Time series project

Project report 2024

Implementation of an ARIMA model to predict the electricity production index in France

Authors : Alquier Matteo - Hotellier Erwann

May 2024

Contents

Here is the link to our github repositorie : https://github.com/malquier/Time-Series—French-electricity-production

I Presentation and preprocessing of the dataset

I.1 Dataset presentation

We have chosen to work on the index of electricity production in mainland France. INSEE provides monthly data on the CVS-CJO index of electricity production in France[1]. The data we have used have been corrected to avoid seasonal variations (CVS) and the number of working days (CJO). This correction allows a better understanding of the fundamental evolution of this index. The published indices have a base year of 2021 and are in base 100. For this study, we will use the Box-Jenkins methodology, which consists of in identifying and estimating an ARIMA model. It involves several stages: stationarization, model identification, parameter estimation and model verification. The series takes monthly values from January 1990 to March 2024. We plot the initial series :

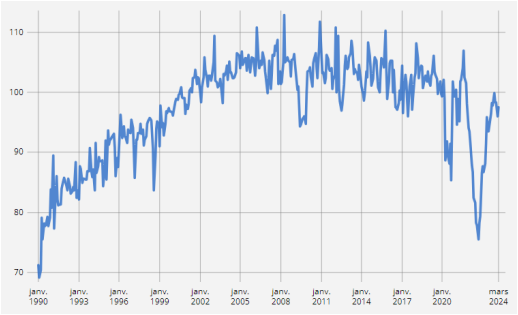


Figure 1: Trend in electricity production index in mainland France

We can see that our series is not stationary. We could use some tests to see if the series is non-stationary. But first, we'd like to know whether a linear trend emerges from this series, even if it seems not, or a quadratic trend. The coefficient for the linear trend (dates) and the coefficient for the constant (intercept) are both non-zero and significant. We'll therefore need to consider the case of unit root tests with constant and trend.

We carry out a test on seasonality to make sure we have no seasonality and can therefore apply an ARIMA model. To do this, we've used an OCSB test to determine whether seasonal differentiation is necessary for the series to be stationary. The test consists in running the following regression:

$$\Delta_S y_t = \delta + \sum_{s=1}^{S-1} \phi_s D_{st} + \eta_t$$

where : $\Delta_S y_t$ is the seasonally differentiated series $(y_t - y_{t-S})$, δ is a constant, D_{st} are seasonality index, ϕ_s are seasonality coefficients, η_t is the error term.

This test then looks at the significance of the linear regression coefficients. If the hypothesis that the coefficients are not equal to zero is accepted at over 95%, then the series need not be seasonally differentiated.

We finally decided that we didn't need to differentiate the series seasonally.

We also decided to apply the augmented Dickey-Fuller (ADF) test in order to see if the serie have an unit root and so is non-stationary. We obtained the following results:

[1]For more information on the time serie, visit https://www.insee.fr/fr/statistiques/serie/010768228

```
        Augmented Dickey-Fuller Test

data:  Data1.ts
Dickey-Fuller = -1.3199, Lag order = 7, p-value = 0.8645
alternative hypothesis: stationary
```

Figure 2: ADF test result for the initial serie

The unit root is not rejected at a threshold of 95% for the current series. The series is therefore not stationary. We'll try to make the series stationary by applying differentiation methods.

I.2 Dataset processing

We will first try to differentiate it at order 1. We obtain the following result:

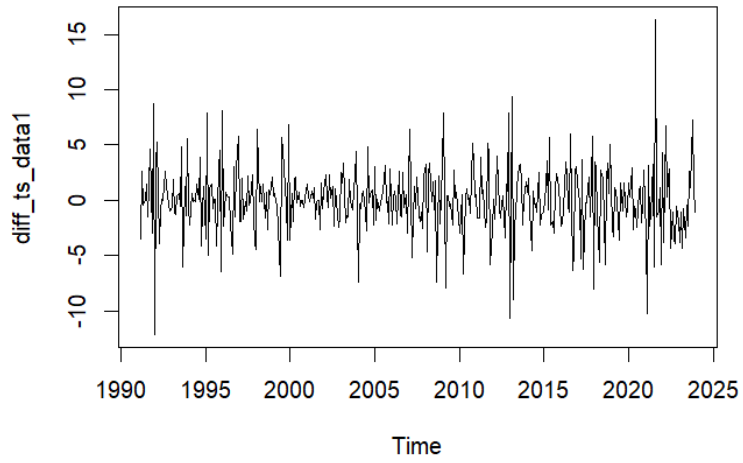


Figure 3: Plot of the differentiated series to order 1

We then noticed that the series appeared to be stationary. To check this, we reused an ADF test and a Phillips-Perron (PP) test. The results of these tests are presented in the appendix, along with a description of how the PP test works. We can see from Figure 7 that the p-value is 0.01, so we reject the hypothesis that the series has a unit root at 95 %. In other words, our series is indeed stationary. We could stop here for differentiation, but we wanted to know whether differentiating once again would result in an even more stationary series, and therefore a lower p-value for the ADF test than that obtained by differentiating to order 1. This was not the case, as we obtained p-values that were not necessarily lower when we differentiated the series to order 2 and 3 (Figure 8 and Figure 9. We have therefore retained the series differentiated at order 1.

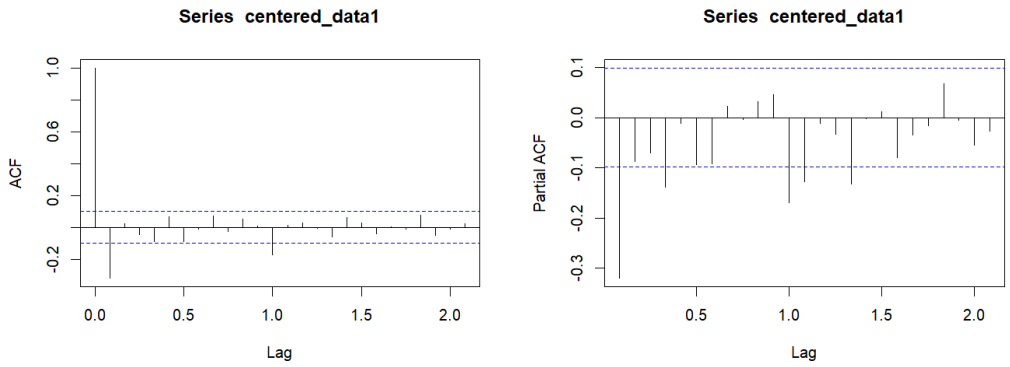
To make the model easier to use, more interpretable and more efficient, we have chosen to center our stationary series, by subtracting the value of each term from the series mean.

II Implementation of the ARIMA model

II.1 ARMA model selection and validation

The aim of this section is to find the values of the maximum orders p and q. To do this, we use the autocorrelation function (ACF) to find q and the partial autocorrelation function (PACF) to find p.

We note that for the ACF Figure4a, only the first delay shows significant total auto-correlation. We'll therefore take $qmax = 1$. What's more, for the PACF Figure4b, we can see that only the fourth delay is significant, so we'll take $pmax = 4$.



(a) Auto-correlation function (b) Partial Auto-correlation function

The most likely models are $p \leq 4$ and $q \leq 1$. We've created an algorithm to choose which combination of p and q would be the best to model the series. To do this, we'll use two tests:

- coefficients nullity test: if the test is validated, we say that the model is well fitted
- the test for absence of auto-correlation of residuals, or the portmanteau test: if the test is validated, we say that the model is valid (here, we wish to accept H_0, so we check that the p-values are greater than 0.1).

