TÉCNICO LISBOA

TIME SERIES

INSTITUTO SUPERIOR TÉCNICO

2ND SEMESTER - 2021/22

Project 1: Atmospheric Pollutants in Portugal - 0zone. A time series study

Project 2: Fitting GARCH-type models to Financial time series

Filipa Costa 92626
Martim Rêgo 86480
Professor:
Manuel Gonzalez Scotto

June 19, 2022

Contents

1 Introduction

2 Air Quality Dataset

The presented dataset corresponds to the values of O_3 particles (hourly ground levels), in micrograms per cubic meter (µg/m3), collected at some stations of the qualar network, including Antas-Espinho, Entrecampos, Estarreja, Laranjeiro-Almada, Mem-Martins, Paio-Pires, Restelo, Sobreiras-Porto and VNTelha-Maia. The values were extracted every hour for the period of 1 year, starting in the beginning of 2020 and ending in the last day of the same year.

The O_3 particles (also known as ozone) are a common air pollutant and cannot be seen with the naked eye. Ozone is a secondary pollutant in the troposphere, the thin layer of air from which we breathe, that results from photochemical reactions involving primary precursor pollutants from industrial activities or transportation, as well as solar radiation.

Pollution episodes caused by high ozone concentrations are most common during the summer months, when the sun is shining brightly, the temperature is high, the wind is calm, and the atmosphere is stable.

Exposure to these pollutants is hazardous to humans, and when the concentration of these particles rises, it can have a detrimental influence on human quality of life and health, particularly at the respiratory system level [1].

3 Initial examination of the data

Firstly, the data was loaded and a date and hour were assigned to each observation as an index. The missing values were checked next, but none were found in all of the cities. The data was then sorted by date and the time series were plotted. The plot of the levels of O_3 particles at Paio-Pires, Portugal, against time, by hours, from the beginning of 2020 to the end of 2020 is shown in Figure 1.

It is vital to note that we will only show how we built a model for the Paio-Pires time series in the main part of the project. However, because the process for each time series ended up being the same, every plot and table for each of the other time series may be found in the appendix.

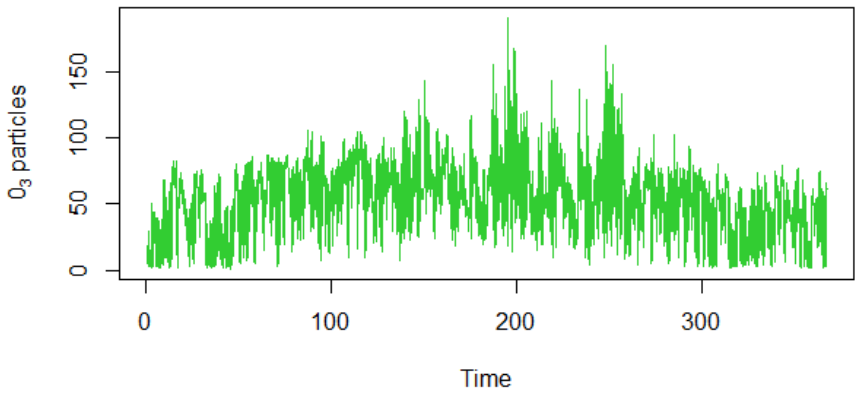

Figure 1: Hourly levels of O_3 particles in $\mu g/m3$ in Paio-Pires since 01/01/2020 until 31/12/2020

Throughout the summer season (from May to middle of August), there are some spikes and a higher variance as it can be seen in Figure 1. The epoch with the highest levels of O_3 occurs around the middle of June; the month with the lowest values of O_3 is January.

To better understand this time series, we tried to look at the boxplots of different time divisions (daily, week daily, monthly).

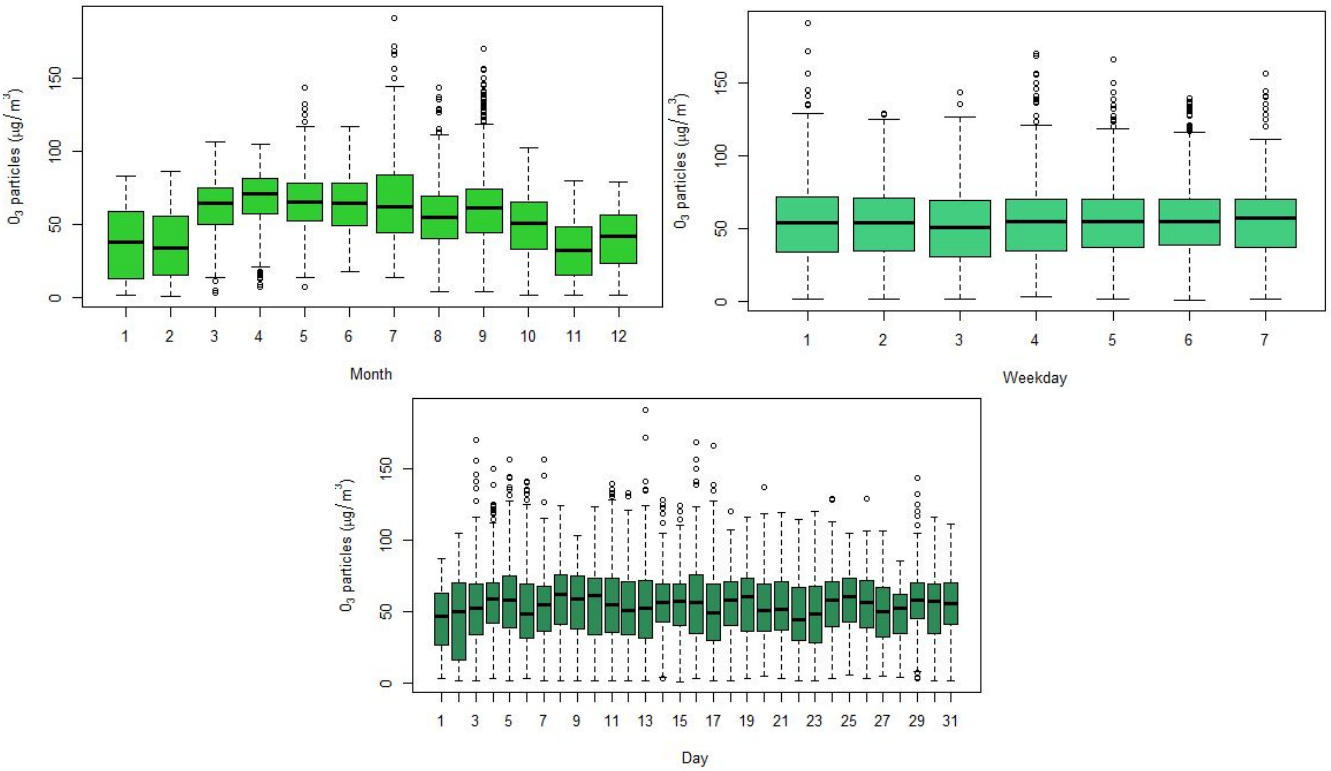

Figure 2: Boxplots of the O_3 particles levels in Paio-Pires with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

By analysing Figure 2, it is possible to point out some patterns.

It is plausible to associate increased variability with the months of May through September, and a higher median into the O_3 particles with the months of April through July. Because pollutants react to heat and sunshine when the weather warms up over the summer months, ground-level ozone pollution rises. With summer wildfires, ozone levels generally rise as well [2]. The months with lower O_3 particles levels medians are January and February.

Furthermore, while looking at a week-period pattern, it's feasible to see a slight increase in the median of the O_3 particles on weekends. This seems contradictory with the fact that the emissions of ozone chemical precursors - from car exhaust fumes and industrial smoke - decrease. Scientists call this effect the "weekend effect" [5].

Finally, the last boxplot does not show any specific trend within a month when looking for a pattern associated to a day of the month. It's also worth noting that since we only have a single year on record, each day of the month is only represented (at most) 12 times, therefore it would probably be unwise to extrapolate any conclusions from such a small sample.

4 Model building

4.1 Decomposition of the time series

Figure 3 shows a plot of the Seasonal-Trend decomposition of the time series (based on Loess), with the period set to 24 hours. This is done to see if there is a clear Trend or Seasonal impact that occurs over time. According to this decomposition, the time series is the sum of the Trend, Seasonal, and Residuals/Random noise components.

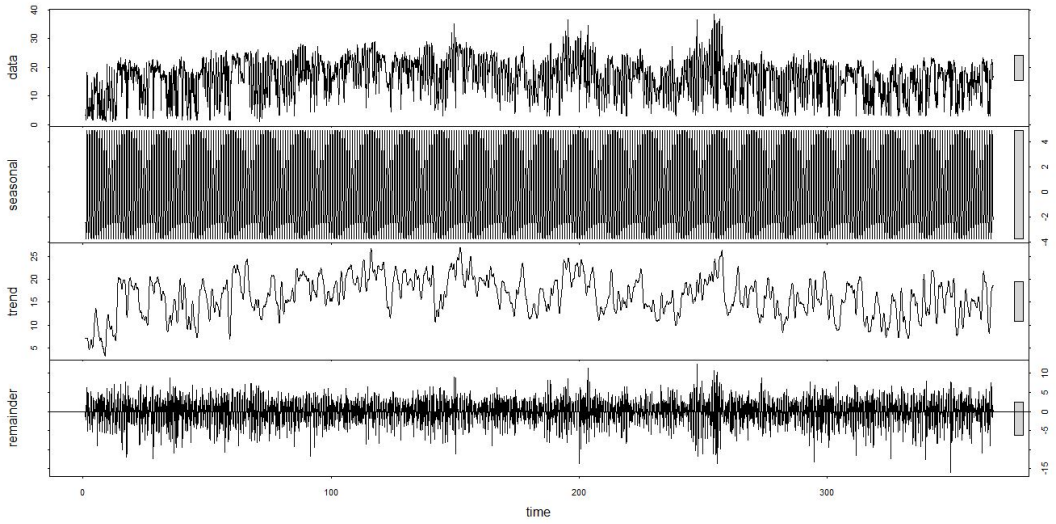

Figure 3: STL decomposition of the original dataset into three components: Seasonal, Trend and Remainder

Because there are so many observations, it is impossible to notice a distinct seasonal or trend pattern in Figure 3.

Because the observations are autocorrelated, the ACF plot 4 of the remainder generated by the stl decomposition shows that the remainder series does not appear stationary. This way, it was decided to perform differences in order to make the series stationary.

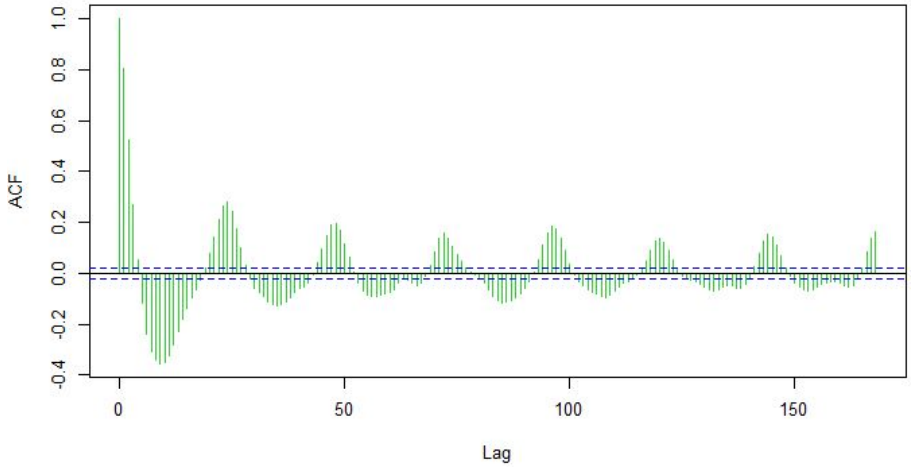

Figure 4: ACF plot of the remainder provided by the STL decomposition (with the Box Cox transformation explained in the next subsection)

4.2 Transformation of the data: Box-Cox

The analysis of an unstable variance supplied the motivation for this log-transformation of the data due to the presence of spikes in the data 1. To solve this problem, the data was transformed using the Box-Cox transformation to obtain a stationary time series. The optimum value of lambda was calculated using the R command BoxCox.lambda ($\lambda = 0.6547766$). In Figure 5, it can be seen that the variance got more stabilised.

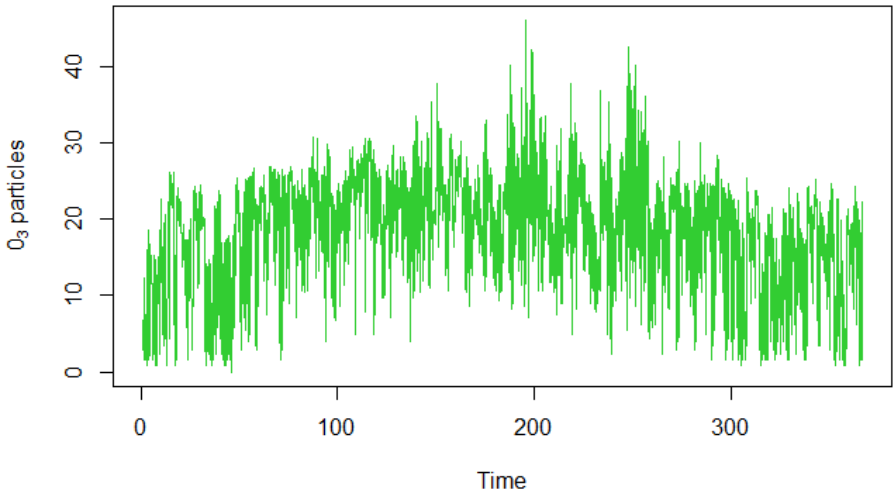

Figure 5: Hourly levels of O_3 particles in $\mu g/m3$ in Paio-Pires since 01/01/2020 until 31/12/2020, after the log transformation of the data

4.3 Stationarity

One of the assumptions for SARIMA analysis is that the time series must be stationary. This means, stabilising the variance, the mean and remove the seasonal and trend component. Once the variance was stabilized, the Augmented Dickey-Fuller test was resorted. The null hypothesis of the time series in hands not being stationary is compared to the time series being stationary in this test. The null hypothesis should be rejected because the p-value was $p < 0.01$, indicating that the time series in question is stationary.

However, the p-value on large samples is not meaningful, since in very large samples, p-values goes quickly to zero, and solely relying on p-values can lead us to claim support for results of no practical significance [3].

Therefore, the next step is to observe the ACF and the PACF plots.

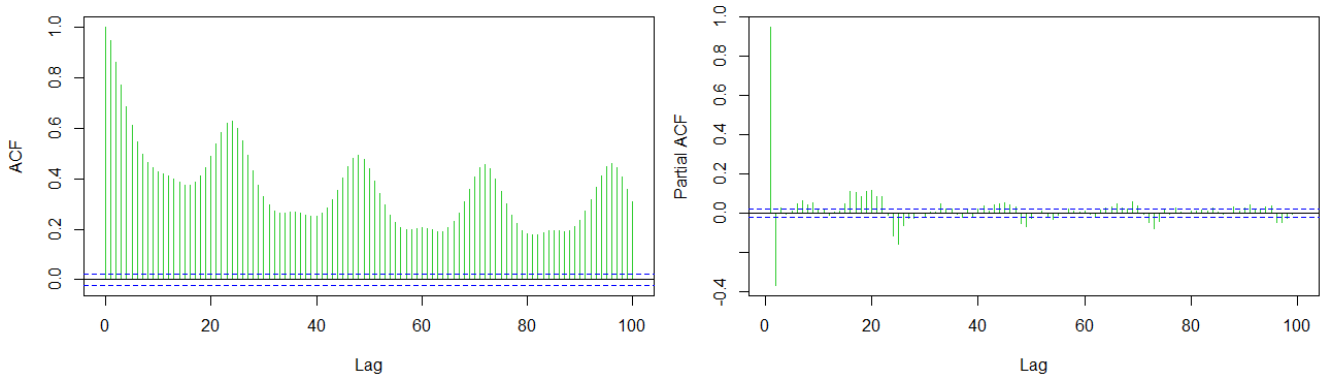

Figure 6: Characteristics of O_3 time series, with Box-Cox transformation, at Paio-Pires station. ACF on the left and PACF on the right

From Figure 6(a) it is clear that the time series is not stationary; the ACF drops slowly without reaching nonsignificant correlation values. The PACF approaches zero as the lags increase in Figure 6(b), yet the partial correlation between observations continues even after a week. As a result, the series must be transformed in order to become stationary.

Since the ACF of Paio Pires station has a correlation peak every 24 hours, a seasonal differencing at lag 24 was conducted, followed by a first differencing at lag $= 1$ to stabilise the mean.

The ACF and PACF of the transformed time series of Y_t ($(1 - B)(1 - B^{24})Y_t$) can be found in Figure 7.

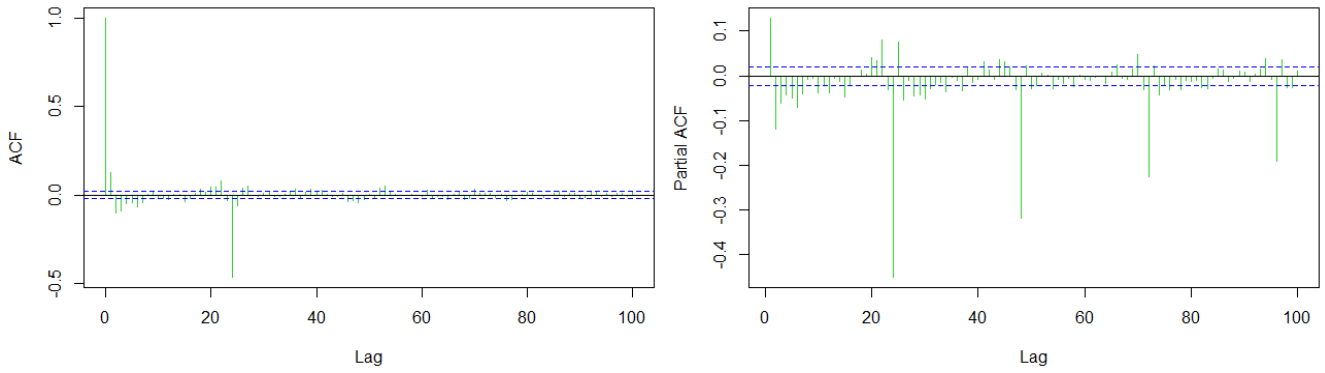

Figure 7: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Paio-Pires station. ACF on the left and PACF on the right

Figure 7 shows that significant lags in the autocorrelation function are only observed up to lag 24, whereas significant

correlations in PACF are found throughout the week, despite the declining trend.

4.4 SARIMA Model

SARIMA is developed by adding extra seasonal terms to ARIMA models: ARIMA(p,d,q)(P, D, Q)S; where (p,d,q) represents the non-seasonal element of the model; p refers to the autoregressive; d is the degree of differencing; and q refers to the moving average. (P,D,Q)S is the seasonal component of the model, where P stands for autoregressive, D for differencing, Q for moving average, and S for seasonality.

To construct the SARIMA model it is needed to find the parameters. For that, the autocorrelation function (ACF) and parcial autocorrelation function (PACF) in Figure 7 will be analysed.

First, from the transformed time series ($(1-B)(1-B^{24})Y_t$) it can be concluded that $d = D = 1$ and $S = 24$. Notice that $S = 24$ is expected since one is dealing with hourly data.

In the next step, it is necessary to determine the possible values of p, P, q and Q. As a starting point to narrow down to a few potential parameters, The ACF will be used to determine the possible values of Q and the PACF to determine the possible values of p.

Values of $0 \leq p, q \leq 5$ will be attempted, because the values for these lags are higher than the significance level in both ACF and PACF plots, and this value appears to decline with higher delays.

In order to decrease the computational complexity and time, as well as to preserve the parsimony of the model, it was decided to set a restriction, $p + q \leq 5$. This is done because when the value of p and q increases there are equally more coefficients to fit hence the time complexity increases. This way, while looking for a decent model to fit the data, this will avoid excessively complicated models, since, as the Occam's razor rule states, the simplest model that can accurately explain the data should be chosen.

The values of P and Q were limited at $0, 1$ and 2.

The supplied data was divided into data train and data test to evaluate the final model's fit. The data used to test was made up of the last 5 observations available in the data because it was requested in the handout to forecast the data into the future up to 5 time periods ahead. It is worth noting that because the time series will be modified using the Box Cox transformation, the test set needs to be transformed as well.

4.4.1 Selection Procedure

Firstly, with the help of the $auto.arima$ command in R, we managed to find the best model. The results obtained are displayed in Table 1:

Model	AIC	BIC
ARIMA(3,0,1)(2,1,0)[24]	37761.41	37810.95

Table 1: Best fitted model by $auto.arima$, for Paio-Pires time series

Then, since we are not sure that we are selecting the best model using the $auto.arima$ command, a matrix with the possible values of p, d, q, P, D, Q, and S explained earlier was first built to begin the building and selection of the best model. Each row represents a different set of the parameters values that will be tested. In total 180 models will be performed. While estimating the models, the command try was added in the loop. Indeed, it was not required for the loop to stop when the estimation algorithm does not converge. It will just provide a *NULL* object.

The models that do not produce white noise residuals are then discarded. The $Ljung - box$ test was used, with the lag fixed at 10, since the *RDocumentation* suggests a better approximation to the null-hypothesis distribution is obtained by setting $lag \geq p + q$. The model is rejected if at least one $p - value$ from the later test is less than the significance level of 5%, because in this case the test provides evidence to reject the null hypothesis, meaning that the model's residuals are not white noise. For this, the command $Box.test$ from $stats$ package will be used.

After confirming which models have white noise residuals, a vector aic is generated that holds the AIC of the computed models. To have a more clear idea, Figure 8 shows some results of the present dataset.

p	d	q	P	D	Q	T	residuals	aic	model
4	1	1	1	1	2	24	y	35776.6625	ARIMA(4,1,1)(1,1,2)[24]
4	1	1	1	1	1	24	y	35776.4798	ARIMA(4,1,1)(1,1,1)[24]
4	1	1	0	1	2	24	y	35775.7264	ARIMA(4,1,1)(0,1,2)[24]
4	1	1	0	1	1	24	y	35801.139	ARIMA(4,1,1)(0,1,1)[24]
3	1	2	0	1	2	24	y	35775.6686	ARIMA(3,1,2)(0,1,2)[24]
3	1	2	0	1	1	24	y	35801.4626	ARIMA(3,1,2)(0,1,1)[24]
3	1	1	1	1	2	24	y	35775.2683	ARIMA(3,1,1)(1,1,2)[24]
3	1	1	1	1	1	24	y	35775.1114	ARIMA(3,1,1)(1,1,1)[24]
3	1	1	0	1	2	24	y	35774.3547	ARIMA(3,1,1)(0,1,2)[24]
3	1	1	0	1	1	24	y	35799.6031	ARIMA(3,1,1)(0,1,1)[24]

Figure 8: Example of the Paio Pires different models, where the residuals column is "y" if the residuals are white noise and "n" if they are not white noise

As it is visible in 8, some models with a smaller AIC than the model selected by $auto.arima$ are obtained. This way, the models with AIC smaller than 37761.41 and with whise noise residuals will then be selected. A column with the AICc and BIC criteria for each model will also be created.

With this modifications, only 38 models are left, and by ordering these models by increasing order of the AIC and BIC criteria. The results of the first 3 best models according to AIC and BIC criteria are disposed in Table 2.

Model	AIC	AICc	BIC
ARIMA(3,0,1)(2,1,0)[24] (auto.arima)	37761.41	37761.42	37810.95
ARIMA(2,1,3)(2,1,1)[24] (1st best AIC)	35773.16	35773.18	35836.86
ARIMA(2,1,3)(1,1,2)[24] (2nd best AIC)	35773.23	35773.25	35836.92
ARIMA(3,1,1)(0,1,2)[24] (3rd best AIC and BIC)	35774.35	35774.37	35823.9
ARIMA(1,1,2)(0,1,2)[24] (1st best BIC)	35775.95	35775.96	35818.42
ARIMA(1,1,2)(1,1,1)[24] (2nd best BIC)	35776.72	35776.73	35819.18

Table 2: Comparison of the best fitted models for Paio Pires time series

According to the AIC (and AICc) value, the best fitted model would be an ARIMA(2,1,3)(2,1,1)[24]. In the same way, if we take into consideration the BIC value, the appropriate model to be chosen would be an ARIMA(1,1,4)(1,1,2)[24].

The AIC seeks to select the model that best describes an unknown, high-dimensional reality by estimating a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model. In this approach, a lower AIC indicates that a model is closer to the truth. BIC, on the other hand, is an estimate of a function of the posterior probability of a model being true that attempts to determine the true model. Furthermore, the BIC penalizes the introduction of new parameters more than the AIC, therefore it favors models with fewer parameters. Since our intention is to select a parsimonious model, we chose the ARIMA(3,1,1)(0,1,2)[24] as our best fitted model because it was the third best model according to both AIC and BIC criteria.

Finally, the last step is to check if the chosen model contains non significant coefficients. To do so, since Arima uses maximum likelihood for estimation, the coefficients are assymptoticaly normal. Hence, we divided the coefficients by their standard errors to get the z-statistics and then calculated the $p - values$ for each coefficient.

The $p - value$ for each coefficient will be:

$$p - value = 2 * (1 - F_{N(0,1)}(|coef|/se)$$

The summary of the estimated coefficients for ARIMA(3,1,1)(0,1,2)[24] model can be observed in Table 3.

ARIMA(3,1,1)(0,1,2)[24]	Coefficient	p-value
ar1	1.1268	0
ar2	-0.2733	0
ar3	0.0578	$1.379e^{-07}$
ma1	-0.9916	0
sma1	-0.895	0
sma2	-0.0573	$1.650e^{-07}$

Table 3: Summary of ARIMA(3,1,1)(0,1,2)[24] coefficients

Because all of the coefficients' p-values were ≤ 0.05, the coefficients were found to be significant, and it is safe to proceed to the residuals diagnostics.

It is important to note that we performed the residual analysis with the best model and the $auto.arima$ model for all the time series, but the residuals of the model resultant from $auto.arima$ were everytime auto-correlated. For example, for the time series Antas-Espinho, the model who showed the lowest measures of AIC and BIC was the one obtained from $auto.arima$, however the residuals diagnostic failed.

4.5 Residual diagnostics

To analyze the residuals, we check if they are independently, identically distributed and driven by a normal distribution. To verify the normality assumption, we should take a look at Figure 9.

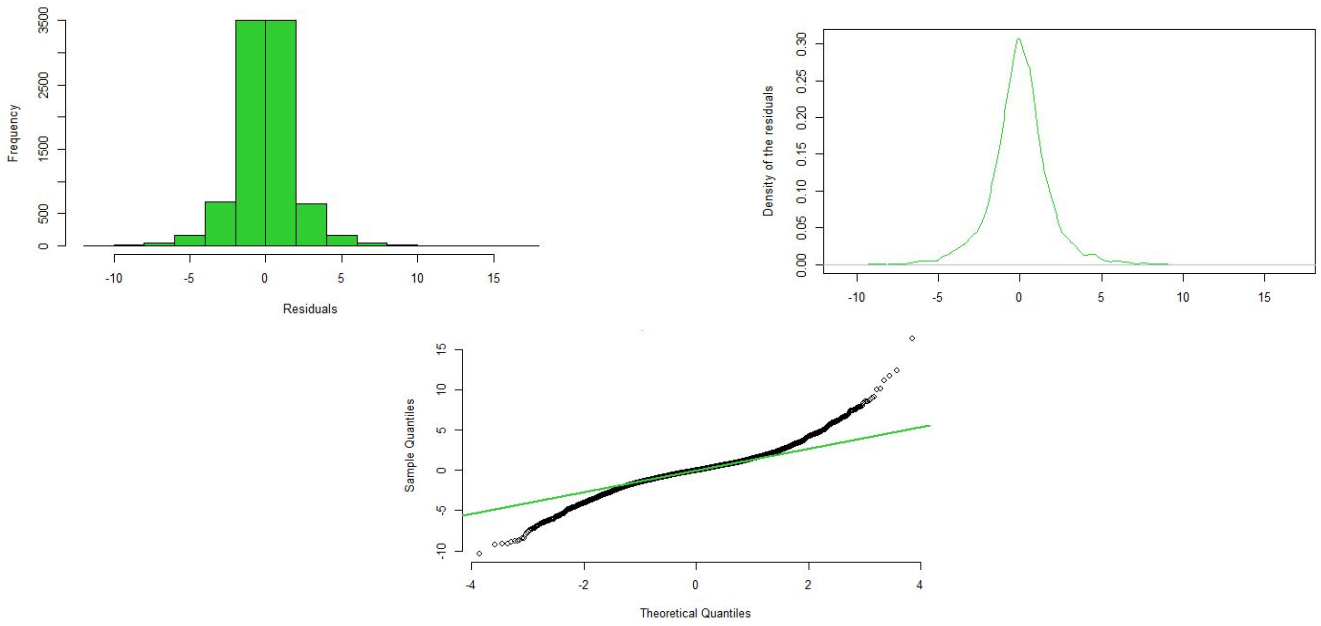

Figure 9: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for the ARIMA(3,1,1)(0,1,2)[24] process

The histogram and density plot, as shown in Figure 9, exhibit a remarkably similar distribution to that of a standard gaussian random variable. This fact is also visible in the QQ-plot of the down side. This way, these representations give an indication that the residuals of the model are approximately normality distributed.

To evaluate the mutual independence of the residuals, we plotted the ACF of the residuals in Figure 10.

