2nd year ENSAE

Linear time series assignment report

Monthly fabrication of french weapons and ammunition

Authors :
Romain Delhommais
Jacques Zhang

May 23, 2024

Contents

I The data

I.1 What does the chosen Series represent?

In this project, we will study the monthly series of the fabrication of french weapons and ammo from 1990 to 2024. This series is accessible on the INSEE website at the following address: `https://www.insee.fr/fr/statistiques/serie/010767975`.
The series is based on a dataset with the base value index of 100 in the year 2021. The series takes values monthly from January 1990 to February 2024. The initial series is plotted on the following graph:

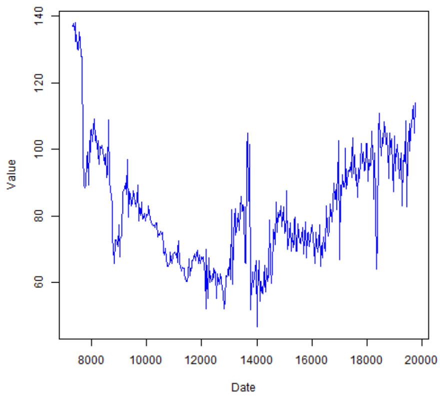

Figure 1: Initial series of the monthly fabrication of french weapons and ammo

We see that the series is not stationary, and it clearly has a trend.

I.2 Transformation of the series

We will differentiate in first order the series to remove the trend. $Diff = X_t - X_{t-1}$
Here is the plot of the differentiated series:

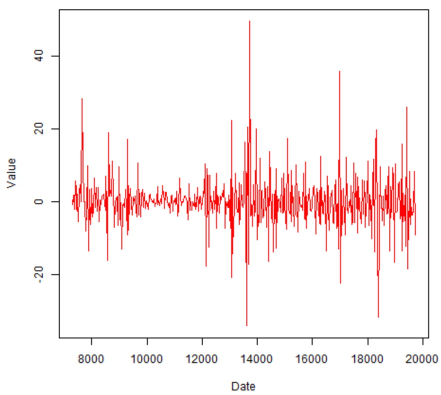

Figure 2: Differentiated series

The differentiated series seems to be stationary. We can already check the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the differentiated series to get an idea of the p and q parameters of the $ARMA$ model.

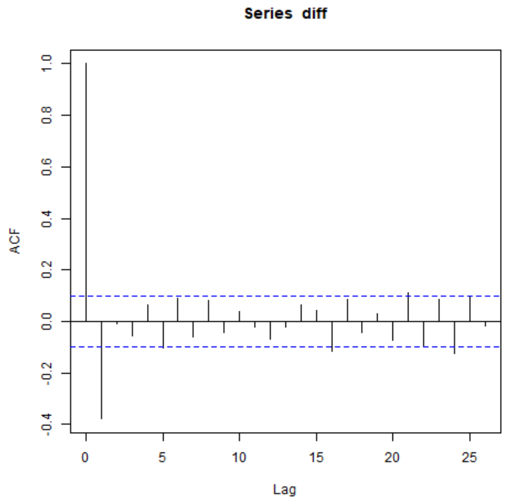

Figure 3: ACF of the differentiated series

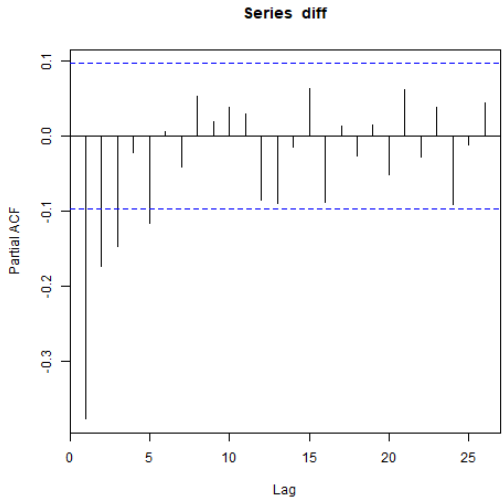

Figure 4: PACF of the differentiated series

To be sure of the stationarity, we will test it using the Augmented Dickey-Fuller (ADF) test, the Phillips-Perron (PP) test and the KPSS stationarity test. We remind that the null hypothesis of the ADF and PP tests is that the series is not stationary and the null hypothesis of the KPSS test is that the series is stationary. The results of these tests are as follows:

Test	Statistic	Lag Order	P-value
Augmented Dickey-Fuller	-19.88518	1	≤ 0.01
Phillips-Perron	-483.79055	5	≤ 0.01
KPSS	0.30456	5	≥ 0.10

Table 1: Summary of Statistical Tests

The ADF and PP tests gave us p-values inferior to 1 %, hence we can reject the null hypothesis at the 1 % level, and the KPSS test gave us a p-value superior to 10 % we cannot reject the null hypothesis at the 10 % level. Hence we can consider that the differentiated series is stationary.