Modèle	p-value Ljung-Box	AR	MA	Intercept
ARIMA(2, 0, 1)	0.1641	0.5159, 0.1800	-0.9084	0.0001
ARIMA(3, 0, 1)	0.1657	0.5149, 0.1845, -0.0131	-0.9052	0.0000
ARIMA(4, 0, 1)	0.205	0.5067, 0.1857, -0.0032, -0.026	-0.8975	-0.0006

Table 1: Selected ARIMA models where ljung-box test and p,q significant

Finally, we drew up tables representing the different BIC and AIC values of all the models, in order to select according to the two penalization criteria. The following tables show that the same model minimizes both criteria.

	$q=0$	$q=1$
$p=0$	2021.115	1974.209
$p=1$	1980.475	1973.107
$p=2$	1979.440	**1966.404**
$p=3$	1979.469	1968.344
$p=4$	1973.715	1970.112

Table 2: AIC values

	$q=0$	$q=1$
$p=0$	2029.068	1987.138
$p=1$	1992.404	1989.012
$p=2$	1995.345	**1986.286**
$p=3$	1999.351	1992.202
$p=4$	1997.573	1997.947

Table 3: BIC values

So we choose the ARIMA(2,0,1) model.

II.2 Express the ARIMA(p,d,q) model for the selected series

Before we can express the ARIMA model followed by the initial series, we need to check that the model we've just selected (ARIMA(2,0,1)) corresponds to a causal model. To do this, we represent the roots of its associated polynomial on the unit disk:

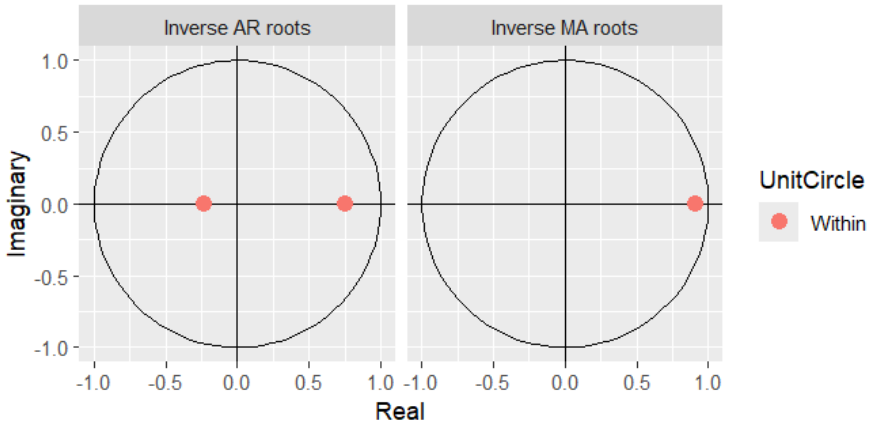


Figure 5: Representation of inverses of roots on the unit circle

The roots are all inside the unit circle, so our starting series follows an ARIMA(2,1,1) model defined by:

The model of our original series is therefore given by an ARIMA(2,1,1) model represented by the following equation:

$$X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \theta_1 \epsilon_{t-1} + \epsilon_t$$
$$X_t = 0.0193 + 0.5159 X_{t-1} + 0.1800 X_{t-2} - 0.9084 \epsilon_{t-1} + \epsilon_t$$

III Forecast

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

Let X_t denote our differentiated series. Let T be its length. X_t follows an ARMA(2,1) model, so it verifies :

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \mu\epsilon_t - \theta_1\epsilon_{t-1}$$

The best linear forecast of X_{t+1} is $\tilde{X}_{t+1}$ such that :

$$\begin{aligned}\tilde{X}_{t+1} &= \mathrm{E}[X_{t+1}|X_t, X_{t-1}, \ldots, X_1] \\ &= \mathrm{E}[\phi_1 X_t + \epsilon_{t+1} - \psi_1\epsilon_t|X_t, X_{t-1}, \ldots, X_1] \\ &= \phi_1 X_t - \psi_1\epsilon_t\end{aligned}$$

