Time Series Project

SOARES DE MELO Nathan, NONNENMACHER Alexandre

Mai 2025

— Manufacture of perfumes and toiletries —

Contents

1 Data

1.1 What does the chosen series represent?

In this project, we will be looking at the manufacture of perfumes and toiletries. This series can be found on the INSEE website via the following link: `https://www.insee.fr/fr/statistiques/serie/010767815`

The series covers the period from January 1990 to February 2025. Measurements are taken on a monthly basis. In the following, we will refer to this series as $(X_t)_{t \in T}$, with $T = \{1, \ldots, t\}$ the set containing all dates when the series is observed. The initial raw series is shown below:

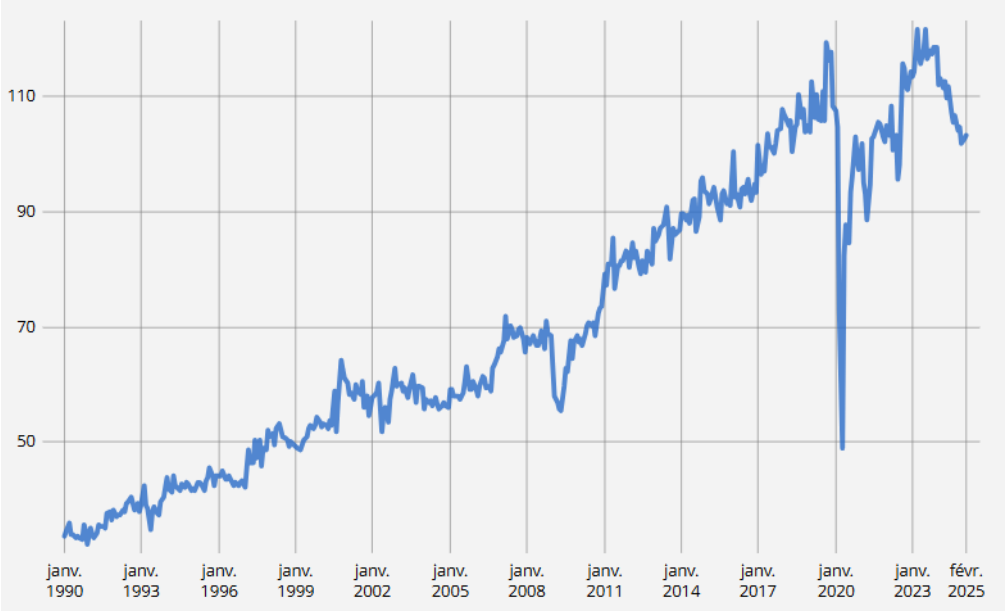

Figure 1: Initial gross series

1.2 Series transformation

To address the presence of abnormal values, most notably a sharp and temporary drop during the COVID-19 crisis, we opted to apply the `tsclean` function from the `forecast` package. This method automatically detects and corrects outliers, fills in missing values, and smooths the time series while preserving its overall structure. By relying on `tsclean`, we avoid the need to explicitly model or remove anomalies caused by exceptional, non-recurring events such as the pandemic. In particular, this approach effectively mitigates the disproportionate influence of the COVID-19 shock, allowing us to work with a cleaner, more regular series for modeling purposes, without discarding valuable information. The newly cleaned series can be found in the appendix.

We observe a clear increasing linear trend in the series. This trend is confirmed by the additive decomposition presented in the appendix. After applying a first-order differencing, the transformed series appears visually stationary, making it suitable for ARIMA modeling.

