, we start by plotting the ACF (Figure A.1) . We observe that the ACF decays too slowly for the series to be considered stationary, suggesting the presence of a linear trend. This can be addressed by applying a first-order differencing : $\Delta X_t = X_t - X_{t-1}$.

We observe (see Figure A.2) that the differenced series exhibits an ACF more consistent with stationarity. A linear regression of the values on time (see Figure A.3) suggests the absence of a linear trend, which supports this assumption. We then perform stationarity tests such as the Augmented Dickey-Fuller test (see Figure A.4), the Phillips-Perron test, and the KPSS test (see Figure A.5). None of these tests contradict the stationarity of the differenced series. We can therefore display the differenced series.

Question 3



(a) Before



(b) After

FIGURE 2 – Comparison Before and After transformation

2 ARMA models

Question 4

Since we now consider the differenced series ΔX_t stationary, we set $d = 1$. We will model an ARMA(p, q) process on the differenced series, which corresponds to an ARIMA$(p, 1, q)$ model on the original series X_t.

The total serial correlation functions are significant (i.e., greater than the bounds $\pm\frac{1.96}{\sqrt{n}}$ of the confidence interval of the test of nullity of the autocorrelation at 95%) until $q^* = 1$, and partial serial correlations until $p^* = 5$ (see Figure A.6). We make a choice to ignore the peaks far after q^* and p^*. If our corrected series follows an ARMA(p, q), then the orders p and q are necessarily such that $p \leq p^*$ and $q \leq q^*$.

We begin by identifying the best model according to the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both criteria indicate the same optimal model (i.e., the one with the lowest value), namely with $p = q = 1$ (see Figure A.7). The coefficient estimates for this model can be seen in Figure 3. All coefficients are statistically significant at all usual levels ; both models are well-adjusted to our corrected series.

Finally, to ensure the validity of these models, we must verify that the residuals are not serially correlated. We conduct a Ljung-Box test, where the null hypothesis is the joint nullity of the autocorrelations up to a given order k. We test for residual autocorrelation for all $k \in \{1, 2, \dots, 24\}$.

The results of the Ljung-Box tests are presented in Figure A.8. For the model, all p-values remain above 0.05, except for lag 24, indicating that the null hypothesis of joint nullity of autocorrelations is not rejected for any $k \leq 23$. Autocorrelation is thus generally not detected. Although the null is rejected at lag 24, the corresponding residual autocorrelation is not substantial based on the residuals' ACF plot. This suggests that the violation is minor and does not invalidate the overall model (see Figure A.9).

The ARIMA$(1, 1, 1)$ model has significant coefficients, minimizes the information criteria, and its residuals are generally not correlated. It therefore appears to be the best model — this is the one we will use going forward.

```
==========================================================
             Estimated coefficients for ARIMA(1,1,1)
--------------------------------------------------------
             Coefficient  Std.Error  t.value  p.value
ar1              0.391      0.072      5.398    0.000
ma1             -0.829      0.046    -18.040    0.000
==========================================================
Note: p-values rounded to 3 decimals. All coefficients significant at the 1% level.
```

FIGURE 3 – Estimated coefficients for ARIMA(1,1,1) model

Question 5

We adopt the convention used in R's `arima()` function, where the ARIMA(p, d, q) model is defined as :

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d X_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t,$$

with B the backshift operator (i.e., $BX_t = X_{t-1}$) and ε_t a white noise error term. Under this convention, AR coefficients enter the characteristic polynomial with negative signs, and MA coefficients with positive signs.

We choose to model the differenced series $\Delta X_t = X_t - X_{t-1}$ as an ARMA(1,1) process. This implies that the original series X_t follows an ARIMA(1,1,1) model.

The theoretical form of the ARMA(1,1) model for ΔX_t is :

$$\Delta X_t = \phi_1 \Delta X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

where ε_t is a white noise error term.