Indeed, $\mathrm{E}[\epsilon_{t+1}|X_t, X_{t-1}, \ldots, X_1] = 0$ because ϵ_t is an innovation and $\mathrm{E}[\epsilon_t|X_t, X_{t-1}, \ldots, X_1] = \epsilon_t$. Similarly, the best linear forecast of X_{t+2} est $\tilde{X}_{t+2}$ such that :

$$\begin{aligned}\tilde{X}_{t+2} &= \mathrm{E}[X_{t+2}|X_t, X_{t-1}, \ldots, X_1] \\ &= \mathrm{E}[\phi_1 X_{t+1} + \epsilon_{t+2} - \psi_1\epsilon_{t+1}|X_t, X_{t-1}, \ldots, X_1] \\ &= \phi_1\tilde{X}_{t+1}\end{aligned}$$

Now let's calculate the prediction errors, defined as follows: $e_{t+h} = X_{t+h} - \tilde{X}_{t+h}$
$e_{t+1} = X_{t+1} - \tilde{X}_{t+1} = \epsilon_{t+1}$
$e_{t+2} = \phi_1(X_{t+1} - \tilde{X}_{t+1}) + \epsilon_{t+2} - \psi_1\epsilon_{t+1} = \phi_1\epsilon_{t+1} + \epsilon_{t+2} - \psi_1\epsilon_{t+1} = (\phi_1 - \psi_1)\epsilon_{t+1} + \epsilon_{t+2}$

The residuals are Gaussian and decorrelated, so we deduce the variance and covariance of the prediction errors:

$$\begin{cases} V(e_{t+1}) = V(\epsilon_{t+1}) = \sigma_\epsilon^2 \\ V(e_{t+2}) = V((\phi_1 - \psi_1)\epsilon_{t+1}) + V(\epsilon_{t+2}) + 2\mathrm{Cov}((\phi_1 - \psi_1)\epsilon_{t+1}, \epsilon_{t+2}) = (1 + (\phi_1 - \psi_1)^2)\sigma_\epsilon^2 \\ \mathrm{Cov}(e_{t+1}, e_{t+2}) = \mathrm{Cov}(\epsilon_{t+1}, (\phi_1 - \psi_1)\epsilon_{t+1} + \epsilon_{t+2}) = (\phi_1 - \psi_1)\sigma_\epsilon^2 \end{cases}$$

Thus, assuming that innovationsϵ are strong white Gaussian noise, we have :

$$\begin{pmatrix} e_{t+1} \\ e_{t+2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right)$$
$$\text{où } \Sigma = \begin{pmatrix} 1 & (\phi_1 - \psi_1) \\ (\phi_1 - \psi_1) & 1 + (\phi_1 - \psi_1)^2 \end{pmatrix}$$

SO $\mathrm{Det}(\Sigma) = \sigma_\epsilon^4 > 0$ according to the assumption made, so Σ is an invertible matrix.

We note $X := \begin{pmatrix} X_{t+1} \\ X_{t+2} \end{pmatrix}$ and $\tilde{X} := \begin{pmatrix} \tilde{X}_{t+1} \\ \tilde{X}_{t+2} \end{pmatrix}$.

The bivariate confidence region of level α is written :

$$R_{1-\alpha} = \{x|(x - \tilde{X})^T\Sigma^{-1}(x - \tilde{X}) \leq q\chi^2(2)(1 - \alpha)\}$$

III.2 Specify the assumptions used to obtain this region

To obtain this confidence region, we made two assumptions:

- Innovations ϵ_t are strong white Gaussian noise: $\epsilon \sim N(0, \sigma_\epsilon^2)$ (iid) with $\sigma_\epsilon^2 > 0$.
- The theoretical model is identified, i.e. the theoretical coefficients are identical to the estimated coefficients. The estimated variance $\hat{\sigma}^2$ is equal to the theoretical variance σ^2.

III.3 Graph this region for $\alpha = 95\%$, comment.

