INSTITUTO SUPERIOR TÉCNICO

TIME SERIES ANALYSIS

2^{nd} SEMESTER - 2021/2022

MECD, & MMAC

Time Series Analysis for Environmental and Financial Data

Group:
Alexandre SILVA - 90004
Ricardo SIMÕES - 93674

Professor:
Manuel Scotto

5 June 2022

Contents

1 Abstract

In this paper we describe and clarify the definitions and the usage of time series models to an environmental problem (O_3 levels) and logarithmic returns for financial assets (companies in the Euronext Lisbon). Then, we applied and compare the results between a wide variety of SARIMA-type models and GARCH-type models for each of the problems at hand. For the ten hourly-ground-levels of Ozone series, we fitted SARIMA-type models and, using a systematic procedure, we found out that the best models used a small set of parameters (between 3 and 4) and, for different series, some models were actually very similar to each other. This is probably a consequence of the close similarities between the series, for example the daily seasonality observed in all examples. For the five financial time series, the GARCH-type models were implemented. We see that simple models often work well, and there is not a big difference between the models, even if we add a lot of parameters. It was recurrent, however, that even when one considers different conditional distributions for the variance, the extreme observations are not well modeled, thus making the residual tests fail.

2 Introduction

This project, developed within the course of Time-series Analysis, approaches two different problems. In one hand, we have one of the rising subjects in today's world: environmental problems. More and more studies are made in order to try to look at long-term climate change variables such as levels of CO_2 or O_3, the last one being the one of the first assignment. This natural time dependence leads us to try to fit this data with SARIMA-type models. Since 10 different time series related to the hourly ground levels of O_3 values are given, collected at different stations of the qualar network, one will be fitted for each of the time series at hand. Then, we will proceed to check how well our model behaves when it tries to forecast the data into the future (up to 5 time periods ahead) and calculate 95% prediction intervals for each of these forecasts.

In the financial area (second assignment) of today an important question arises: how one defines and measures the risk of financial assets such as stock? Furthermore, it is not enough only to measure risks they also need to be compared to help us make decisions on different financial questions. Because of these comparability reasons one uses instead of the prices the returns of an asset. In this paper we will be dealing the logarithmic returns, and we aim to model this kind of data through a family of models: the GARCH-type models. In the first part of the theoretical background we will remind some of the definitions and some stylized facts about financial data. Then, we will try to find the best model that encompasses all of these characteristics in it, trying to validate our final model with residuals checking.

3 Time Series of hourly ground levels of Ozone values

3.1 Plan of Approach

The general procedure for each one of the ten time series can be characterized in the following steps:

1. Check for Missing Values (MVs) and Outliers;
2. Apply transformations (if needed) in order to impute Missing values and remove outliers;
3. Analyse the stationarity of the time series (Check for Trend, Seasonality and Variability);
4. Apply transformations (if needed) in order to turn the series stationary;
5. Identify the possible seasonal and regular component through the ACF and PACF of the residuals of the seasonal model;
6. Compare different models using Information Criteria (AIC, and BIC) and forecasting performance indicators (MAE and RMSE);
7. Analyse the residuals of the best model(s);
8. Forecast the future (5 time periods ahead) and calculate 95% prediction intervals, for the best model;

3.2 Plan Execution

3.2.1 Important Remarks

Since we will be analysing ten time series, we will only explain the important details for the first one. The remaining time series will be analysed with the same rationale, but with a high-level explanation.

3.2.2 Antas-Espinho

For the Antas-Espinho region, the hourly-ground-levels of O_3 values time series had no missing values. In terms of outliers, there was no significant odd value among the ones presented in the time series, which is something that can be visually verified in Figure 1.

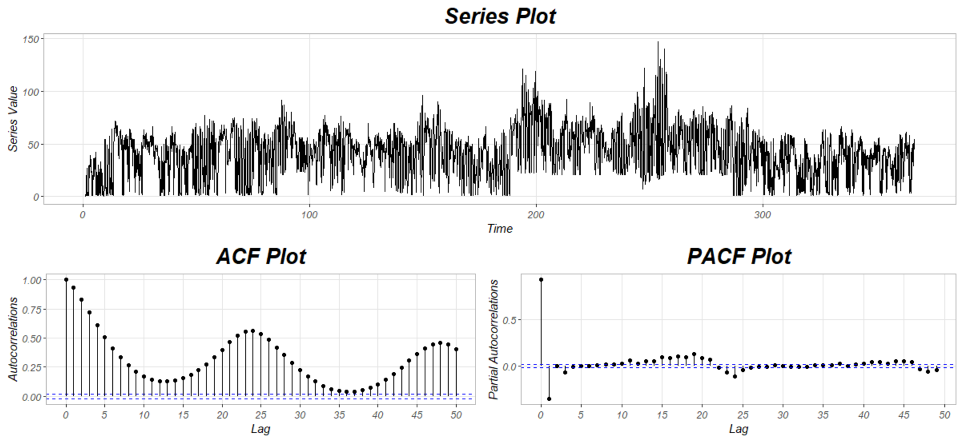


Figure 1: Antas-Espinho time series, ACF and PACF

Besides the series plot of the time-series we also obtained the ACF and PACF plots. In particular, the ACF shows a slow decay indicating a non-stationary behavior. For this reason, in order to fit a $SARIMA(p, d, q)(P, D, Q)_s$ model to our series, we will first have to transform the series in a stationary one (for example, with difference operators, related with the parameters d and D), and then we will proceed to find the autoregressive and moving average terms of the model (p, q, P and Q).

In terms of trend (d), the series does not seem to assume a clear upwards or downwards trend. For this reason, we will assert d=0. In terms of seasonal behaviour (D), it is not clear if exists one from the plot of the series, but ACF shows significant autocorrelations at lags 24, 48, etc. This gives us an idea that maybe the seasonality is 24 and, consequentially, maybe D=1. We decided to further investigate this using boxplots that aggregate the values by hour of the day. Let us consider the plot from Figure 2.

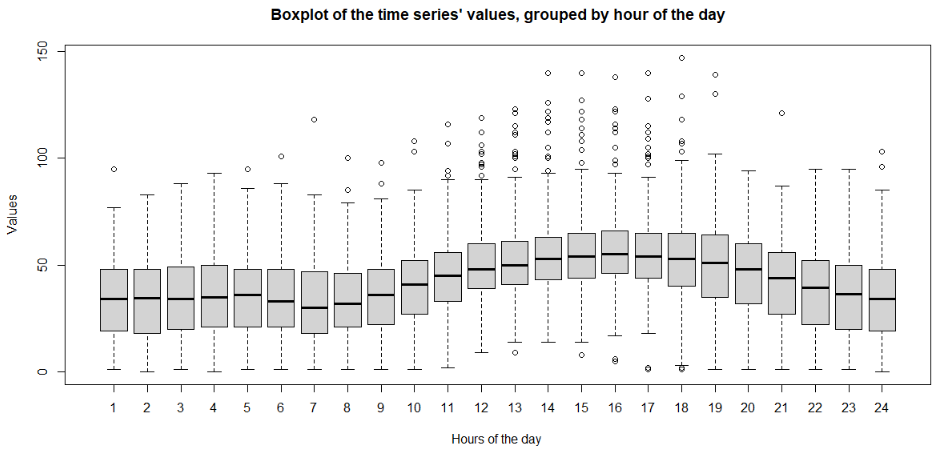


Figure 2: Antas-Espinho time series observations grouped by hour of the day

We can see that in the beginning of the day, the O_3 levels are stable around a specific level, then they tend to fall around 6/7 AM. Throughout the afternoon they rise until they achieve a peak at 4 PM, and then they fall to the level, where it started, and then the patterns repeats. This is another clear evidence that the seasonality is, in fact, 24 (hours). Some studies on the patterns of ground-level 0_3 were already conducted in Portugal (see [4]), although the regions of the study were not the same as the ones we are studying in this report. One important takeaway of that study is that there is a daily pattern, characterized exactly the same way as we saw it in the boxplot. Since, that study already concluded that this pattern exists in different regions of Portugal, than it is possible that our final models may not vary between regions, in terms of structure (i.e, number of autoregressive and moving average coefficients in ordinary and seasonal parts).

Having identified the order of the trend and seasonal difference operators, and applied them to our original series, we should now observe a stationary time series. Let's consider the difference series in figure 3.