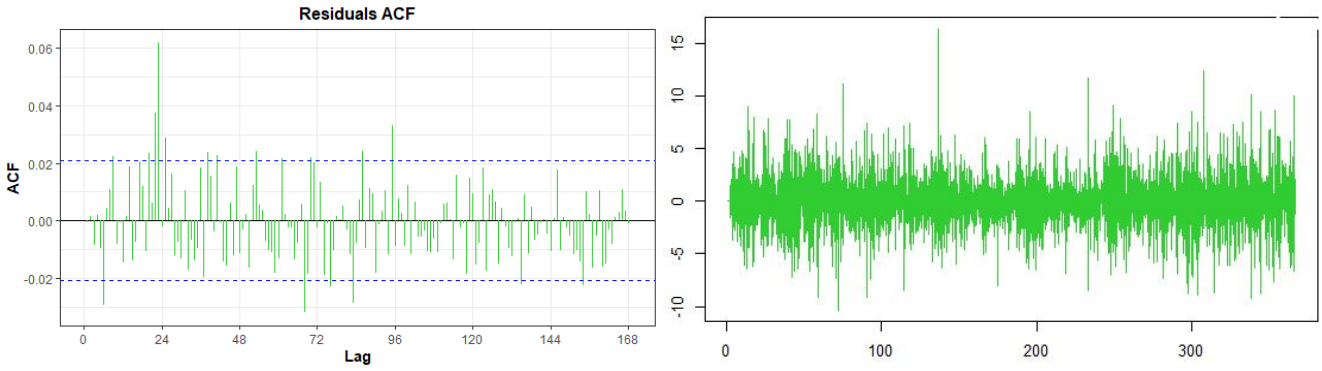

Figure 10: ACF (left) and plot (right) of the residuals for the ARIMA(3,1,1)(0,1,2)[24] process

We can see that by rejecting the 24th lag, we obtained extremely good results because the values are very low. We can also see that the ACF of the residuals has no evident correlation, hence we can assume that the residuals are independent. As stated in the selection procedure, the Ljung-Box Test yields a p-value greater than 0.05, allowing us to not reject the null hypothesis and conclude that the residuals are uncorrelated.

The residuals appear to be white noise with a normal distribution.

Finally, the SARIMA model for Paio Pires time series will be:

$$X_t = (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)(1 - B)(1 - B^{24})X_t = (1 + \theta_1 B)(1 + \Theta_1 B^{24} + \Theta_2 B^{48})Z_t$$

And substituting by the estimated coefficients, we obtain:

$$X_t = (1 - 1.1268B + 0.2733B^2 - 0.0578B^3)(1 - B)(1 - B^{24})X_t = (1 - 0.9916B)(1 - 0.895B^{24} - 0.0573B^{48})Z_t$$

where Z_t is a white noise.

4.6 Forecasting

As requested, the values acquired for the forecast of the following 5 time periods ahead, corresponding to the last 5 days of 2020, as well as their respective 95% prediction intervals, are given in Table 4.

Date	Lower Bound	Predicted Value	Upper Bound	True Value	Relative Error	MAPE
31/12/2020, 19h	11.1015	14.7467	18.3919	18.5167	20%	
31/12/2020, 20h	8.6360	14.1506	19.6652	20.2779	30%	
31/12/2020, 21h	7.7022	14.3414	20.9806	21.4910	33%	30%
31/12/2020, 22h	6.9614	14.3649	21.7684	21.2511	32%	
31/12/2020, 23h	6.2274	14.1973	22.1672	21.2512	33%	

Table 4: Predicted values, Confidence intervals and MAPE of ARIMA(3,1,1)(0,1,2)[24] prediction

The MAPE (absolute percentage error) value of the model is equal to 30%.

This does not sound like the best score, but it is also important to note that O_3 particles level is correlated with weather (what we concluded in the first examination) and the weather prediction problem is still very complex and unsolved **verificar no final**.

In Figure 11, we can observe each time period forecast, as well as the original data and the fitted data. Dark green and light green correspond to a 80% and 95% confidence interval, respectively.

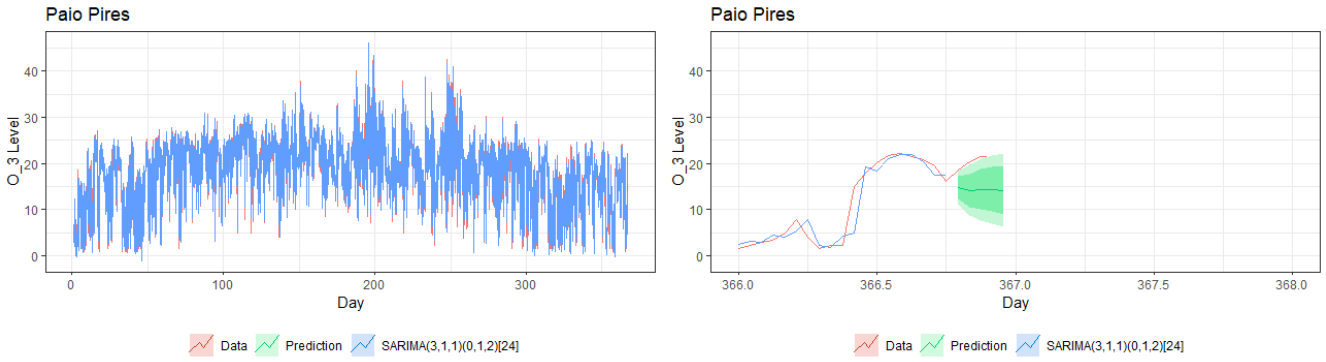

Figure 11: Original time series (red) of the year 2020 with forecasted values for the last five hours of December 2020 (green) (left). Zoomed plot (right)

5 Assignment 1 - Conclusion

We test our model out-of-sample, which means it had not seen the data it predicted. One thing that could be further improved is the methodology of real value tests: we compute MAPE on just 5 time period ahead, thus we train/test split with test split equal to 5 and train to the rest of the data set. We could undertake some kind of cross validation to improve the dependability of our results. This would assist us in drawing conclusions about the generalization of our model.

Besides, the study of residuals derived for these models, as well as the Ljung-Box test, reveal that the models are well-fitting to the data.

Furthermore, the residuals' ACF and PACF show some small correlations at several lags. This means that there is information on the data that the model is not able to explain. A combination of ARIMA and GARCH family models, for example, could be used to solve this problem, with the GARCH model modeling the conditional variance.

Moreover, we can observe that the forecast predictions are consistent, and therefore we conclude that the models provide useful information that may be used to anticipate future values of the 0_3 particle levels.

However, it is possible to see that the predicted values are lower than the true values for the vast majority of time series. This could be due to New Year's Eve; because we only have one year of data, the time series misses the pattern, and the conditions deviate from regularity. Adding an exogenous variable to our model to predict the effects of the holiday might be a smart idea in this case.

6 Part 2

6.1 The datasets

For this part of the assignment we were given 5 datasets referring to stocks of 5 different companies: EDP, NOS, GALP, MOTAENGIL and NOVABASE from March 20 2020 up to March 11 2022. Each dataset had 9 variables:

- **Date**

- **Open** The opening value of the stock for that day
- **Close** The closing value of the stock for that day
- **High** The highest value of the stock that day
- **Low** The lowest value of the stock that day
- **Number of shares**
- **Number of trades**
- **Turnover** a measure of stock liquidity
- **vwap** The Volume Weighted Average Price (the average price of a stock weighted by the total trading volume)

However, in the scope of this project, we only care about daily closing values, therefore only the variables **Date** and **Close** will be used.

6.2 First examination of the data

First of all, we must calculate the log return time series for each stock. Remember that the log-return is defined as

$$X_t = \log P_t - \log P_{t-1}$$

Where P_t is the price of the stock at time *t*.
We can now plot the log-return series for the closing values

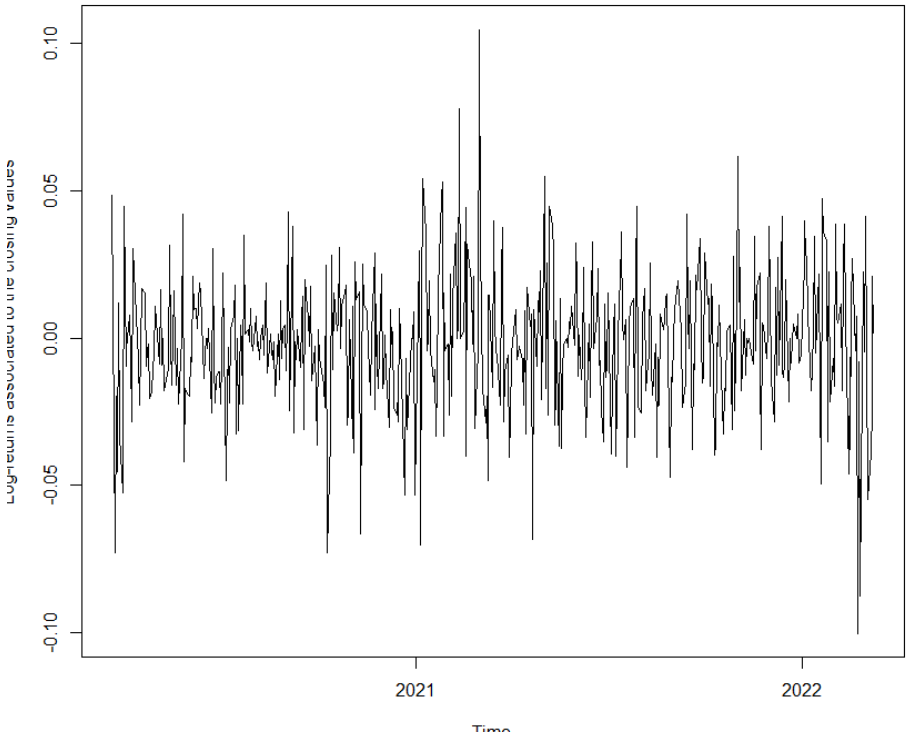

Figure 12: Plot of the log-returns for the closing values of EDP stock

We can spot some volatility around the beginning (March 2020), in the first few months of 2021, and again towards the end of the plot (March 2022). We can deduce that the first instance of high volatility might be related with the beginning of the COVID pandemic, while the last instance is probably related with the invasion of Ukraine by Russia, two events that have had a strong economical impact that is naturally going to be reflected in the stock market. This trend is spotted in all the other stocks, there is high volatility throughout the year following the beginning of the pandemic, and in some cases, in the beginning of the war in Ukraine as well. These can be seen in the Appendix.

Since we're evaluating a log-returns time series, we expect that the sample mean is close to 0, the variance should be around the order of 10^{-4} or smaller, the ACF should be negligible at all lags except the first, the distribution should have long tails, the sample ACF of the absolute and square values should be positive for most lags and the response to volatility should be asymmetric. Here's what we got for the EDP stock

Mean	Variance	Kurtosis
-0.0008648083	0.0003078225	6.940005

Table 5: Properties of the log-return time series for the EDP stock

We see that the mean is indeed close to 0, the variance is of the order 10^{-4} and the kurtosis is well above 3, the kurtosis of a normal distribution, meaning it does have heavy tails. Below is the boxplot of the log-returns for the EDP stock

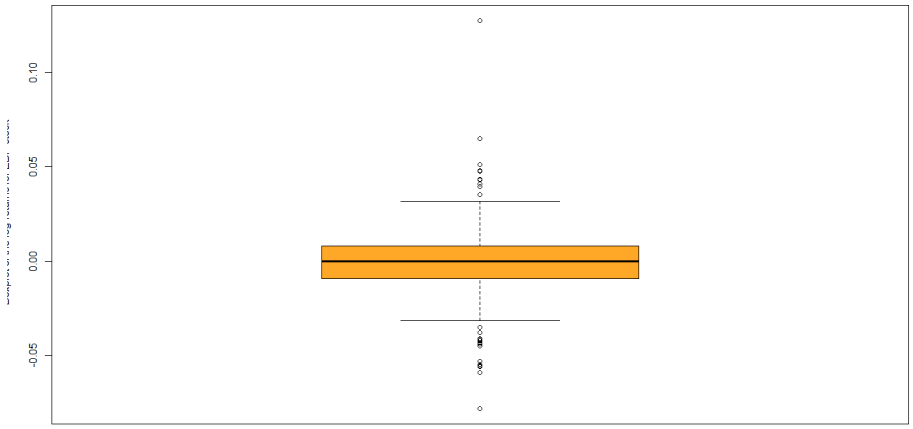

Figure 13: Boxplot of the log-returns for the closing values of EDP stock

Now we need to check the ACF for the log-returns, as well as the square and absolute values, to see if they also behave as expected for a financial time series.

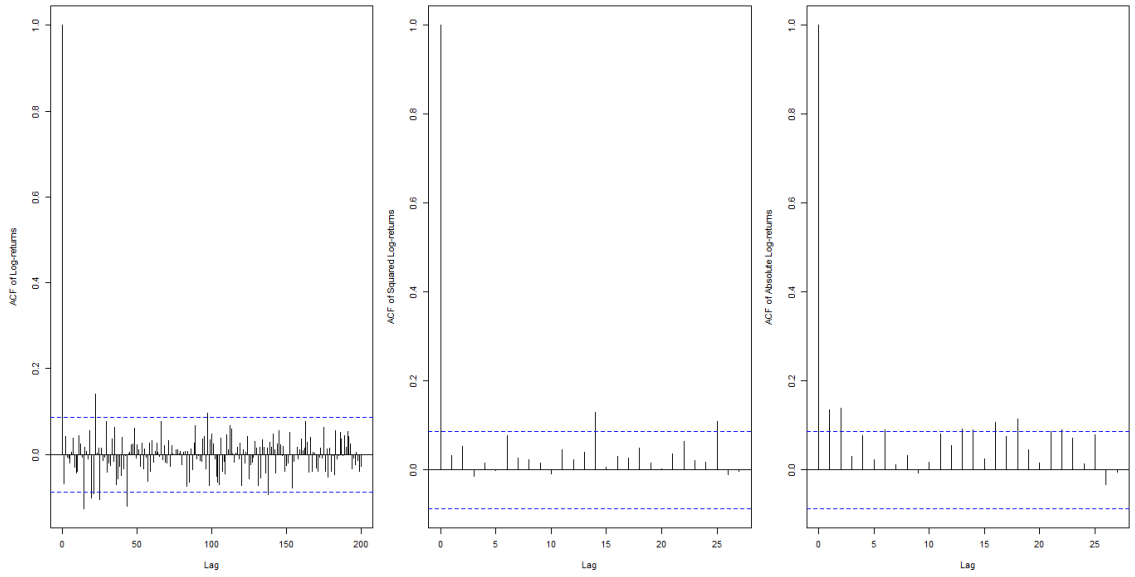

Figure 14: ACF of the log-returns for the closing values of EDP stock, as well as their square and absolute values

As expected, for most lags we are within the 95% confidence interval, and for the squared and absolute values, the ACF is positive in almost all lags.

6.3 The GARCH family

For this assignment we were asked to use a GARCH-type model to fit the series of log-returns. Unlike a SARIMA process, a GARCH (Generalized Autoregressive conditional heteroskedasticity) model can capture sudden bursts and time-varying volatility. As such, they are more appropriate to be used in a financial time series, which is the case here. Before introducing the various types of GARCH models used, we must first introduce the ARCH model, which is a simpler version that also models volatility. In the ARCH model, returns are uncorrelated, but dependent, and that dependence can be modelled by a quadratic function of its lagged values. The returns are modelled as such:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i X_{t-i}^2 \end{cases}$$

Where Z_t is standard Gaussian white noise and σ_t is a time-dependent standard deviation.

To implement our models, we'll be using the *rugarch* package. This package allows us to fit a variety of GARCH-type models to our log-returns series. We'll be using 4 of the 5 types learned in class, since fiGARCH is not supported by this package. Those are:

- GARCH
- iGARCH
- GARCH-M
- apARCH

This package also allows us to use a variety of conditional distributions. We'll be using the Normal distribution (norm), the Student distribution (std) and the Generalized Error distribution (ged).

6.3.1 GARCH

The GARCH model is the ARMA equivalent of the ARCH model. Here the σ_t term takes the form of:

$$\sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i X_{t-i}^2 + \sum_{i=j}^{q} b_j \sigma_{t-j}^2$$

The GARCH model's advantage towards the ARCH model is that it is more parsimonious than the ARCH model and therefore is a better fit for modeling time series data when the data exhibits heteroskedacisticity and volatility clustering. [4]

6.3.2 iGARCH

The iGARCH model has the same form as the GARCH model but adds the condition

$$\sum_{i=1}^{p} a_i + \sum_{i=j}^{q} b_j = 1$$

Thus importing a unit root to the GARCH process.

6.3.3 GARCH-M

The GARCH-M model adds a heteroskedasticity term into the mean equation. That is

$$Y_t = u + \delta\sigma_t^2 + X_t$$

where u and δ are constants. A positive δ indicates that the return is positively related to its past volatility.

6.3.4 apARCH

The main advantage of the apARCH model is that, unlike other GARCH-type models, it can deal with asymmetric response of the volatility for positive and negative shocks. This can be convenient in the context of a financial time series, since they tend to react more to negative shocks than positive ones. In this model the volatility has the following form:

$$\sigma_t^\delta = a_0 + \sum_{i=1}^{p} a_i \left(|X_{t-i}| - \gamma_i X_{t-i}\right)^\delta + \sum_{j=1}^{q} b_j \sigma_{t-j}^\delta$$

where $\delta \geq 0$ represents the parameter for the power term and $-1 \leq \gamma_i \leq 1$ is the leverage parameter.

6.4 Model Selection

To select our model for each of the 5 time series, we wanted to try out all 4 types of models listed before, and for a lot of different combinations of parameters. As such, we created the function *select.GARCH* that receives as inputs *model, data, p, q* and *dist*. The variable *model* refers to one of the 4 types of models described above, *data* is in our case, the log-return time series, *p* and *q* are limits for the p and q parameters and *dist* is one of the 3 conditional distributions described above. For each distribution we ran all 4 model types with the limit of 5 for both p and q, therefore generating 100 models. We then can choose the best model based on the AIC and BIC criteria, that are both part of the output of our function. It's worth noting that while sometimes there is a model that is the best performer in both criteria, but other times it might not be obvious which model is better. While the function outputs both AIC and BIC, the latter won't be considered for model selection, since we found that it penalizes apARCH models a lot more than others. We also must

mention that we were unable to get a apARCH model for NOS and NOVABASE stock, as the function did not converge. The 5 following tables will show the best models we got for each stock

Distribution	Model	AIC	BIC
norm	GARCH-M(1,1)	-4.6847	-4.6514
std	GARCH-M(1,1)	-4.7026	-4.6610
ged	GARCH-M(1,1)	-4.7040	-4.6624

Table 6: Comparison of the best models for the log-return time series of the daily closing values for the EDP stock

Distribution	Model	AIC	BIC
norm	apARCH(2,4)	-4.7056	-4.6140
std	apARCH(2,4)	-4.7315	-4.6316
ged	apARCH(2,4)	-4.7305	-4.6306

Table 7: Comparison of the best models for the log-return time series of the daily closing values for the GALP stock

Distribution	Model	AIC	BIC
norm	apARCH(1,4)	-4.8185	-4.7435
std	apARCH(1,3)	-4.8747	-4.7998
ged	apARCH(1,4)	-4.8654	-4.7822

Table 8: Comparison of the best models for the log-return time series of the daily closing values for the MOTAENGIL stock

Distribution	Model	AIC	BIC
norm	iGARCH(3,4)	-5.3515	-5.2849
std	iGARCH(1,1)	-5.4979	-5.4646
ged	iGARCH(1,2)	-5.4679	-5.4262

Table 9: Comparison of the best models for the log-return time series of the daily closing values for the NOS stock

Distribution	Model	AIC	BIC
norm	iGARCH(1,5)	-5.3421	-5.2837
std	iGARCH(1,1)	-5.4586	-5.4252
ged	iGARCH(1,1)	-5.4725	-5.4392

Table 10: Comparison of the best models for the log-return time series of the daily closing values for the NOVABASE stock

6.4.1 Parameter estimation

For the EDP stock we obtained a GARCH-M(1,1) model. We can estimate its parameters using the *ugarchfit* function and we obtain:

$$a_0 = 3.860905e-05,\ a_1 = 0.09588424,\ b_1 = 0.08383353,\ u = -1.845947e-03,\ \delta = 14.74516$$

A positive δ indicates that the return is positively related to its past volatility.

For the GALP stock we got a apARCH(2,4) model, the most complex of the 5 we obtained. The parameters are:

$$a_0 = 3.101330e-04,\ a_1 = 0.07994460,\ a_2 = 0.05472751\ b_1 = 1.213591e-20,\ b_2 = 0.06142236,\ b_3 = 0.3428542,$$
$$b_4 = 0.3664335,\ \gamma_1 = 0.9999967,\ \gamma_2 = -0.9999632,\ \delta = 14.67475$$

Since $\gamma_1 > 0$, we can deduce that negative shocks have a higher influence on volatility than positive shocks, which is normal behaviour for a financial time series.

For the MOTAENGIL stock we got a apARCH(1,3) model with the following parameters:

$$a_0 = 1.214551e-05,\ a_1 = 0.1518232,\ b_1 = 0.125059,\ b_2 = 0.1236193,\ b_3 = 0.3886196,\ \gamma_1 = -0.7345449,$$
$$\delta = 23.67479$$

Here we get a negative value for γ_1, which means that positive shocks have a higher influence on volatility than negative shocks, which is suprising in a financial time series.

For the NOS stock we get a iGARCH(1,1) model, and the parameters are:

$$a_0 = 7.046636e-07,\ a_1 = 0.01769066,\ b_1 = 0.9823093$$

We can observe that indeed $a_1 + b_1 = 1$

Finally, for the NOVABASE stock we also get a iGARCH(1,1) model. We got the following parameters:

$$a_0 = 7.046636e-07,\ a_1 = 0.07672307,\ b_1 = 0.9232769$$

6.5 Residual Diagnostics

In figure 15 below we can see a plot of the ACF for the standard residuals, squared standard residuals and absolute value of standard residuals for the EDP model. We will only show this plot for the EDP model for convenience.

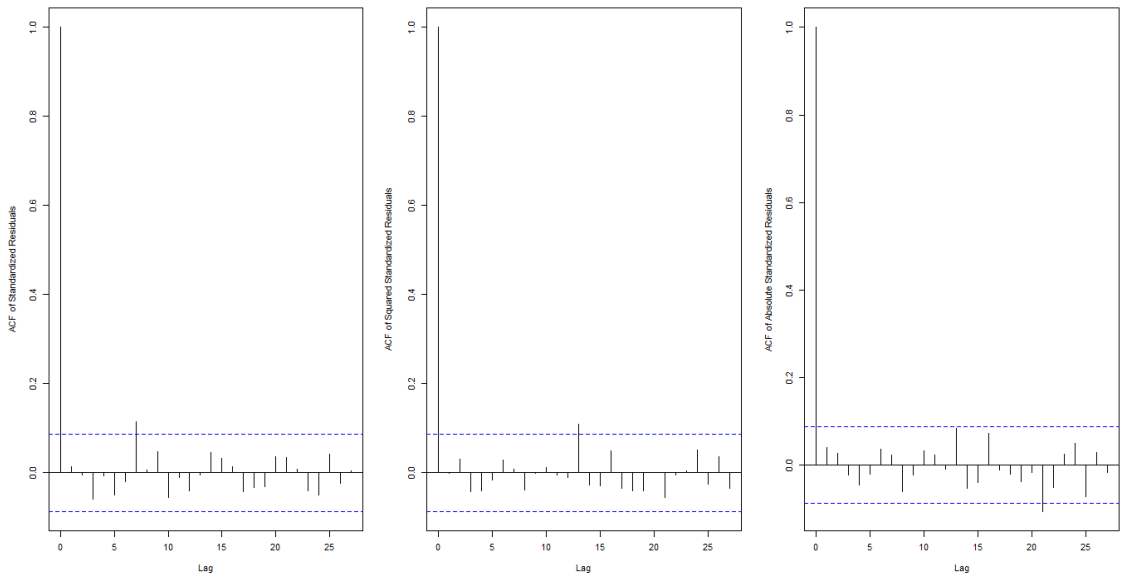

Figure 15: Plot of the ACF of the standard residuals, squared standard residuals, and absolute value of standard residuals, respectively, for the EDP model

We can see that for almost all lags we are within the 95% confidence interval. This is good news, it means that the residuals have constant volatility, and the volatility is completely explained by the model. We obtain similar plots for all of the 5 models.

Next, we'll be performing a number of tests. The *ARCH LM* test, which is a test to detect autoregressive conditional heteroscedasticity, the *Ljung-Box* test, which tests whether the standardized residuals are correlated, and the *Shapiro-Wilk* test, to determine the normality of the standardized residuals. We can see the results in the table below.

Stock	ARCH LM 1	ARCH LM 2	ARCH LM 3	Ljung-Box lag 10	Ljung-Box lag 20	Ljung-Box lag 30
EDP	0.3514 (lag 3)	0.5658 (lag 5)	0.7319 (lag 7)	0.2287	0.5667	0.8162
GALP	0.2966 (lag 7)	0.5840 (lag 9)	0.3084 (lag 11)	0.8592	0.9097	0.9178
MOTAENGIL	0.4866 (lag 5)	0.7001 (lag 7)	0.6519 (lag 9)	0.8047	0.5273	0.6799
NOS	0.4164 (lag 3)	0.7529 (lag 5)	0.9150 (lag 7)	0.2725	0.05461	0.1544
NOVABASE	0.4504 (lag 3)	0.6675 (lag 5)	0.8167 (lag 7)	0.3979	0.2238	0.01802

Table 11: p-values for various tests for all the different stocks

We can see that for all standard significance levels, the *ARCH LM* tests are not rejected. This means the residuals have constant volatility, as intended. For the *Ljung-Box* tests, the null hypothesis states that the residuals are uncorrelated. This hypothesis is not rejected for almost every p-value in this table. The only 2 exceptions are for the NOS stock, at lag 20, for the 10% significance level, and for the NOVABASE stock at lag 30 for the significance levels of 5% and 10%. It's worth noting here that these are the 2 stocks for which we were unable to fit an apARCH model, which may have ended up being more appropriate. Overall we do not reject the null hypothesis and can assume that the residuals are uncorrelated, as intended.

Next we want to evaluate the normality of the standardized residuals, and for this we use the *Shapiro-Wilk test*. We obtain the following table

Stock	Shapiro-Wilk test
EDP	0.008226
GALP	1.725e-05
MOTAENGIL	3.264e-08
NOS	<2.2e-16
NOVABASE	6.118e-11

Table 12: p-values for *Shapiro-Wilk* test for all the different stocks

These results are incredibly bad. We realized that the *Shapiro-Wilk* test is not appropriate to apply in this context, as it is highly sensitive to outliers. As such, we've resorted to a simple QQ-Plot. Below are the 5 QQ-Plots.