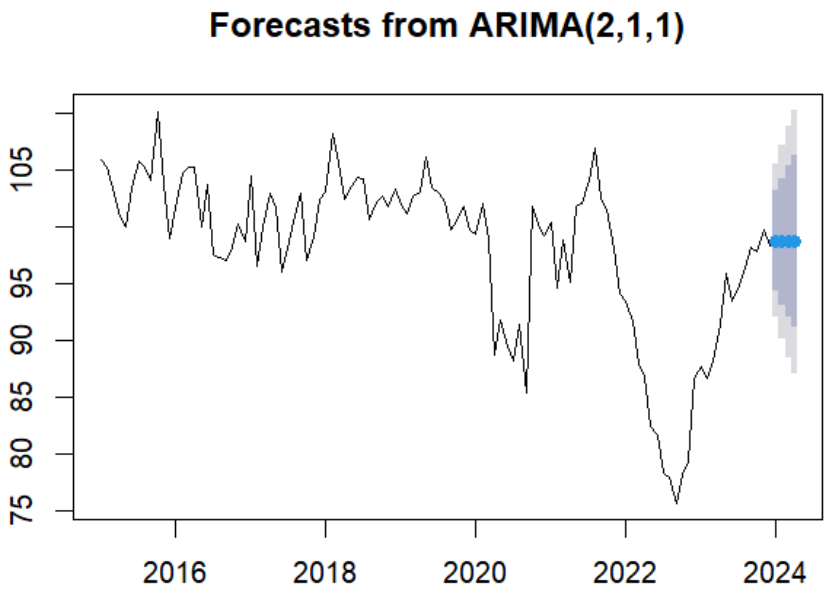


Figure 6: Forecast for non-differentiated series

We've represented here (with the blue dots) the prediction made over the next 4 months of the series, assuming it follows an ARIMA(2,1,1) model. The 95% confidence region corresponds to the grey area surrounding these 4 points.

III.4 Open question : Let Y_t be a stationary series available from $t = 1$ to T. It is assumed that Y_{T+1} is available more quickly than X_{T+1}. Under what condition(s) can this information improve the forecast of X_{T+1}? How would you test it?

If we assume that Y_{T+1} is available faster than X_{T+1}, then we can use the information from Y_{T+1} to improve the prediction of X_{T+1} if and only if the variable Y_t instantaneously causes, in Granger's sense, the variable X_t.

In our case, Y_t causes X_t in the Granger sense if and only if :

$$\mathbb{E}[X_{T+1}|((X_T, Y_T), (X_{T-1}, Y_{T-1}), \ldots) \cup Y_{T+1}] \neq \mathbb{E}[X_{T+1}|((X_T, Y_T), (X_{T-1}, Y_{T-1}), \ldots)]$$

To do this, we can perform a Granger test, which compares the predictions of X_t with and without Y_{T+1}.

IV Appendix

IV.1 Explication of Augmented Dickey-Fuller test

The Augmented Dickey-Fuller (ADF) test checks if a time series X_t is non-stationary due to a unit root. The test uses the following regression:

$$\Delta X_t = c + bt + \beta X_{t-1} + \sum_{l=1}^{k} \phi_l \Delta X_{t-l} + \epsilon_t$$

where:

- $\Delta X_t = X_t - X_{t-1}$ (first difference of X_t)
- c is a constant
- bt is a trend term
- βX_{t-1} is the lagged level
- $\sum_{l=1}^{k} \phi_l \Delta X_{t-l}$ are lagged differences
- ϵ_t is the error term

Hypothesis:

- Null (H_0): $\beta = 0$ (unit root, non-stationary)
- Alternative (H_1): $\beta < 0$ (stationary)

The test statistic is:

$$\tau = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$$

Compare τ to critical values. If τ is less than the critical value, reject H_0, indicating stationarity.

IV.2 Explication of Phillips-Perron test

The Phillips-Perron (PP) test checks for a unit root in a time series, accounting for serial correlation and heteroskedasticity in the error terms. The test uses the regression model:

$$\Delta X_t = \beta X_{t-1} + \epsilon_t$$

where:

- $\Delta X_t = X_t - X_{t-1}$ (first difference)
- ϵ_t is the error term

Hypotheses:

- Null (H_0): $\beta = 0$ (unit root, non-stationary)
- Alternative (H_1): $\beta < 0$ (stationary)

The test statistic is:

$$Z_\alpha = \frac{\hat{\beta}}{\mathrm{SE}(\hat{\beta})}$$

where $\mathrm{SE}(\hat{\beta})$ is adjusted for serial correlation and heteroskedasticity using the Newey-West estimator. Compare Z_α to critical values. If Z_α is less than the critical value, reject H_0, indicating stationarity.