Here are the series before and after transformation:

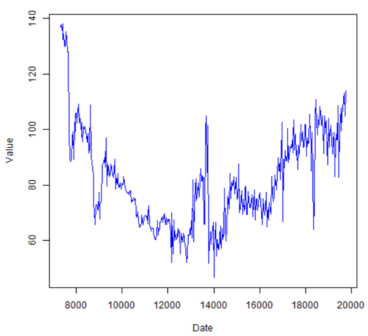

Figure 5: Initial series

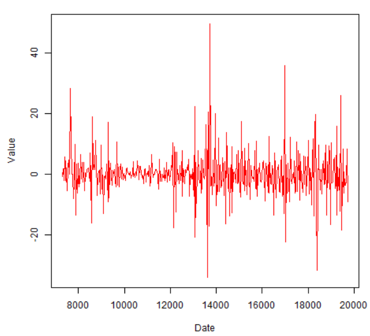

Figure 6: Transformed series

II $ARMA$ models

II.1 $ARMA$ Model Selection

We see that the tests of the stationarity consider a maximum lag of 5. Hence we will select our p and q parameters in the $ARMA$ model between 0 and 6. To get the best model, we will use the AIC and BIC criteria.

q \ p	0	1	2	3	4	5	6
0	2821, 2829	2760, 2772	2749, 2765	2742, 2763	2744, 2768	2741, 2769	2743, 2775
1	2738, 2750	2739, 2755	2741, 2761	2742, 2767	2744, 2772	2743, 2775	2739, 2775
2	2739, 2755	2741, 2761	2740, 2764	2742, 2770	2743, 2775	2744, 2780	2737, 2777
3	2741, 2761	2740, 2764	2735, 2763	2735, 2767	2745, 2781	2743, 2783	2742, 2786
4	2742, 2767	2741, 2769	2735, 2767	2737, 2774	2739, 2779	2745, 2789	2740, 2788
5	2744, 2772	2742, 2774	2738, 2774	2740, 2780	2734, 2778	2740, 2788	2749, 2801
6	2743, 2775	2744, 2780	2745, 2786	2745, 2789	2740, 2788	2738, 2790	2741, 2798

Table 2: AIC and BIC values for different combinations of p and q

For a model (p, q), it's considered a significant model if the absolute value of the fraction of the estimated coefficient and its standard error is greater than 1.96 for the highest coefficient associated to autoregressive and moving-average terms. In the table we represent this information with YES and NO as the first coordinate of our table. Furthermore, a model is considered as valid if the p value of the Ljung-Box test is superior to 0.05. We will represent this information with YES and NO as the second coordinate of our table.

q \ p	0	1	2	3	4	5	6
0	NO, NO	NO, YES	NO, YES	NO, YES	NO, YES	NO, YES	NO, YES
1	NO, NO	NO, YES	NO, YES	NO, YES	NO, YES	NO, YES	NO, YES
2	NO, NO	NO, YES	NO, YES	YES, YES	NO, YES	NO, YES	NO, YES
3	NO, YES	NO, YES	NO, YES	NO, YES	YES, YES	NO, YES	NO, YES
4	NO, YES	NO, YES	NO, YES	NO, YES	NO, YES	YES, YES	NO, YES
5	NO, YES	NO, YES	NO, YES	YES, YES	NO, YES	NO, YES	YES, YES
6	NO, YES	YES, YES	YES, YES	NO, YES	NO, YES	NO, YES	NO, YES

Table 3: Validation Results for $ARMA(p, q)$ Models

Valid models are highlighted in red.