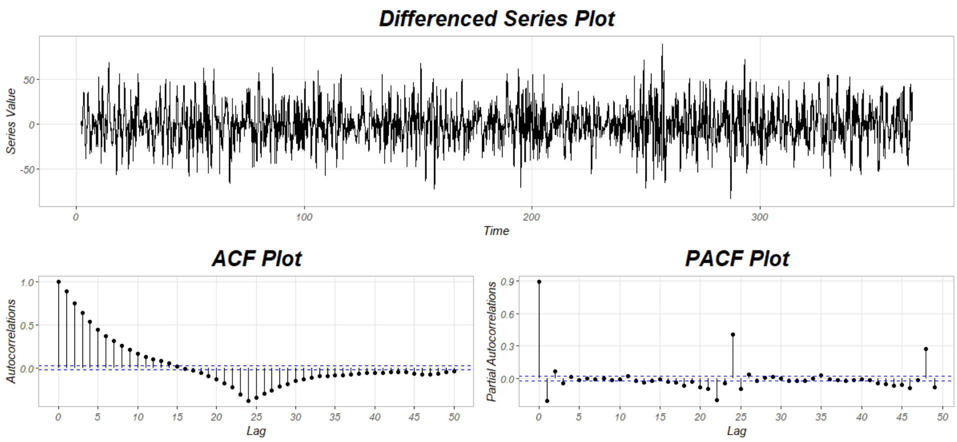


Figure 3: Differenced Time Series of Antas-Espinho with ACF and PACF

The time series seems to be stationary, having a ACF decaying exponentially to zero, with a mean value of zero and constant variance. There are some periods where variance seems to increase slightly, which is something that can be minimized with a linear transformation like a Box-Cox transformation, defined as:

$$BoxCox(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ ln(x), & \lambda = 0 \end{cases}$$

The *Arima* function we used to fit the models finds the optimal value for lambda (λ), which was 0.7575 for this particular time series, and applies the transformation to the original series. The optimal lambda value is the one which results in the best approximation of a normal distribution curve.

Now we only have to find the autoregressive and moving average terms for the seasonal and non-seasonal components of the SARIMA model (p, q, P and Q). To do that, we will take a look at the ACF and PACF plots from above and adopt a step-by-step approach. In other words, we will choose one of the four parameters by inspection and we will pick the most obvious one. After that, we will plot the ACF and PACF of the residuals series with that extra parameter added to the model and pick the next most obvious parameter until the residuals are white noise.

One important aspect of this analysis is the significance thresholds of the correlations. By default, the ones that appear are the 5% critical values of the auto-correlation, and they are influenced by the number of observations the series has (see [1]). The more observations the series has, the lower the thresholds will be. Since we have several thousands of observations for each series, we will despise low correlation values (< 0.1).

This process, however, can be somewhat subjective. For instance, looking at the PACF when could pick $p = 1$ or $p = 2$, or even $P = 2$ since there is a significant spike at lags $h = 1$, a smaller one at $h = 2$, but also at $h = 24$ and $h = 48$. Therefore, we will consider all the possible models that may appear when adding new parameters, which may lead us to different models in the end. We will present this step-by-step approach for three different final SARIMA-type models in the appendix section (Tabs. 67, 68, 69). One important thing to mention is that using this process, we obtained the exact same eight SARIMA-type models with the same configurations, for all the series. These "paths" that allowed to achieve models whose residuals are white noise can be seen in the Figure 70, in the Appendix section.

In order to evaluate the models, we picked four criteria that were appropriate for our particular problem, and we will use them to select the best models according to a rigorous procedure. For instance, AIC and BIC are factor-based criteria that discourage the fitting of models with too many parameters. The last one tends to penalize more the introduction of new parameters than the other. This is an important things to take into consideration, since we want our models to have the less amount of parameters possible, while being capable of generalize for the future. For the AIC criterion, we will first find the model with the smaller AIC value, and then we will compute the differences of the minimum AIC with value for all the other models. For the BIC the rationale is the same. In fact, for these two information criteria we should always compute differences. Starting with AIC:

$$\Delta_i^{AIC} = AIC_i - AIC_{min}$$

Then, we will use a rule of thumb outlined in [3], which is:

1. if $\Delta_i < 2$, then there is substantial support for the i-th model (or the evidence against it is worth only a bare mention), and the proposition that it is a proper description is highly probable;
2. if $2 < \Delta_i < 4$, then there is strong support for the i-th model;
3. if $4 < \Delta_i < 7$, then there is considerably less support for the i-th model;

4. models with $\Delta_i > 10$ then the i-th model has essentially no support;

For BIC the rationale is similar:

$$\Delta_i^{BIC} = BIC_i - BIC_{min}$$

We will use Raftery's approach (see [2]):

1. if $\Delta_i \leq 2$, then smaller BIC does not indicate to be better than the other;
2. if $2 < \Delta_i \leq 6$, then the smaller BIC value is weakly likely to be better;
3. if $6 < \Delta_i \leq 10$, then the smaller BIC value is strongly likely to be better;
4. models with $\Delta_i > 10$, then the smaller BIC value is considerable better, with a very strong likelihood;

As for the remaining perfomance indicators, MAE and RMSE, they are defined according to the following formulas:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{n} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}}$$

There are quite a few performance indicators that can be used, but we decided to pick two that are simple to understand. There is no big difference between both indicators we decided to use. The MAE is preferable, because from an interpretability point of view, it's easier to understand its mathematical expression and what it is actually measuring. We decided to also measure the RMSE. It penalizes more the large errors, since the errors are squared before they are averaged. In consequence, models with very large errors will be more penalized. A good example of this can be found in [5].

We will now present Table 1, which contains all the models obtained from the step-by-step analysis from above, with their respective values for the four key performance indicators. All the models were fitted using all the data from the series.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	43272.01	43300.32	4.67	6.66
$(1,0,1)(1,1,1)_{24}$	43253.78	43289.17	4.66	6.65
$(1,0,1)(0,1,2)_{24}$	43254.34	43289.73	4.66	6.65
$(1,0,1)(1,1,2)_{24}$	43253.62	43296.09	4.66	6.65
$(2,0,0)(0,1,1)_{24}$	43304.6	43332.91	4.67	6.67
$(2,0,0)(1,1,1)_{24}$	43287.82	43323.21	4.66	6.66
$(2,0,0)(0,1,2)_{24}$	43288.39	43323.21	4.67	6.66

Table 1: Metrics of the best models for Antas-Espinho

Since the model $(1,0,1)(1,1,1)_{24}$ is the one with the lower AIC and BIC, at an integer place level, let's consider now Table 2.

Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	18	11
i=2	$(1,0,1)(1,1,1)_{24}$	0	0
i=3	$(1,0,1)(0,1,2)_{24}$	1	1
i=4	$(1,0,1)(1,1,2)_{24}$	0	7
i=5	$(2,0,0)(0,1,1)_{24}$	51	44
i=6	$(2,0,0)(1,1,1)_{24}$	34	34
i=7	$(2,0,0)(0,1,2)_{24}$	35	35

Table 2: AIC and BIC comparisons for the models of Antas-Espinho

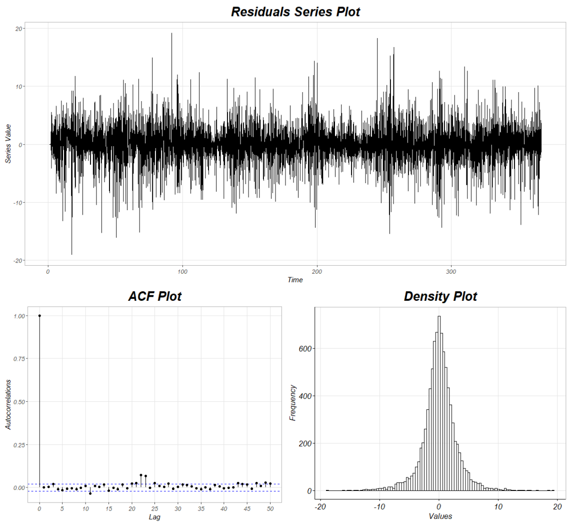
Looking at the MAE of Table 1, the forecast's distance from the true value is, on average, a value between 4.66 and 4.67. We can see that these differences are practically null, especially because the time series is composed by integer numbers. This rationale is similar for the RMSE, although this last one varies between 6.66 and 6.67. These errors don't seem to be really high, especially because the time series in questions has a range of 0 to 147 µg/m^3. Since there seems to be no apparent different in terms of the error indicators, we will analyse the information criterion differences of Table 2. From the point of view of the information criteria, the winner model is $i = 2$. Using the two rules of thumb defined above for these two criteria, it seems that there is also no significant difference with the model $i = 3$. Model $i = 4$ can also be considered, although it's BIC value indicates a strong evidence that it is worse than the other two. For the remaining ones, the differences are very large, which is something that goes against these remaining models, so we will discard them.

We will now present the residuals of these three models, which is "what is left" after we fit the model. They will be useful to check if the models captured all information of the time series.