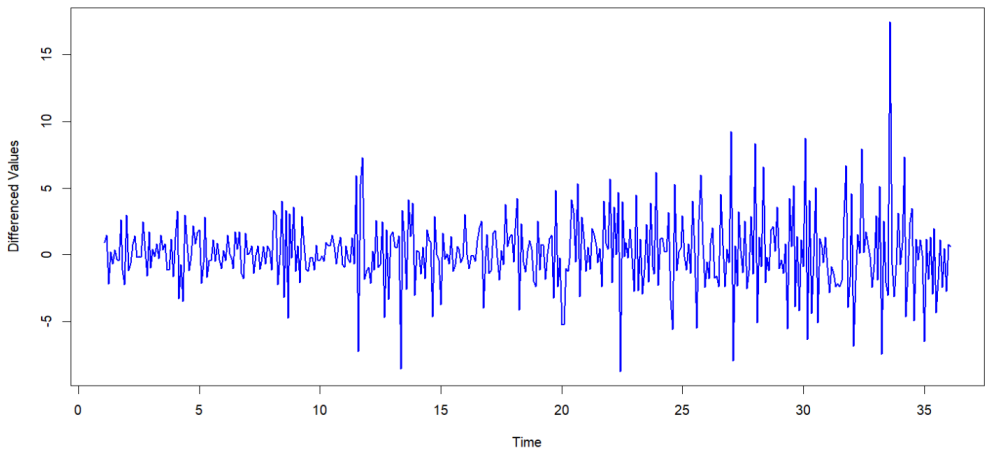

Figure 2: Differentiated series to order 1

We now wish to confirm our intuition that the differentiated series is stationary. To do this, we perform the classic unit root tests: Augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and the KPSS stationarity test. For the first two tests, the null hypothesis is the presence of a unit root (i.e. non-stationarity of the series), while for the KPSS test, the null hypothesis is stationarity of the series.

The two unit root tests (ADF and PP) gave us a p-value of less than 0.01, allowing us to reject H_0 at the 1% level in favor of the alternative hypothesis of series stationarity. Concerning the KPSS test, because the p-value is greater than 0.1 we fail to reject the null hypothesis at the 0.1 level, so we decide not to reject it. Further differentiation is therefore unnecessary. For the rest of the study, we can therefore assume that our series is stationary.

Table 1: ADF and PP test results

Type	Lag order	t-value	p-value
ADF	7	-7.9322	< 0.01
PP	5	-520.06	< 0.01

Table 2: KPSS test result

Type	Lag order	Statistique	p-value
KPSS	5	0.040977	> 0.1

1.3 Graphical representation of the transformed series

Finally, we can compare our two series: the initial raw series with linear trend and the stationary differentiated series. These are shown in Figure 4 below.