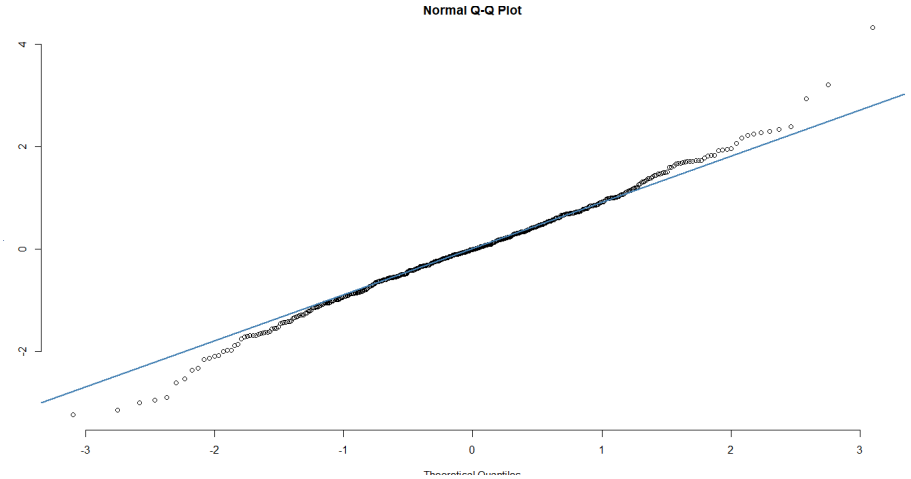

Figure 16: QQ-Plot of the standardized residuals for the EDP model

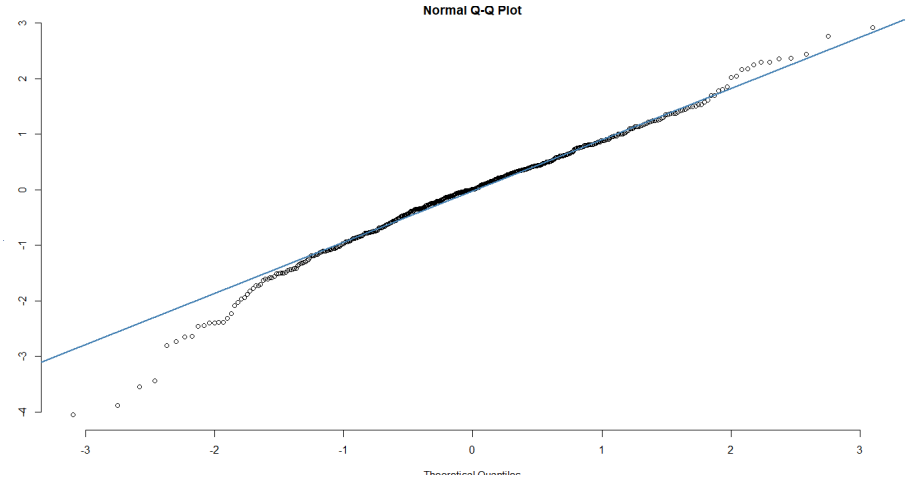

Figure 17: QQ-Plot of the standardized residuals for the GALP model

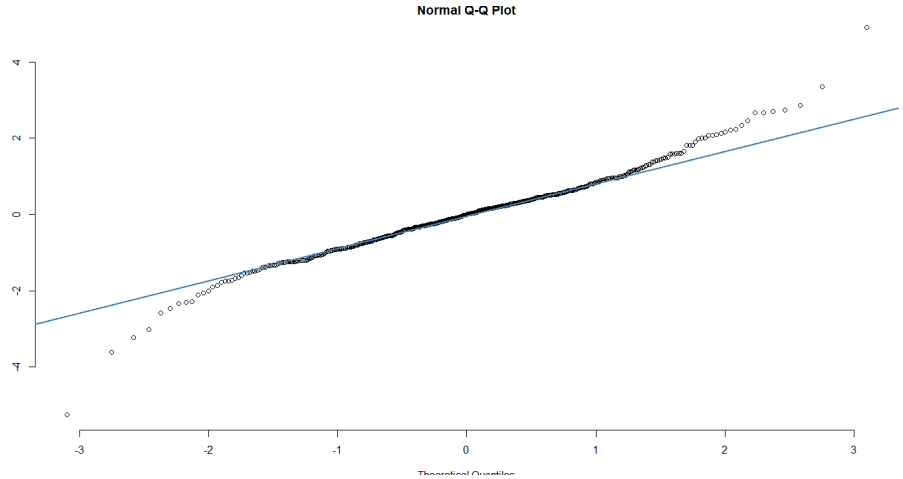

Figure 18: QQ-Plot of the standardized residuals for the MOTAENGIL model

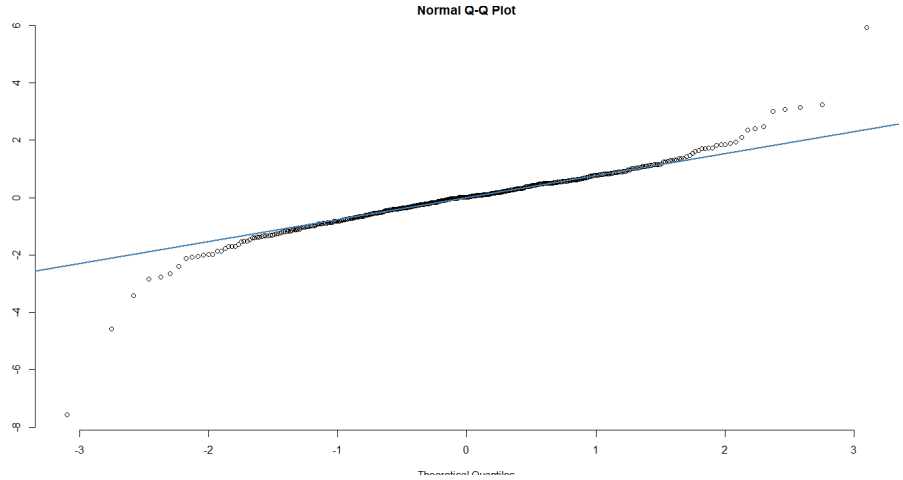

Figure 19: QQ-Plot of the standardized residuals for the NOS model

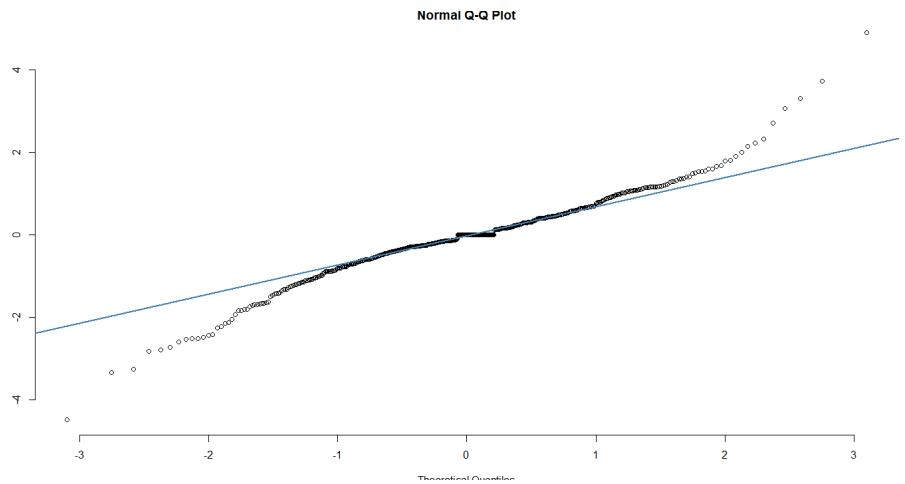

Figure 20: QQ-Plot of the standardized residuals for the NOVABASE model

Looking at these plots, the assumption of normality seems like a reasonable one for all models, except the model for the NOVABASE stock. It's worth reminding again that we were unable to fit an apARCH model to the NOVABASE time series, which could have yielded better results.

6.6 Assignment 2 - Conclusion

We were tasked with fitting a GARCH-type model to each of the 5 time series referring to 5 different companies. We have found that GARCH-type models seem to be a great fit for financial time series in order to predict volatility. Not only that, we can conclude that there is no single best model for financial time series. This is illustrated by the fact that we have found that for some companies, the apARCH model was the better fit, while for others we got other models, such as iGARCH and GARCH-M. In general, we got good performing models that also performed just as expected for a financial time series. It is worth noting that the model for NOVABASE wasn't entirely well-behaved, as it failed to get normally distributed standardized residuals, and failed the *Ljung-Box* test for 2 of the standard significance levels. Not only that, the model that we fitted to the MOTAENGIL log-return series did not behave as anticipated, as we concluded that the volatility was more influenced by positive shocks rather than negative ones as is usual in financial time series. Our models only predict volatility and not actual stock prices. If we wanted to go that extra step, we could do so by building a ARIMA-GARCH hybrid model, but that was outside the scope of this assignment.

7 Appendix

7.1 Project 1

7.1.1 Restelo

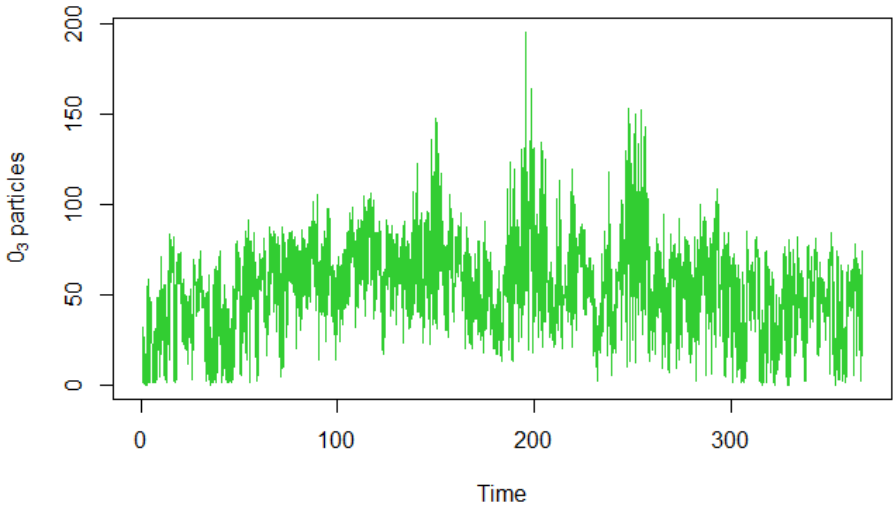

Figure 21: Hourly levels of O_3 particles in $\mu g/m3$ in Restelo since 01/01/2020 until 31/12/2020

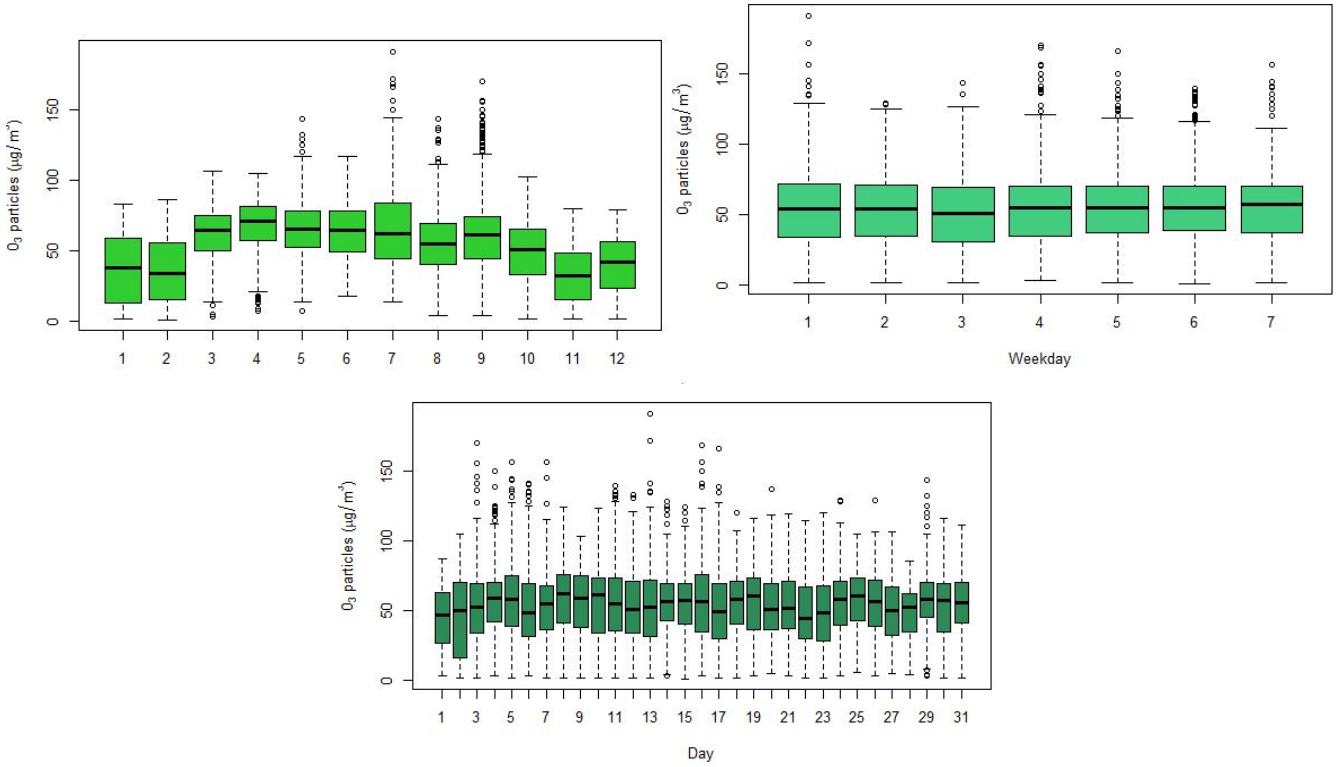

Figure 22: Boxplots of the O_3 particles levels in Restelo with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

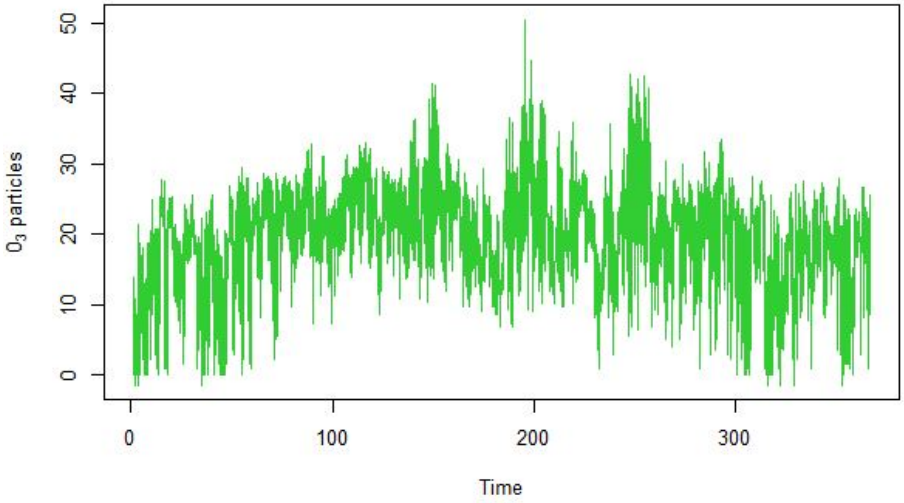

Figure 23: Hourly levels of O3 particles in $\mu g/m3$ in Restelo since 01/01/2020 until 31/12/2020, after the log transformation of the data

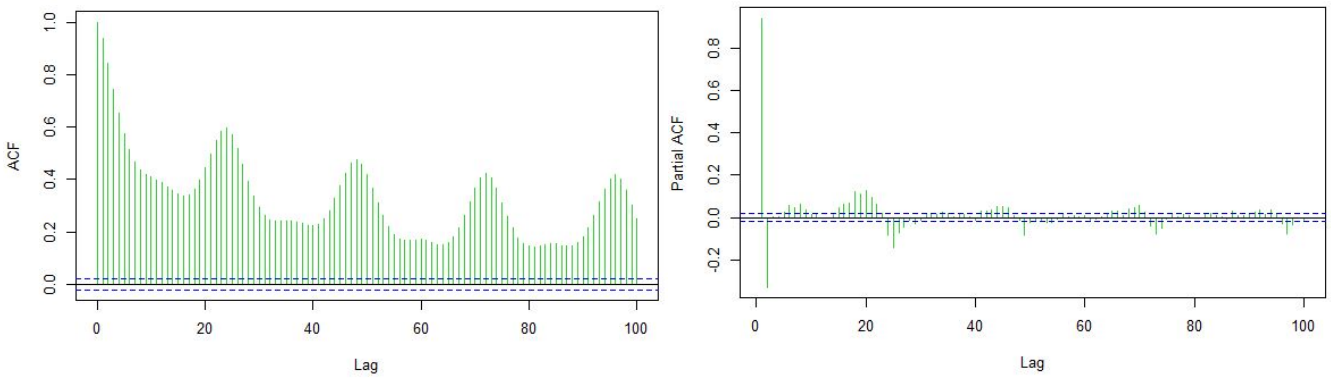

Figure 24: Characteristics of O_3 time series, with Box-Cox transformation, at Restelo station. ACF on the left and PACF on the right

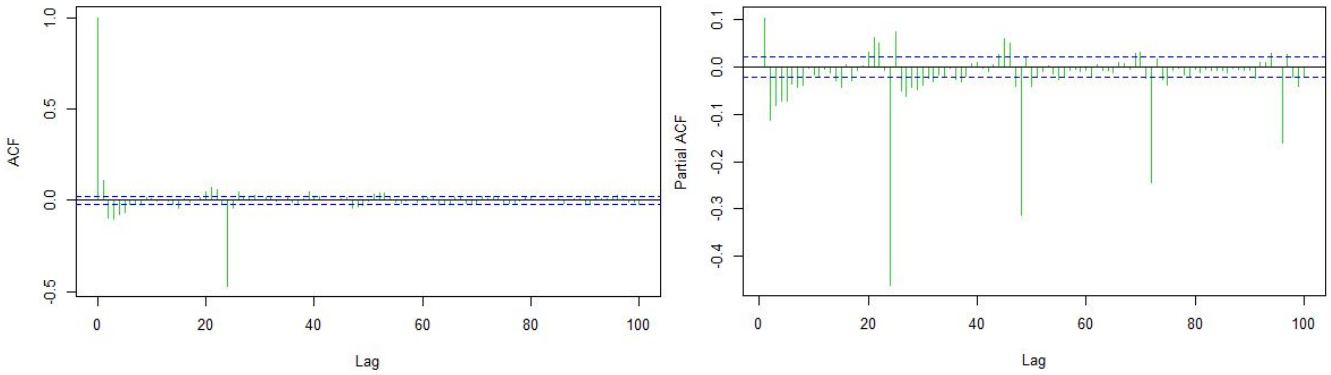

Figure 25: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Restelo station. ACF on the left and PACF on the right

Restelo			
Model	**AIC**	**AICc**	**BIC**
ARIMA(4,1,1)(1,0,0)[24] (auto.arima)	39052.12	39052.14	39101.68
ARIMA(3,1,2)(0,1,2)[24] (1st best AIC and 3rd best BIC)	37927.61	37927.62	37984.22
ARIMA(1,1,2)(0,1,2)[24] (1st best BIC)	37938.05	37938.06	37980.51
ARIMA(1,1,2)(1,1,1)[24] (2nd best BIC)	37938.09	37938.1	37980.56

Table 13: Comparison of the best fitted models for Restelo time series

Restelo		
ARIMA(3,1,2)(0,1,2)[24]	**Coefficient**	**p-value**
ar1	1.7349	0
ar2	-0.9348	0
ar3	0.1663	0
ma1	-1.6389	0
ma2	0.6428	0
sma1	-0.9059	0
sma2	-0.0490	6.330e-06

Table 14: Summary of ARIMA(3,1,2)(0,1,2)[24] coefficients

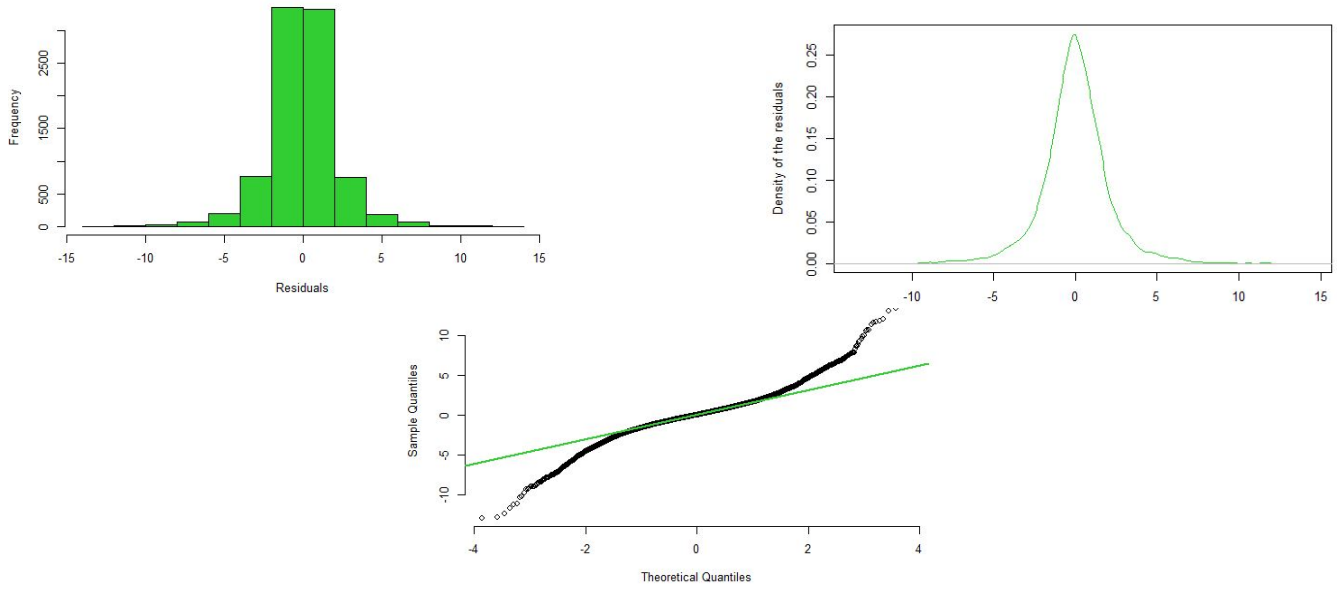

Figure 26: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Restelo

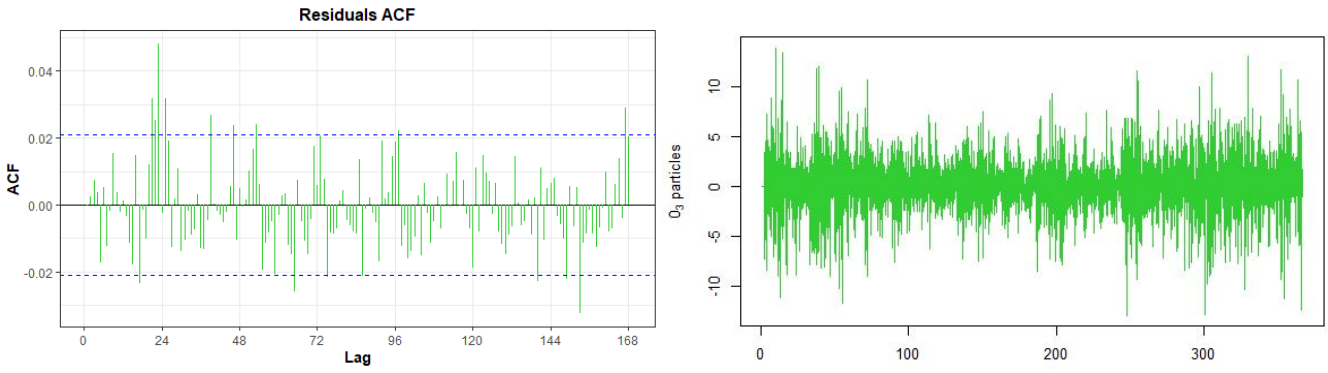

Figure 27: ACF (left) and plot (right) of the residuals for Restelo

7.1.2 Sobreiro

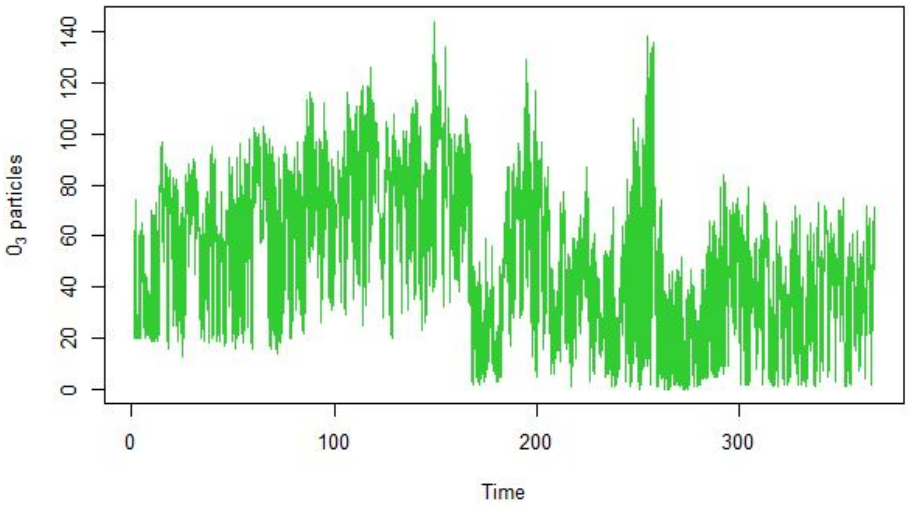

Figure 28: Hourly levels of O_3 particles in $\mu g/m3$ in Sobreiro since 01/01/2020 until 31/12/2020

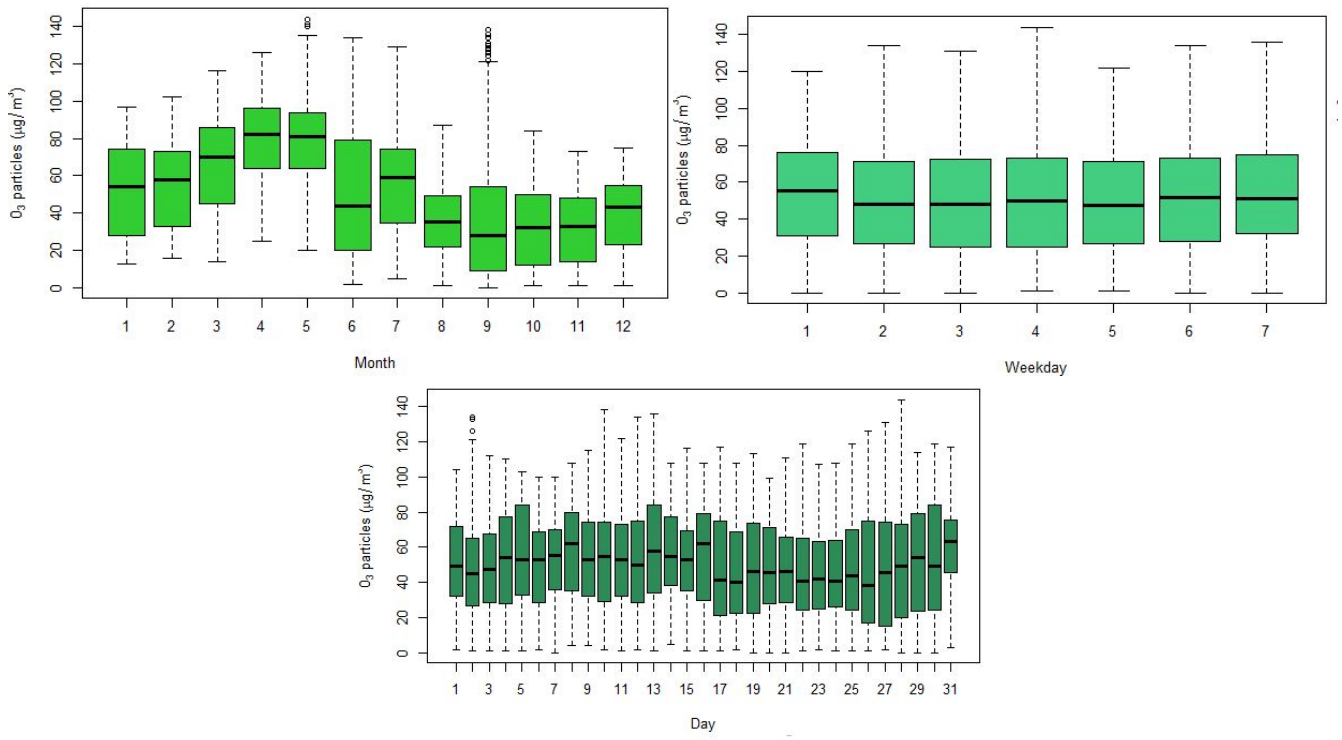

Figure 29: Boxplots of the O_3 particles levels in Sobreiro with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

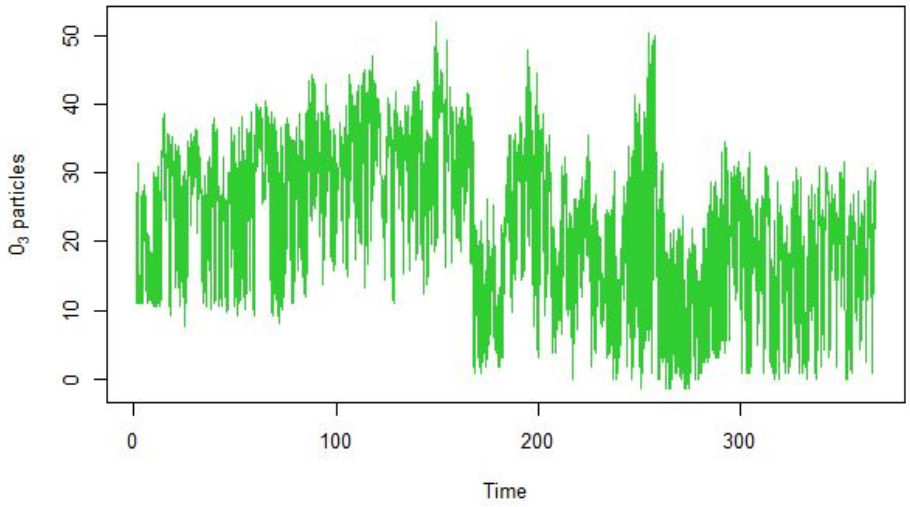

Figure 30: Hourly levels of O3 particles in $\mu g/m3$ in Sobreiro since 01/01/2020 until 31/12/2020, after the log transformation of the data

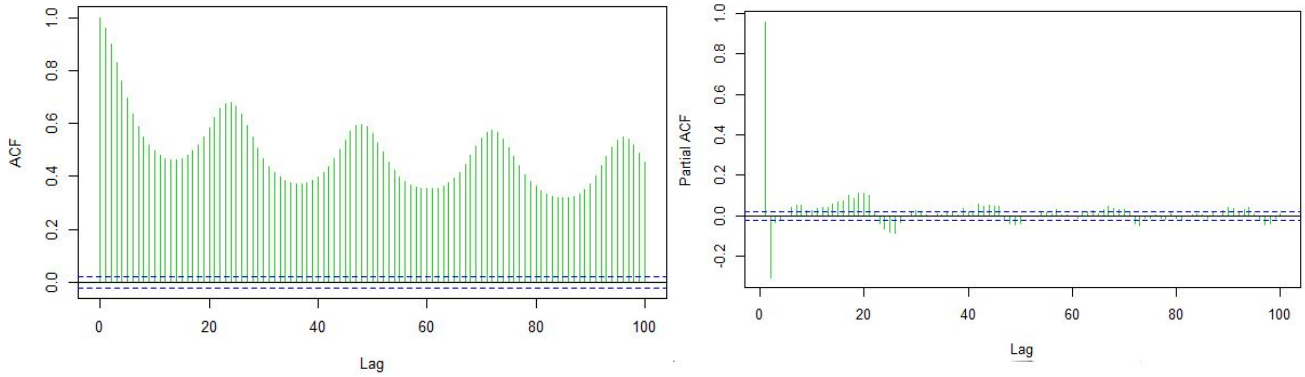

Figure 31: Characteristics of O_3 time series, with Box-Cox transformation, at Sobreiro station. ACF on the left and PACF on the right