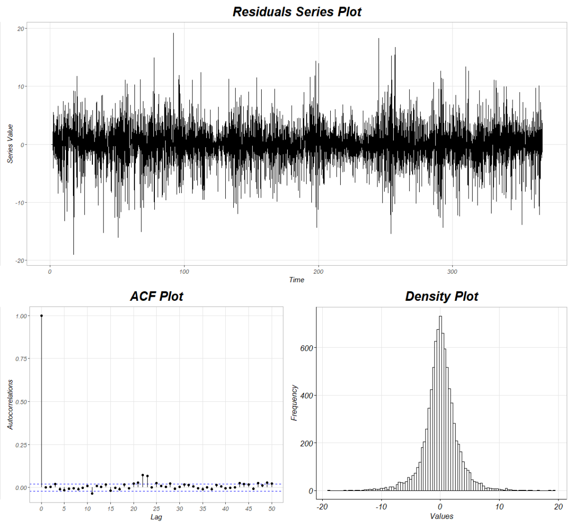
There a couple important properties that the residuals should exhibit:

- Uncorrelated between each other, otherwise not all information was captured;
- Stationary, with zero-mean;
- Normally distributed;

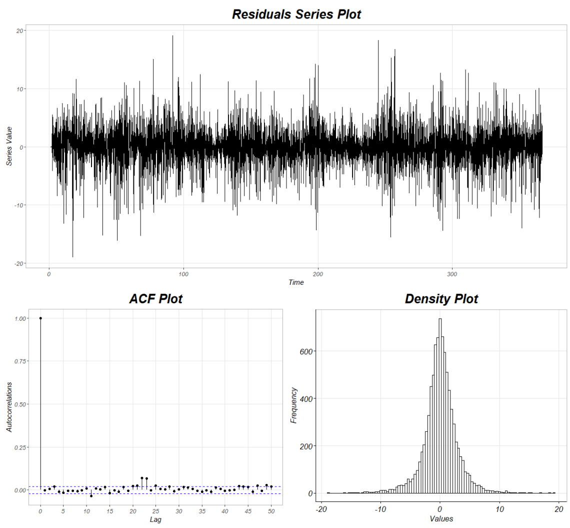
In Figure 4 we can visualize if these properties are being satisfied, for the three models. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 20, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 2$, $i = 3$ and $i = 4$ were, respectively: 0.0165, 0.0163 and 0.0185.



(a) $(1,0,1)(1,1,1)_{24}$ model



(b) $(1,0,1)(0,1,2)_{24}$ model



(c) $(1,0,1)(1,1,2)_{24}$ model

Figure 4: Residuals of the final models for Antas-Espinho

By inspection, we can see that all properties from above are being satisfied. It could be argued that for the three models, there is still is a spike in the ACF around lag $h = 24$, however, the correlation value is quite low, so all information seems to be accurately captured by the models. Since these two models only differ in the AIC and BIC values, we will pick $i = 2$ because it is the one that minimizes these indicators. The results seen in Figure 5 and Table 3 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

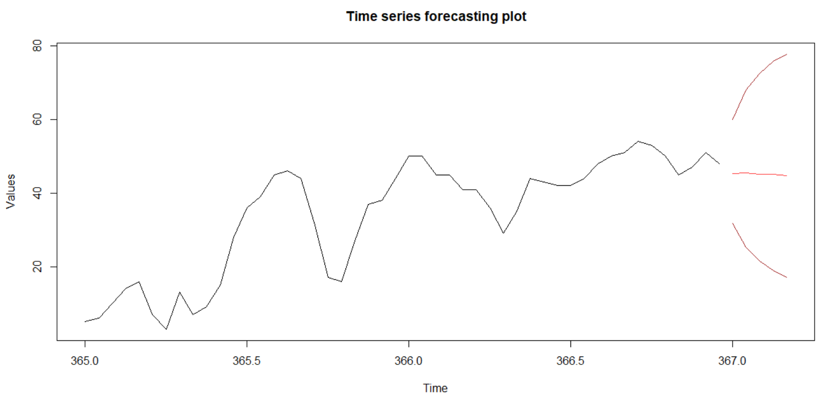


Figure 5: Time Series forecasting plot for Antas-Espinho

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	45.33	31.82	59.91
2nd hour	45.39	25.26	68.03
3rd hour	45.14	21.38	72.54
4th hour	45.15	19.01	75.81
5th hour	44.78	17.07	77.71

Table 3: Time Series forecasting values for Antas-Espinho

For a general expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 0.8793$, $\theta_1 = 0.2739$, $\Phi_1 = 0.0526$ and $\Theta_1 = -0.9397$. The step-by-step approach for this particular model can be found in Table 67.

3.2.3 Entrecampos

Figure 6 shows the plot of the time series of Entrecampos with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.7912. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

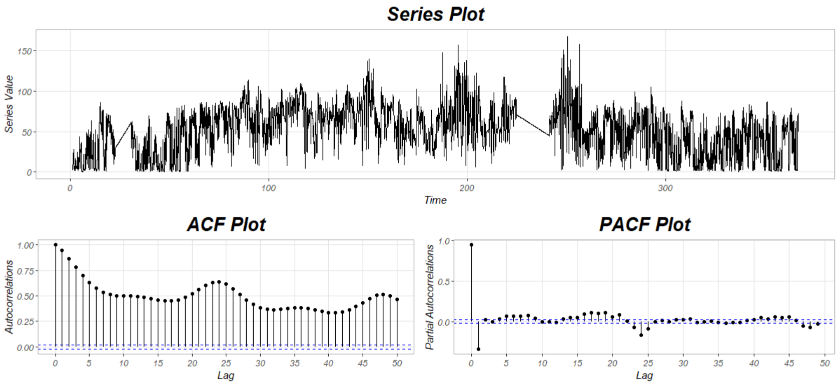


Figure 6: Entrecampos Time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 7).

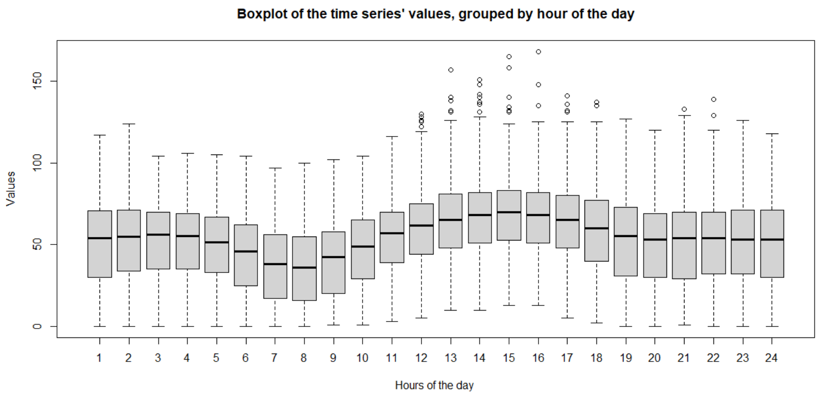


Figure 7: Entrecampos time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 8) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

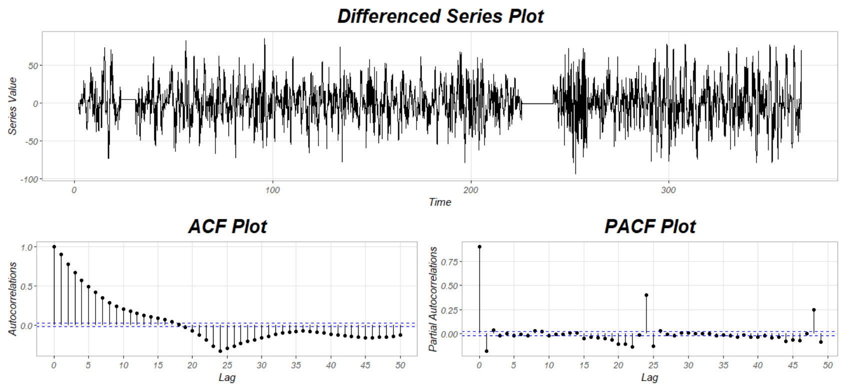


Figure 8: Differenced Time Series of Entrecampos with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 4 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	46541.98	46570.3	5.13	7.46
$(1,0,1)(1,1,1)_{24}$	46505.00	46540.39	5.12	7.44
$(1,0,1)(0,1,2)_{24}$	46506.05	46541.44	5.12	7.44
$(1,0,1)(1,1,2)_{24}$	46506.38	46548.85	5.12	7.44
$(2,0,0)(0,1,1)_{24}$	46565.67	46593.98	5.14	7.47
$(2,0,0)(1,1,1)_{24}$	46528.49	46563.88	5.12	7.45
$(2,0,0)(0,1,2)_{24}$	46529.55	46563.88	5.12	7.45

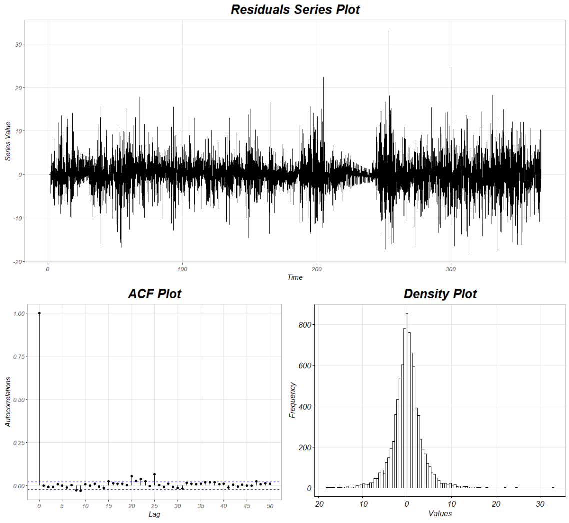
Table 4: Metrics of the best models for Entrecampos

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the model $i = 2$ has the minimum AIC and BIC (see Table 5).