By minimizing the AIC and BIC criteria on the valid models, we find that the model $ARMA(4, 5)$ is the best model according to the AIC criteria and the model $ARMA(2, 3)$ is the best model according to BIC criteria. We will test the significance of the coefficients of models $ARMA(4, 5)$ and $ARMA(2, 3)$ by a student test. We will have as null hypothesis that the coefficients of the model are equal to 0.

Looking at the p-value of the test (table 4 and 5), the model $ARMA(4, 5)$ is well adjusted as the coefficients are significant at the 1 % level. But the model $ARMA(2, 3)$ is not well adjusted

	P-value
ar1	< 0.01
ar2	< 0.01
ar3	< 0.01
ar4	< 0.01
ma1	< 0.01
ma2	< 0.01
ma3	< 0.01
ma4	< 0.01
ma5	< 0.01
intercept	0.7019237

Table 4: P-values for significance test of $ARMA(4, 5)$

Parameter	P-value
ar1	< 0.01
ar2	< 0.01
ma1	< 0.01
ma2	0.7786925
ma3	< 0.01
intercept	0.7133136

Table 5: P-values for significance test of $ARMA(2, 3)$

as the coefficients of $MA(2)$ are not significant at usual level. Therefore, we choose the model $ARMA(4, 5)$ as the best model for this series.

II.2 Residual normality

In order to obtain the most reliable prediction, it's good to have normal residuals. We can check using QQ plot if the residuals are normally distributed, i.e. if the residuals are aligned with the normal distribution.

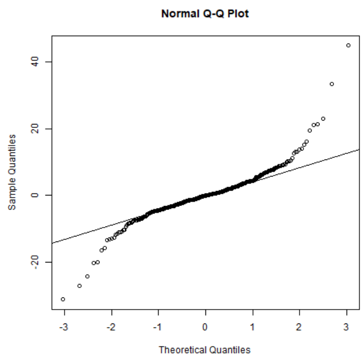

Figure 7: QQ plot of the residuals

We can see that our residuals are not normally distributed mostly in the extremes. We can confirm that using the Jarque-Bera test which is a test of the residuals for normality. The null hypothesis is that the residuals are normally distributed.

The p-value is less than 1 %, hence we reject the null hypothesis and conclude that the residuals are not normally distributed.

Finally, for our differentiated series we selected the model $ARMA(4, 5)$ with non normal resid-

χ^2	df	p-value
1394.8	2	$< 2.2 \times 10^{-16}$

Table 6: Jarque-Bera Test Results

uals. Here are our estimated coefficients:

$$\begin{aligned} X_t = 0.06 \\ &- 0.96X_{t-1} - 0.22X_{t-2} - 0.91X_{t-3} - 0.88X_{t-4} \\ &+ 0.47\epsilon_{t-1} - 0.26\epsilon_{t-2} + 0.86\epsilon_{t-3} + 0.46\epsilon_{t-4} - 0.53\epsilon_{t-5} \end{aligned}$$

III Prediction

In the following part, we denote as T the total length of the series and we assume that the redisuals are gaussian: $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.
We have an $ARMA(4,5)$ model and can thus write that:

$$X_t = \epsilon_t + \sum_{i=1}^{4} \phi_i X_{t-i} + \sum_{i=1}^{5} \theta_i \epsilon_{t-i}$$

III.1 α-level Confidence Region

Knowing that $\forall h > 0$, $\mathbb{E}[\epsilon_{T+h}|X_T, X_{T-1}, \dots] = 0$, we know that the optimal predictions in T are given by:

$$\begin{cases} \hat{X}_{T+1|T} = \sum_{i=1}^{4} \phi_i X_{T+1-i} + \sum_{i=1}^{5} \theta_i \epsilon_{T+1-i} \\ \hat{X}_{T+2|T} = \phi_1 \hat{X}_{T+1|T} + \sum_{i=2}^{4} \phi_i X_{T+2-i} + \sum_{i=2}^{5} \theta_i \epsilon_{T+2-i} \end{cases}$$

Denoting $\hat{X} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix}$ and $X = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix}$, we have that:

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

As ϵ_{T+1} and ϵ_{T+2} are gaussian variables, we can calculate the variance and we have that:
$$\begin{cases} \mathbb{V}[\hat{X}_{T+1|T} - X_T] = \mathbb{V}[\epsilon_{T+1}] & = \sigma^2 \\ \mathbb{V}[\hat{X}_{T+2|T} - X_T] = \mathbb{V}[\epsilon_{T+2} + \epsilon_{T+1}(\phi_1 + \theta_1)] & = \sigma^2(1 + (\phi_1 + \theta_1)^2) \end{cases}$$