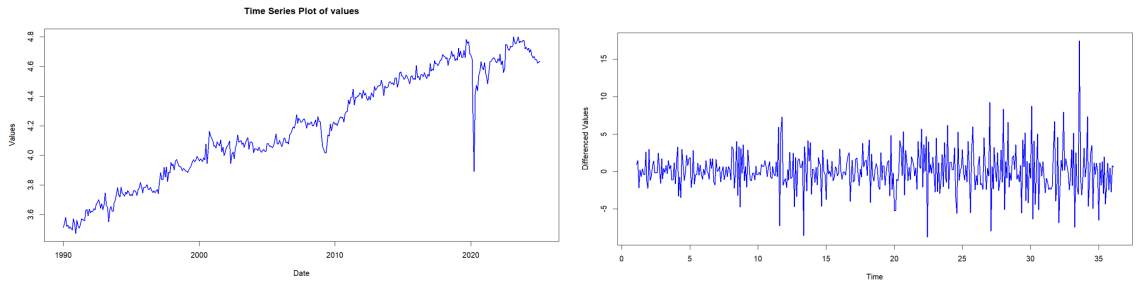

Figure 3: Comparison of the initial raw series and the stationary series

2 ARIMA Model

2.1 Model selection

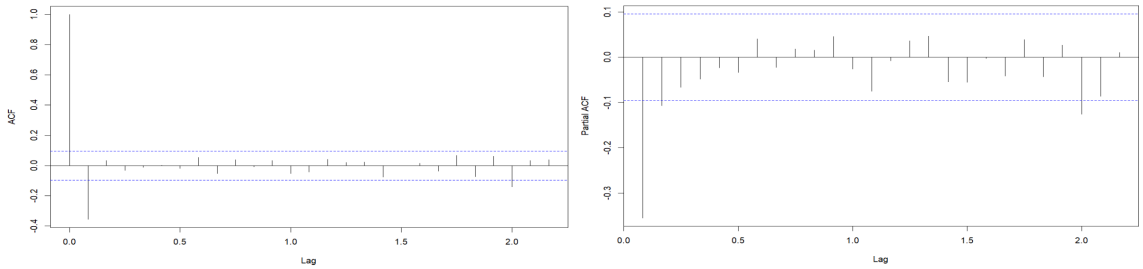

Figure 4: ACF and PACF of the differentiated series

We use the ACF and PACF graphs to identify the (p, q) orders. We note that when $p > 2$, the ACF peaks are mostly no longer significantly different from 0. Similarly, when $q > 1$, the PACF peaks are no longer significantly different from 0. From these results, we can build all the ARIMA(p, 1, q) models with $p \leq 2$ and $q \leq 1$ to find the one that minimizes the AIC and BIC criteria.

Based on the comparison of AIC and BIC values across previously determined combinations of p and q, we selected the ARIMA(0,1,1) model as the most appropriate specification. This model yielded the following estimated coefficient:

Coefficient	Estimate	Standard Error
ma1	-0.3874	0.0446

Table 3: Estimation of ARIMA coefficients and their standard errors

In this model, the ma1 coefficient is estimated at -0.3874 with a standard error of 0.0064, yielding a z-value of approximately 8.686. This value exceeds the typical critical threshold (e.g., -1.96 for a 5% significance level), indicating that the coefficient is statistically significant.

2.2 Checking model validity

In order to have a model that is valid and useful for prediction, a number of assumptions must be verified, particularly concerning residuals.

Firstly, our model must be perfectly identified, i.e. the coefficients found previously must be the real coefficients (or at least the estimators must converge to the real values).

Secondly, the innovation process must be Gaussian i.i.d. We can test both the normality of the process and its independence.

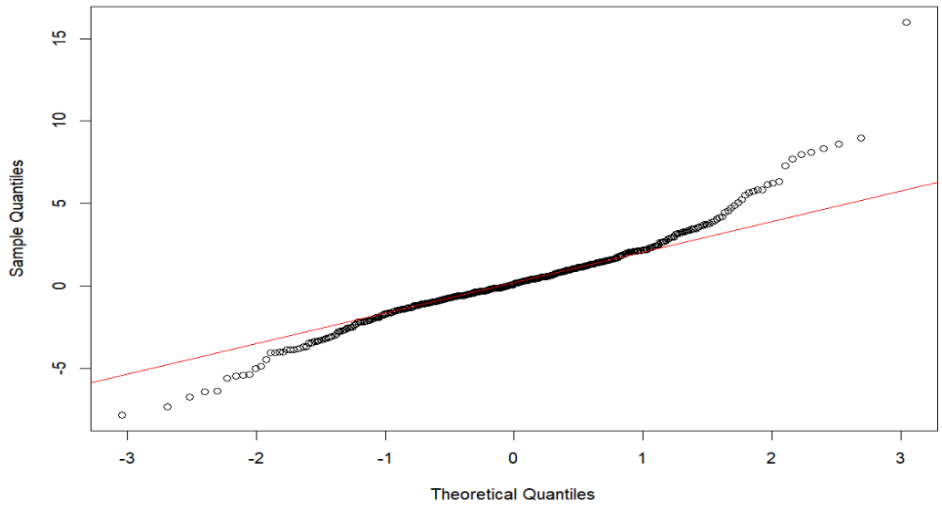

Figure 5: QQplot

On the QQplot in Figure 5, we can see that our residuals do not look normal, especially at the extreme quantiles, which do not seem to line up. To check our visual intuition concerning normality and the independence of the residuals, we will perform two tests (see table 4):

Ljung-Box Test: This test is used to assess the presence of autocorrelation in the residuals up to a given number of lags. Here, the test statistic is $Q^* = 22.507$ with 23 degrees of freedom, and the corresponding p-value is 0.4899. Since the p-value is much greater than common significance levels, we fail to reject the null hypothesis that the residuals are white noise. This suggests that our ARIMA(0,1,1) model adequately captures the autocorrelation structure of the data.