With the estimated coefficients $\hat{\phi}_1 = 0.391$ and $\hat{\theta}_1 = -0.829$, the model can be explicitly written as :

$$\Delta X_t = 0.391\,\Delta X_{t-1} + \varepsilon_t - 0.829\,\varepsilon_{t-1}$$

Therefore, the original series X_t follows the ARIMA(1,1,1) model, which can be expressed as :

$$(1 - \phi_1 B)(1 - B)X_t = (1 + \theta_1 B)\varepsilon_t,$$

or, substituting the estimated coefficients,

$$(1 - 0.391B)(1 - B)X_t = (1 - 0.829B)\varepsilon_t,$$

where B is the backshift operator defined by $BX_t = X_{t-1}$.

We can display (see Figure A.10) the original time series along with the fitted values from the ARIMA(1,1,1) model.

3 Predictions

Question 6

To construct a confidence region of level $1 - \alpha$ for the pair of future values (X_{T+1}, X_{T+2}) predicted by the ARIMA(1,1,1) model, we assume that the forecast errors follow a bivariate normal distribution centered at zero with covariance matrix Σ.

Let $\widehat{X}_{T+1}$ and $\widehat{X}_{T+2}$ denote the forecasts of X_{T+1} and X_{T+2} respectively. Then, the confidence region of level $1 - \alpha$ is given by the set of points $(x_1, x_2) \in \mathbb{R}^2$ such that :

$$\begin{bmatrix} x_1 - \widehat{X}_{T+1} \\ x_2 - \widehat{X}_{T+2} \end{bmatrix}^\top \Sigma^{-1} \begin{bmatrix} x_1 - \widehat{X}_{T+1} \\ x_2 - \widehat{X}_{T+2} \end{bmatrix} \le \chi^2_{2,1-\alpha}$$

This equation defines an elliptical region in the plane, centered at the vector of forecasts $(\widehat{X}_{T+1}, \widehat{X}_{T+2})$, with shape and orientation determined by the covariance matrix Σ, which is given by :

$$\Sigma = \begin{bmatrix} \mathrm{Var}(\widehat{e}_{T+1}) & \mathrm{Cov}(\widehat{e}_{T+1}, \widehat{e}_{T+2}) \\ \mathrm{Cov}(\widehat{e}_{T+1}, \widehat{e}_{T+2}) & \mathrm{Var}(\widehat{e}_{T+2}) \end{bmatrix}$$

The quantile $\chi^2_{2,1-\alpha}$ corresponds to the $1 - \alpha$ level of the chi-squared distribution with 2 degrees of freedom.

In practice, the elements of Σ are computed from the model's structure and estimated parameters, taking into account the accumulation of uncertainty as we forecast further into the future.

This confidence region allows us to assess the joint uncertainty associated with the predictions for X_{T+1} and X_{T+2}, and to determine whether future observations fall within a statistically expected range.

Question 7

The construction of the confidence region for the future values (X_{T+1}, X_{T+2}) is based on the following key assumptions :

1. **Correct model specification :** The ARIMA(1,1,1) model correctly describes the dynamics of the observed time series. This means the model structure and its estimated parameters accurately capture the dependence in the data.
2. **Gaussianity of residuals :** The innovations (or error terms) of the model are assumed to be independently and identically distributed (i.i.d.) according to a normal distribution :

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

   This ensures that forecast errors are also normally distributed and allows us to use the chi-squared distribution in defining the elliptical confidence region.
3. **Linearity of the forecasting equations :** The future values X_{T+1} and X_{T+2} are linear combinations of past observations and error terms, which enables analytical computation of the mean vector and covariance matrix of forecast errors.
4. **Stationarity of the differenced series :** The original time series X_t is assumed to be integrated of order 1, i.e., $\Delta X_t = X_t - X_{t-1}$ is stationary. This is required for the ARIMA(1,1,1) model to be valid.
5. **Asymptotic normality of forecast errors :** Even if the errors are not strictly normal, the forecast errors are assumed to be asymptotically normal for large samples due to the central limit theorem. This justifies the use of the chi-squared distribution in finite-sample inference.