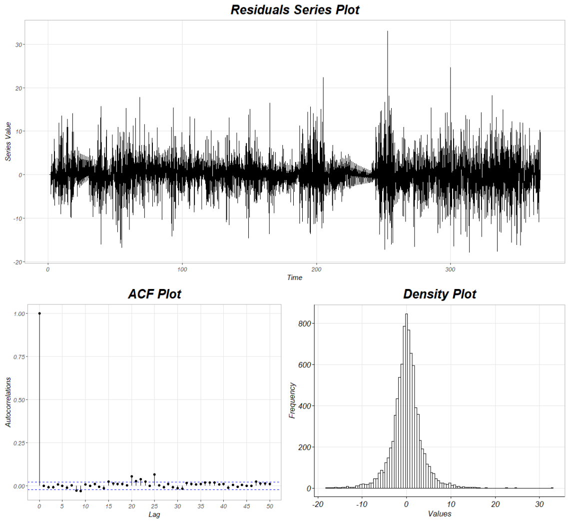
Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	37	30
i=2	$(1,0,1)(1,1,1)_{24}$	0	0
i=3	$(1,0,1)(0,1,2)_{24}$	1	1
i=4	$(1,0,1)(1,1,2)_{24}$	1	8
i=5	$(2,0,0)(0,1,1)_{24}$	61	54
i=6	$(2,0,0)(1,1,1)_{24}$	23	23
i=7	$(2,0,0)(0,1,2)_{24}$	25	25

Table 5: AIC and BIC comparisons for the models of Entrecampos

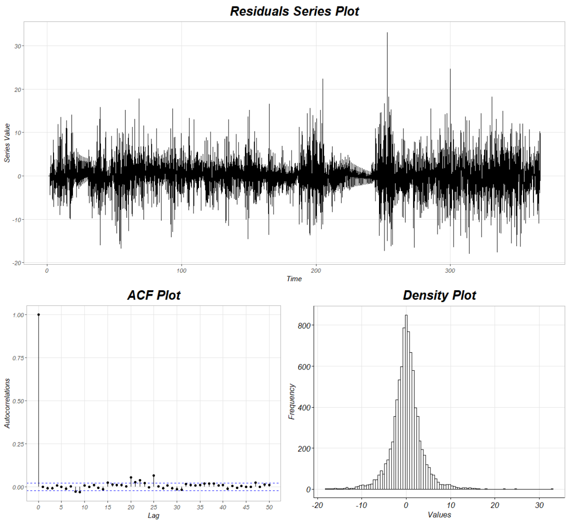
According the rules we defined above for comparing AIC and BIC values, models $i = 2$ and $i = 3$ are practically identical. The model $i = 4$ will also be considered, although it has a higher BIC value. All the other models are significantly worse, so we will discard them. Figure 9 presents the residuals series, ACF and distribution plots of these three models selected.



(a) $(1,0,1)(1,1,1)_{24}$ model



(b) $(1,0,1)(0,1,2)_{24}$ model



(c) $(1,0,1)(1,1,2)_{24}$ model

Figure 9: Residuals of the final models for Entrecampos

The plots from Figure 9 show that the residuals from the three models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 19, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 2$, $i = 3$ and $i = 4$ were, respectively: 0.0405, 0.0409 and 0.0410.

The three models were viable for forecasting, but we decided to go ahead with model $i = 2$ since it's the one with the lowest AIC and BIC values.

The results seen in Figure 10 and Table 6 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

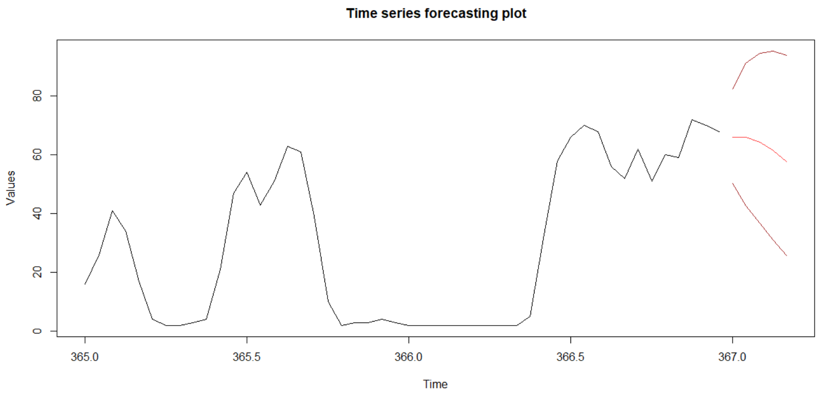


Figure 10: Time Series forecasting plot for Entrecampos

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	65.99	50.31	82.49
2nd hour	66.00	42.78	91.10
3rd hour	64.23	36.68	94.57
4th hour	61.39	31.12	95.28
5th hour	57.60	25.67	93.93

Table 6: Time Series forecasting values for Entrecampos

For a general expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 0.9039$, $\theta_1 = 0.2190$, $\Phi_1 = 0.0724$ and $\Theta_1 = -0.9489$.

3.2.4 Estarreja

Figure 11 shows the plot of the time series of Estarreja with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.5383. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

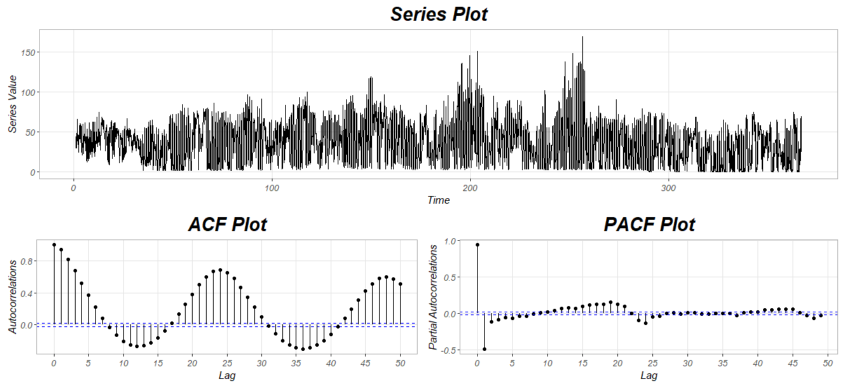


Figure 11: Estarreja time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 12).

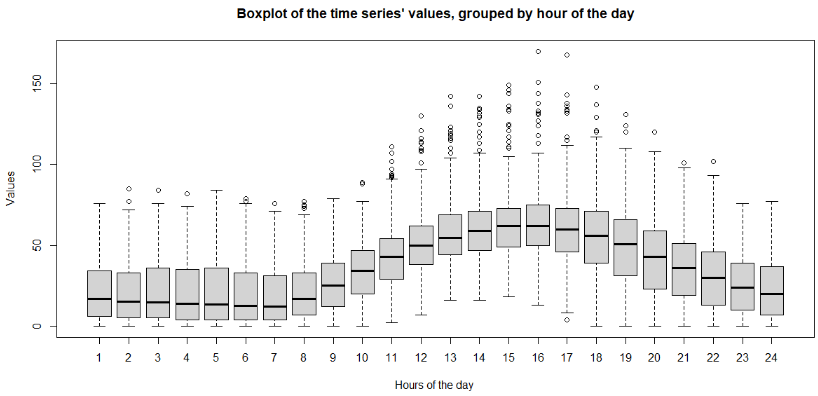


Figure 12: Estarreja time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 13) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

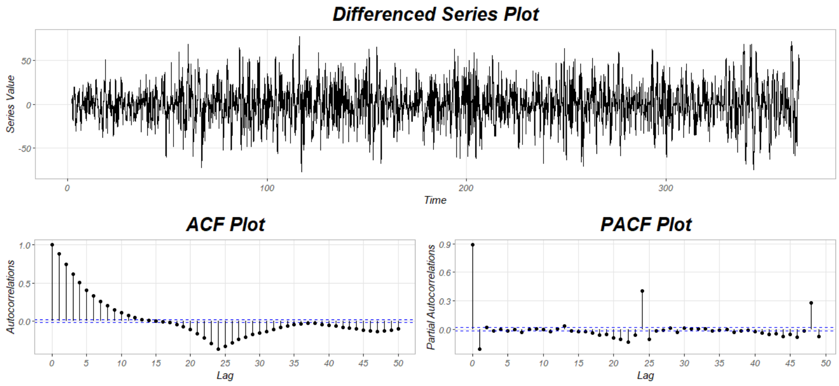


Figure 13: Differenced Time Series of Estarreja with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 7 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	32057.42	32085.73	4.95	7.19
$(1,0,1)(1,1,1)_{24}$	32013.66	32049.05	4.94	7.15
$(1,0,1)(0,1,2)_{24}$	32012.88	32048.27	4.94	7.15
$(1,0,1)(1,1,2)_{24}$	32014.62	32057.09	4.94	7.15
$(2,0,0)(0,1,1)_{24}$	32007.4	32035.71	4.93	7.14
$(2,0,0)(1,1,1)_{24}$	31967.21	32002.6	4.91	7.12
$(2,0,0)(0,1,2)_{24}$	31966.51	32002.6	4.91	7.12