Thus, the prediction errors follow a normal distribution of parameters $\mu = 0$ and Σ, meaning $X - \hat{X} \sim \mathcal{N}(0, \Sigma)$ with the variance matrix $\Sigma = \sigma^2 \begin{pmatrix} 1 & \phi_1 + \theta_1 \\ \phi_1 + \theta_1 & 1 + (\phi_1 + \theta_1)^2 \end{pmatrix}$

Since $det(\Sigma) = \sigma^4 * [(1 + (\phi_1 + \theta_1)^2 - (\phi_1 + \theta_1)^2] = \sigma^4$, the variance matrix is inversible as we assumed that $\sigma^2 > 0$. By the properties of multivariate normal distributions and the χ^2 distribution, we have that: $(X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \sim \chi^2(2)$. Therefore, we can deduce an α-level confidence region by taking

$$\{X \in \mathbb{R}^2 | (X - \hat{X})^T \Sigma^{-1} (X - \hat{X}) \leq q_{1-\alpha}\}$$

where $\forall u \in [0,1]$, q_u is the u-th quantile of the $\chi^2(2)$ distribution.

III.2 Hypotheses

To obtain the previous results, it should be noted that we made some assumptions:

- The model is perfectly known.
- The coefficients obtained previously (part. 2) are the true coefficients of the models (they are not random).
- The residuals follow a normal distribution $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.
- The variance matrix Σ is invertible, i.e. that $\sigma^2 \neq 0$.

It should also be noted that, if σ^2 is unknown, it is necessary to estimate it before doing anything, and that the resulting confidence interval will thus change form (widening its tails and becoming less precise).

III.3 Graphical representation

The following forecast graph shows the estimated values with their confidence intervals highlighted. We can see that the predictions are very uncertain, emphasizing the capability of the model to predict the next value but not really the ones after that.

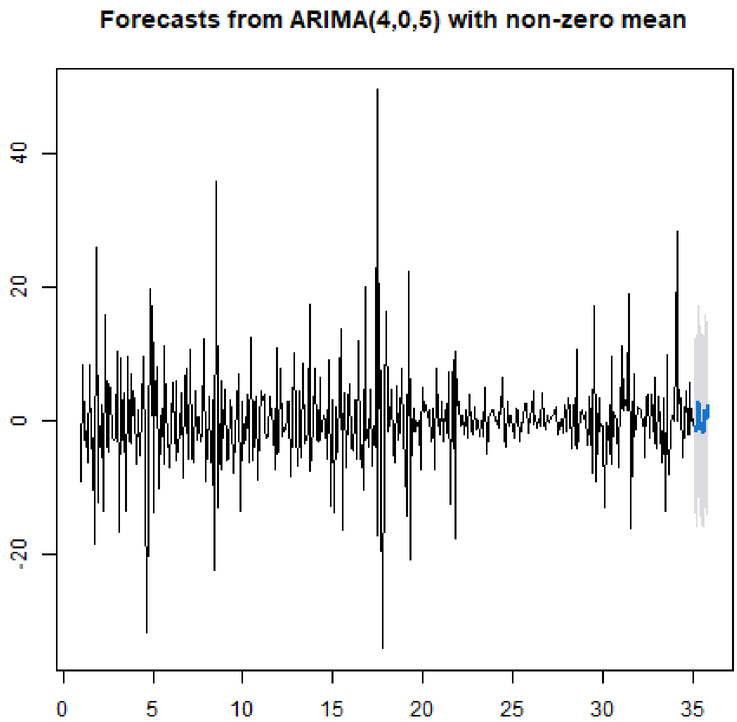

Figure 8: Forecasted values with their confidence region

III.4 Open question

If Y_{T+1} influences X_{T+1} in a way consistent with Granger causality, it could enhance the accuracy of the X_{T+1} forecast.

Granger causality implies that knowing the past and current values of Y_T helps reduce the mean squared error (MSE) of predicting X_{T+1}, compared to predictions made without the knowledge of Y_T. This condition can be tested using a Wald-type test, using for example the R function *causality* from the package *vars*.