Sobreiro		
ARIMA(2,1,1)(0,1,1)[24]	**Coefficient**	**p-value**
ar1	1.0833	0
ar2	-0.1858	0
ma1	-0.9821	0
sma1	-0.9643	0

Table 16: Summary of ARIMA(2,1,1)(0,1,1)[24] coefficients

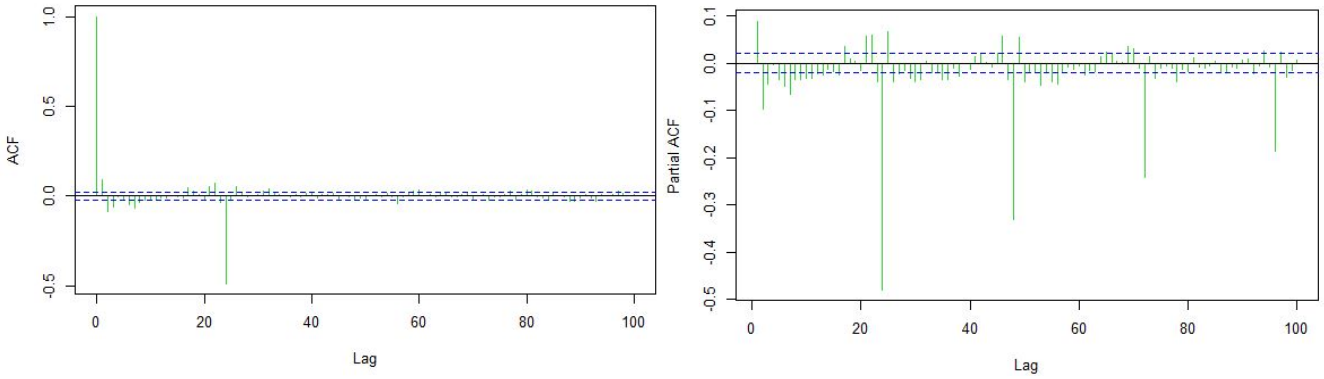

Figure 32: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Sobreiro station. ACF on the left and PACF on the right

Sobreiro			
Model	**AIC**	**AICc**	**BIC**
ARIMA(4,1,4)(2,0,0)[24] (auto.arima)	42774.83	42774.86	42852.71
ARIMA(2,1,3)(0,1,1)[24] (1st best AIC)	41994.48	41994.49	42044.02
ARIMA(2,1,3)(2,1,1)[24] (4th best AIC)	41996.97	41996.99	42060.67
ARIMA(2,1,1)(0,1,1)[24] (5th best AIC and 1st best BIC)	41998.43	41998.44	42033.82

Table 15: Comparison of the best fitted models for Sobreiro time series

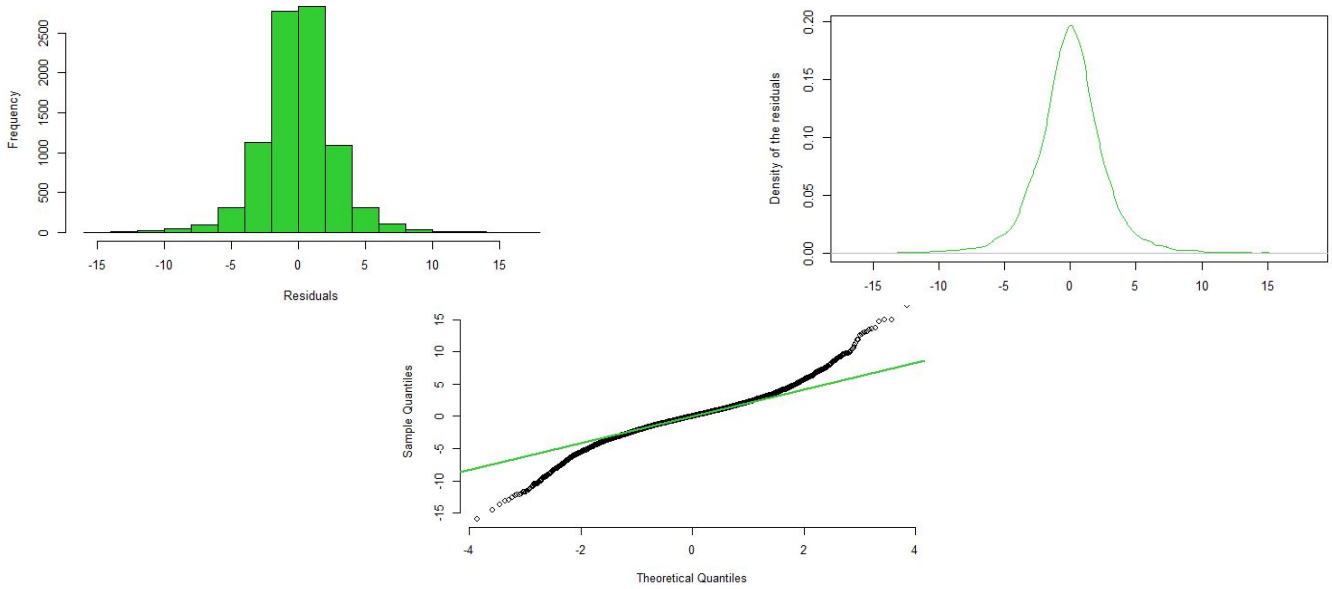

Figure 33: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Sobreiro

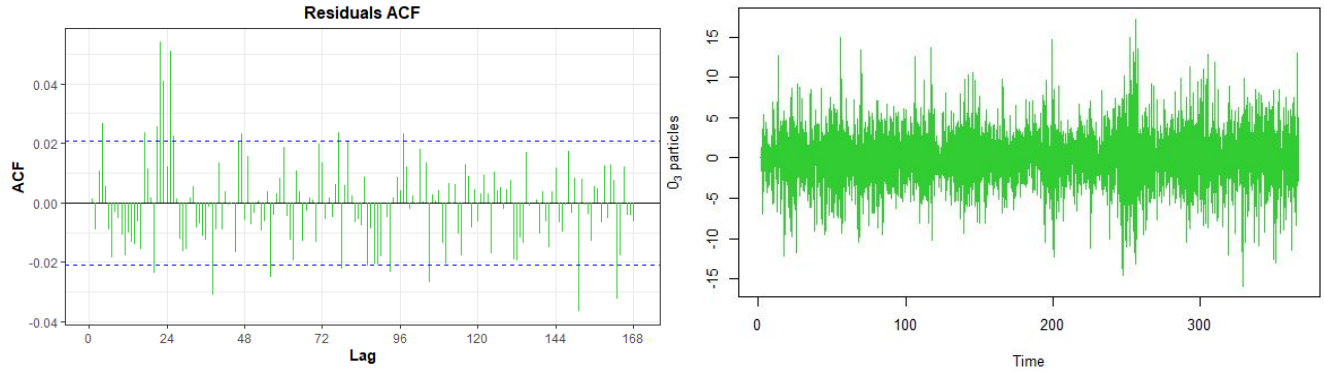

Figure 34: ACF (left) and plot (right) of the residuals for Sobreiro

7.1.3 VN Telha-Maia

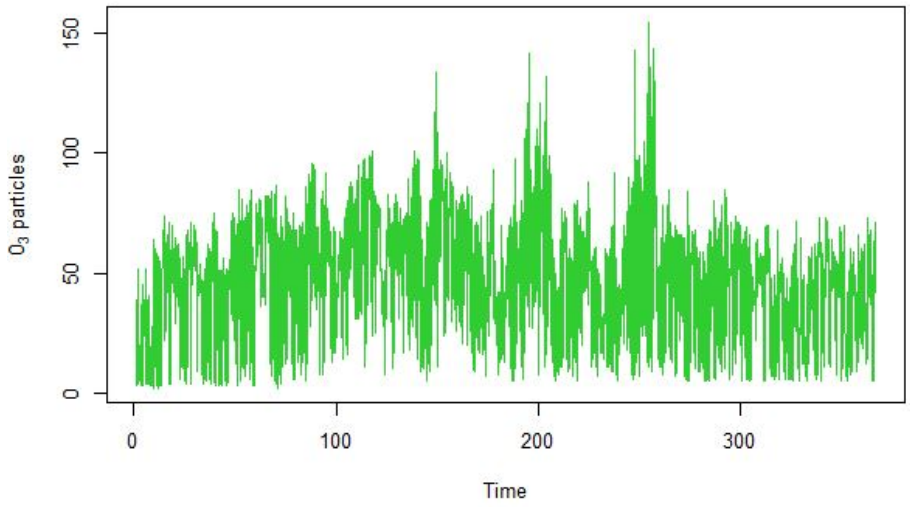

Figure 35: Hourly levels of O_3 particles in $\mu g/m3$ in VN Telha-Maia since 01/01/2020 until 31/12/2020

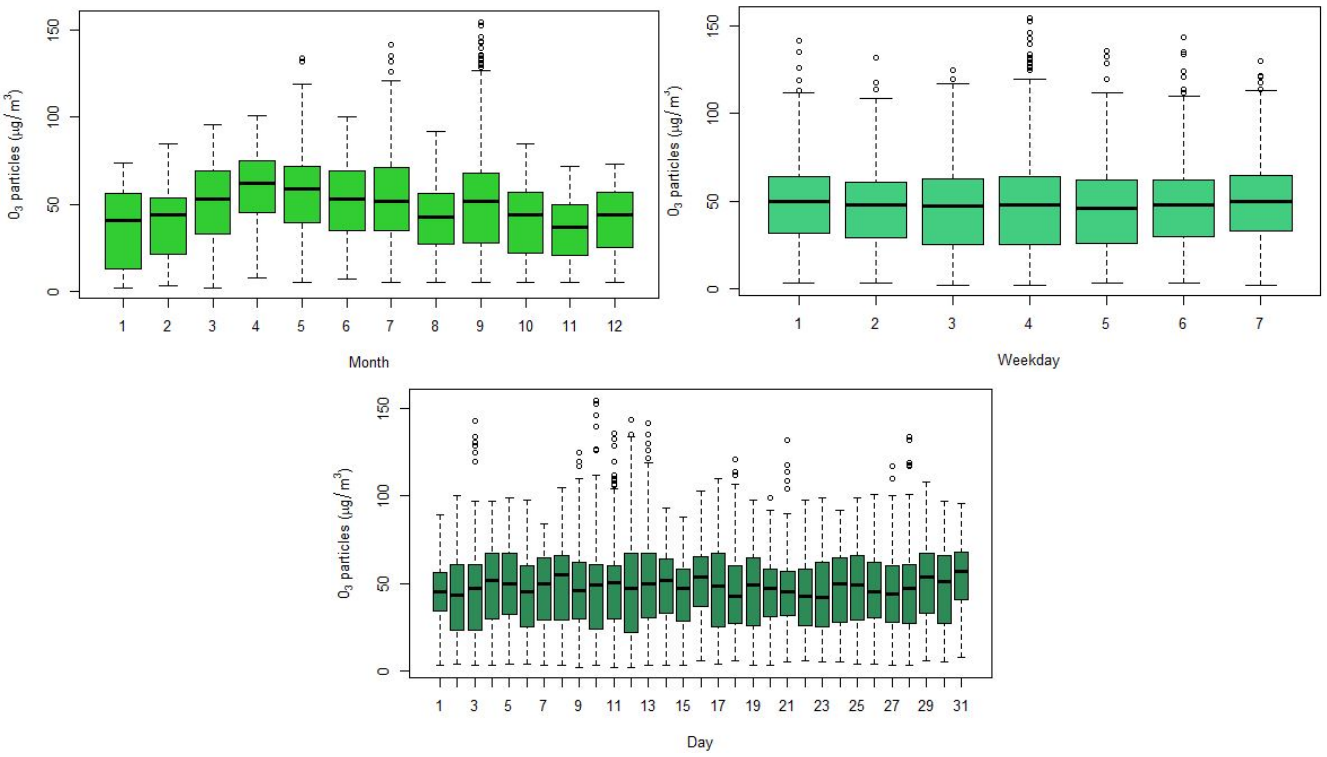

Figure 36: Boxplots of the O_3 particles levels in VN Telha-Maia with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

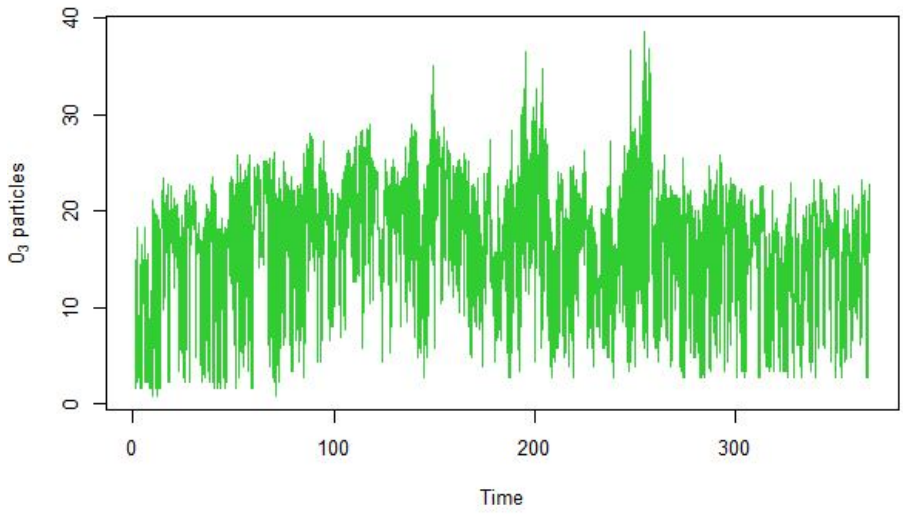

Figure 37: Hourly levels of O3 particles in $\mu g/m3$ in VN Telha-Maia since 01/01/2020 until 31/12/2020, after the log transformation of the data

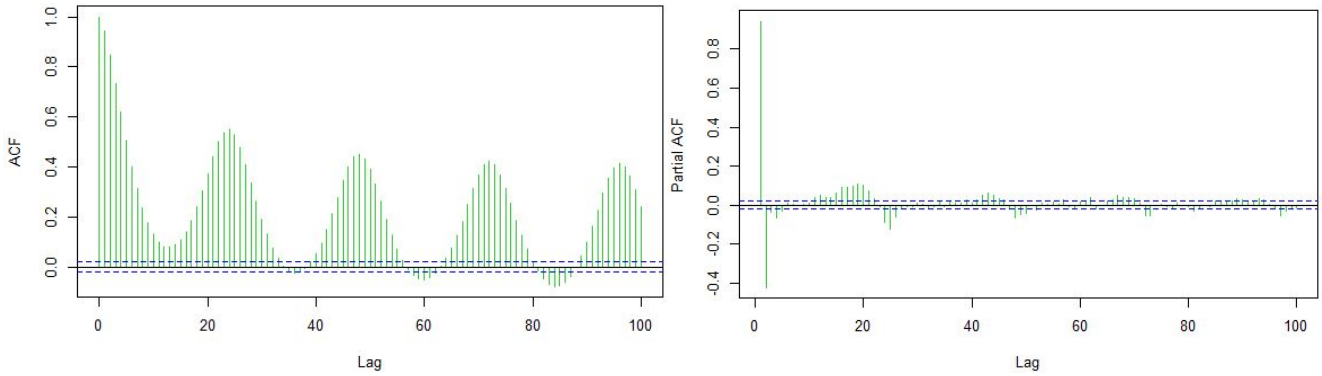

Figure 38: Characteristics of O_3 time series, with Box-Cox transformation, at VN Telha-Maia station. ACF on the left and PACF on the right

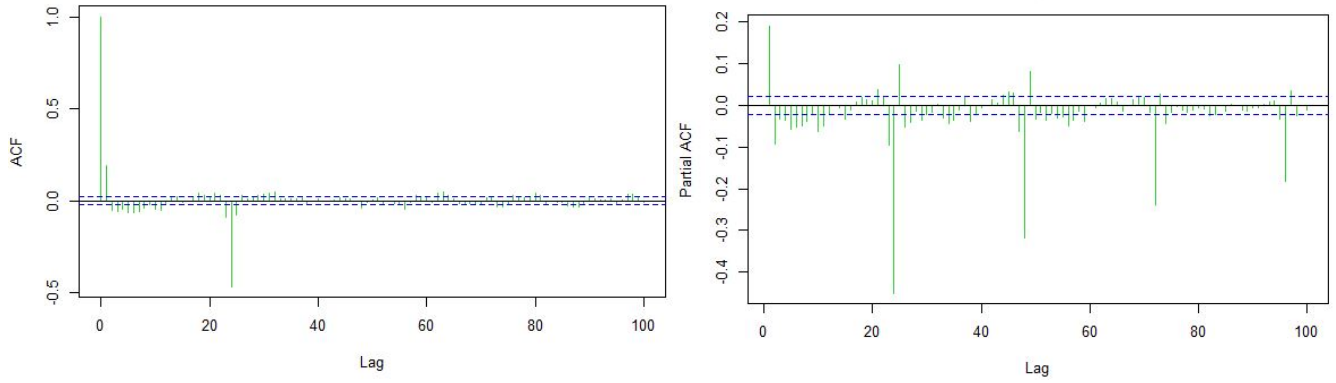

Figure 39: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at VN Telha-Maia station. ACF on the left and PACF on the right

VN Telha-Maia			
Model	**AIC**	**AICc**	**BIC**
ARIMA(5,0,1)(2,1,0)[24] (auto.arima)	36642.49	36642.52	36713.27
ARIMA(4,1,1)(1,1,2)[24] (1st best AIC)	34662.63	34662.65	34726.32
ARIMA(2,1,3)(0,1,2)[24] (2nd best AIC and 3rd best BIC)	34663.52	34663.53	34720.13
ARIMA(2,1,2)(0,1,2)[24] (1st best BIC)	34668.99	34669	34718.53

Table 17: Comparison of the best fitted models for VN Telha-Maia time series

VN Telha-Maia		
ARIMA(2,1,3)(0,1,2)[24]	**Coefficient**	**p-value**
ar1	1.0553	0
ar2	-0.1575	8.070724e-05
ma1	-0.8666	0
ma2	-0.1262	2.987737e-03
sma1	-0.9094	0
sma2	-0.0444	5.153165e-05

Table 18: Summary of ARIMA(2,1,3)(0,1,2)[24] coefficients

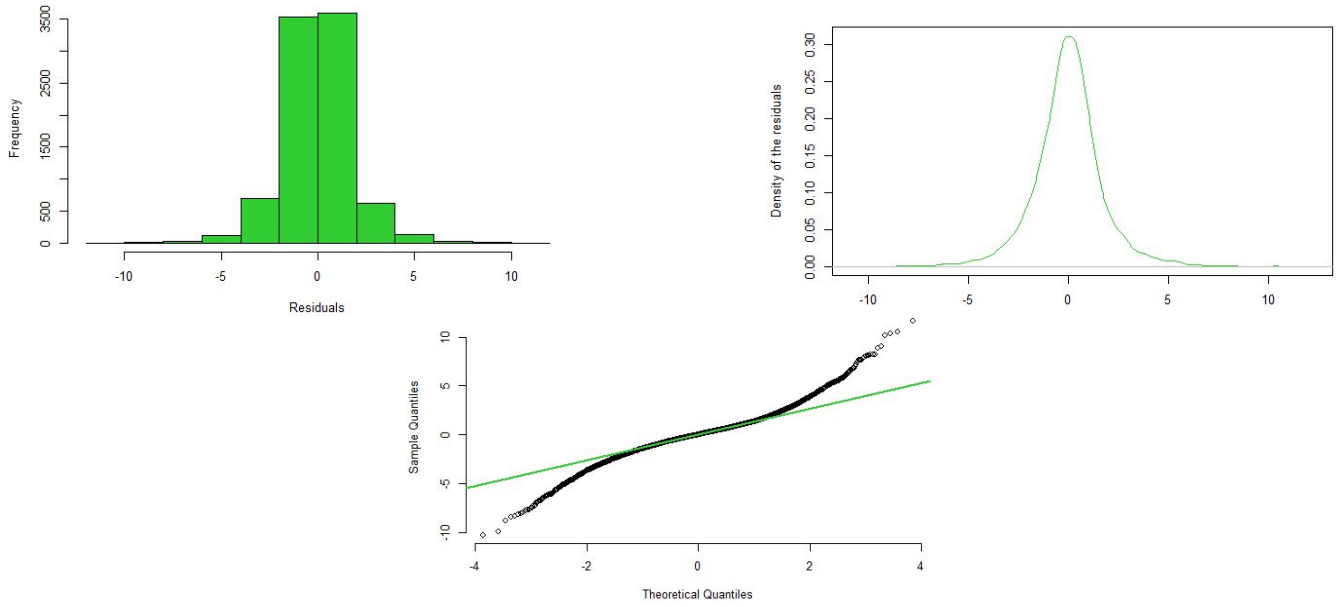

Figure 40: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for VN Telha-Maia

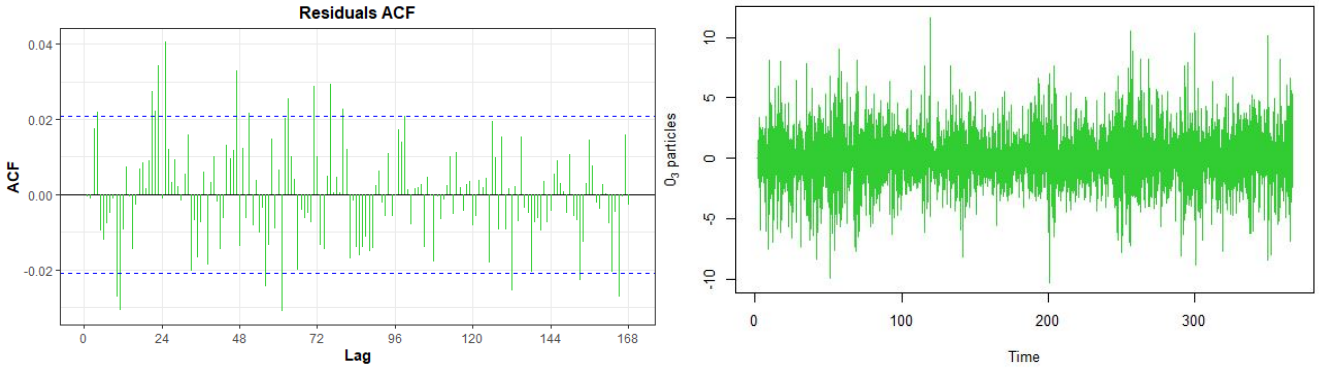

Figure 41: ACF (left) and plot (right) of the residuals for VN Telha-Maia

7.1.4 Antas-Espinho

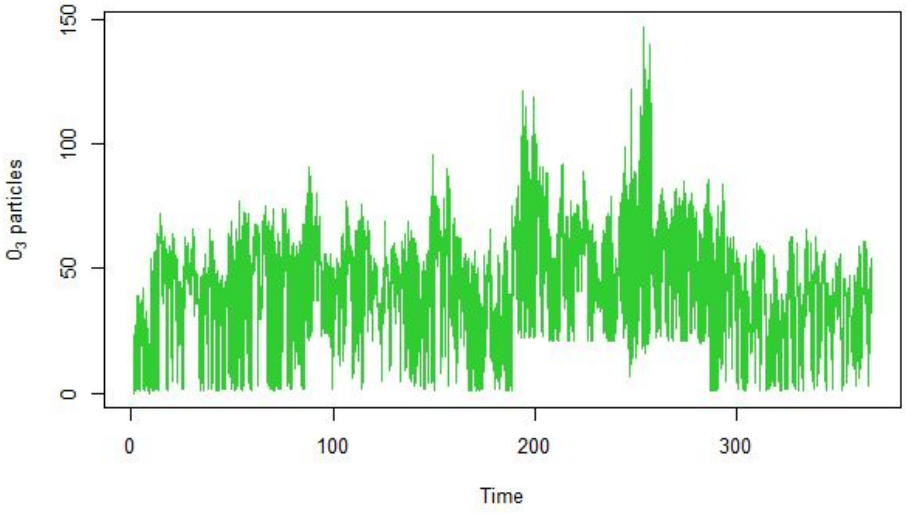

Figure 42: Hourly levels of O_3 particles in $\mu g/m3$ in Antas-Espinho since 01/01/2020 until 31/12/2020

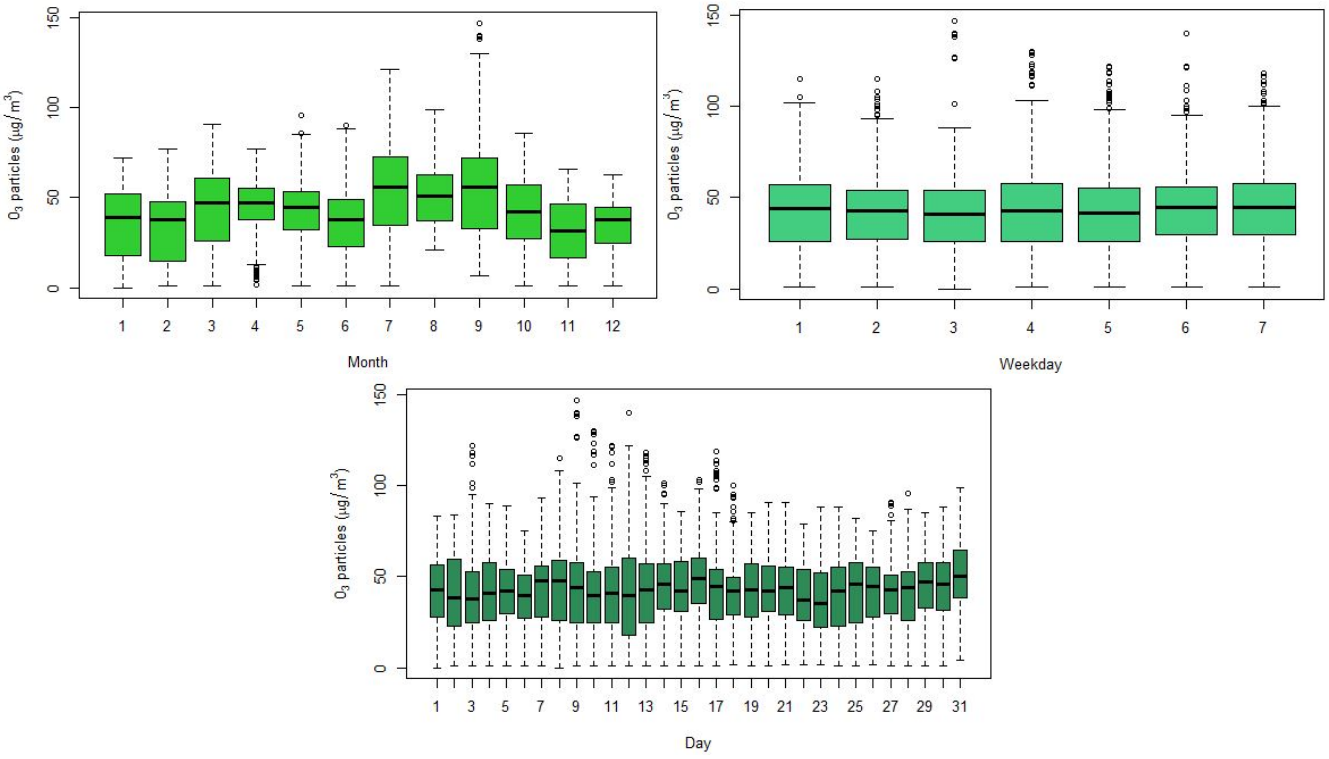

Figure 43: Boxplots of the O_3 particles levels in Antas-Espinho with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

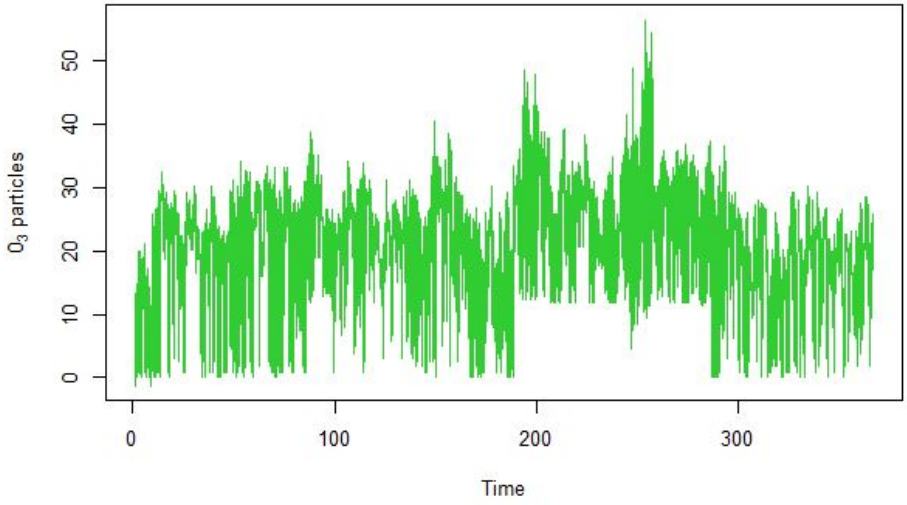

Figure 44: Hourly levels of O3 particles in $\mu g/m3$ in Antas-Espinho since 01/01/2020 until 31/12/2020, after the log transformation of the data