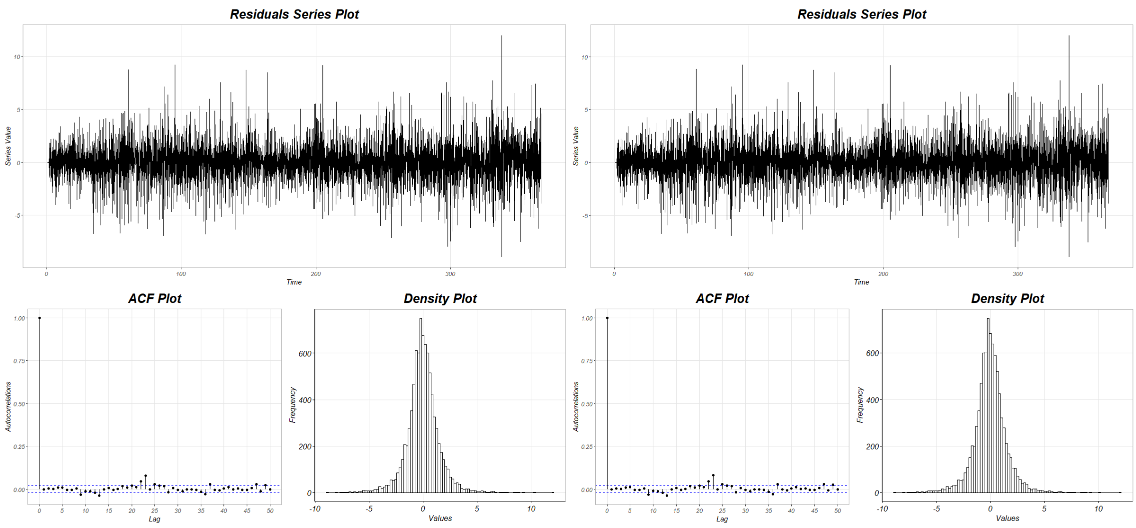
Table 7: Metrics of the best models for Estarreja

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the model $i = 7$ has the minimum AIC and BIC (see Table 8).

	Models	Δ_i^{AIC}	Δ_i^{BIC}
i = 1	$(1,0,1)(0,1,1)_{24}$	91	84
i=2	$(1,0,1)(1,1,1)_{24}$	47	47
i=3	$(1,0,1)(0,1,2)_{24}$	46	46
i=4	$(1,0,1)(1,1,2)_{24}$	48	55
i=5	$(2,0,0)(0,1,1)_{24}$	41	34
i=6	$(2,0,0)(1,1,1)_{24}$	1	1
i=7	$(2,0,0)(0,1,2)_{24}$	0	0

Table 8: AIC and BIC comparisons for the models of Estarreja

According the rules we defined above for comparing AIC and BIC values, we conclude that the difference between the models $i = 7$ and $i = 6$ is negligible. All the other models are significantly worse, so we will discard them. Figure 14 presents the residuals series, ACF and distribution plots of the two best models selected.



(a) $(2,0,0)(1,1,1)_{24}$ model (b) $(2,0,0)(0,1,2)_{24}$ model

Figure 14: Residuals of the final models for Estarreja

The plots from Figure 14 show that the residuals from both models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 19, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 6$ and $i = 7$ were both equal to 0.0133.

Both models were viable for forecasting, but we decided to go ahead with model i=7 since it's the one with the lowest AIC and BIC values.

The results seen in Figure 15 and Table 9 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

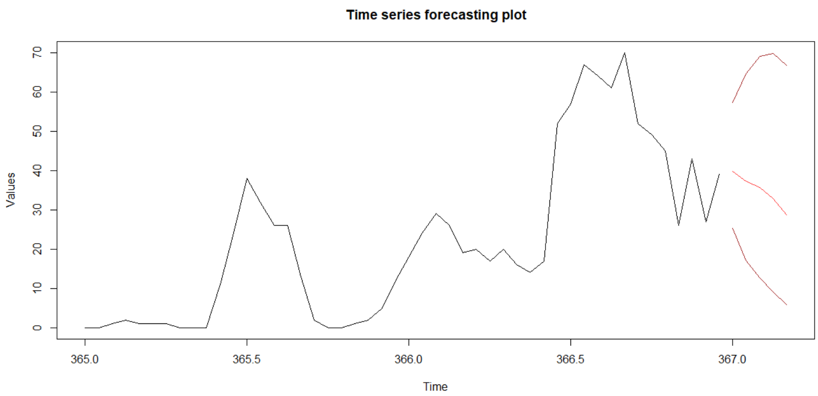


Figure 15: Time Series forecasting plot for Estarreja

	Predicted Value	**Lower Bound**	**Upper Bound**
1^{st} hour	39.84	25.28	57.38
2^{nd} hour	37.36	17.16	64.54
3^{rd} hour	35.65	12.62	69.11
4^{th} hour	32.88	9.07	69.84
5^{th} hour	28.68	5.81	66.82

Table 9: Time Series forecasting values for Estarreja

For a general expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 1.1479$, $\phi_2 = -0.2536$, $\Theta_1 = -0.8634$ and $\Theta_2 = 0.0716$. The step-by-step approach for this particular time series can be found in Table 68.

3.2.5 Laranjeiro-Almada

Figure 16 shows the plot of the time series of Entrecampos with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.6286. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

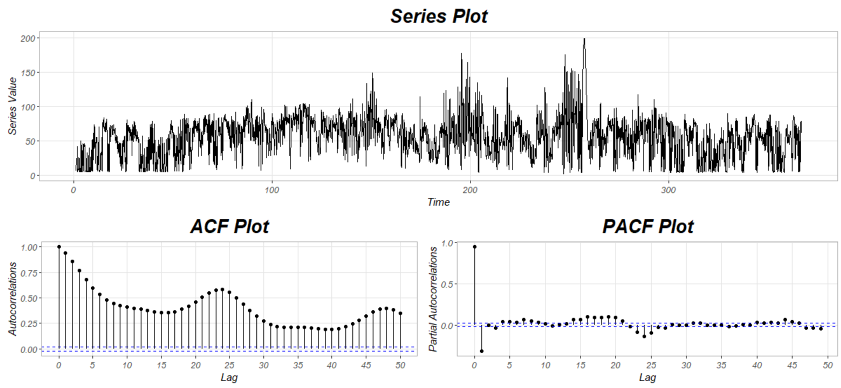


Figure 16: Laranjeiro-Almada time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 17).

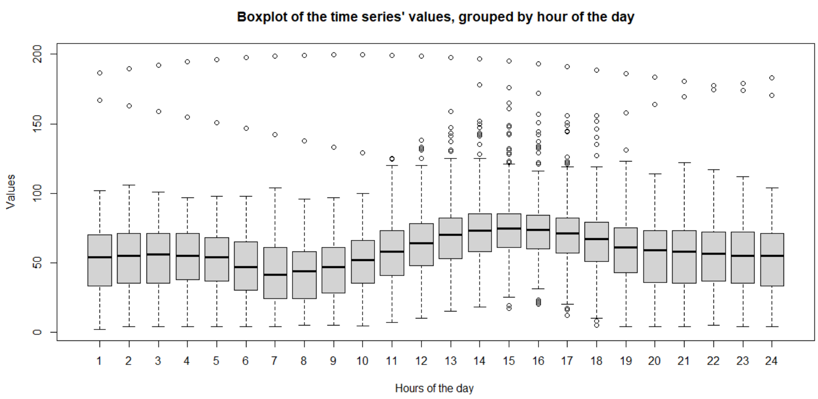


Figure 17: Laranjeiro-Almada time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 18) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

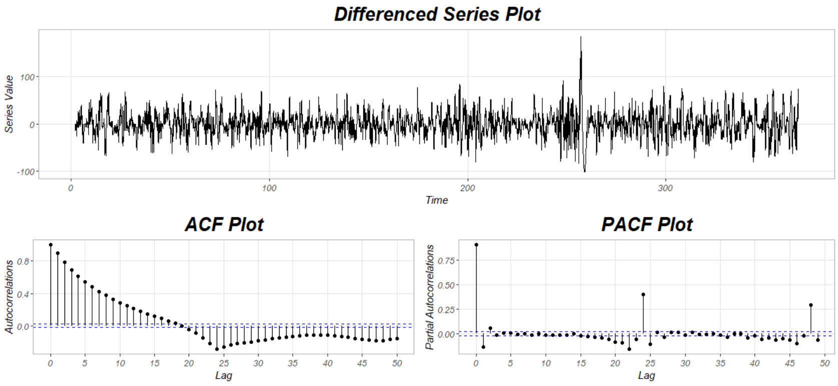


Figure 18: Differenced Time Series of Laranjeiro-Almada with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 10 have residuals whose ACF and PACF values are very low.

