Time Series Analysis of Tool Manufacturing in France from 2002 to 2024

Nils Peyrouset and Malo Evain
ENSAE Paris

May 2024

Abstract

In this report, we will study the behavior and prediction of a time series, specifically the manufacturing of tools in France over more than 10 years. Our work will be divided into three parts. First, we will analyze and transform the data to make the series stationary. Next, we will attempt to fit ARMA and ARIMA models to the series under study. Finally, we will conclude with a prediction task based on past data. Throughout this work, we will validate our results using various statistical tests such as the ADF, KPSS, and autocorrelation of residuals tests.

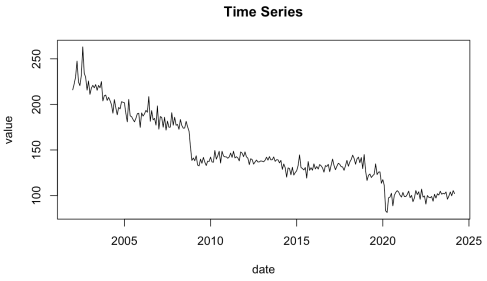
Contents

1 The Data

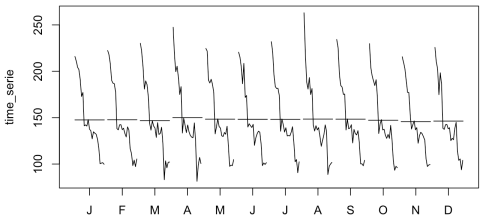
1.1 What do the data represent?

The series we are studying in this report represents the industrial production index related to tool manufacturing. The data were extracted from the INSEE website, which you can find at this link. The industrial production indices allow us to track the monthly evolution of industrial activity in France. The studied series includes 267 observations from January 2002 to March 2024 with a monthly frequency.

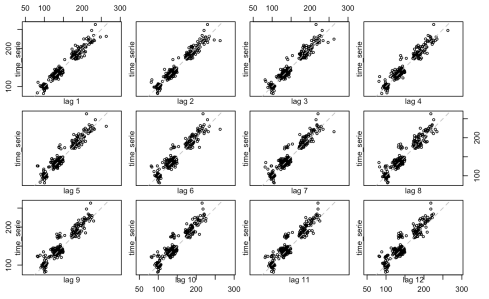
The first task we performed was to plot graphs to analyze the series.



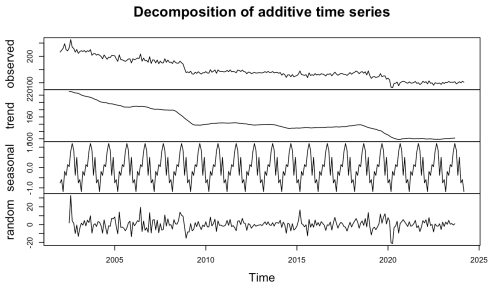
(a) Evolution of the series over time



(b) Month plot



(c) Lag plot

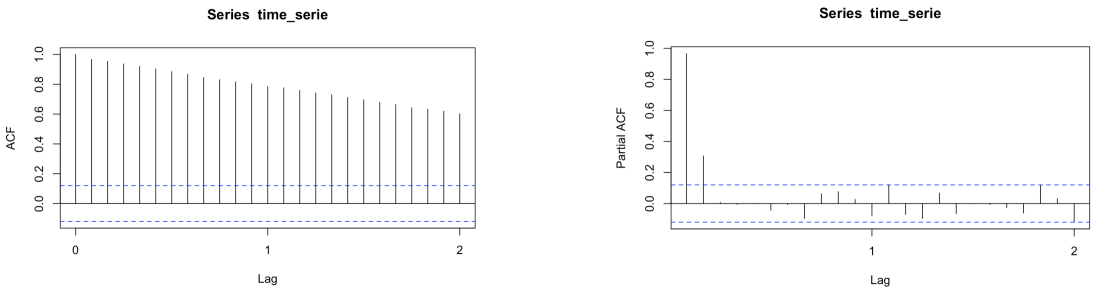


(d) Decomposition of the series

Figure 1: Graphical analysis of the time series

- Figure(a): By plotting the series over time, we notice that there is likely a negative trend. However, the variance does not seem to depend on time, which may guide us in differencing the model to achieve a stationary series. There also does not seem to be any seasonality.
- Figure(b): The month plot shows 12 similar monthly patterns, suggesting an absence of seasonality.
- Figure(c): The lag plot shows a strong correlation between the variables.
- Figure(d): The error represented by the decomposition seems to be constant over time, which confirms the additive model.