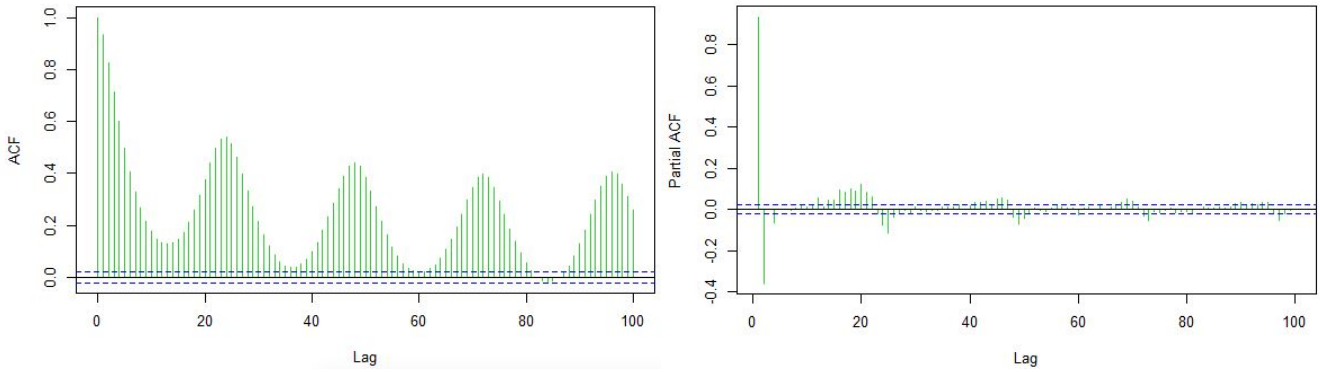

Figure 45: Characteristics of O_3 time series, with Box-Cox transformation, at Antas-Espinho station. ACF on the left and PACF on the right

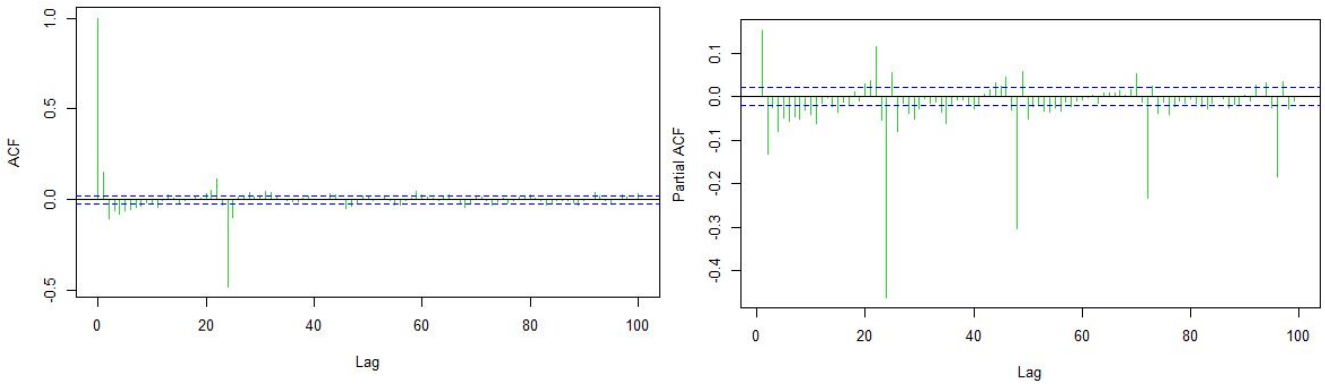

Figure 46: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Antas-Espinho station. ACF on the left and PACF on the right

Antas-Espinho			
Model	**AIC**	**AICc**	**BIC**
ARIMA(2,1,0)(2,0,0)[24] (auto.arima)	44446.35	44446.35	44481.75
ARIMA(1,1,2)(2,1,2)[24] (1st best AIC)	46443.06	46443.07	46499.67
ARIMA(1,1,3)(2,1,2)[24] (2nd best AIC)	46444.79	46444.81	46508.49
ARIMA(1,1,3)(1,1,1)[24] (1st best BIC and 3rd best AIC)	46445.25	46445.27	46494.8

Table 19: Comparison of the best fitted models for Antas-Espinho time series

Antas-Espinho		
ARIMA(1,1,3)(1,1,1)[24]	**Coefficient**	**p-value**
ar1	0.8768	0
ma1	-0.7559	8.070724e-05
ma2	-0.2260	0
ma3	-0.0063	6.066401e-01
sar1	0.0755	5.134537e-11
sma1	-0.9593	0

Table 20: Summary of ARIMA(1,1,3)(1,1,1)[24] coefficients

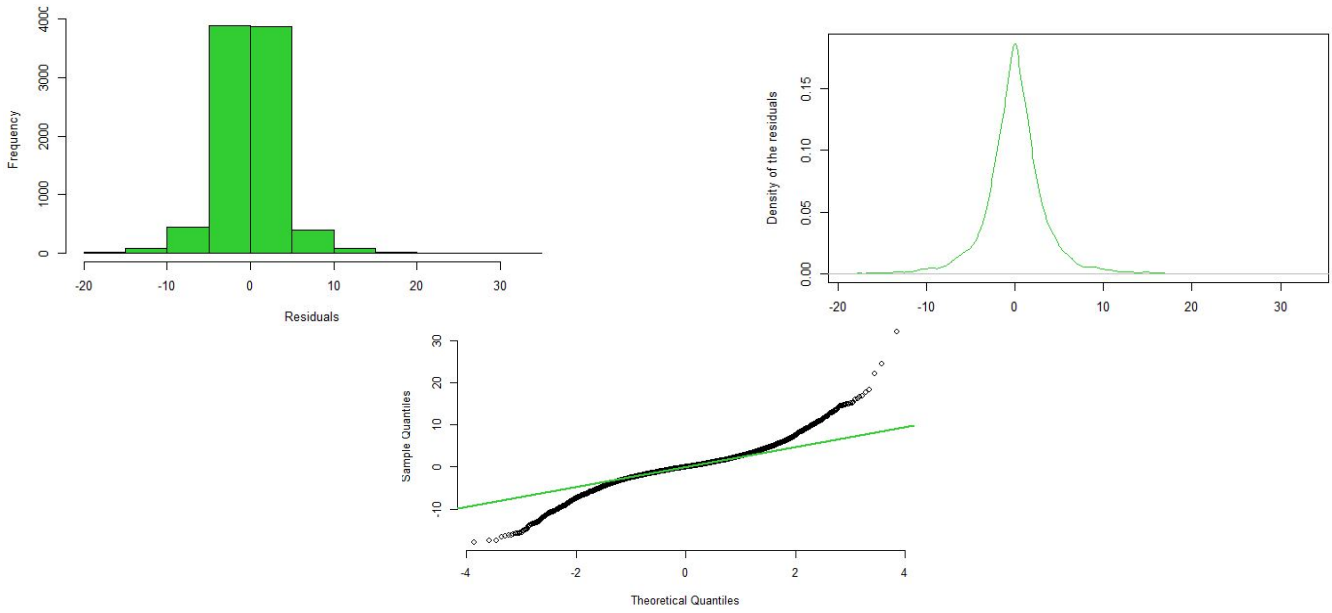

Figure 47: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Antas-Espinho

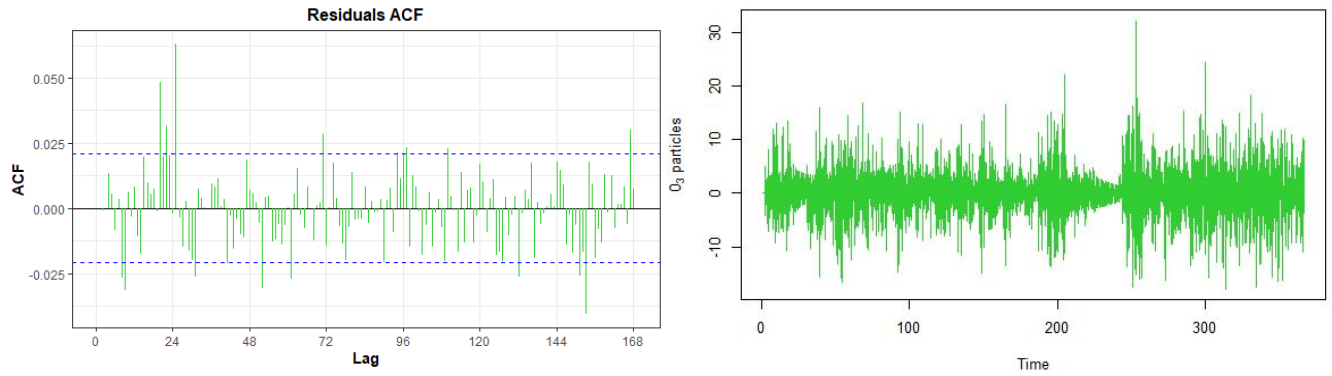

Figure 48: ACF (left) and plot (right) of the residuals for Antas-Espinho

7.1.5 Antas-Espinho

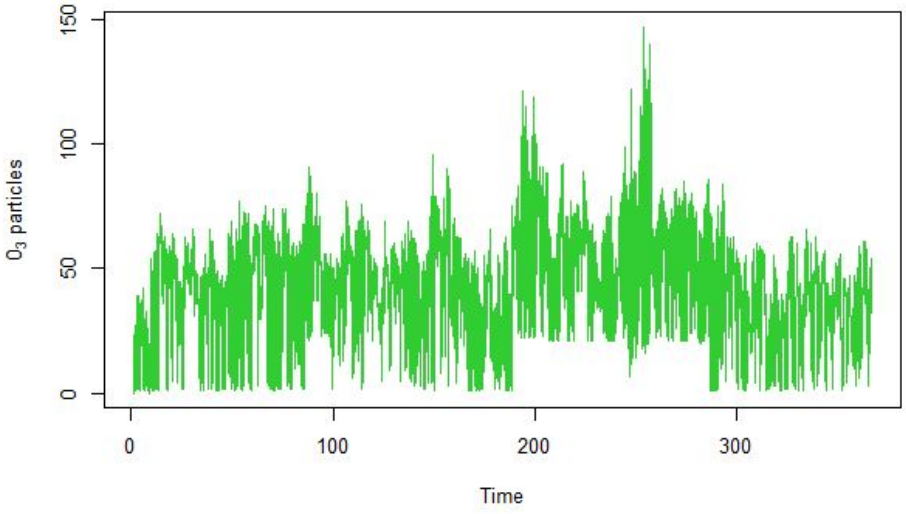

Figure 49: Hourly levels of O_3 particles in $\mu g/m3$ in Antas-Espinho since 01/01/2020 until 31/12/2020

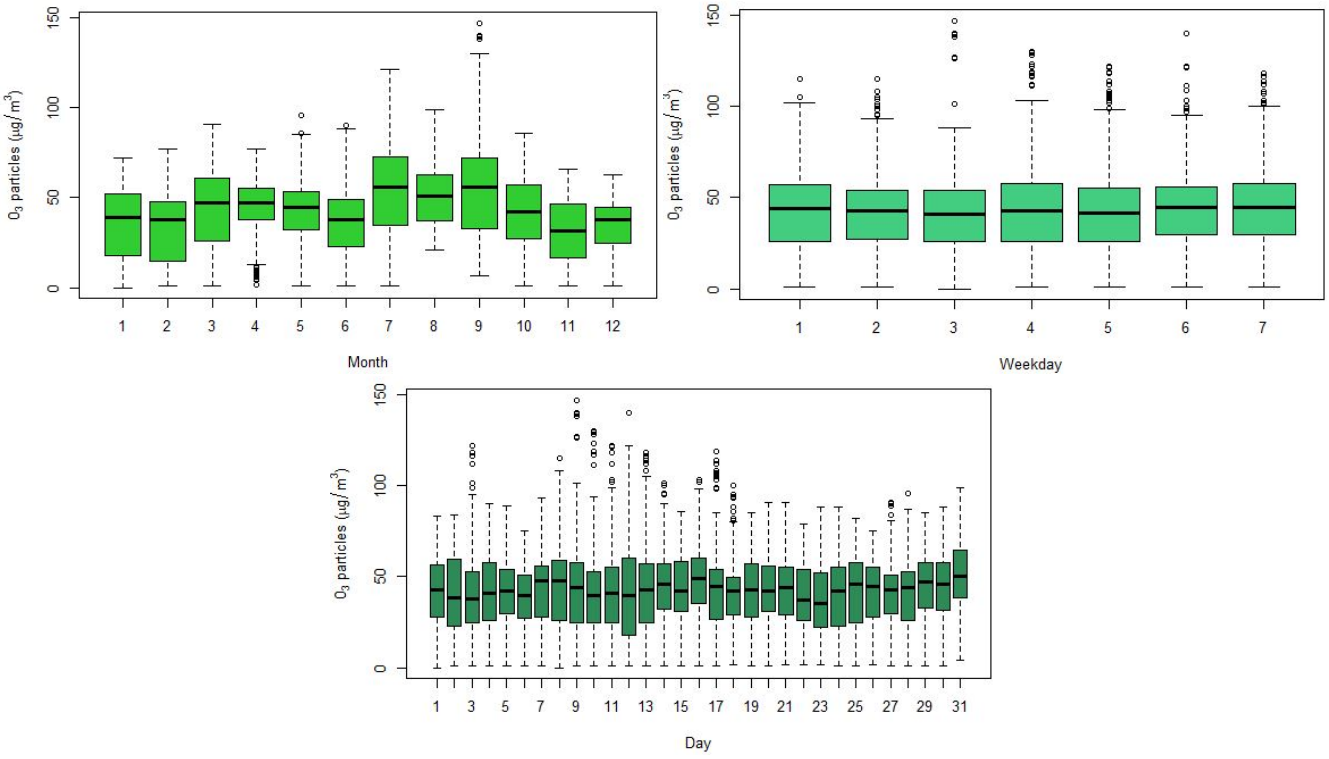

Figure 50: Boxplots of the O_3 particles levels in Antas-Espinho with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

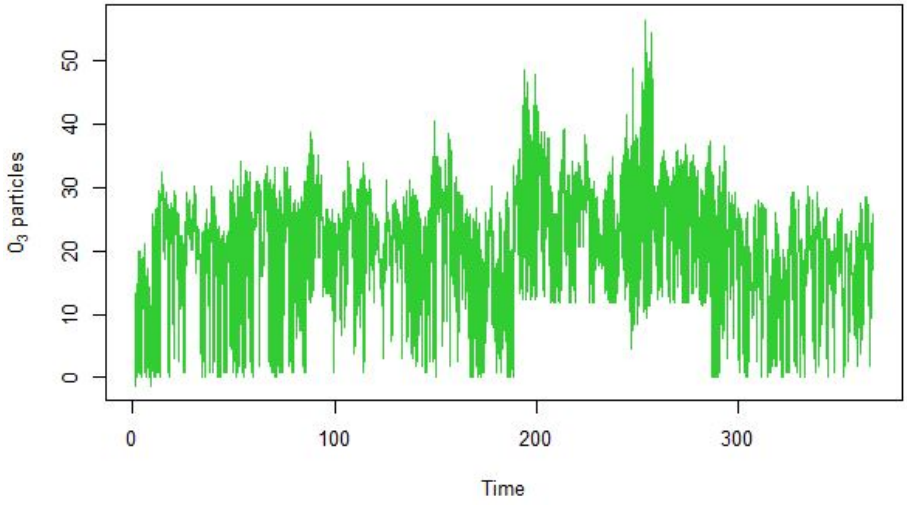

Figure 51: Hourly levels of O3 particles in $\mu g/m3$ in Antas-Espinho since 01/01/2020 until 31/12/2020, after the log transformation of the data

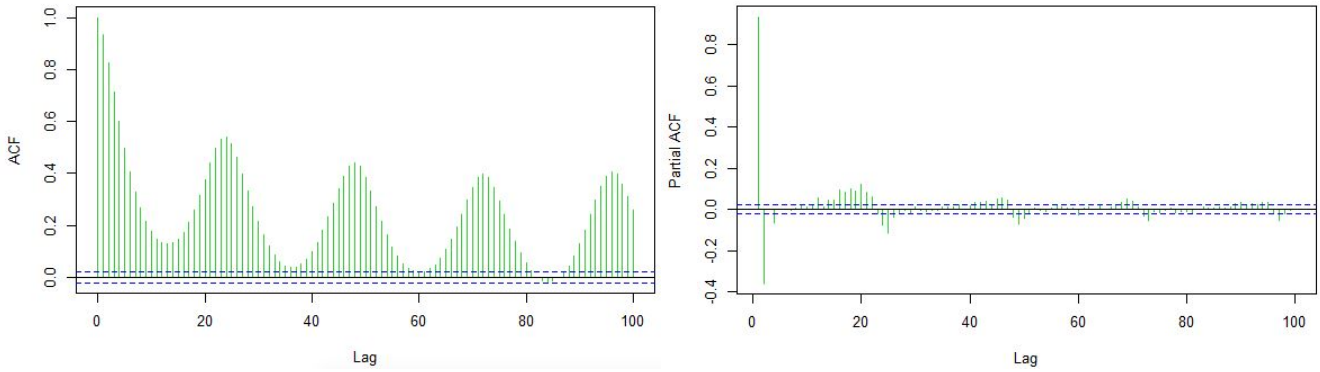

Figure 52: Characteristics of O_3 time series, with Box-Cox transformation, at Antas-Espinho station. ACF on the left and PACF on the right

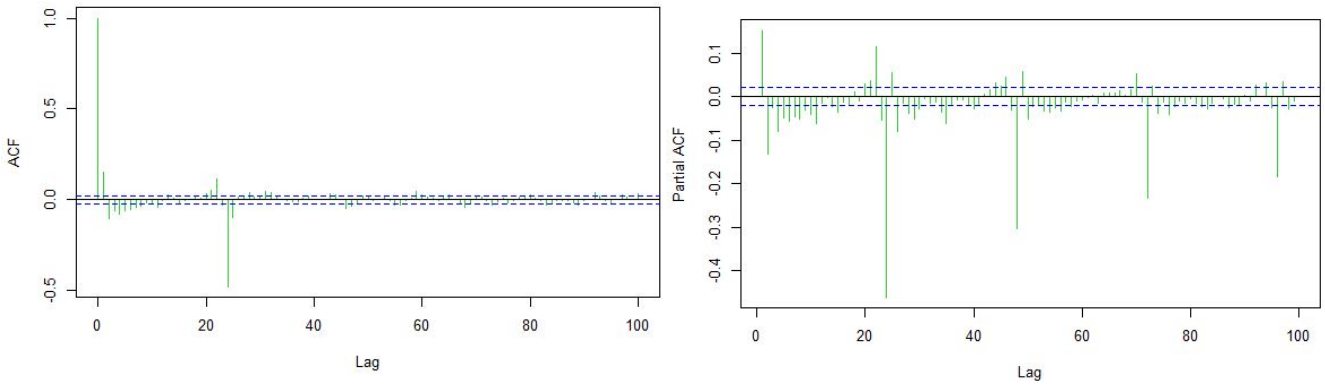

Figure 53: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Antas-Espinho station. ACF on the left and PACF on the right

Antas-Espinho			
Model	**AIC**	**AICc**	**BIC**
ARIMA(2,1,0)(2,0,0)[24] (auto.arima)	44446.35	44446.35	44481.75
ARIMA(1,1,2)(2,1,2)[24] (1st best AIC)	46443.06	46443.07	46499.67
ARIMA(1,1,3)(2,1,2)[24] (2nd best AIC)	46444.79	46444.81	46508.49
ARIMA(1,1,3)(1,1,1)[24] (1st best BIC and 3rd best AIC)	46445.25	46445.27	46494.8

Table 21: Comparison of the best fitted models for Antas-Espinho time series

Antas-Espinho		
ARIMA(1,1,3)(1,1,1)[24]	**Coefficient**	**p-value**
ar1	0.8768	0
ma1	-0.7559	8.070724e-05
ma2	-0.2260	0
ma3	-0.0063	6.066401e-01
sar1	0.0755	5.134537e-11
sma1	-0.9593	0

Table 22: Summary of ARIMA(1,1,3)(1,1,1)[24] coefficients

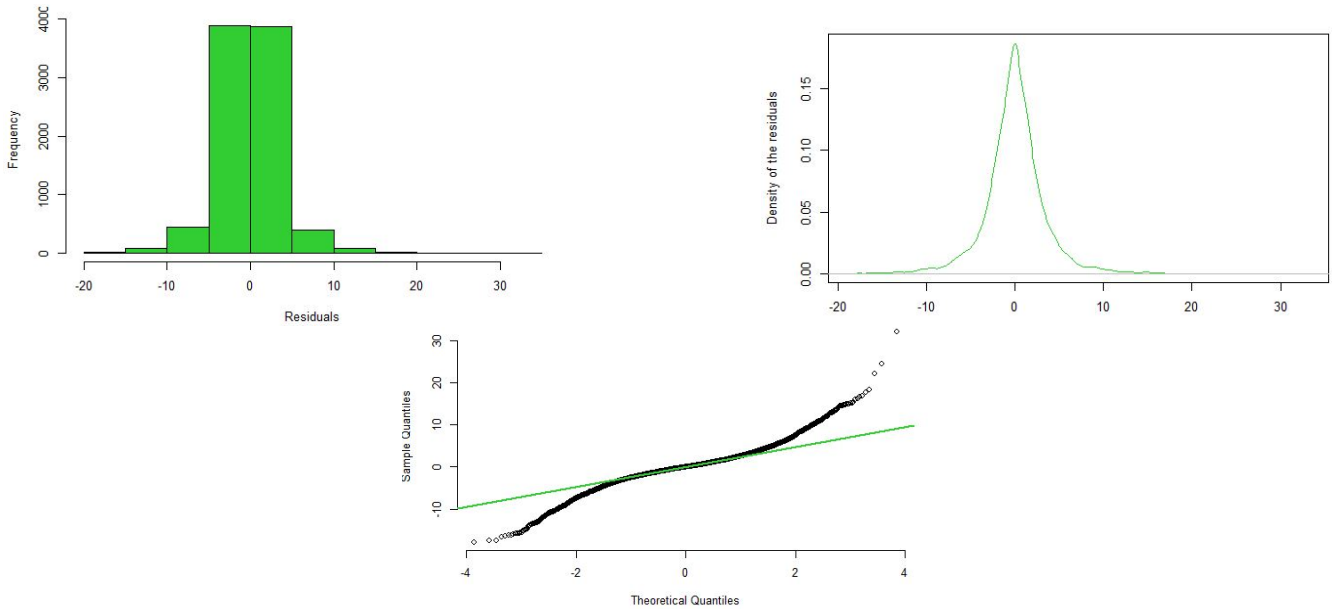

Figure 54: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Antas-Espinho

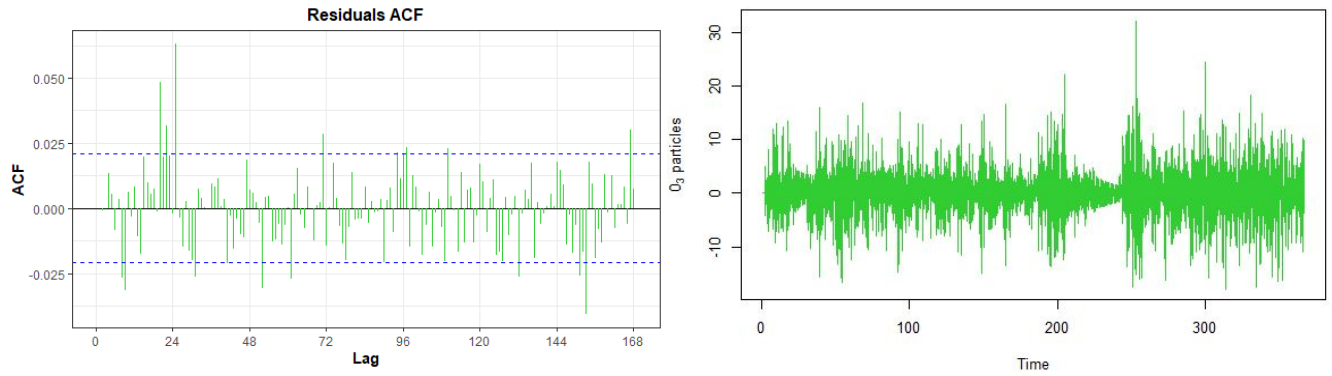

Figure 55: ACF (left) and plot (right) of the residuals for Antas-Espinho

7.1.6 Laranjeiro-Almada

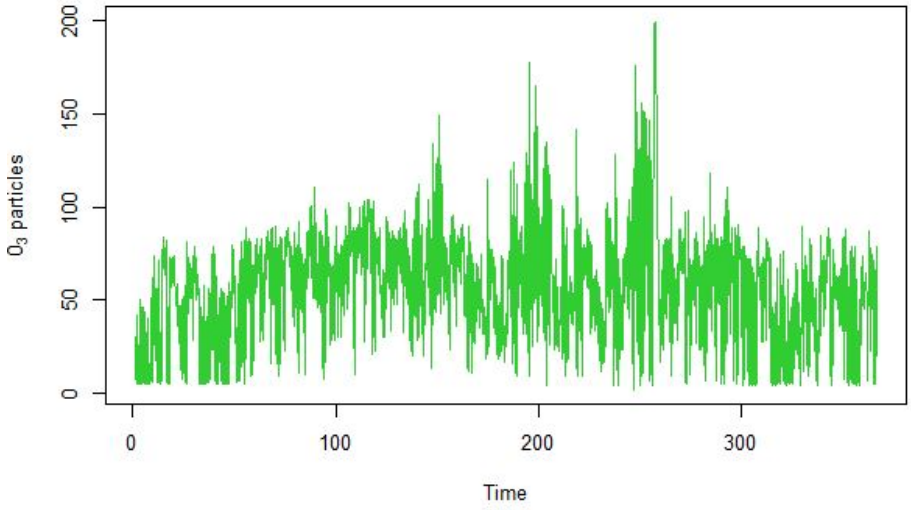

Figure 56: Hourly levels of O_3 particles in $\mu g/m3$ in Laranjeiro-Almada since 01/01/2020 until 31/12/2020

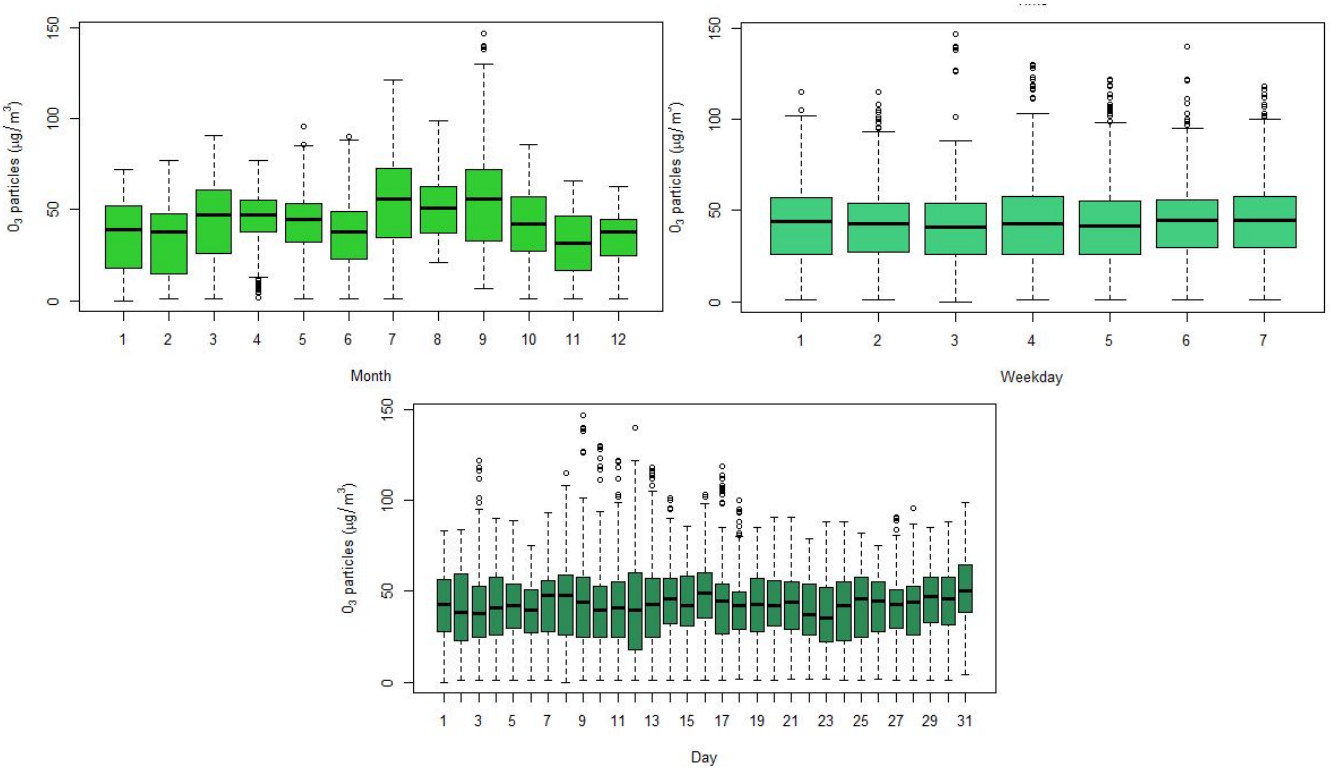

Figure 57: Boxplots of the O_3 particles levels in Laranjeiro-Almada with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