Models	AIC	BIC	MAE	RMSE
$(1,0,1)(0,1,1)_{24}$	36744.77	36773.08	5.37	7.92
$(1,0,1)(1,1,1)_{24}$	36725.08	36760.47	5.36	7.90
$(1,0,1)(0,1,2)_{24}$	36724.04	36759.43	5.36	7.90
$(1,0,1)(1,1,2)_{24}$	36723.09	36765.56	5.36	7.90
$(2,0,0)(0,1,1)_{24}$	36765.84	36794.16	5.38	7.93
$(2,0,0)(1,1,1)_{24}$	36745.49	36780.88	5.37	7.92
$(2,0,0)(0,1,2)_{24}$	36744.44	36780.88	5.37	7.92

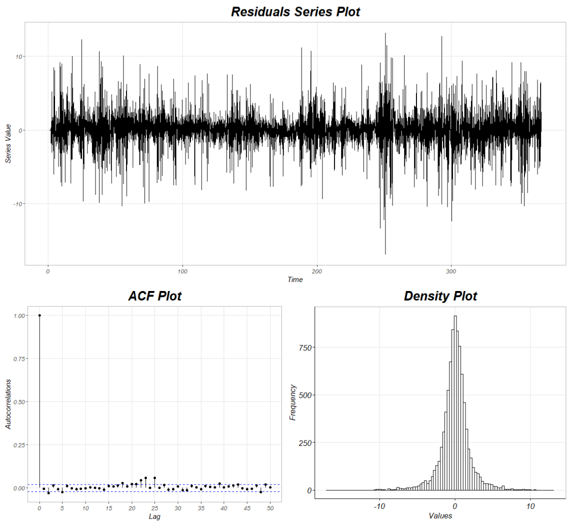
Table 10: Metrics of the best models for Laranjeiro-Almada

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the model $i = 4$ has the minimum AIC and model $i = 3$ has the lower BIC (see Table 11).

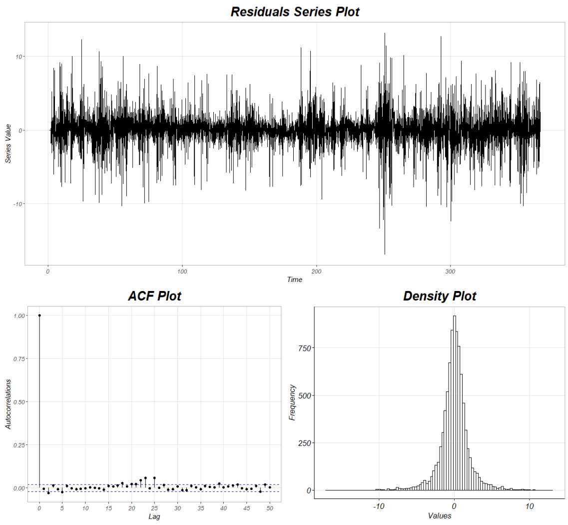
Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	22	14
i=2	$(1,0,1)(1,1,1)_{24}$	2	1
i=3	$(1,0,1)(0,1,2)_{24}$	1	0
i=4	$(1,0,1)(1,1,2)_{24}$	0	6
i=5	$(2,0,0)(0,1,1)_{24}$	43	35
i=6	$(2,0,0)(1,1,1)_{24}$	22	21
i=7	$(2,0,0)(0,1,2)_{24}$	21	20

Table 11: AIC and BIC comparisons for the models of Laranjeiro-Almada

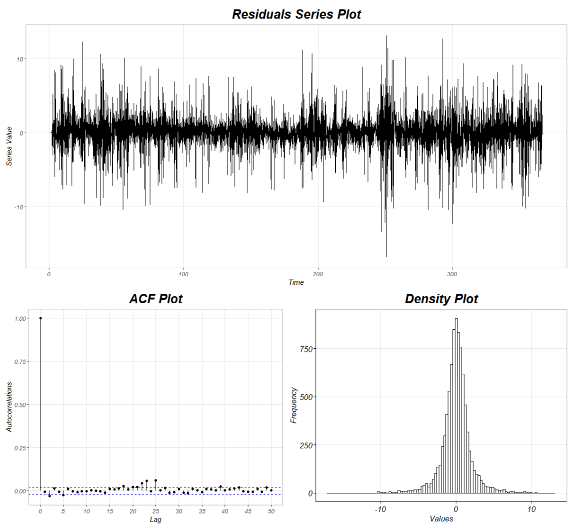
According the rules we defined above for comparing AIC and BIC values, we conclude that the second and third models ($i = 2$ and $i = 3$) are practically identical. The fourth model will also be considered, although it has a higher BIC value. All the other models are significantly worse, so we will discard them. Figure 19 presents the residuals series, ACF and distribution plots of all the three models selected.



(a) $(1,0,1)(1,1,1)_{24}$ model



(b) $(1,0,1)(0,1,2)_{24}$ model



(c) $(1,0,1)(1,1,2)_{24}$ model

Figure 19: Residuals of the final models for Laranjeiro-Almada

The plots from Figure 19 show that the residuals from the three models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 20, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 2$, $i = 3$ and $i = 4$ were, respectively: 0.0210, 0.0215 and 0.0241.

The three models were viable for forecasting, but we decided to go ahead with model i=3 since it's the one with the lowest BIC value, and an AIC only one unit higher the model with the lowest AIC value.

The results seen in Figure 20 and Table 12 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

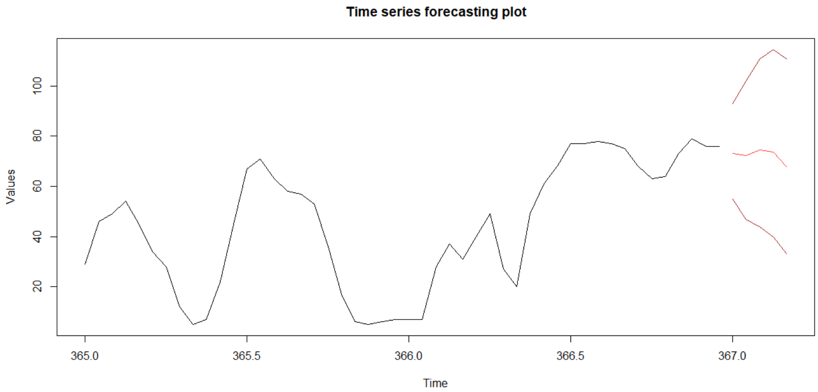


Figure 20: Time Series forecasting plot for Laranjeiro-Almada

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	73.16	55.16	92.98
2nd hour	72.40	46.74	102.01
3rd hour	74.60	43.93	110.91
4th hour	73.75	40.04	114.54
5th hour	67.87	33.18	111.01

Table 12: Time Series forecasting values for Laranjeiro-Almada

For a general expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 0.9049$, $\theta_1 = 0.1707$, $\Theta_1 = -0.8983$ and $\Theta_2 = -0.058$. The step-by-step approach for this particular time series can be found in Table 69.

3.2.6 Mem-Martins

Figure 21 shows the plot of the time series of Mem-Martins with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.9328. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

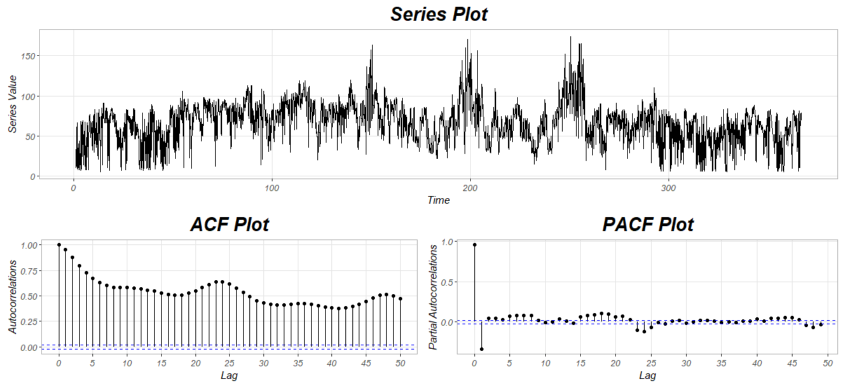


Figure 21: Mem-Martins time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 22).

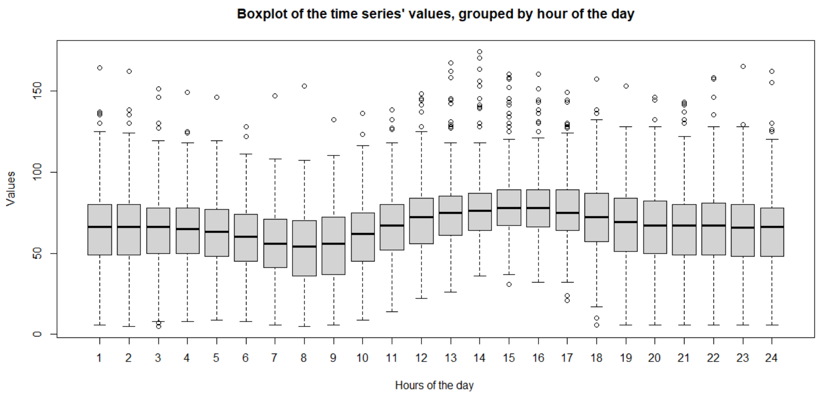


Figure 22: Mem-Martins time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 23) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