Moreover, by regressing the series on its past values, we obtain a coefficient of -0.45, which confirms our hypothesis that the series has a decreasing trend. However, we cannot confirm the significance of this coefficient because the test is not valid in the presence of possibly autocorrelated residuals. Therefore, we will consider a constant and trend scenario for performing the ADF tests.



(a) ACF of the series (b) PACF of the series

Figure 2: ACF and PACF of the series

Here, we can see that:

- The PACF does not show any repeated pattern, particularly no repetition of significant values. Thus, the series does not seem to exhibit seasonality.
- The autocorrelations decrease very gradually, and the first-order partial autocorrelation is close to 1. Therefore, the series does not seem to be stationary.

The graphical analysis informs us about the nature of the series; however, it is far from sufficient. That is why we will perform the KPSS and ADF tests to

confirm our intuition that the series is not stationary. The results are reported in the following table 1.

Table 1: Results of the different tests on the series

Test	Stats	Lag	p-value
KPSS	0.602632	5	≤ 0.01
ADF	-2.4001	21	0.4074

The KPSS test allows us to test the stationarity of the series (the null hypothesis). Here, the results show a small p-value. Therefore, we reject the stationarity hypothesis at the 1% level. The ADF test allows us to show the existence of a unit root in the case with a trend, and therefore the non-stationarity of the series. The large p-value does not allow us to reject the unit root hypothesis at the 5% level.

1.2 Making the Series Stationary

To make the series stationary, we will start by eliminating the trend by differencing: $X_t = \Delta S_t = S_t - S_{t-1}$, where S_t is our original series. This time, by regressing X_t on X_{t-1}, we obtain a coefficient of 0.0019, which is a good sign for a zero linear trend. However, once again, we cannot rely on this regression, so we will perform the two previous tests (KPSS and ADF).

We obtain the results in Table 2 2.

Table 2: Results of the different tests on the differenced series

Test	Stats	Lag	p-value
KPSS	0.024977	5	$0.1 \geq 0.1$
ADF	-3.8349	21	0.017

The p-value of the KPSS test is well above 0.05, which allows us not to reject the stationarity hypothesis with 95% confidence. Additionally, the p-value of the ADF test is well below

0.05, which allows us to reject the unit root hypothesis at the 5% level. So we can confirm that the differenced series is stationary.

1.3 Before-After treatment

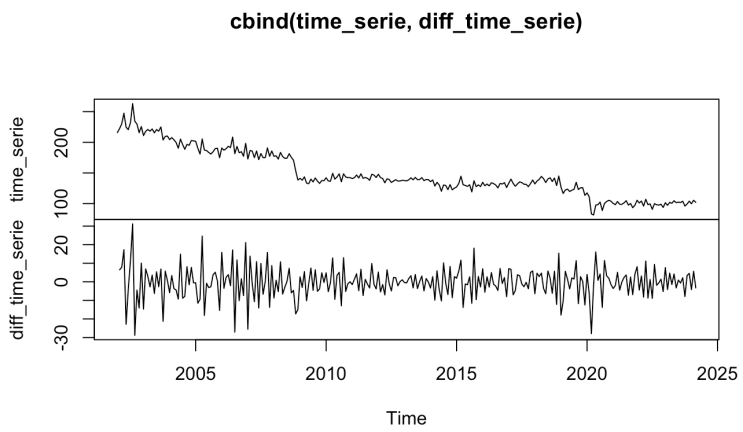
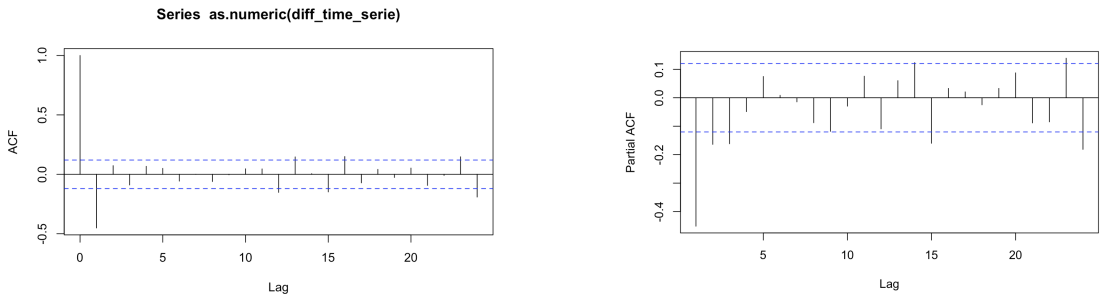