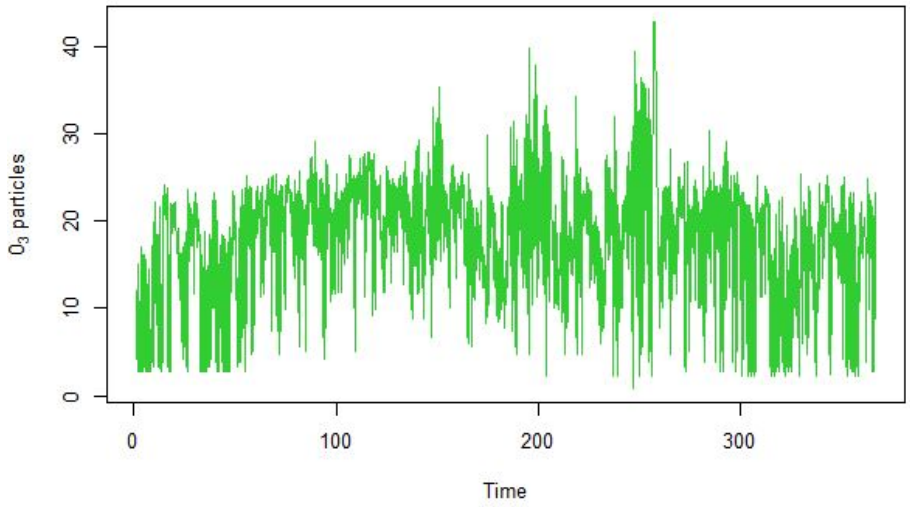

Figure 58: Hourly levels of O3 particles in $\mu g/m3$ in Laranjeiro-Almada since 01/01/2020 until 31/12/2020, after the log transformation of the data

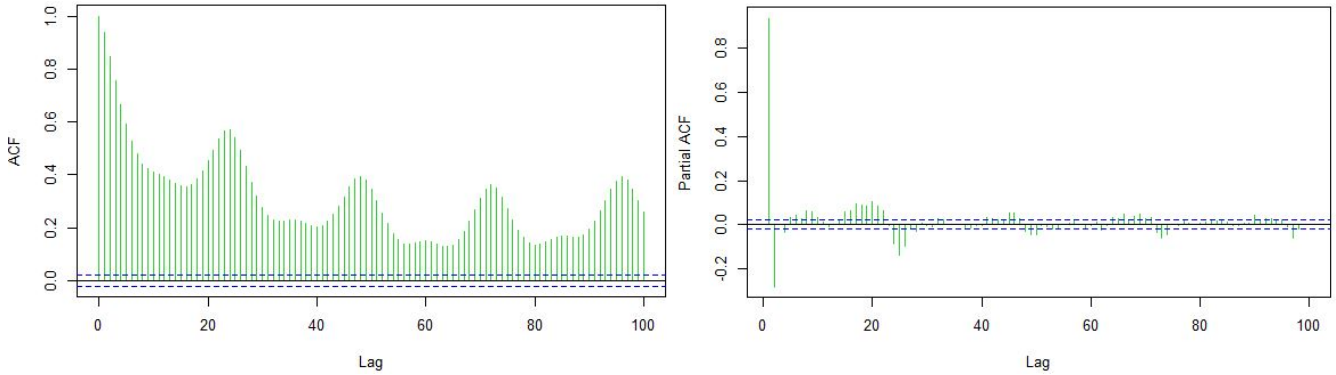

Figure 59: Characteristics of O_3 time series, with Box-Cox transformation, at Laranjeiro-Almada station. ACF on the left and PACF on the right

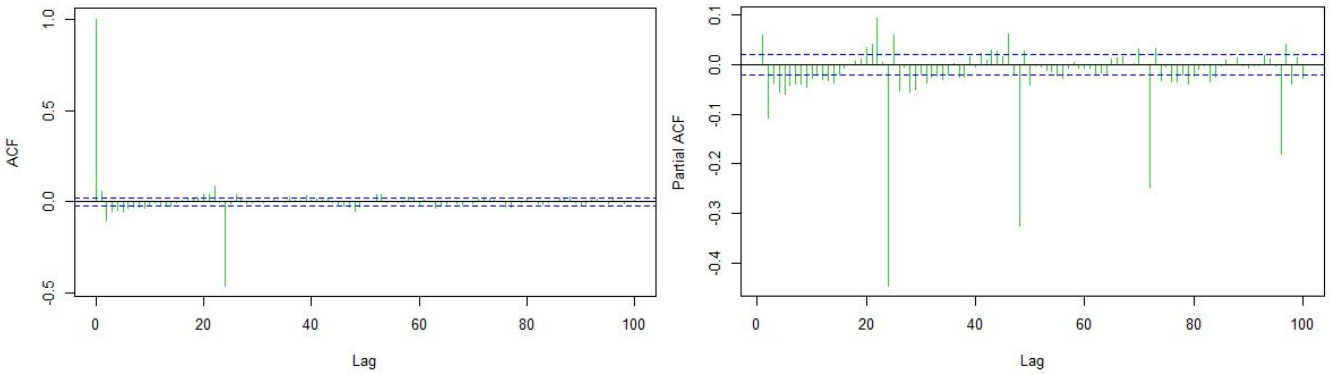

Figure 60: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Laranjeiro-Almada station. ACF on the left and PACF on the right

Laranjeiro-Almada			
Model	**AIC**	**AICc**	**BIC**
ARIMA(2,1,0)(1,0,0)[24] (auto.arima)	38142.33	38142.34	38177.73
ARIMA(1,1,3)(2,1,1)[24] (1st best AIC)	36678.01	36678.02	36734.63
ARIMA(1,1,2)(0,1,2)[24] (1st best BIC)	36684.34	36684.34	36726.8
ARIMA(2,1,2)(0,1,2)[24] (3rd best BIC and 8th best AIC)	36680.16	36680.18	36729.71

Table 23: Comparison of the best fitted models for Laranjeiro-Almada time series.

Laranjeiro-Almada		
ARIMA(2,1,2)(0,1,2)[24]	**Coefficient**	**p-value**
ar1	0.7264	0
ar2	0.1457	9.208505e-03
ma1	-0.6573	0
ma2	-0.3319	7.127441e-09
sma1	-0.9021	0
sma2	-0.0550	7.789101e-07

Table 24: Summary of ARIMA(2,1,2)(0,1,2)[24] coefficients

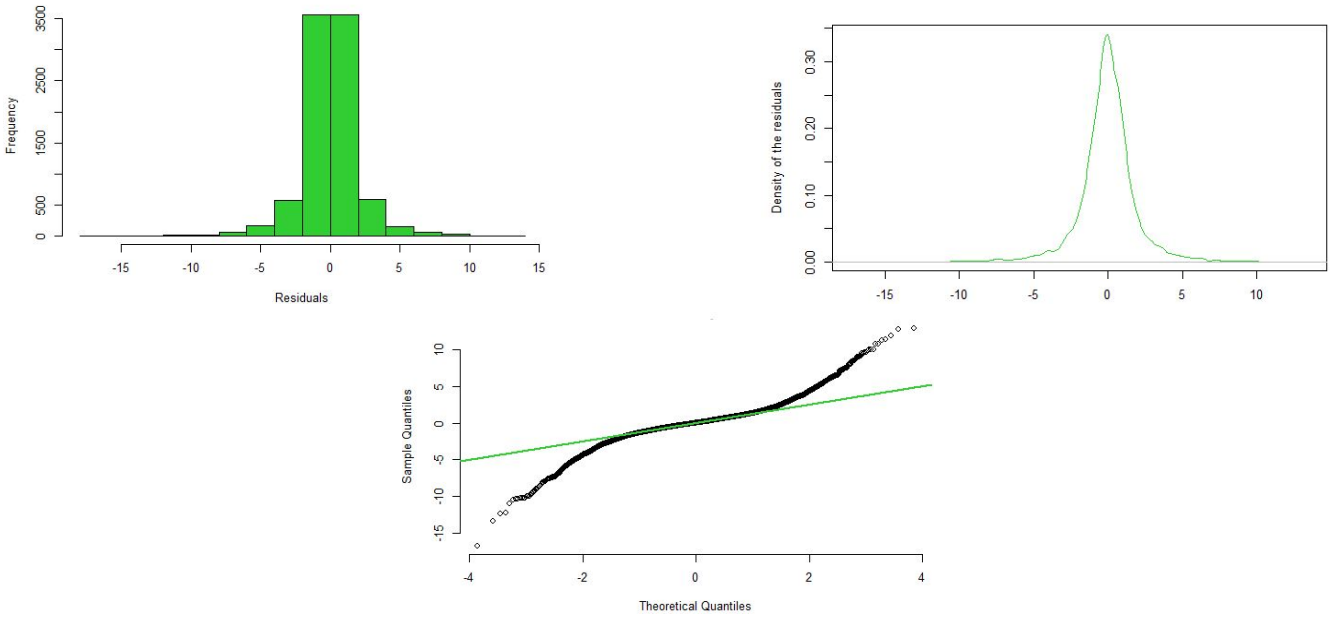

Figure 61: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Laranjeiro-Almada

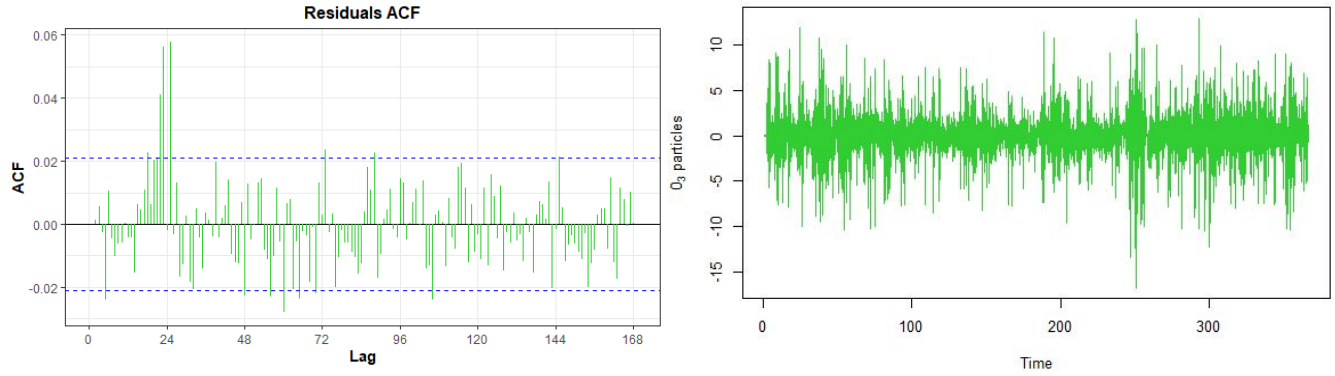

Figure 62: ACF (left) and plot (right) of the residuals for Laranjeiro-Almada

7.1.7 Estarreja

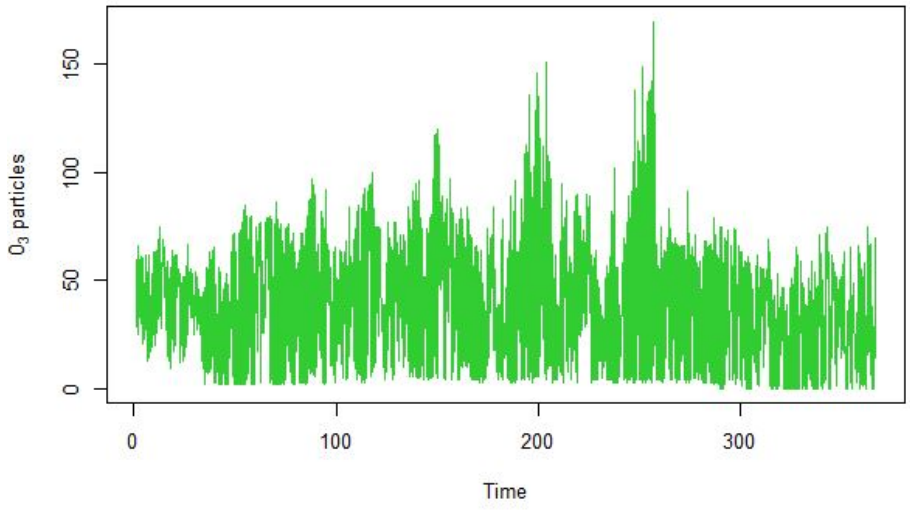

Figure 63: Hourly levels of O_3 particles in $\mu g/m3$ in Estarreja since 01/01/2020 until 31/12/2020

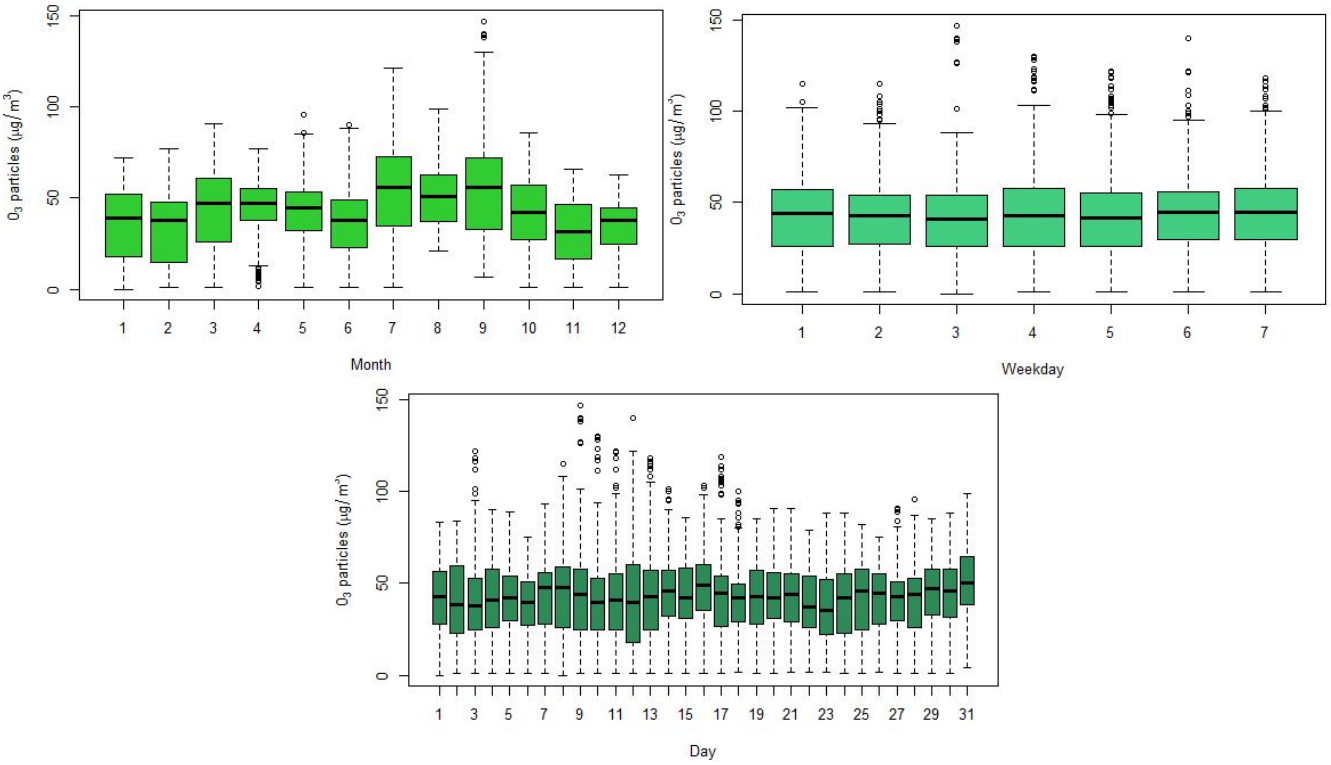

Figure 64: Boxplots of the O_3 particles levels in Estarreja with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

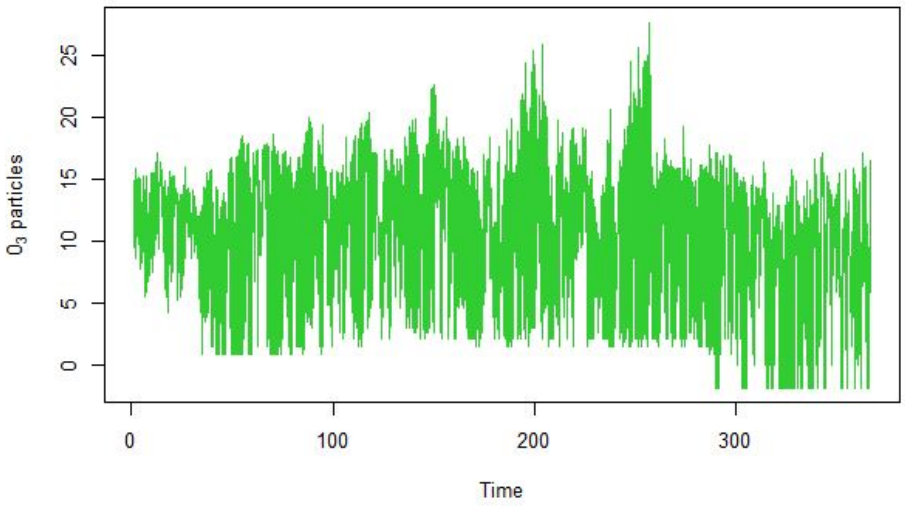

Figure 65: Hourly levels of O3 particles in $\mu g/m3$ in Estarreja since 01/01/2020 until 31/12/2020, after the log transformation of the data

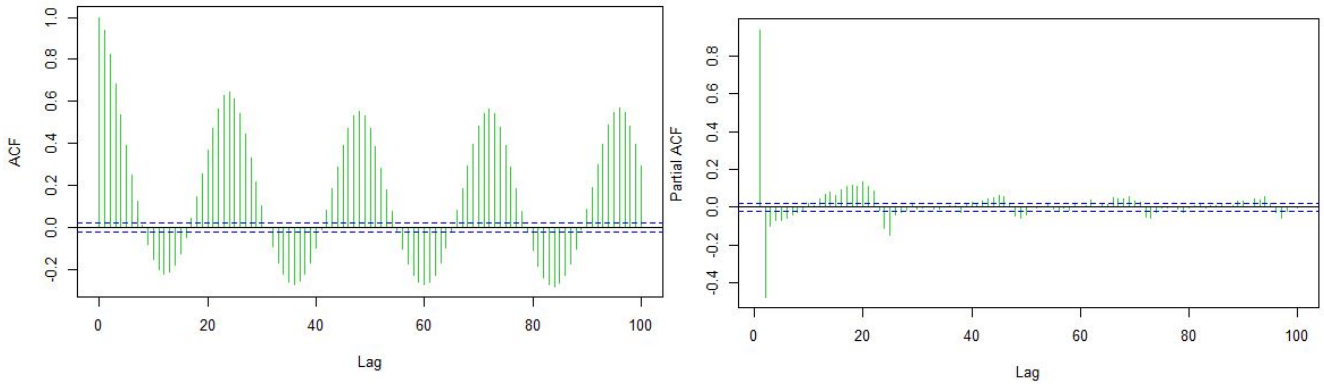

Figure 66: Characteristics of O_3 time series, with Box-Cox transformation, at Estarreja station. ACF on the left and PACF on the right

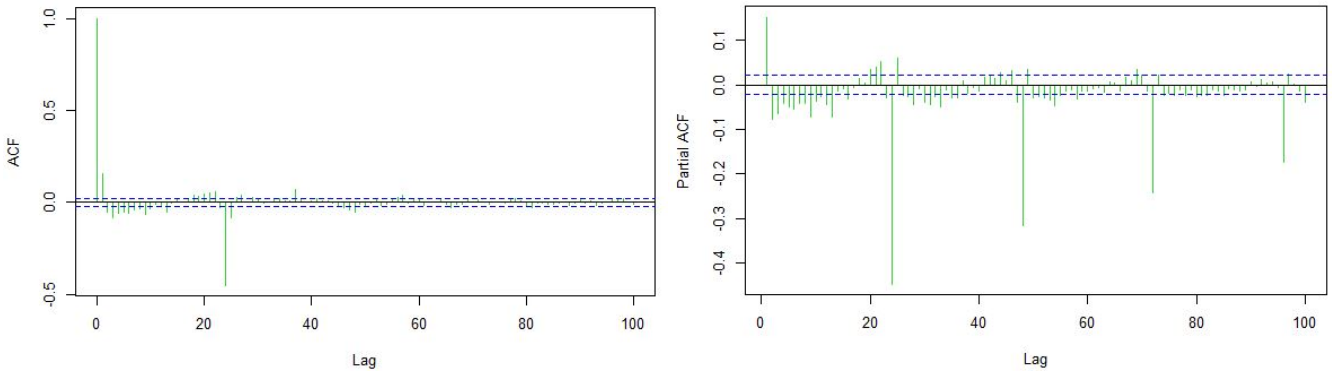

Figure 67: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Estarreja station. ACF on the left and PACF on the right

Estarreja			
Model	**AIC**	**AICc**	**BIC**
ARIMA(2,0,1)(2,1,0)[24] (auto.arima)	33764.33	33764.34	33806.8
ARIMA(2,1,3)(0,1,2)[24] (1st best AIC)	31934.7	31934.72	31991.32
ARIMA(2,1,1)(0,1,2)[24] (1st best BIC and 3rd best AIC)	31935.9	31935.91	31978.36

Table 25: Comparison of the best fitted models for Estarreja time series.

Estarreja		
ARIMA(2,1,1)(0,1,2)[24]	**Coefficient**	**p-value**
ar1	1.1492	0
ar2	-0.2574	0
ma1	-0.9978	0
sma1	-0.8653	0
sma2	-0.0717	4.168221e-11

Table 26: Summary of ARIMA(2,1,1)(0,1,2)[24] coefficients

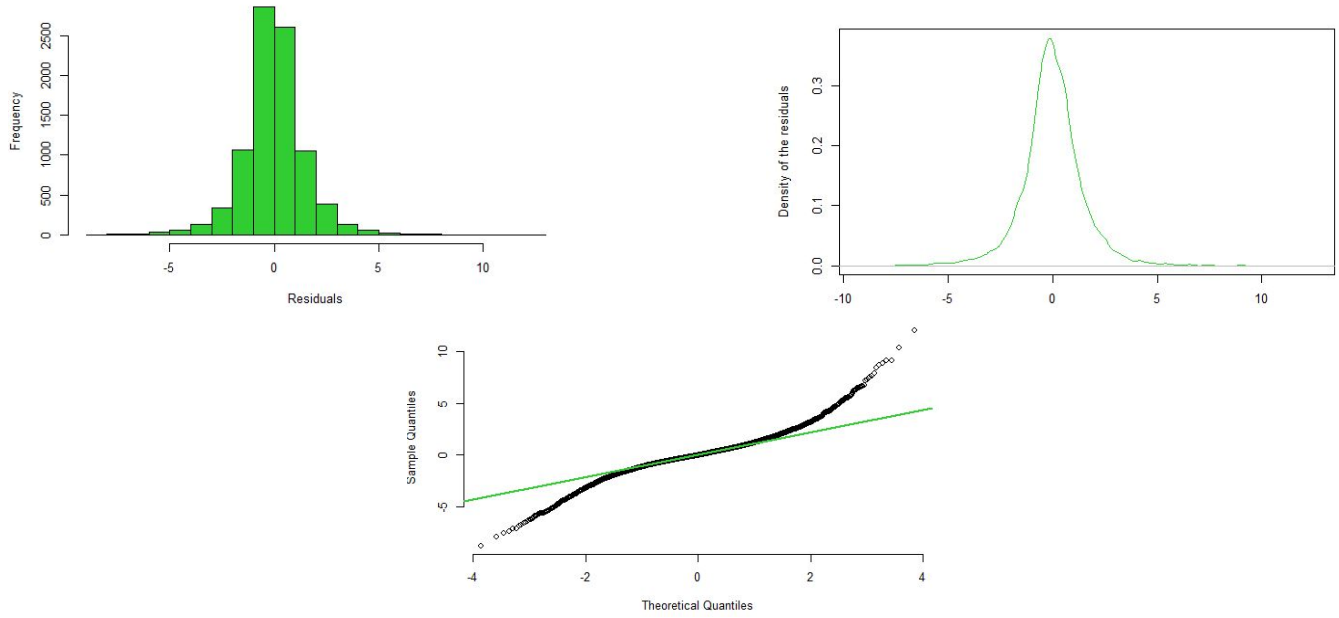

Figure 68: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Estarreja

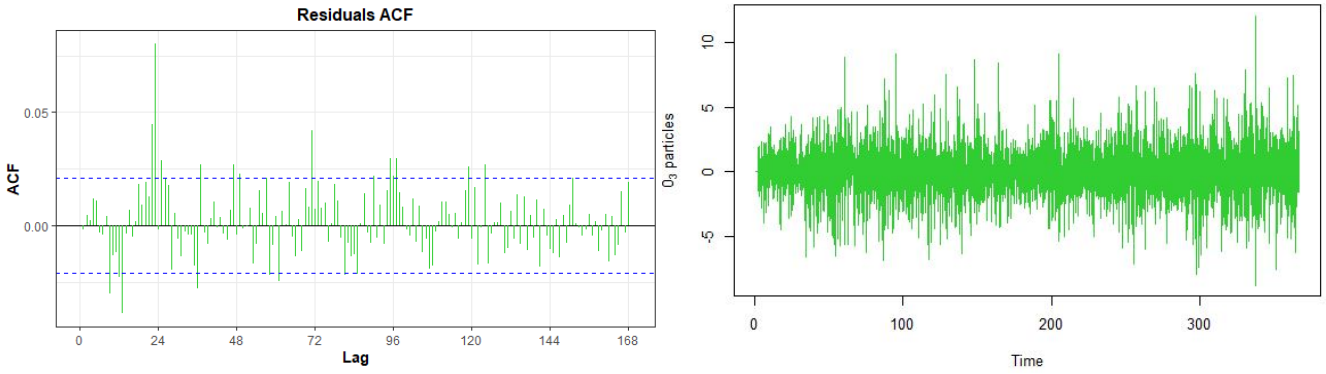

Figure 69: ACF (left) and plot (right) of the residuals for Estarreja

7.1.8 Entrecampos

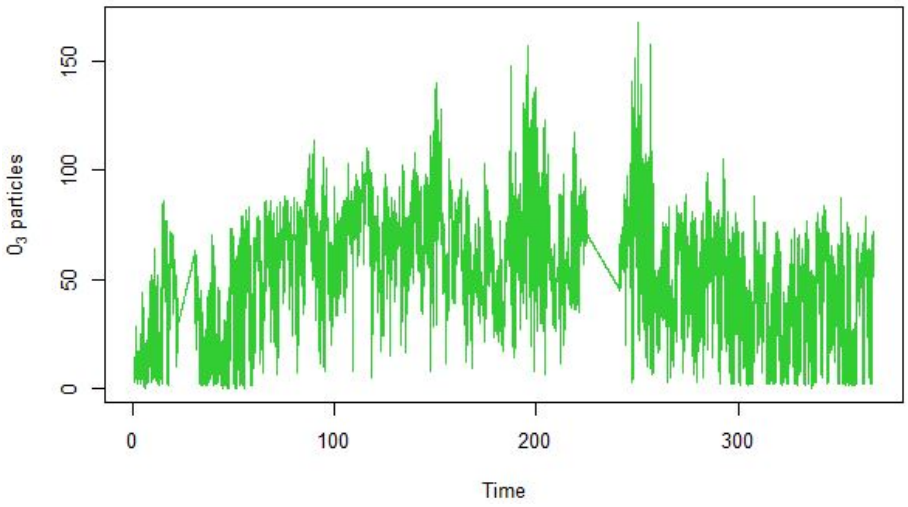

Figure 70: Hourly levels of O_3 particles in $\mu g/m3$ in Entrecampos since 01/01/2020 until 31/12/2020

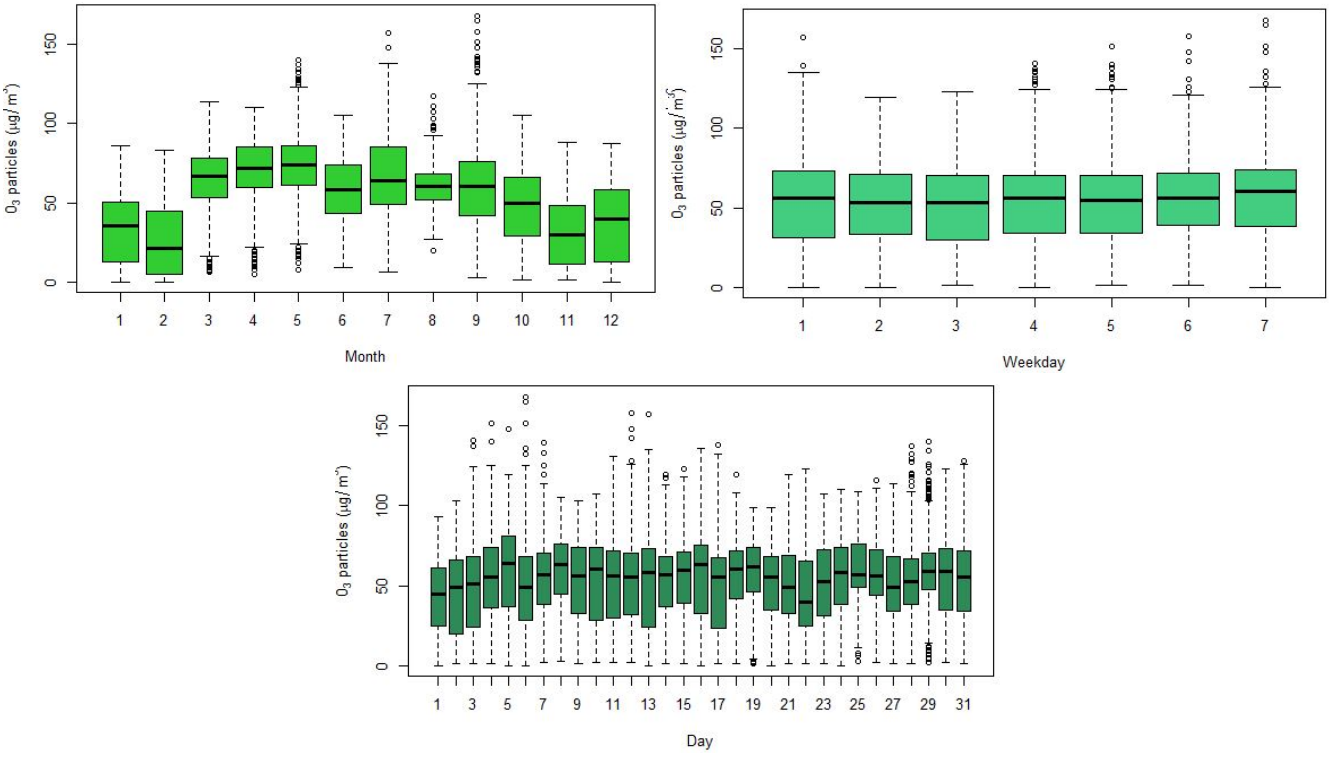

Figure 71: Boxplots of the O_3 particles levels in Entrecampos with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