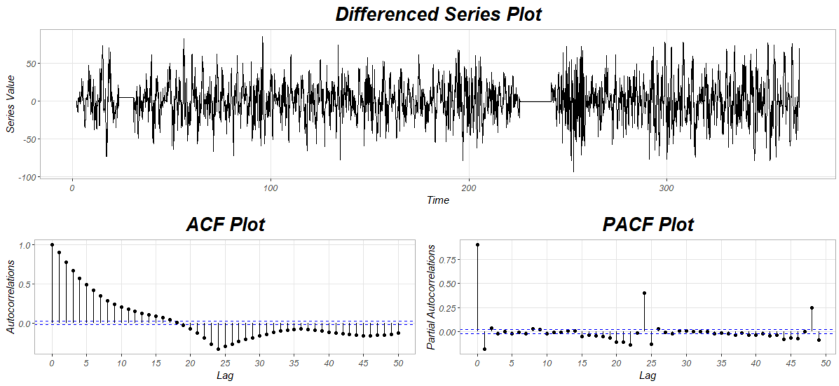


Figure 23: Differenced Time Series of Mem-Martins with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 13 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	52693.38	52721.69	4.24	6.38
$(1,0,1)(1,1,1)_{24}$	52680.24	52715.63	4.24	6.38
$(1,0,1)(0,1,2)_{24}$	52680.68	52716.07	4.24	6.38
$(1,0,1)(1,1,2)_{24}$	52684.33	52726.8	4.24	6.38
$(2,0,0)(0,1,1)_{24}$	52714.43	52742.74	4.24	6.39
$(2,0,0)(1,1,1)_{24}$	52699.62	52735.01	4.24	6.38
$(2,0,0)(0,1,2)_{24}$	52700.13	52735.01	4.24	6.39

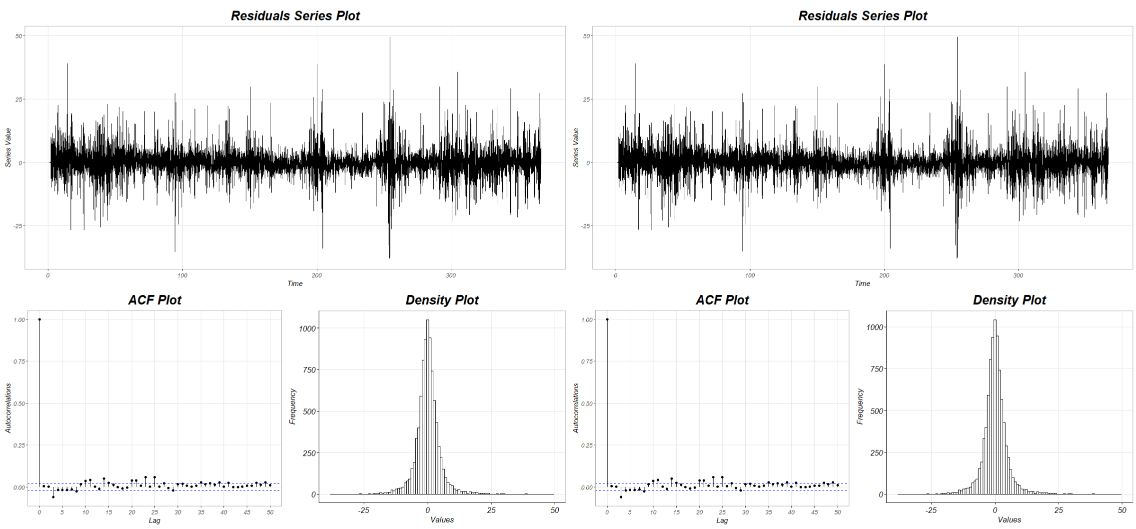
Table 13: Metrics of the best models for Mem-Martins

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the models $i = 2$ and $i = 3$ have the lowest AIC and BIC values (see Table 14).

	Models	Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	13	6
i=2	$(1,0,1)(1,1,1)_{24}$	0	0
i=3	$(1,0,1)(0,1,2)_{24}$	0	0
i=4	$(1,0,1)(1,1,2)_{24}$	4	11
i=5	$(2,0,0)(0,1,1)_{24}$	34	27
i=6	$(2,0,0)(1,1,1)_{24}$	19	19
i=7	$(2,0,0)(0,1,2)_{24}$	20	20

Table 14: AIC and BIC comparisons for the models of Mem-Martins

According the rules we defined above for comparing AIC and BIC values, we conclude that the second and third models ($i = 2$ and $i = 3$) are practically identical. The fourth model could also be considered, but its BIC value is more than 10 units larger than the BIC value from models $i = 2$ and $i = 3$. All the other models are significantly worse, so we will discard them. Figure 24 presents the residuals series, ACF and distribution plots of the two models selected.



(a) $(1,0,1)(1,1,1)_{24}$ model (b) $(1,0,1)(0,1,2)_{24}$ model

Figure 24: Residuals of the final models for Mem-Martins

The plots from Figure 24 show that the residuals from both models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 2, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 2$ and $i = 3$ were, respectively: 0.9539 and 0.9539.

Both were viable for forecasting, but we decided to go ahead with model $i = 2$ since it's the one with the lowest AIC and BIC values (at a decimal place level).

The results seen in Figure 25 and Table 15 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

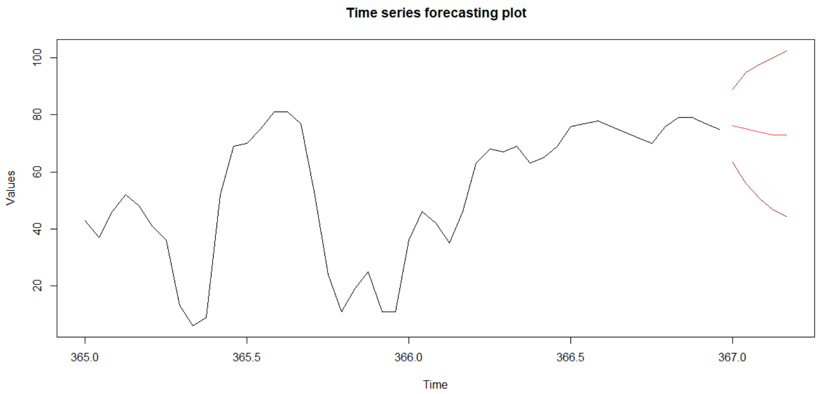


Figure 25: Time Series forecasting plot for Mem-Martins

	Predicted Value	**Lower Bound**	**Upper Bound**
1^{st} *hour*	76.17	63.46	89.03
2^{nd} *hour*	75.30	56.02	94.93
3^{rd} *hour*	73.86	50.45	97.80
4^{th} *hour*	73.05	46.66	100.12
5^{th} *hour*	73.05	44.35	102.54

Table 15: Time Series forecasting values for Mem-Martins

For a general expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 0.9228$, $\theta_1 = 0.2272$, $\Phi_1 = 0.0453$ and $\Theta_1 = -0.9488$.

3.2.7 Paio-Pires

Figure 26 shows the plot of the time series of Mem-Martins with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series.

The optimal lambda parameter for the Box-Cox transformation is: 0.6548. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

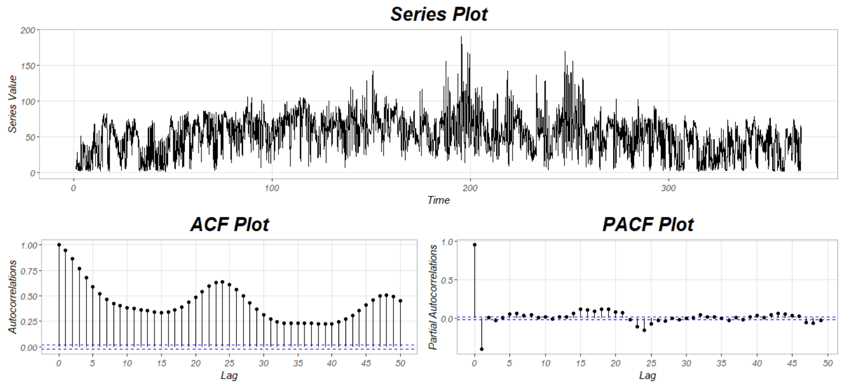


Figure 26: Paio-Pires time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 27).