Figure 3: series before and after treatment

2 ARMA and ARIMA Models

2.1 ARMA Model

We now seek to identify the best models to fit our time series.



(a) ACF of the differenced series

(b) PACF of the differenced series

Figure 4: ACF and PACF of the differenced series

According to Figure 4, the ACF is significant up to lag 1 and the PACF up to lag 4, so we set $q_{\max} = 1$ and $p_{\max} = 4$.

To determine which parameters to choose, we minimize the two well-known

information criteria:

$$\text{AIC}(p,q) = \log(\hat{\sigma}^2) + \frac{2(p+q)}{n}$$
$$\text{BIC}(p,q) = \log(\hat{\sigma}^2) + \frac{(p+q)\log(n)}{n}$$

with $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\epsilon}_t^2$.

	q=0	q=1
p=0	1889.854	1822.247
p=1	1832.432	1823.378
p=2	1827.750	1825.342
p=3	1822.851	1824.722
p=4	1824.559	1825.582

Table 3: AIC

	q=0	q=1
p=0	FALSE	TRUE
p=1	FALSE	FALSE
p=2	FALSE	FALSE
p=3	FALSE	FALSE
p=4	FALSE	FALSE

Table 4: min AIC

	q=0	q=1
p=0	1893.437	1829.414
p=1	1839.599	1834.128
p=2	1838.501	1839.676
p=3	1837.185	1842.640
p=4	1842.476	1847.083

Table 5: BIC

	q=0	q=1
p=0	FALSE	TRUE
p=1	FALSE	FALSE
p=2	FALSE	FALSE
p=3	FALSE	FALSE
p=4	FALSE	FALSE

Table 6: min BIC

According to the AIC and BIC, the model to retain is an ARMA(0,1), which minimizes errors for both criteria. Thus, we have $X_t = \epsilon_t + \theta_1 \epsilon_{t-1}$.

For θ_1, we obtain:

Table 7: Values and significance of θ_1

Value	Standard Error	Significance
-0.5196	0.0497	-10.4

Since the significance is well below -1.96 (the 5

For the adjusted R^2, we find $\boxed{0.226526}$.

As shown in Table 8 in the appendix, we can reject the correlation of the residuals for the ARMA(0,1) model.

Thus, the ARMA(0,1) model meets all selection criteria, making it a perfect candidate for forecasting.

2.2 ARIMA Model

We differenced the initial series once to obtain the series X_t. So, $d = 1$. Thus, the model corresponding to our series is ARIMA(0,1,1). The model coefficients are exactly the same as in Table 7.