IV.3 ADF tests Results - Section II

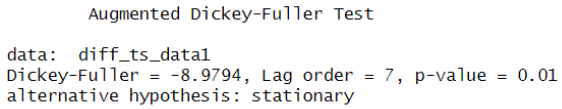


Figure 7: ADF test result on our differentiated series in order 1

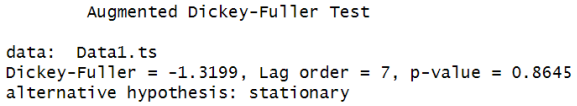


Figure 8: Results of the ADF test on our series differentiated by order 2

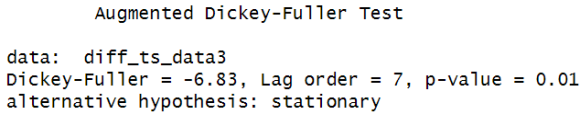


Figure 9: Results of the ADF test on our series differentiated by order 3

IV.4 Residue distribution

We wish to verify the assumption made in section III.1, i.e. that the residuals are Gaussian. To do this, we have plotted the density of the residuals and the density of the normal distribution. We can see that the residuals do not necessarily follow a normal distribution. To find out whether the distribution of the residuals is close to a normal distribution, we use the Jarque and Bera test, which checks whether the data follow a normal distribution. We have the following results, from which we can deduce that we reject the null hypothesis that the data have a normal distribution by 1%. We can therefore question the hypothesis made in Part III.

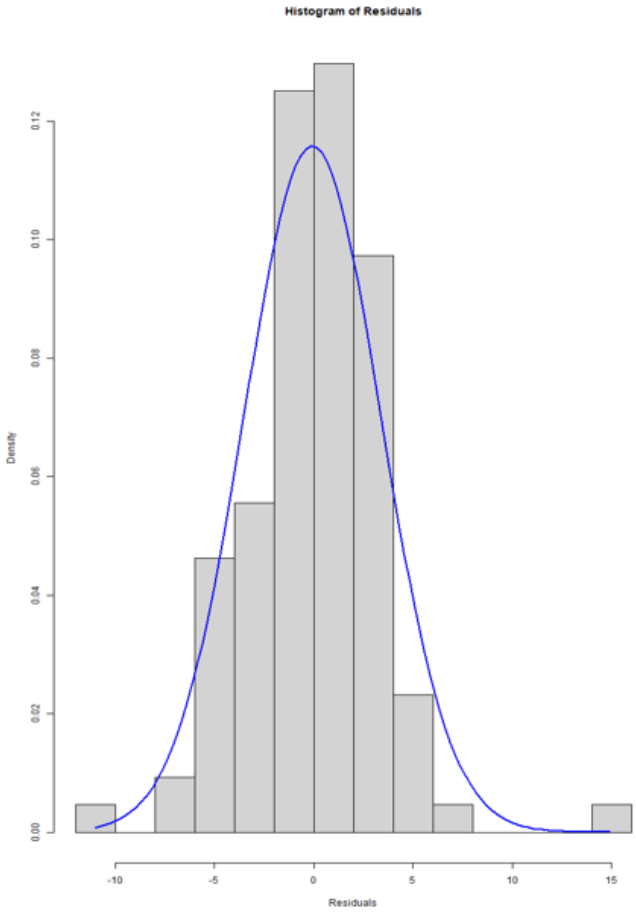


Figure 10: Residuals density

```
        Jarque Bera Test

data:  modele$residuals
X-squared = 37.289, df = 2, p-value = 7.993e-09
```

Figure 11: Jarque and Bera test