Shapiro-Wilk Test: This test assesses whether the residuals follow a normal distribution. The test yields a statistic $W = 0.94593$ and a p-value less than 2.2×10^{-16}, which is far below any standard significance level. Therefore, we reject the null hypothesis of normality. This strong departure from normality is likely explained by the presence of extreme values in the data, notably during the COVID-19 period, as previously discussed. Even after applying the `tsclean` transformation, these anomalies still significantly impact the distribution of the residuals.

In summary, while the residuals appear uncorrelated (which is desirable for model adequacy), they do not follow a normal distribution, likely due to external shocks such as the COVID-19 crisis.

Test	**Statistic**	**df / Param.**	**p-value**
Ljung-Box Test	$Q^* = 22.507$	df = 23	0.4899
Shapiro-Wilk Test	$W = 0.94593$	—	$< 2.2 \times 10^{-16}$

Table 4: Diagnostic tests on residuals from ARIMA(1,1,1)

2.3 Model equation

The equation of our ARIMA(0,1,1) model is written as follows:

$$\nabla X_t = X_t - X_{t-1} = Y_t$$
$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

3 Prediction

For the rest of this project, we assume that the residuals of the series are Gaussian, i.e., $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

3.1 Confidence region of level α

Given that $\mathbb{E}[\varepsilon_{T+h} \mid Y_T, Y_{T-1}, \dots] = 0$ for all $h > 0$, we know from theory that the optimal forecasts at time T are:

$$\hat{Y}_{T+1|T} = \theta_1 \varepsilon_T$$

$$\hat{Y}_{T+2|T} = \hat{Y}_{T+1|T} +$$

We now compute the prediction errors $Y_{T+1} - \hat{Y}_{T+1|T}$ and $Y_{T+2} - \hat{Y}_{T+2|T}$.
Let:

$$\hat{\mathbf{Y}} = \begin{pmatrix} \hat{Y}_{T+1|T} \\ \hat{Y}_{T+2|T} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_{T+1} \\ Y_{T+2} \end{pmatrix}$$

Then:

$$\mathbf{Y} - \hat{\mathbf{Y}} = \begin{pmatrix} Y_{T+1} - \hat{Y}_{T+1|T} \\ Y_{T+2} - \hat{Y}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \varepsilon_{T+1} \\ \varepsilon_{T+2} + (\theta_1 + \phi_1)\varepsilon_{T+1} \end{pmatrix}$$

We then compute the variances of the prediction errors:

$$\mathrm{Var}(Y_{T+1} - \hat{Y}_{T+1|T}) = \mathrm{Var}(\varepsilon_{T+1}) = \sigma^2$$

$$\mathrm{Var}(Y_{T+2} - \hat{Y}_{T+2|T}) = \mathrm{Var}(\varepsilon_{T+2} + (\theta_1 + \phi_1)\varepsilon_{T+1}) = \sigma^2 \left(1 + (\theta_1 + \phi_1)^2\right)$$

Therefore, $\mathbf{Y} - \hat{\mathbf{Y}} \sim \mathcal{N}(0, \Sigma)$, where the variance-covariance matrix Σ is:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \theta_1 + \phi_1 \\ \theta_1 + \phi_1 & 1 + (\theta_1 + \phi_1)^2 \end{pmatrix}$$

Since $\det(\Sigma) = \sigma^2 > 0$, the matrix Σ is invertible under our assumption $\sigma^2 > 0$.
From theory, we know:

$$(\mathbf{Y} - \hat{\mathbf{Y}})^\top \Sigma^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}) \sim \chi^2(2)$$

Hence, the confidence region of level α is given by:

$$\left\{ \mathbf{Y} \in \mathbb{R}^2 \;\middle|\; (\mathbf{Y} - \hat{\mathbf{Y}})^\top \Sigma^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}) \leq q_{1-\alpha}^{\chi^2(2)} \right\}$$

where $q_{1-\alpha}^{\chi^2(2)}$ is the $(1-\alpha)$-quantile of the chi-squared distribution with 2 degrees of freedom.