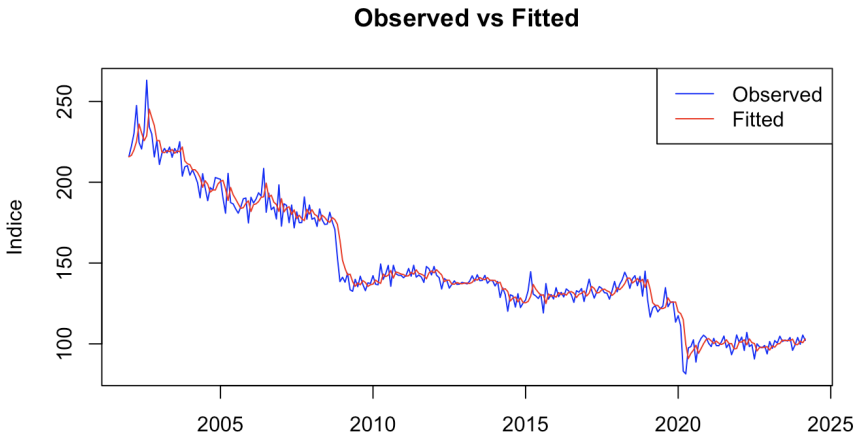


Figure 5: Forecast of the ARIMA(0,1,1) model

When representing the observed values and the predicted values, it seems that the chosen model fits our case well.

3 Prediction

3.1 Confidence Regions at Level α

We assume that the residuals are Gaussian, i.e., $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. We have an MA(1) model that we can write:

$$X_t = \epsilon_t - \theta_1 \epsilon_{t-1} \tag{1}$$

We have $E[\epsilon_{T+h}|X_T, X_{T-1}, \ldots] = 0 \ \forall h > 0$, and according to the course, we know that the optimal forecast at time T is given by:

$$\begin{cases} \hat{X}_{T+1|T} = -\theta_1 \epsilon_T \\ \hat{X}_{T+2|T} = 0 \end{cases}$$

Now, let's calculate the prediction errors $X_{T+1} - \hat{X}_{T+1|T}$ and $X_{T+2} - \hat{X}_{T+2|T}$. We have:

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

Thus:

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

Thus, we have $X - \hat{X} \sim \mathcal{N}(0, \Sigma)$ where Σ is the variance-covariance matrix such that:

$$\Sigma = \sigma_\epsilon^2 \begin{pmatrix} 1 & -\theta_1 \\ -\theta_1 & 1 + \theta_1^2 \end{pmatrix}$$

Since $\det(\Sigma) = \sigma_\epsilon^2$, the variance-covariance matrix is invertible if and only if $\sigma_\epsilon^2 > 0$, an assumption we have assumed true. According to the course, we finally obtain $(X - \hat{X})^\top \Sigma^{-1} (X - \hat{X}) \sim \chi^2(2)$. Thus, the confidence region is:

$$(X - \hat{X})^\top \Sigma^{-1} (X - \hat{X}) \leq \chi^2_{1-\alpha}(2)$$

Now, the 95

$$(X - \hat{X})^\top \Sigma^{-1} (X - \hat{X}) \leq \chi^2_{0.95}(2) \approx 5.99$$

Thus, we have:

$$\frac{1}{\sigma_\epsilon^2} (\epsilon_{T+1}^2 + \frac{(\epsilon_{T+2} - \theta_1 \epsilon_{T+1})^2}{1 + \theta_1^2}) \leq 5.99$$

In practice, this allows us to predict the values within a certain confidence level, taking into account the variability of our prediction errors.

4 Conclusion

In conclusion, through rigorous analysis and fitting of ARMA and ARIMA models, we've been able to identify the optimal model for our time series. The ARIMA(0,1,1) model, in particular, has shown significant predictive power and reliability, which was validated through both AIC and BIC criteria as well as confidence region calculations. These methods allowed us to create a model that not only fits the historical data but also provides accurate forecasts, crucial for informed decision-making.

5 Annexe

5.1 Figures

lag	pval
1	NA
2	NA
3	NA
4	NA
5	NA
6	0.027164068
7	0.047538145
8	0.031509912
9	0.060936930
10	0.084598640
11	0.129579925
12	0.079931840
13	0.036770046
14	0.058203939
15	0.031974955
16	0.019071280
17	0.028780238
18	0.039893737
19	0.057283065
20	0.079864115
21	0.049602350
22	0.065846794
23	0.064072732
24	0.006437511

Table 8: test de corrélation des résidus pour ARMA(0,1)