Violations of these assumptions may lead to confidence regions that do not have the correct coverage probability, hence reducing the reliability of inference.

Question 8

Figure A.11 displays the 95% joint confidence region for the predicted values $(\hat{X}_{T+1}, \hat{X}_{T+2})$ produced by the ARIMA(1,1,1) model. The region takes the shape of an ellipse centered on the forecast point. Its geometry reflects the model's assumptions and the joint uncertainty surrounding short-term predictions. In particular, the narrow width of the ellipse suggests that the model's predictive power is strong at this short horizon, and the near-circular shape implies a weak correlation between the one- and two-step ahead forecast errors.

Figure A.12 complements this analysis by showing the observed series in black and the forecasted values $\hat{X}_{T+1}$ and $\hat{X}_{T+2}$ in blue, along with their individual 95% confidence intervals. The transition from observed to predicted values is smooth, indicating that the model extends the existing pattern without structural discontinuity. The slightly downward slope of the forecasted segment reflects a recent short-term trend captured by the model. The confidence bands widen slightly from X_{T+1} to X_{T+2}, which is consistent with the expected increase in forecast uncertainty.

Together, these two figures illustrate the coherence of the ARIMA model's short-term forecasting and the reasonable bounds of its uncertainty. The consistency between the individual and joint representations confirms the model's adequacy and the reliability of the inferential procedures used.

Question 9

To determine whether the value of Y_{T+1} helps improve the prediction of X_{T+1}, we need to assess whether Y_t contains predictive information about X_t. A standard way to test this is through Granger causality.

If Y_t Granger-causes X_t, then the past values of Y_t improve the prediction of X_t beyond what is achieved using the past of X_t alone. In this case, having access to Y_{T+1} would allow us to form a better forecast for X_{T+1}, for instance by including it as an exogenous regressor in an ARIMAX model.

To test for Granger causality, we can run a joint regression including lagged values of both X_t and Y_t, and check whether the coefficients of the Y_t lags are jointly significant.

Annexes

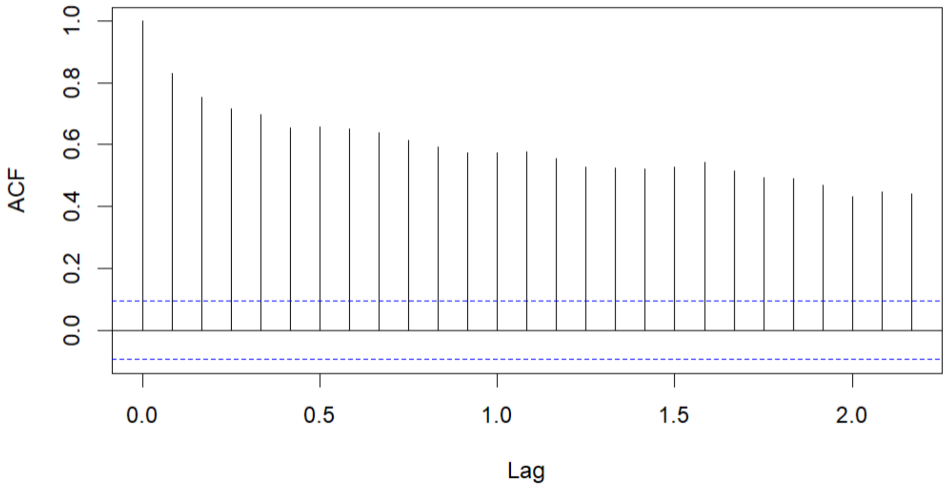


Figure A.1 : Autocorrelation function (ACF) of the original series

click here.