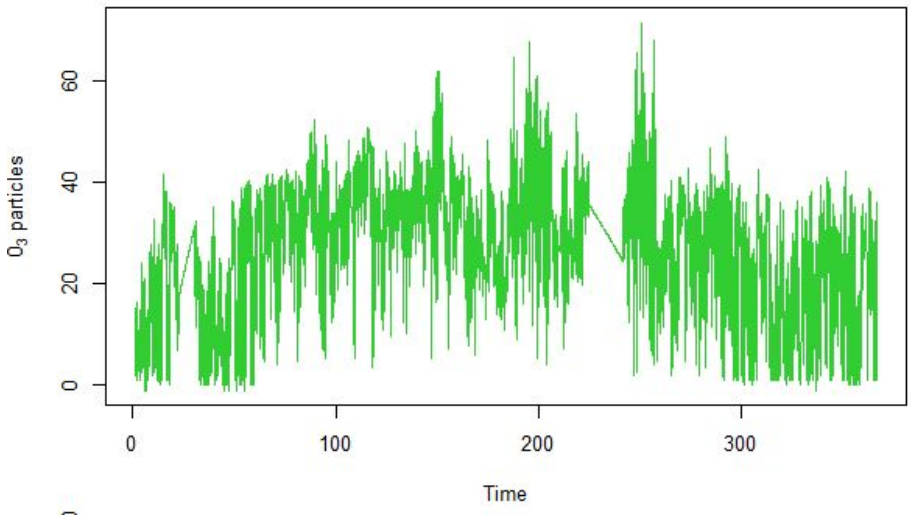

Figure 72: Hourly levels of O3 particles in $\mu g/m3$ in Entrecampos since 01/01/2020 until 31/12/2020, after the log transformation of the data

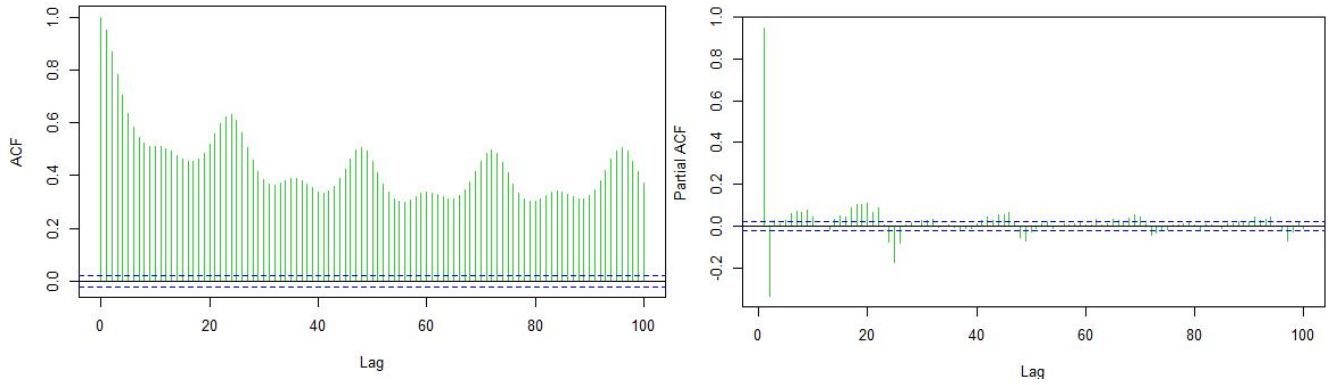

Figure 73: Characteristics of O_3 time series, with Box-Cox transformation, at Entrecampos station. ACF on the left and PACF on the right

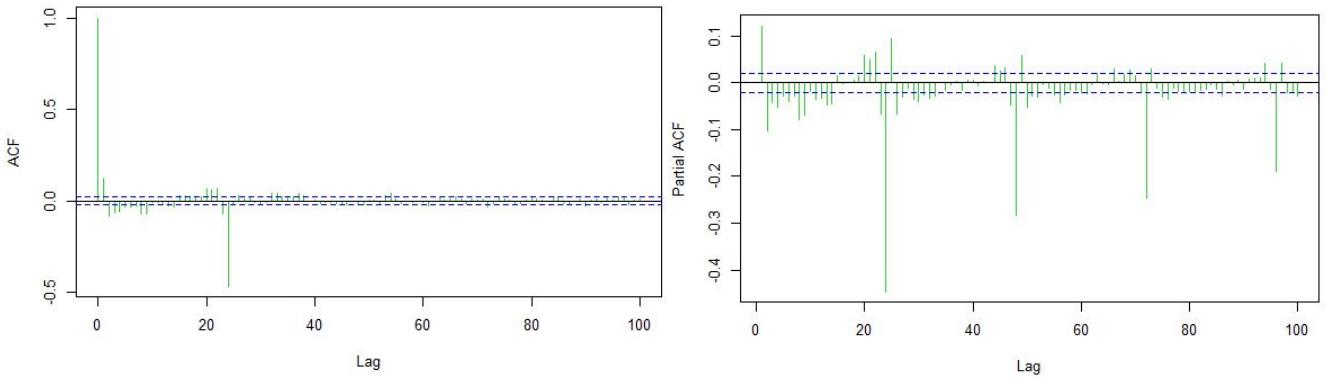

Figure 74: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Entrecampos station. ACF on the left and PACF on the right

Estarreja			
Model	**AIC**	**AICc**	**BIC**
ARIMA(5,1,0)(2,0,0)[24] (auto.arima)	47725.43	47725.45	47782.07
ARIMA(1,1,2)(2,1,2)[24] (1st best AIC)	46443.06	46443.07	46499.67
ARIMA(1,1,3)(1,1,1)[24] (1st best BIC and 3rd best AIC)	46445.25	46445.27	46494.8

Table 27: Comparison of the best fitted models for Entrecampos time series.

Estarreja		
ARIMA(1,1,3)(1,1,1)[24]	**Coefficient**	**p-value**
ar1	0.8768	0
ma1	-0.7559	0
ma2	-0.2260	0
ma3	-0.0063	6.066401e-01
sar1	0.0755	5.134537e-11
sma1	-0.9593	0

Table 28: Summary of ARIMA(1,1,3)(1,1,1)[24] coefficients

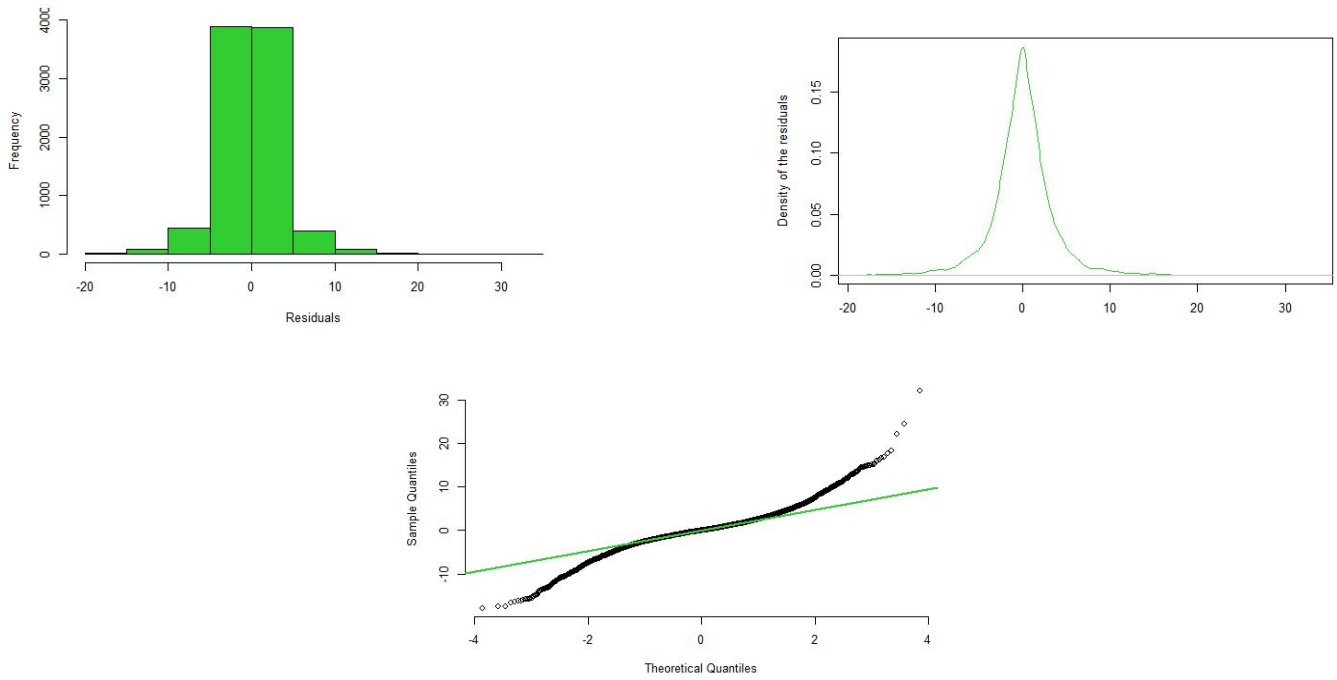

Figure 75: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Entrecampos

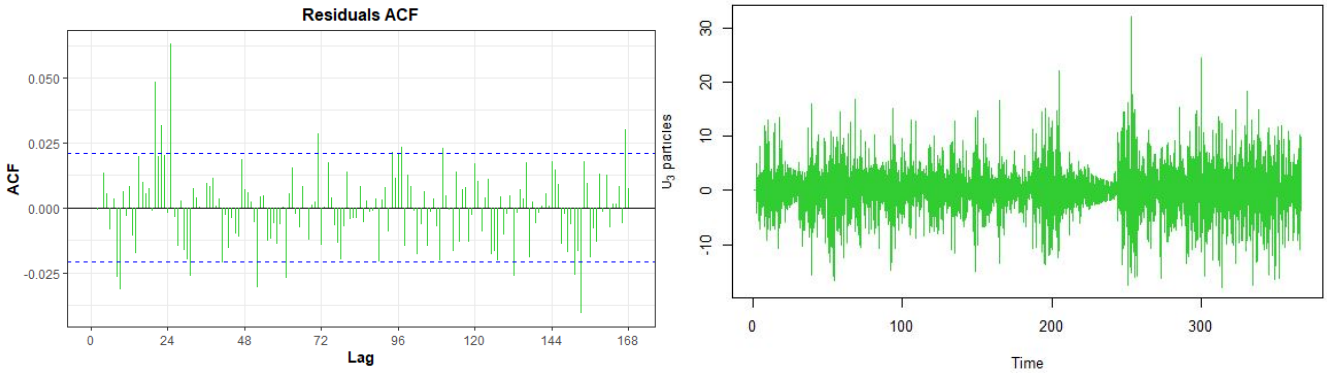

Figure 76: ACF (left) and plot (right) of the residuals for Entrecampos

7.1.9 Mem-Martins

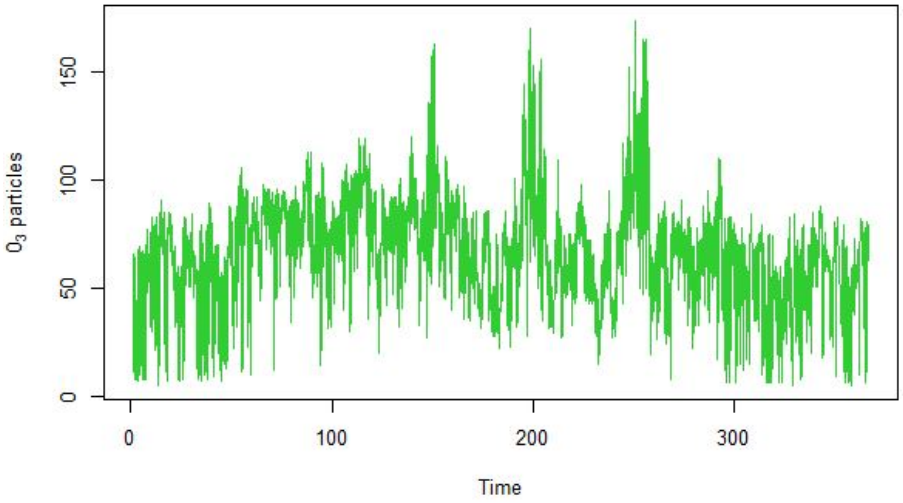

Figure 77: Hourly levels of O_3 particles in $\mu g/m3$ in Mem-Martins since 01/01/2020 until 31/12/2020

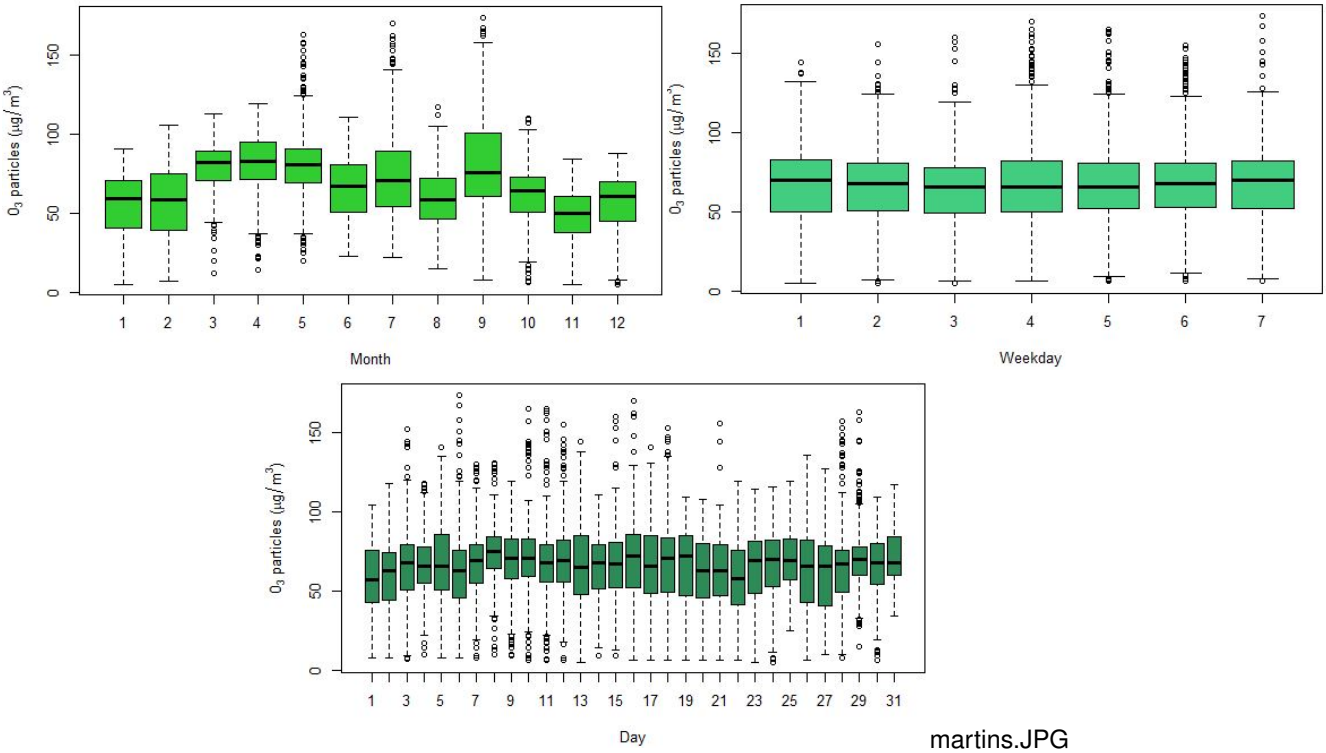

Figure 78: Boxplots of the O_3 particles levels in Mem-Martins with respect to different time periods splitting: daily, monthly, and week daily (where 1 corresponds to Monday and 7 to Sunday)

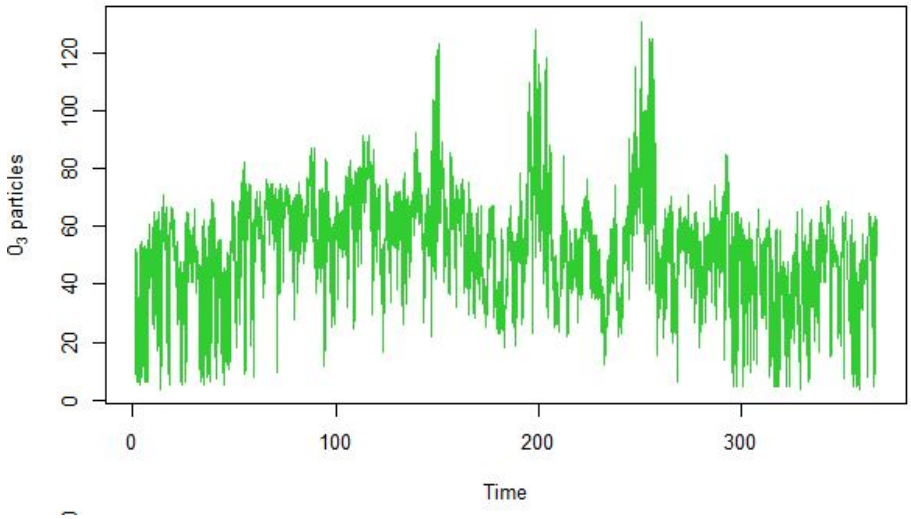

Figure 79: Hourly levels of O3 particles in $\mu g/m3$ in Mem-Martins since 01/01/2020 until 31/12/2020, after the log transformation of the data

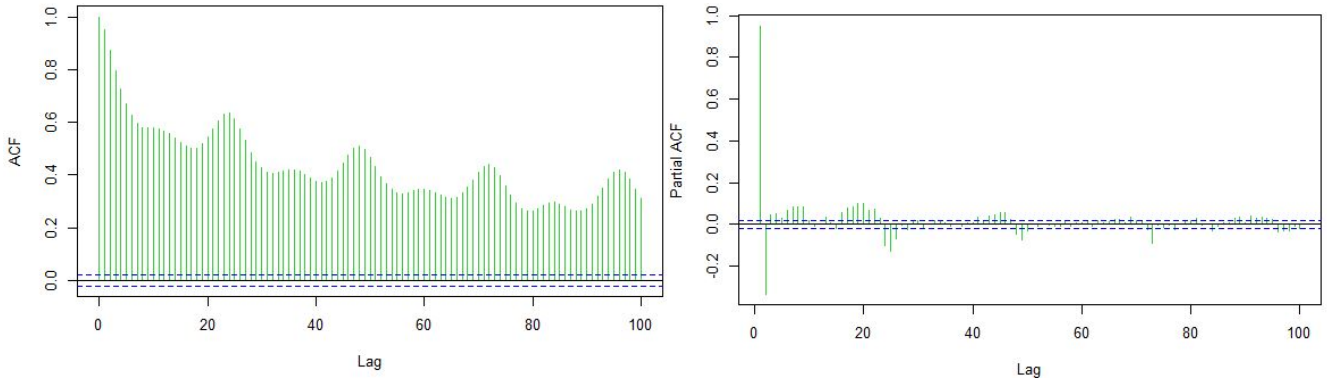

Figure 80: Characteristics of O_3 time series, with Box-Cox transformation, at Mem-Martins station. ACF on the left and PACF on the right

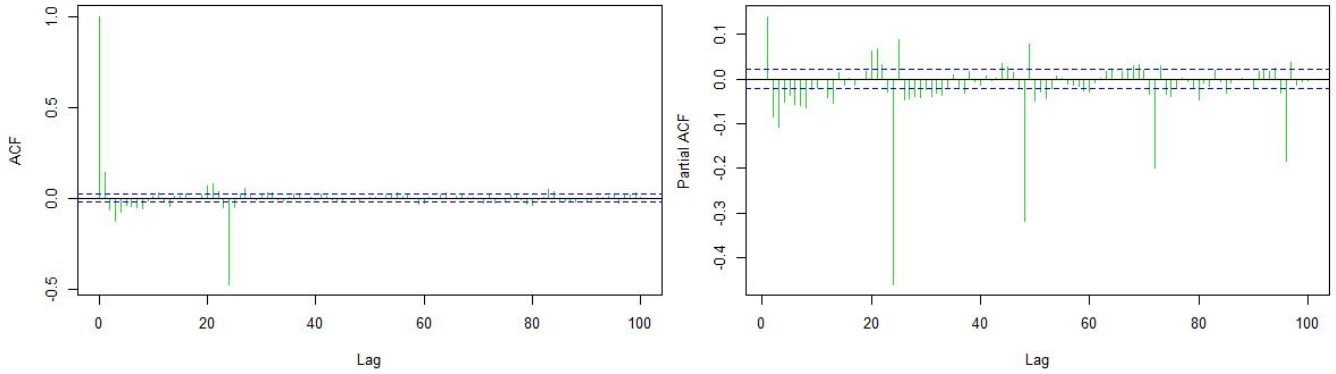

Figure 81: Characteristics of O_3 levels with Box-Cox transformation and differenced time series, at Mem-Martins station. ACF on the left and PACF on the right

Mem-Martins			
Model	**AIC**	**AICc**	**BIC**
ARIMA(4,1,3)(2,0,0)[24] (auto.arima)	53404.25	53404.28	53482.13
ARIMA(1,1,3)(2,1,2)[24] (1st best AIC)	52548.86	52548.88	52612.55
ARIMA(1,1,3)(0,1,1)[24] (1st best BIC)	52559.61	52559.62	52602.07
ARIMA(1,1,3)(0,1,2)[24] (3rd best AIC an 2nd best BIC)	52553.53	52553.54	52603.07

Table 29: Comparison of the best fitted models for Mem-Martins time series.

Mem-Martins		
ARIMA(1,1,3)(0,1,2)[24]	**Coefficient**	**p-value**
ar1	0.8114	0
ma1	-0.6633	0
ma2	-0.2142	0
ma3	-0.0669	4.174254e-07
sma1	-0.9219	0
sma2	-0.0314	4.502311e-03

Table 30: Summary of ARIMA(1,1,3)(0,1,2)[24] coefficients

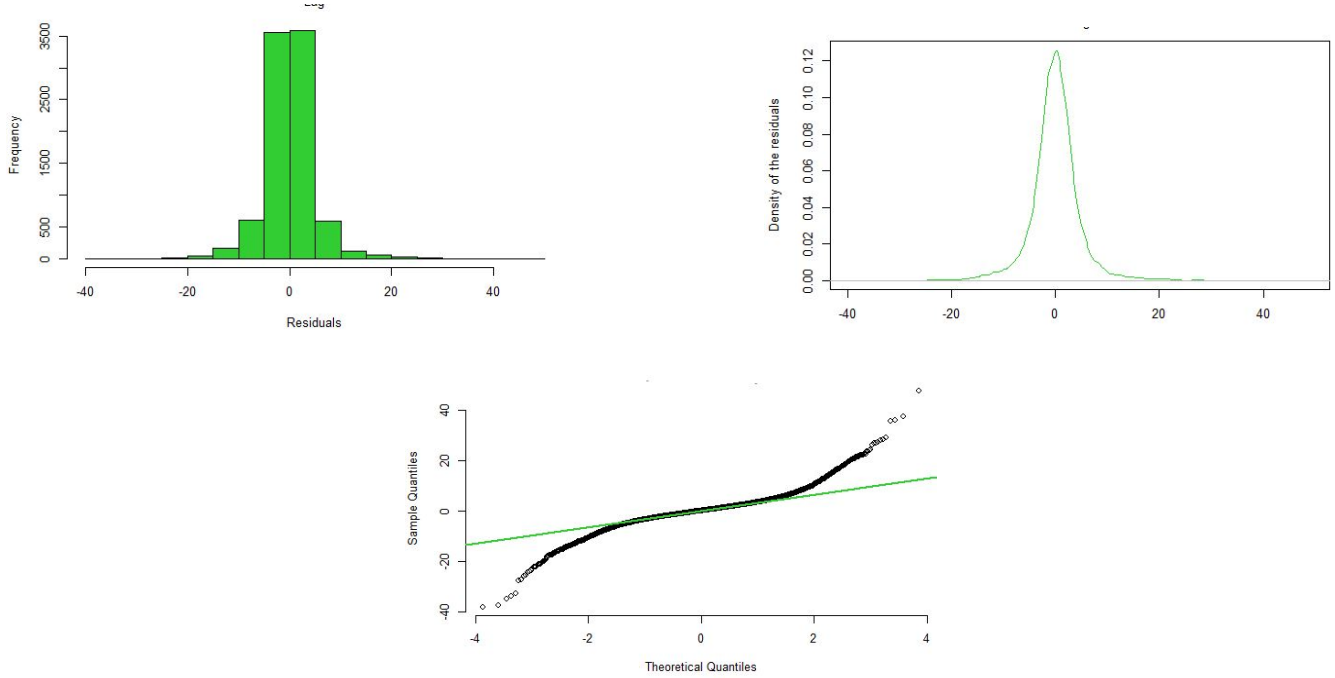

Figure 82: Histogram (left), Density function (right) and QQ-Plot (down) of the residuals for Mem-Martins

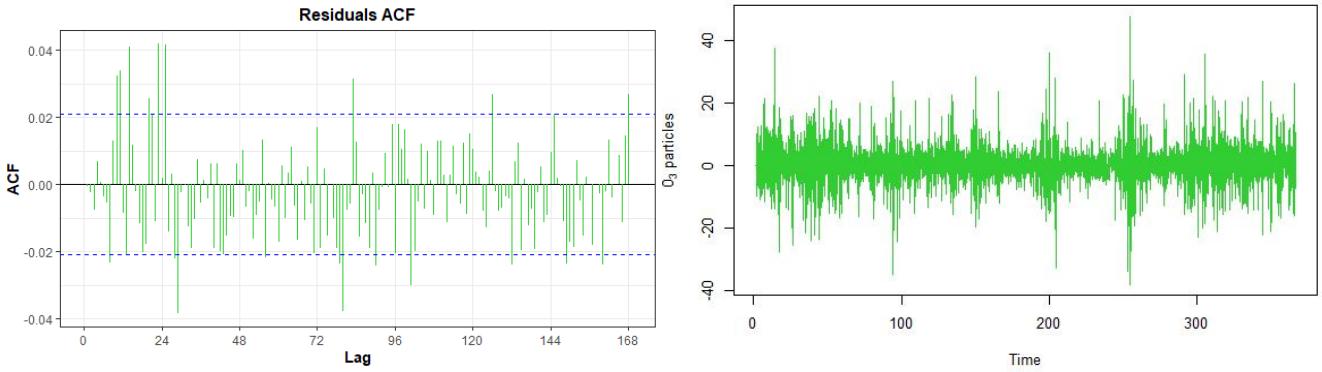

Figure 83: ACF (left) and plot (right) of the residuals for Mem-Martins

7.1.10 Forecast

	Date	Lower Bound	Predicted Value	Upper Bound	True Value	MAPE
Restelo	31/12/2020, 19h	15.90939	20.03125	24.15312	23.01197	
	31/12/2020, 20h	13.16806	19.28365	25.39923	24.53777	
	31/12/2020, 21h	11.89174	19.19977	26.50781	25.53132	18%
ARIMA(3,1,1)(0,1,2)[24]	31/12/2020, 22h	11.80495	19.87337	27.94178	24.28652	
	31/12/2020, 23h	11.64853	20.24261	28.83669	23.52555	
Sobreiro	31/12/2020, 19h	19.59471	24.79471	29.99471	24.86733	
	31/12/2020, 20h	15.89721	23.63225	31.36728	24.15844	
	31/12/2020, 21h	13.74738	23.14119	32.53501	25.91825	4%
ARIMA(2,1,1)(0,1,1)[24]	31/12/2020, 22h	12.02262	22.57341	33.12420	24.51374	
	31/12/2020, 23h	10.69839	22.09429	33.49019	21.98841	
VN Telha-Maia	31/12/2020, 19h	16.78889	20.21075	23.63261	20.22349	
	31/12/2020, 20h	14.12416	19.43972	24.75528	19.75220	
	31/12/2020, 21h	12.02031	18.54205	25.06380	18.30238	6%
ARIMA(2,1,2)(0,1,2)[24]	31/12/2020, 22h	11.18498	18.52682	25.86866	15.74584	
	31/12/2020, 23h	10.76295	18.68958	26.61621	16.79187	
Antas-Espinho	31/12/2020, 19h	16.866398	23.57126	30.27612	24.24466	
	31/12/2020, 20h	12.960857	23.03240	33.10394	22.28351	
	31/12/2020, 21h	11.597948	23.70054	35.80314	23.07400	4%
ARIMA(1,1,3)(1,1,1)[24]	31/12/2020, 22h	9.439788	22.93189	36.42399	24.63106	
	31/12/2020, 23h	7.650367	22.14913	36.64788	23.46618	
Laranjeiro-Almada	31/12/2020, 19h	14.632915	18.47113	22.30934	20.13431	
	31/12/2020, 20h	12.105779	17.72444	23.34311	22.00751	
	31/12/2020, 21h	11.364731	18.02766	24.69059	23.20875	19%
ARIMA(2,1,2)(0,1,2)[24]	31/12/2020, 22h	10.527395	17.93657	25.34575	22.61253	
	31/12/2020, 23h	9.612822	17.57556	25.53829	22.61253	
Estarreja	31/12/2020, 19h	9.097583	12.02597	14.95435	12.559163	
	31/12/2020, 20h	6.820734	11.28670	15.75266	8.873118	
	31/12/2020, 21h	5.040174	10.49241	15.94464	12.210647	14%
ARIMA(2,1,1)(0,1,2)[24]	31/12/2020, 22h	4.272857	10.37163	16.47040	9.093341	
	31/12/2020, 23h	3.481471	10.01443	16.54738	11.490362	
Entrecampos	31/12/2020, 19h	16.866398	23.57126	30.27612	30.99155	
	31/12/2020, 20h	12.960857	23.03240	33.10394	30.56546	
	31/12/2020, 21h	11.597948	23.70054	35.80314	35.99692	31%
ARIMA(2,1,1)(0,1,2)[24]	31/12/2020, 22h	9.439788	22.93189	36.42399	35.17559	
	31/12/2020, 23h	7.650367	22.14913	36.64788	34.34935	