3.2 Assumptions

The derivation of the previous results relies on several key assumptions:

- The ARMA model structure is correctly specified.
- The parameters estimated in Section 2 are assumed to be the true parameters of the underlying process.
- The innovations (white noise) follow a normal distribution: $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.
- The noise variance σ^2 is strictly positive: $\sigma^2 > 0$.

The canonical form of the ARMA model, with all roots lying outside the unit circle and no common roots between the AR and MA polynomials, ensures that the innovations are linearly independent. This justifies the explicit computation of the forecasts $\widehat{X}_{T+1|T}$ and $\widehat{X}_{T+2|T}$.

Moreover, the assumption of normality of the residuals is critical and was tested in Section 2. We also presume that the residual variance σ^2 is known.

If the true parameters of the ARMA model are also unknown, then the uncertainty in the estimated covariance matrix $\widehat{\Sigma}$ increases further. It now reflects both the error in estimating the residual variance and the uncertainty associated with the estimation of the model's parameters.

3.3 Graphical Representation

We are now in a position to visualize the confidence region at the 95% level. The graphical display (Figure 6) shows this region in light gray, representing the 95% confidence region for the joint forecast of future values. The dark blue point corresponds to the predicted values of X_{T+1} and X_{T+2}.

3.4 Open Question

We now address the following open question: let Y_t a stationary time series available from $t = 1$ to T. We assume that Y_{T+1} is available faster than X_{T+1}. Under which condition(s) does this information allow you to improve the prediction of X_{T+1} ? How would you test it/them ?

The information provided by Y_{T+1} can help improve the forecast of X_{T+1} if and only if there exists an instantaneous Granger causality from Y_t to X_t.

This means that, conditionally on the past values of X_t, the knowledge of Y_{T+1} provides additional predictive power for X_{T+1}, i.e., it reduces the mean squared prediction error compared to a model that does not incorporate Y_t.

To test for this condition, one can use a regression-based Granger causality test. Specifically, estimate two models:

- A restricted model: regress X_{T+1} on past values of X_t only;
- An unrestricted model: regress X_{T+1} on both past values of X_t and the contemporaneous value Y_{T+1}.

Then, perform an F-test (or likelihood ratio test) to assess whether the inclusion of Y_{T+1} significantly improves the model. A statistically significant result supports the hypothesis that Y_t Granger-causes X_t, and thus that using Y_{T+1} can enhance the prediction of X_{T+1}.