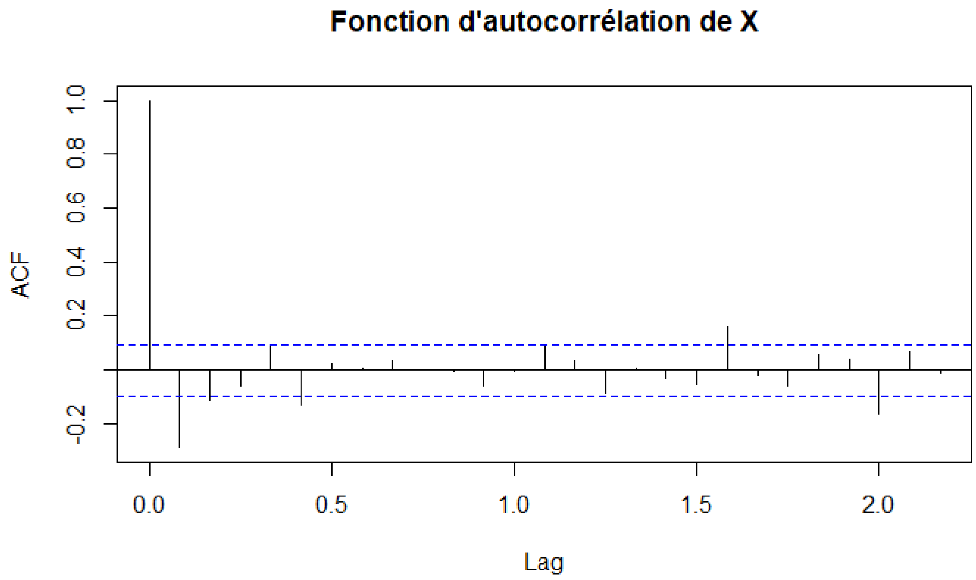


Figure A.2 : Autocorrelation function (ACF) of the differenced series

click here.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.1505	0.2690	0.559	0.576
Time	-0.0065	0.0133	-0.485	0.628

Residuals :
Min = -9.3525 1Q = -1.4436 Median = 0.0588 3Q = 1.3827 Max = 19.0583

Residual standard error : 2.761 on 418 degrees of freedom

Multiple R^2 **:** 0.00056 **Adjusted** R^2 **:** -0.00183

F-statistic : 0.235 on 1 and 418 DF ***p*-value :** 0.628

Table A.3 : Results of linear regression `valeur` $\sim$ `time`

click here.

```
Augmented Dickey-Fuller Test
Alternative hypothesis: stationary

Type 1: no drift, no trend
 lag   ADF     p.value
  0   -27.5     0.01
  1   -20.4     0.01
  2   -17.3     0.01
  3   -13.6     0.01
  4   -13.5     0.01
  5   -12.3     0.01

Type 2: with drift, no trend
 lag   ADF     p.value
  0   -27.5     0.01
  1   -20.4     0.01
  2   -17.3     0.01
  3   -13.5     0.01
  4   -13.5     0.01
  5   -12.3     0.01

Type 3: with drift and trend
 lag   ADF     p.value
  0   -27.5     0.01
  1   -20.4     0.01
  2   -17.3     0.01
  3   -13.6     0.01
  4   -13.5     0.01
  5   -12.3     0.01

Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Figure A.4 : Augmented Dickey-Fuller Test results (types 1, 2 and 3)

click here.

```
Phillips-Perron Unit Root Test
Alternative hypothesis: stationary

Type 1: no drift, no trend
 lag  Z_rho  p.value
  5    -441     0.01
-----

Type 2: with drift, no trend
 lag  Z_rho  p.value
  5    -441     0.01
-----

Type 3: with drift and trend
 lag  Z_rho  p.value
  5    -440     0.01
---------------
Note: p.value = 0.01 means p.value <= 0.01