Table 31: Predicted values, Confidence intervals and MAPE of models prediction

	Date	Lower Bound	Predicted Value	Upper Bound	True Value	MAPE
Mem-Martins	31/12/2020, 19h	44.00963	53.51535	63.02107	59.83239	
	31/12/2020, 20h	38.04575	52.51810	66.99046	62.07205	
	31/12/2020, 21h	34.13477	51.73620	69.33763	62.07205	16%
ARIMA(1,1,3)(0,1,2)[24]	31/12/2020, 22h	29.36561	48.98046	68.59532	60.57959	
	31/12/2020, 23h	27.67180	48.69007	69.70833	59.08453	

Table 32: Predicted values, Confidence intervals and MAPE of models prediction

7.2 Project 2

7.2.1 GALP

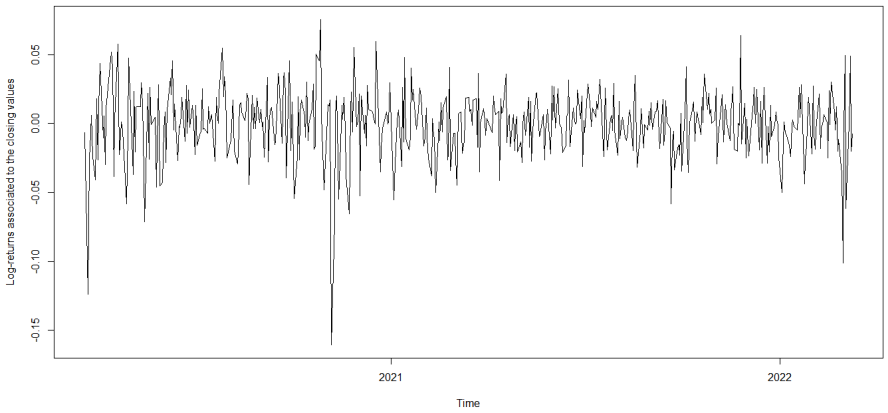

Figure 84: Plot of the log-returns for the closing values of GALP stock

Mean	Variance	Kurtosis
-0.0006110983	0.0005860637	5.130331

Table 33: Properties of the log-return time series for the GALP stock

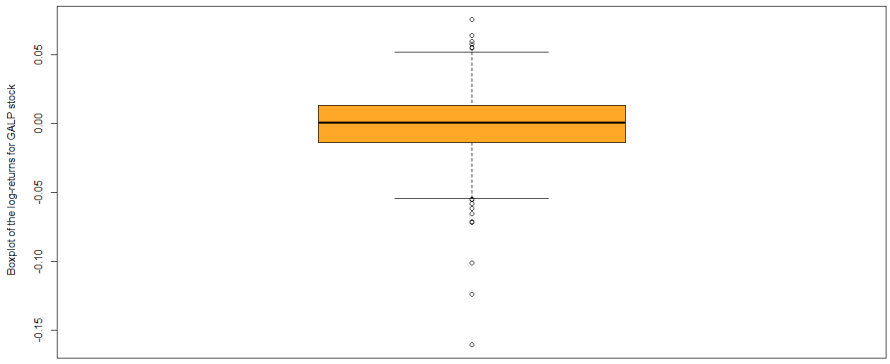

Figure 85: Boxplot of the log-returns for the closing values of GALP stock

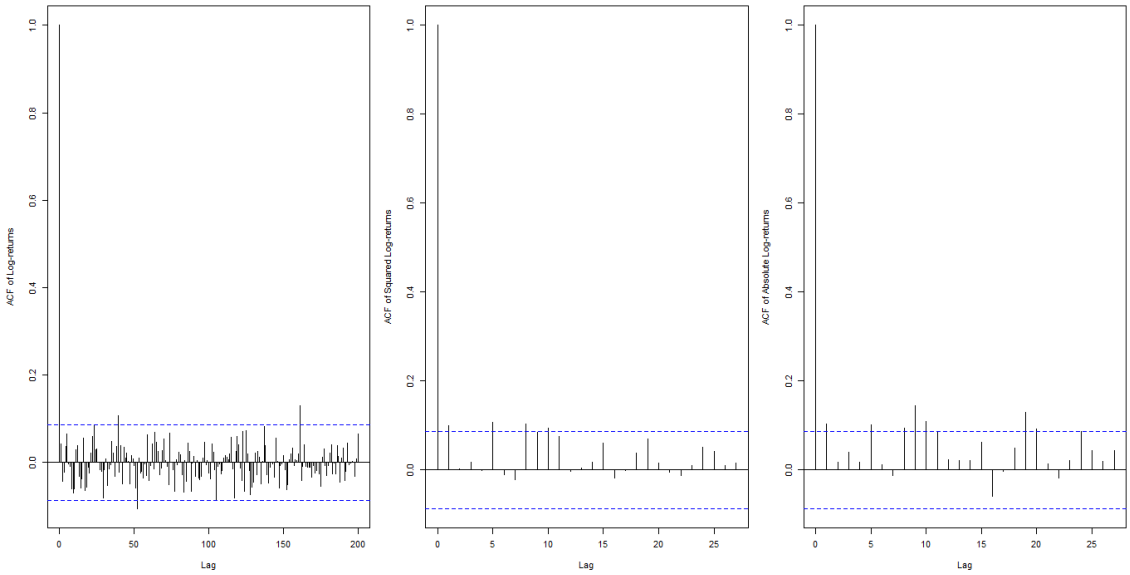

Figure 86: ACF of the log-returns for the closing values of GALP stock, as well as their square and absolute values

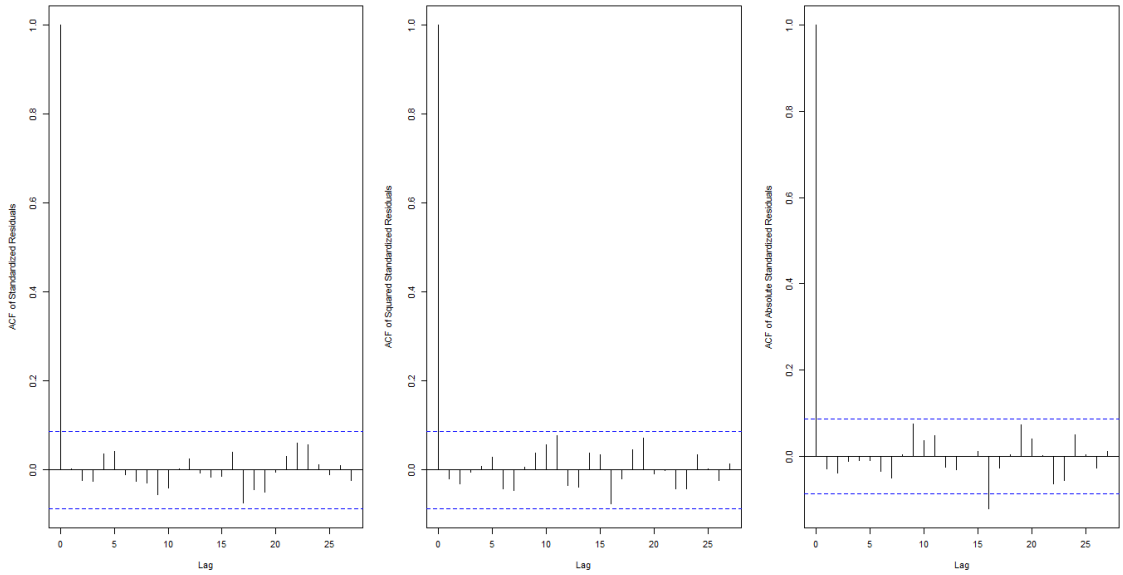

Figure 87: Plot of the ACF of the standard residuals, squared standard residuals, and absolute value of standard residuals, respectively, for the GALP model

7.2.2 MOTAENGIL

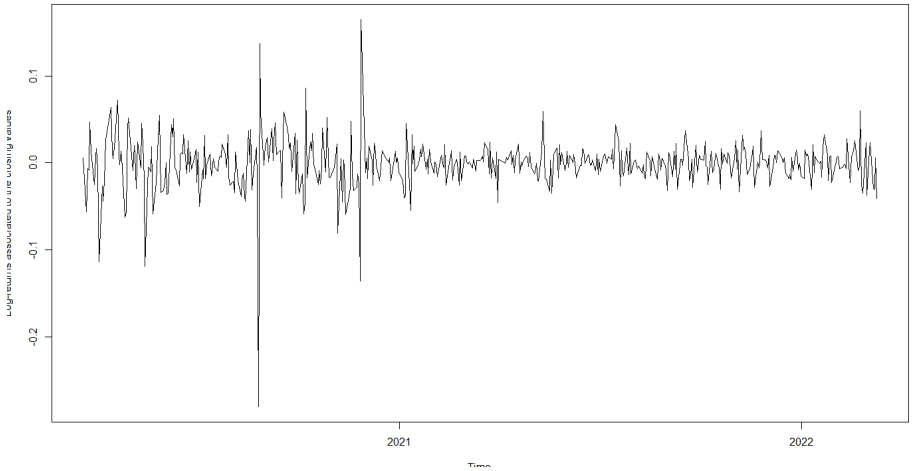

Figure 88: Plot of the log-returns for the closing values of MOTAENGIL stock

Mean	Variance	Kurtosis
-0.0004095451	0.0007938765	23.14985

Table 34: Properties of the log-return time series for the MOTAENGIL stock

Figure 89: Boxplot of the log-returns for the closing values of MOTAENGIL stock

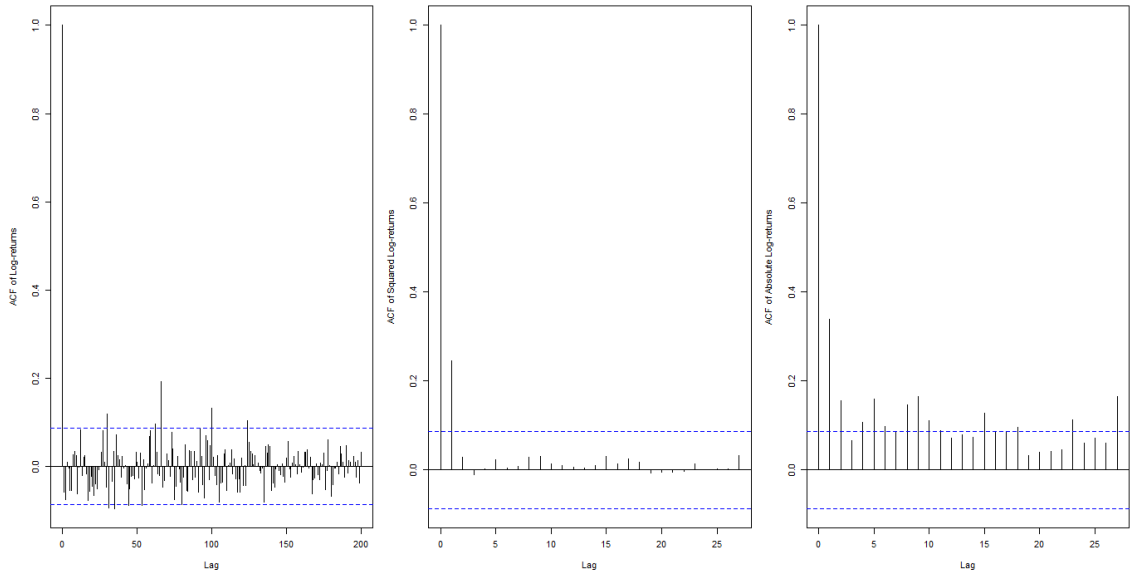

Figure 90: ACF of the log-returns for the closing values of MOTAENGIL stock, as well as their square and absolute values

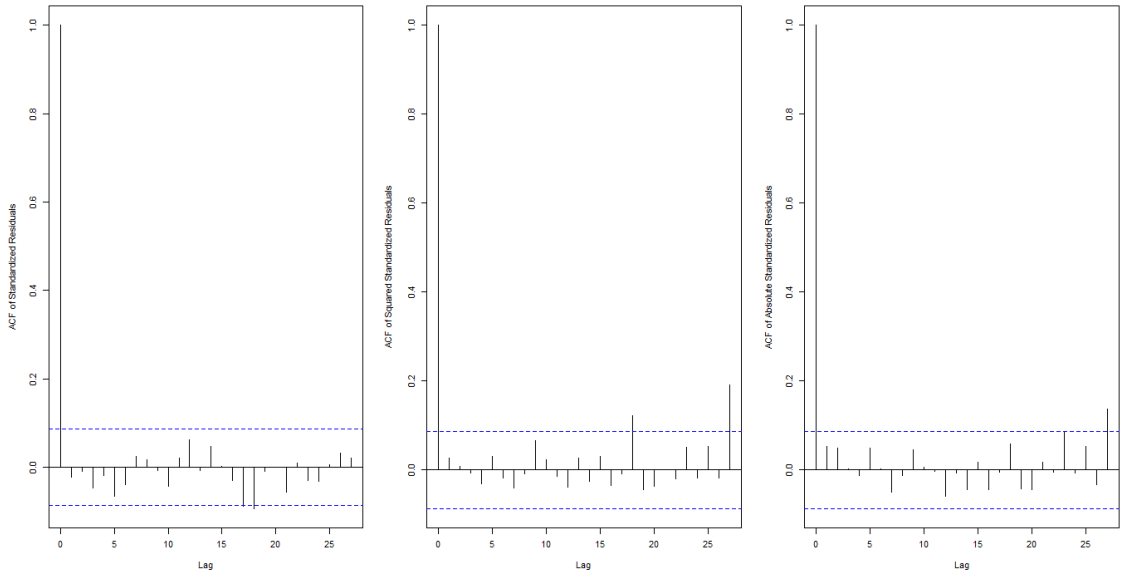

Figure 91: Plot of the ACF of the standard residuals, squared standard residuals, and absolute value of standard residuals, respectively, for the MOTAENGIL model

7.2.3 NOS

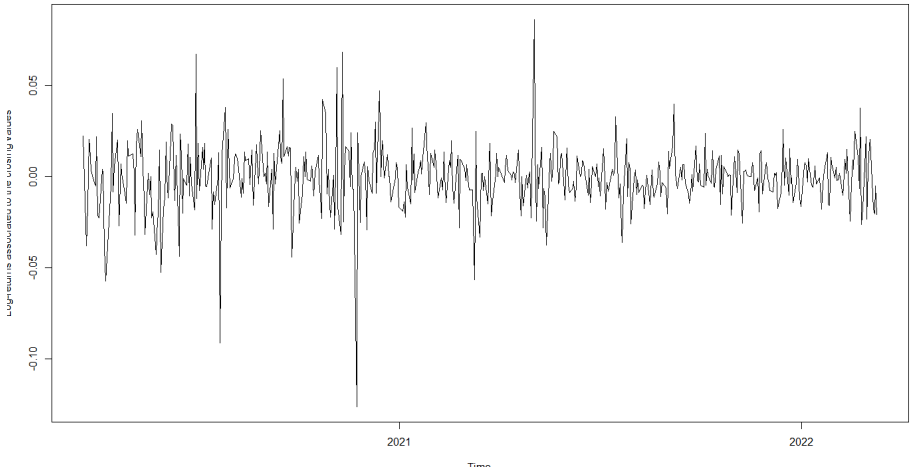

Figure 92: Plot of the log-returns for the closing values of NOS stock

Mean	Variance	Kurtosis
-0.0002705415	0.0003107792	8.053498

Table 35: Properties of the log-return time series for the NOS stock

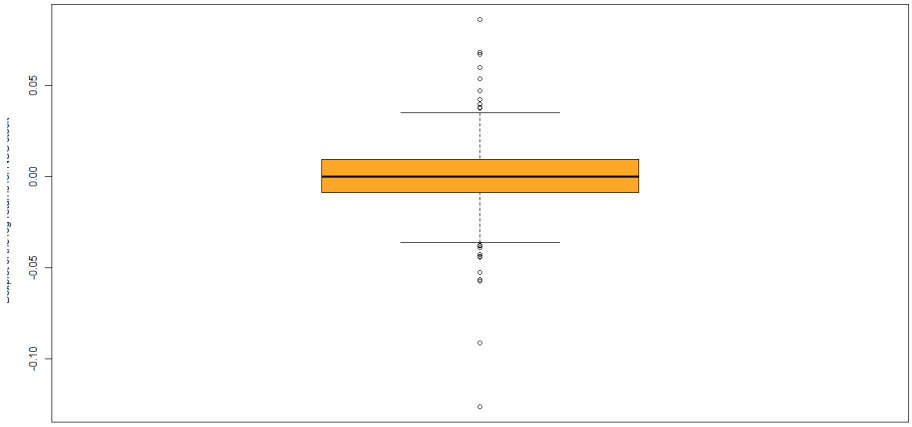

Figure 93: Boxplot of the log-returns for the closing values of NOS stock

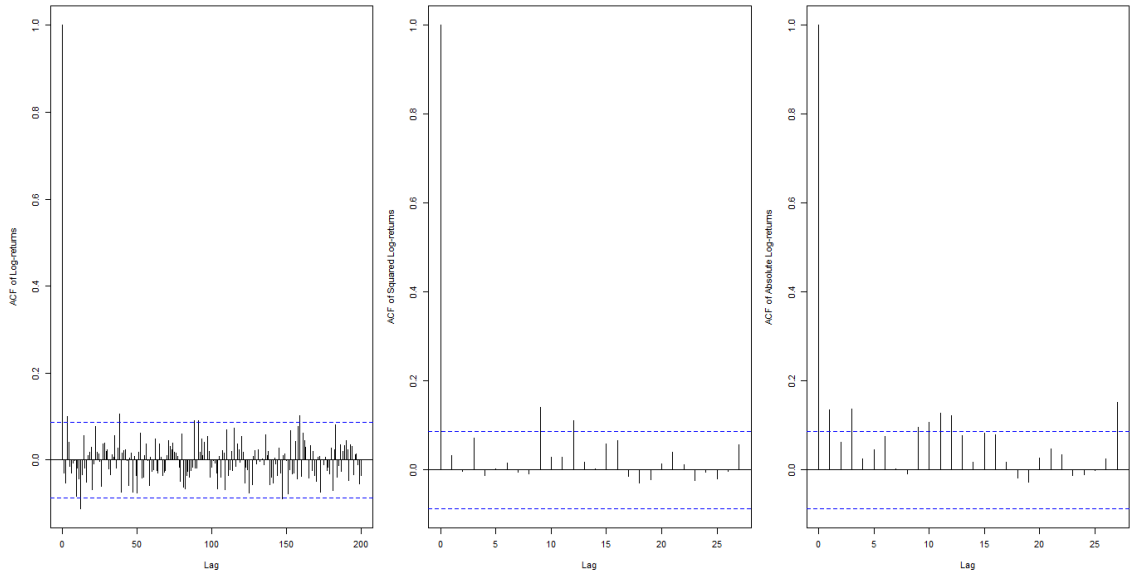

Figure 94: ACF of the log-returns for the closing values of NOS stock, as well as their square and absolute values

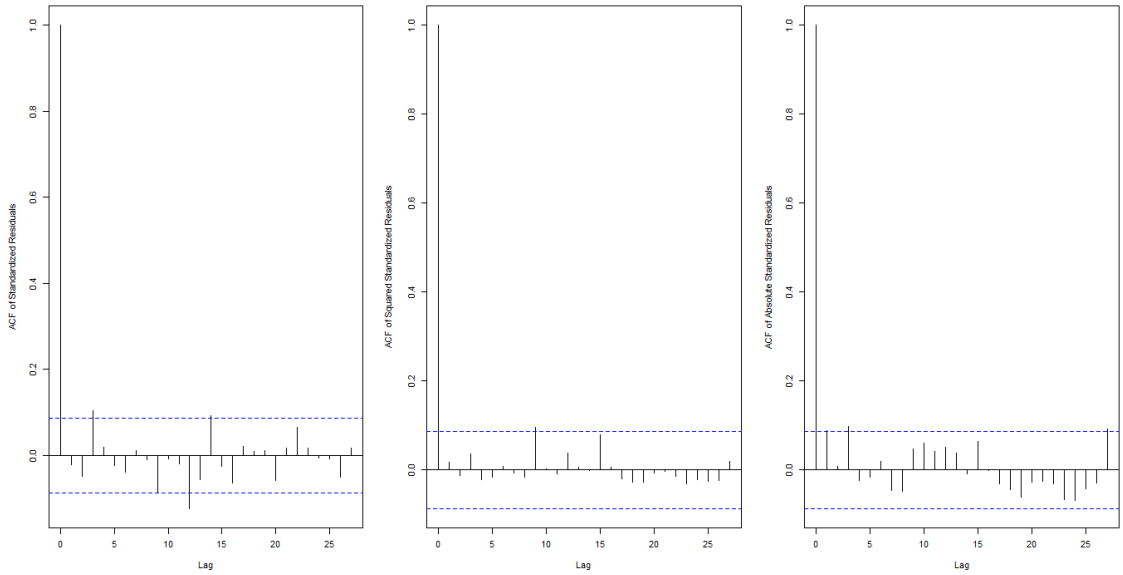

Figure 95: Plot of the ACF of the standard residuals, squared standard residuals, and absolute value of standard residuals, respectively, for the NOS model

7.2.4 NOVABASE

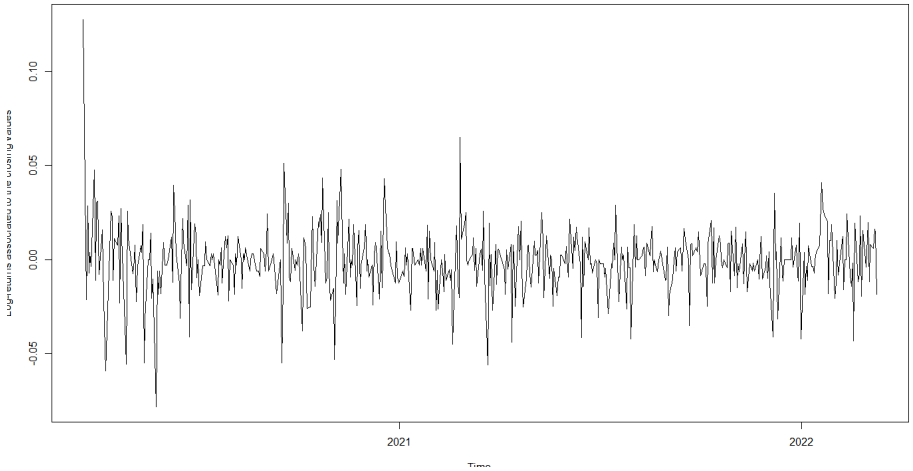

Figure 96: Plot of the log-returns for the closing values of NOVABASE stock

Mean	Variance	Kurtosis
-0.0008648083	0.0003078225	6.940005

Table 36: Properties of the log-return time series for the NOVABASE stock

Figure 97: Boxplot of the log-returns for the closing values of NOVABASE stock

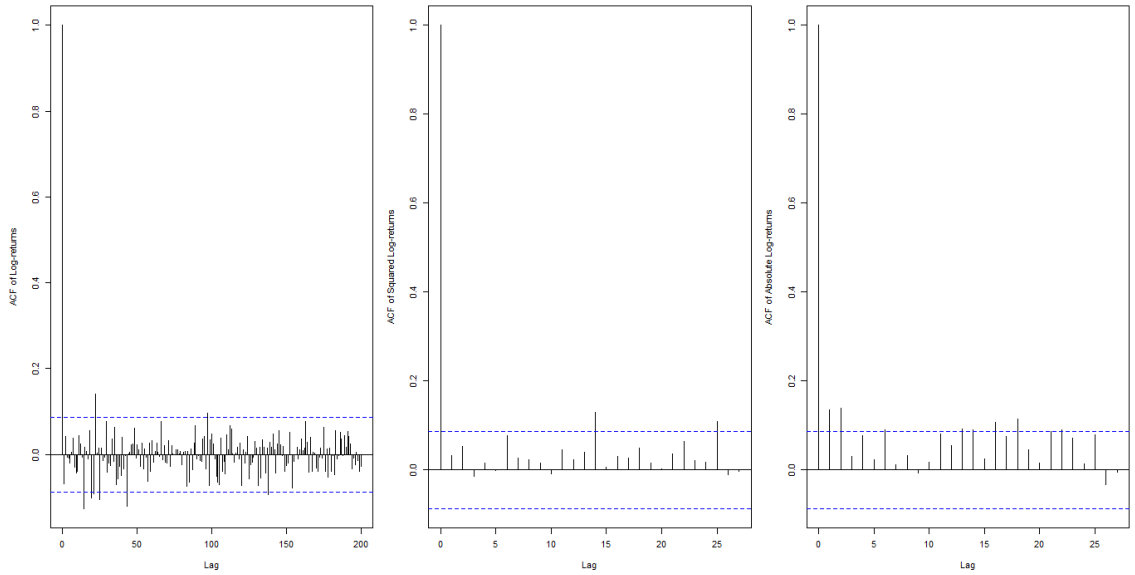

Figure 98: ACF of the log-returns for the closing values of NOVABASE stock, as well as their square and absolute values

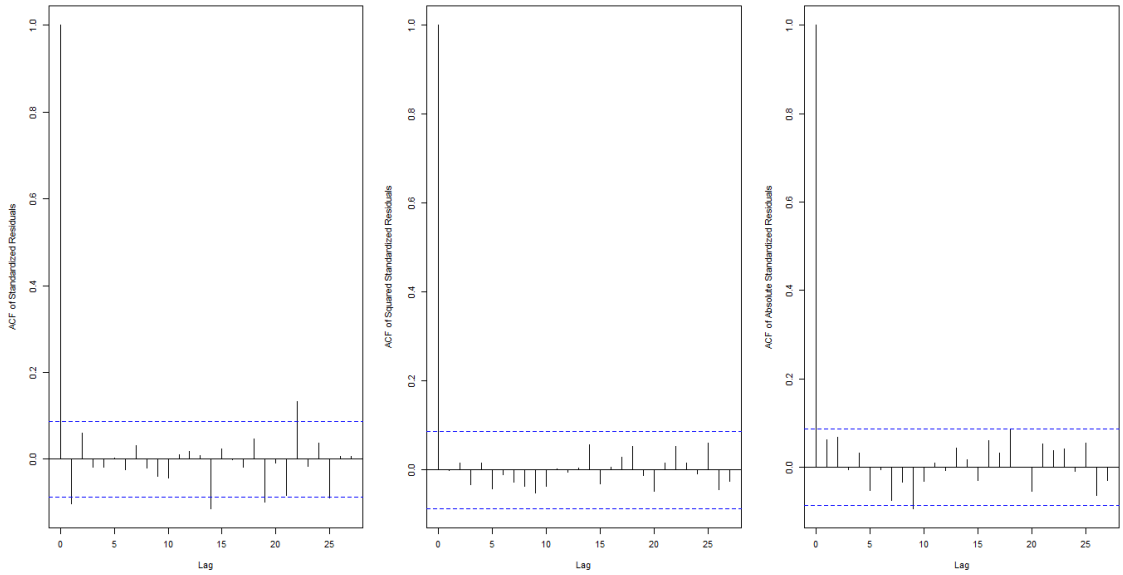

Figure 99: Plot of the ACF of the standard residuals, squared standard residuals, and absolute value of standard residuals, respectively, for the NOVABASE model

References

[1] Agência Portuguesa do Ambiente. [n.d.]. *Ozono (O3)*. https://apambiente.pt/ar-e-ruido/ozono-o3

[2] Harvard. [n.d.]. *The complex relationship between heat and ozone*. https://news.harvard.edu/gazette/story/2016/04/the-complex-relationship-between-heat-and-ozone/

[3] Mingfeng Lin, Henry C Lucas Jr, and Galit Shmueli. 2013. Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research* 24, 4 (2013), 906–917.

[4] Refk Selmi. [n.d.]. *What is the difference between ARCH and GARCH models?* https://www.researchgate.net/post/What-is-the-difference-between-GARCH-and-ARCH

[5] Chemistry World. [n.d.]. *The weekend effect*. https://www.chemistryworld.com/news/the-weekend-effect/3003987.article