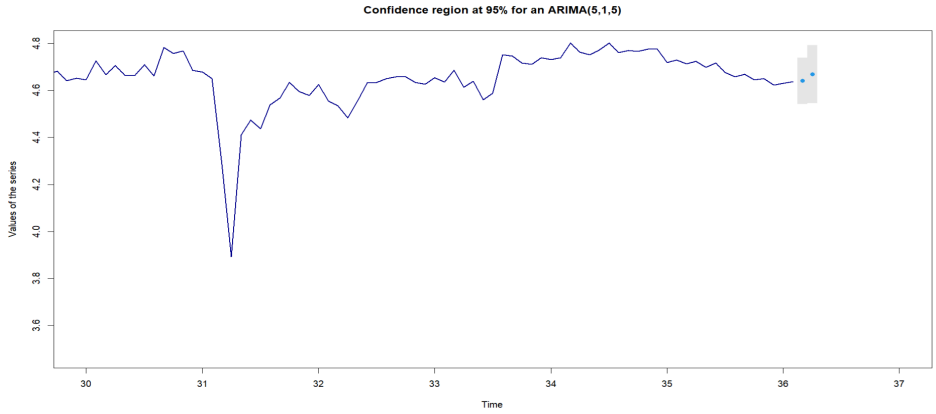

Figure 6: Predictions for T+1 and T+2

4 Appendix

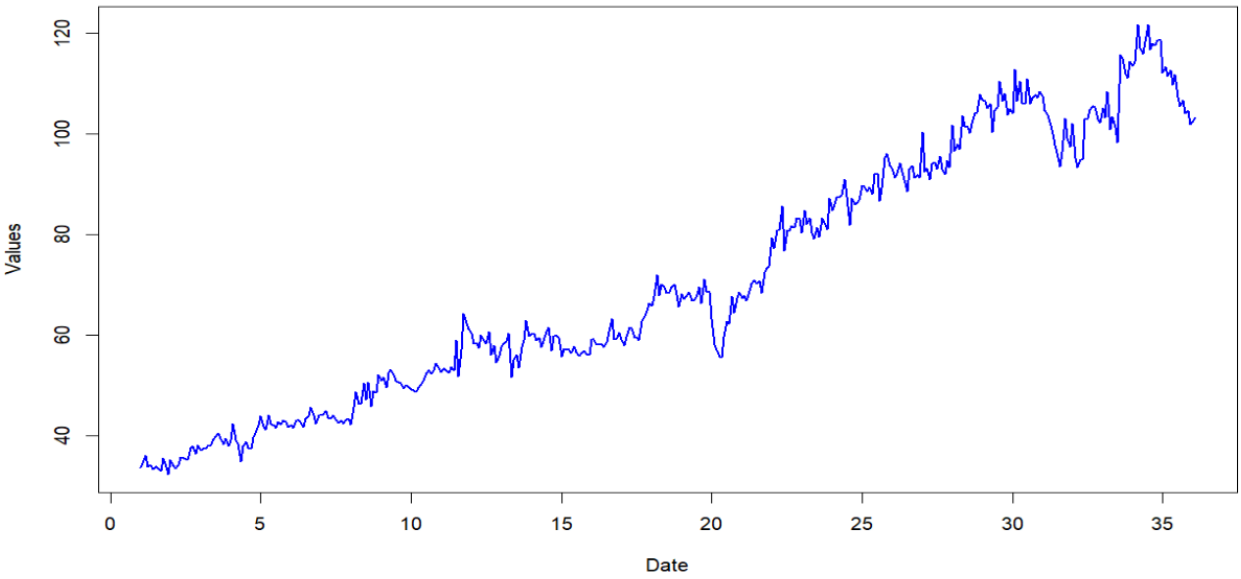

Figure 7: Cleaned series

	p <int>	q <int>	AIC <dbl>	BIC <dbl>
2	0	1	1993.117	2001.203
8	0	1	1993.117	2001.203
4	1	1	1994.999	2007.127
10	1	1	1994.999	2007.127
5	2	0	1996.467	2008.595
11	2	0	1996.467	2008.595
6	2	1	1996.653	2012.824
12	2	1	1996.653	2012.824
3	1	0	1998.656	2006.741
9	1	0	1998.656	2006.741
1	0	0	2051.768	2055.811
7	0	0	2051.768	2055.811