KPSS Unit Root Test
Alternative hypothesis: nonstationary

Type 1: no drift, no trend
 lag   stat   p.value
  4    0.225     0.1
-----

Type 2: with drift, no trend
 lag   stat   p.value
  4   0.0724     0.1
-----

Type 3: with drift and trend
 lag   stat   p.value
  4   0.0189     0.1
-----------
Note: p.value = 0.01 means p.value <= 0.01
    : p.value = 0.10 means p.value >= 0.10
```

Figure A.5 : Phillips-Perron and KPSS unit root test results (types 1, 2 and 3)

click here.

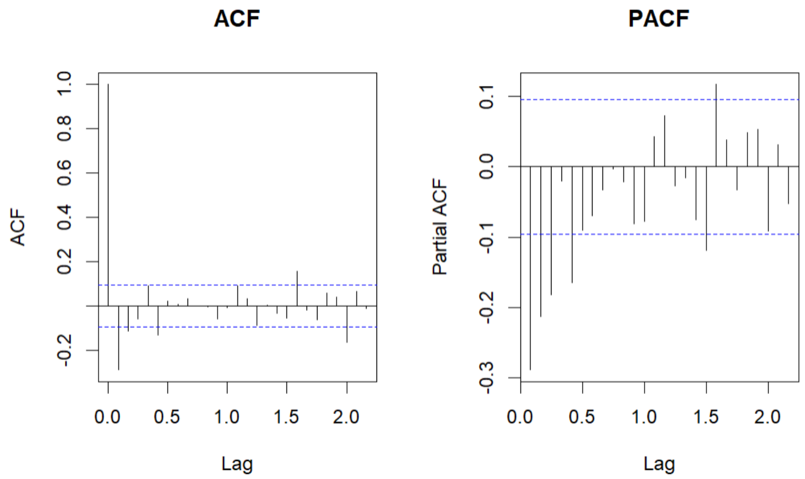


Figure A.6 : ACF and PACF for the differenced series ΔX_t

click here.

```
AIC criterion:
          q=0      q=1      q=2      q=3      q=4      q=5
p=0 2045.462 1983.588 1968.588 1967.244 1969.184 1965.056
p=1 2011.006 1964.947 1966.930 1968.815 1970.623 1966.388
      q=0   q=1   q=2   q=3   q=4   q=5
p=0 FALSE FALSE FALSE FALSE FALSE FALSE
p=1 FALSE  TRUE FALSE FALSE FALSE FALSE

BIC criterion:
          q=0      q=1      q=2      q=3      q=4      q=5
p=0 2049.502 1991.669 1980.708 1983.405 1989.385 1989.298
p=1 2019.086 1977.067 1983.091 1989.016 1994.865 1994.670
      q=0   q=1   q=2   q=3   q=4   q=5
p=0 FALSE FALSE FALSE FALSE FALSE FALSE
p=1 FALSE  TRUE FALSE FALSE FALSE FALSE
```

Figure A.7 : AIC and BIC values for different ARMA(p, q) models

click here.

```
Ljung-Box test p-values for lags 1 to 24:

Lag_01 : 0.00000000    Lag_02 : 0.94188521    Lag_03 : 0.99703204
Lag_04 : 0.41079314    Lag_05 : 0.15285415    Lag_06 : 0.23872020
Lag_07 : 0.33532772    Lag_08 : 0.39669175    Lag_09 : 0.50288035
Lag_10 : 0.59720562    Lag_11 : 0.57529584    Lag_12 : 0.66337166
Lag_13 : 0.48262655    Lag_14 : 0.53417371    Lag_15 : 0.41283486
Lag_16 : 0.47723302    Lag_17 : 0.50691200    Lag_18 : 0.54521655
Lag_19 : 0.14108331    Lag_20 : 0.17939486    Lag_21 : 0.19398615
Lag_22 : 0.21299843    Lag_23 : 0.25789910    Lag_24 : 0.03742285
```

Figure A.8 : Ljung-Box test p-values for residual autocorrelation up to lag 24

click here.

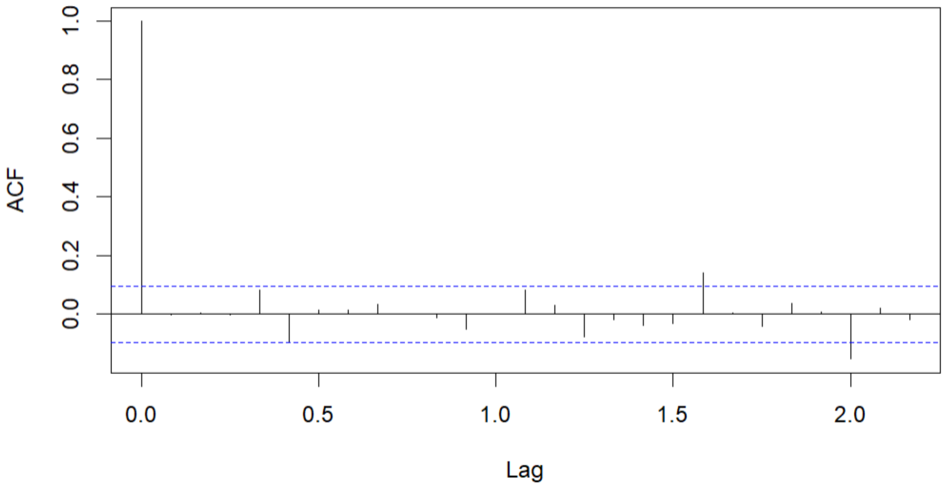


Figure A.9 : Autocorrelation function (ACF) the residuals of ARIMA(1,1,1)

click here.

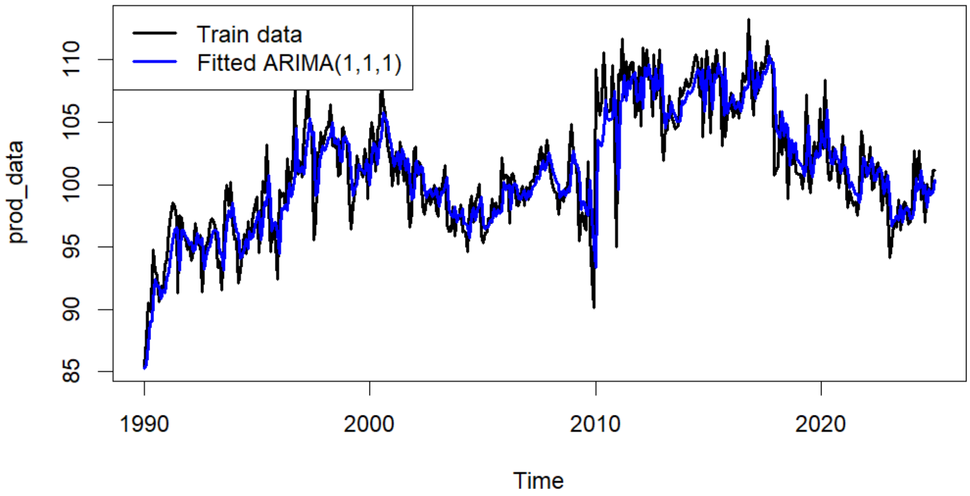


Figure A.10 : Time series and ARIMA(1,1,1) fitted values

click here.

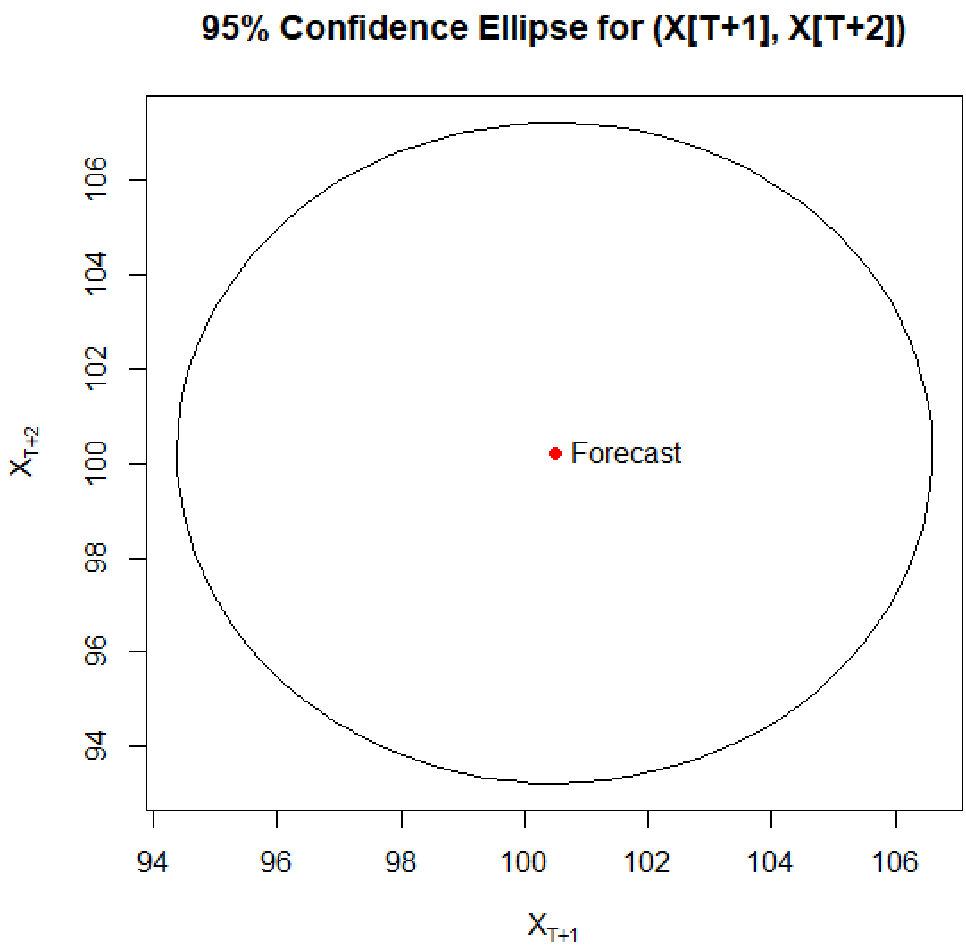


Figure A.11 : Joint 95% confidence region for (X_{T+1}, X_{T+2})

click here.

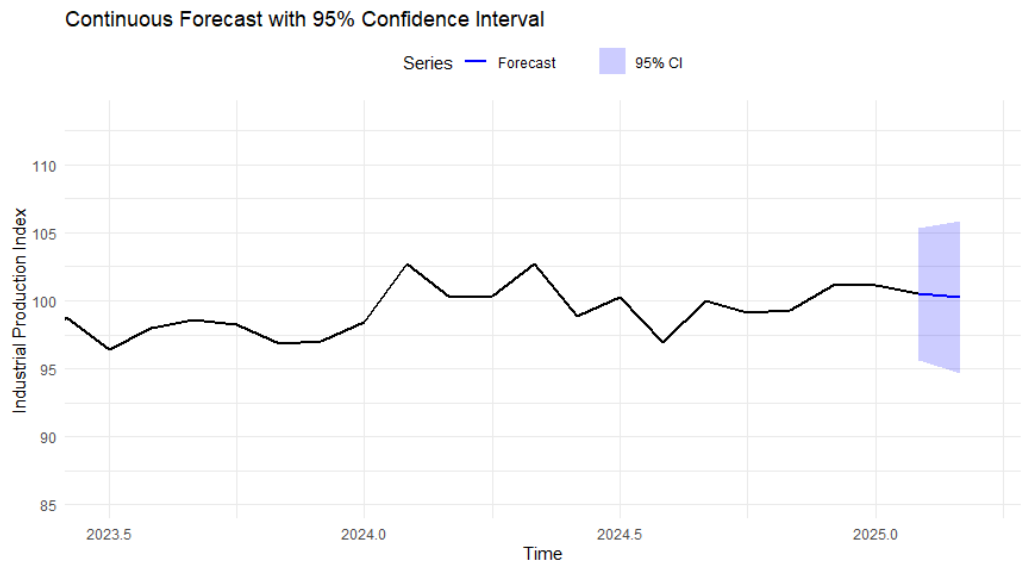


Figure A.12 : Observed series with ARIMA(1,1,1) forecasts and 95% confidence intervals

click here.