Figure 8: AIC and BIC of the series

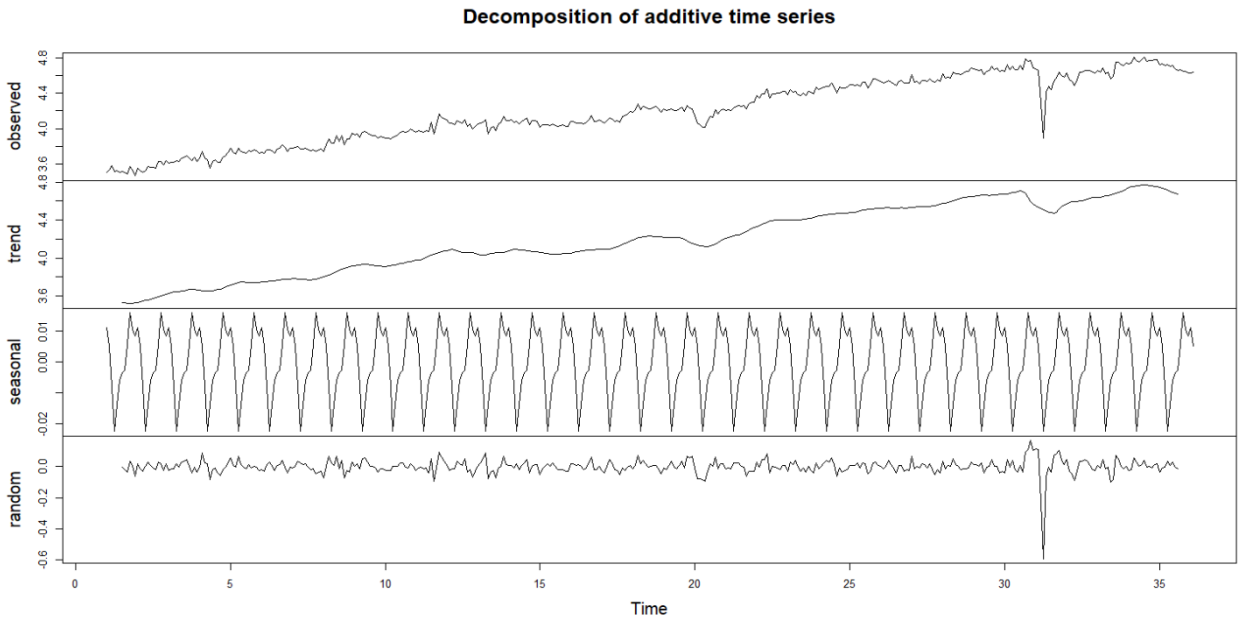

Figure 9: Additive decomposition of the series

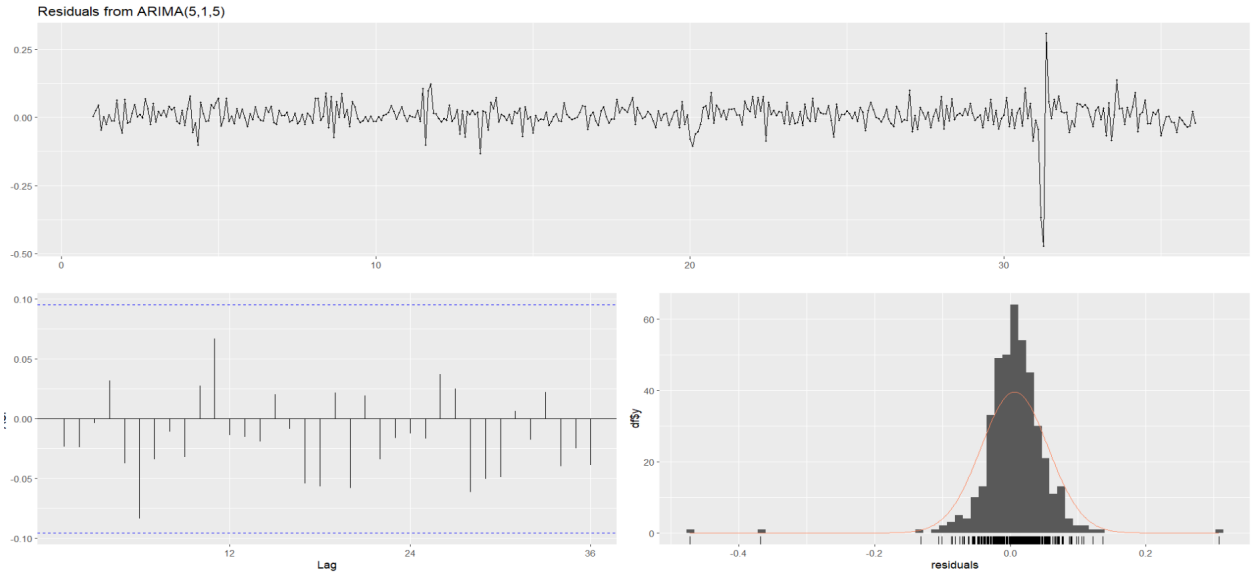

Figure 10: Results of the *checkresiduals* function applied to the time series