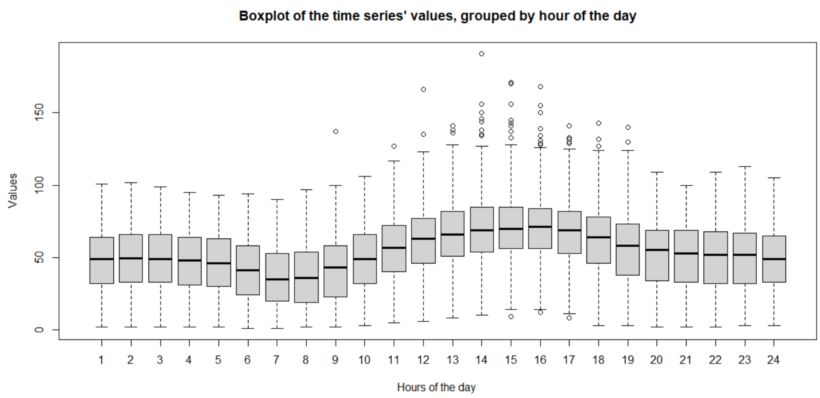


Figure 27: Paio-Pires time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 28) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

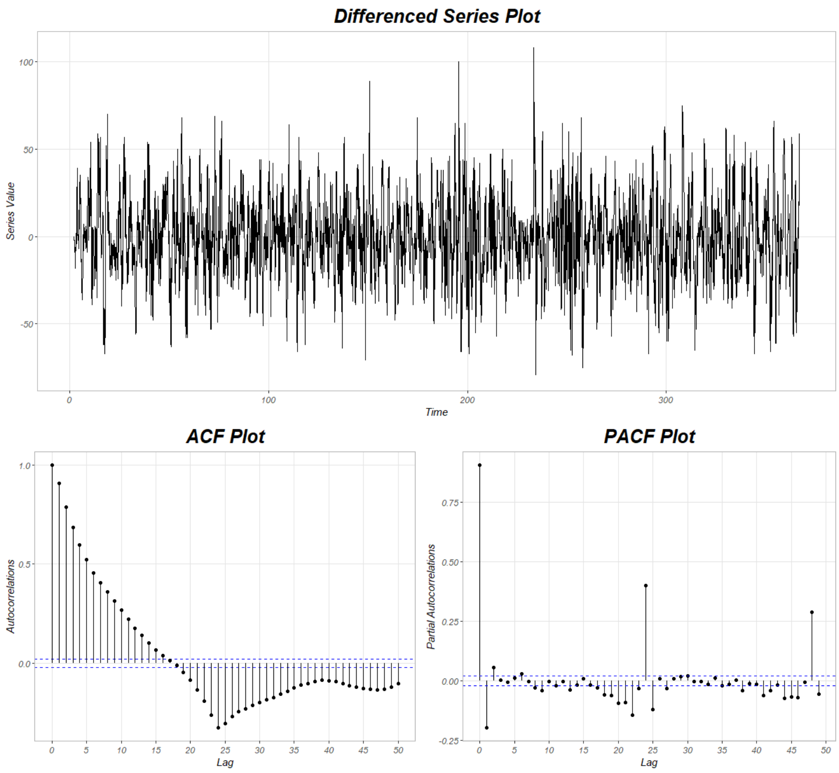


Figure 28: Differenced Time Series of Paio-Pires with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 16 have residuals whose ACF and PACF values are very low.

Models	AIC	BIC	MAE	RMSE
$(1,0,1)(0,1,1)_{24}$	35836.67	35864.99	4.81	6.77
$(1,0,1)(1,1,1)_{24}$	35813.02	35848.41	4.80	6.76
$(1,0,1)(0,1,2)_{24}$	35812.44	35847.83	4.80	6.76
$(1,0,1)(1,1,2)_{24}$	35813.75	35856.22	4.80	6.76
$(2,0,0)(0,1,1)_{24}$	35867.46	35895.78	4.81	6.78
$(2,0,0)(1,1,1)_{24}$	35844.05	35879.44	4.80	6.77
$(2,0,0)(0,1,2)_{24}$	35843.5	35879.44	4.80	6.77

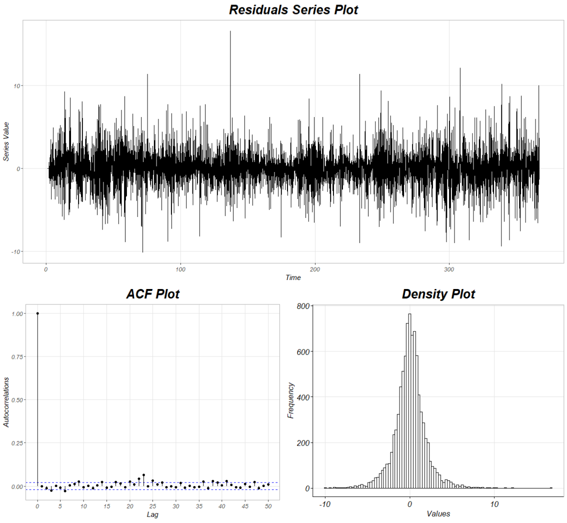
Table 16: Metrics of the best models for Paio-Pires

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the model i=3 has the minimum AIC and BIC values (see Table 17).

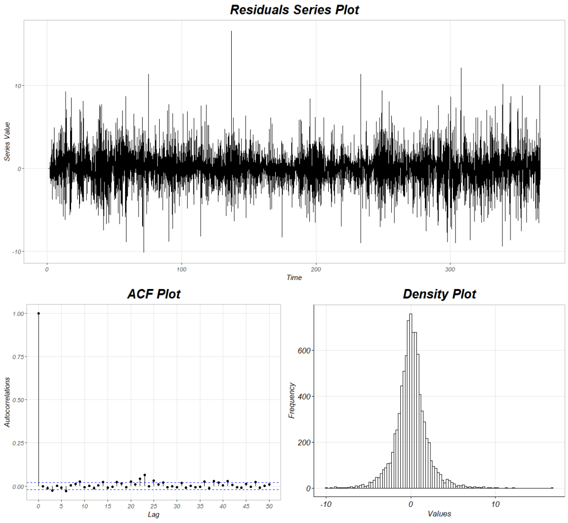
Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	24	17
i=2	$(1,0,1)(1,1,1)_{24}$	1	1
i=3	$(1,0,1)(0,1,2)_{24}$	0	0
i=4	$(1,0,1)(1,1,2)_{24}$	1	8
i=5	$(2,0,0)(0,1,1)_{24}$	55	48
i=6	$(2,0,0)(1,1,1)_{24}$	32	32
i=7	$(2,0,0)(0,1,2)_{24}$	31	31

Table 17: AIC and BIC comparisons for the models of Paio-Pires

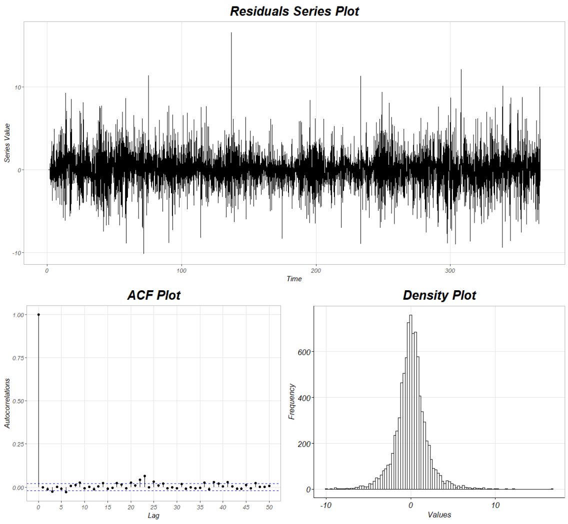
According the rules we defined above for comparing AIC and BIC values, we conclude that the second and third models ($i = 2$ and $i = 3$) are practically identical. The model $i = 4$ will also be considered, although it has a higher BIC value. All the other models are significantly worse, so we will discard them. Figure 29 presents the residuals series, ACF and distribution plots of all the three models selected.



(a) $(1,0,1)(1,1,1)_{24}$ model



(b) $(1,0,1)(0,1,2)_{24}$ model



(c) $(1,0,1)(1,1,2)_{24}$ model

Figure 29: Residuals of the final models for Paio-Pires

The plots from Figure 29 show that the residuals from the three models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 16, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 2$, $i = 3$ and $i = 4$ were, respectively: 0.0158, 0.0157 and 0.0157.

The three models were viable for forecasting, but we decided to go ahead with model $i = 3$ since it's the one with the lowest AIC and BIC values.

The results seen in Figure 30 and Table 18 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

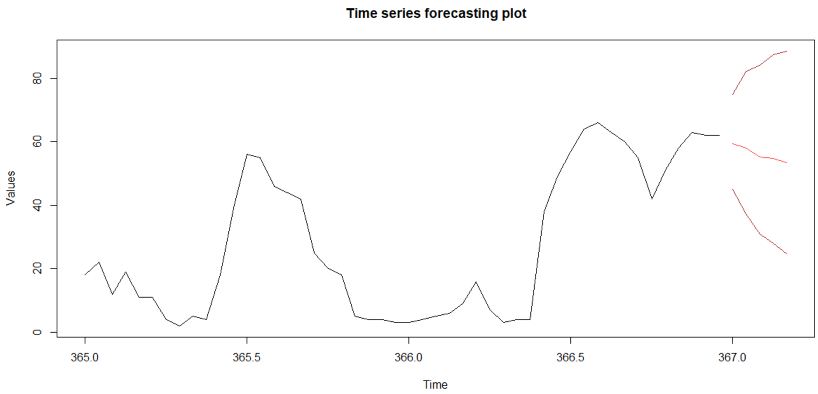


Figure 30: Time Series forecasting plot for Paio-Pires

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	59.41	45.12	74.99
2nd hour	58.17	37.26	82.08
3rd hour	55.29	30.92	84.14
4th hour	54.89	27.94	87.51
5th hour	53.35	24.84	88.60

Table 18: Time Series forecasting values for Paio-Pires

For a genereal expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 0.9082$, $\theta_1 = 0.2287$, $\Theta_1 = -0.8909$ and $\Theta_2 = -0.0556$

3.2.8 Restelo

Figure 31 shows the plot of the time series of Restelo with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series.

The optimal lambda parameter for the Box-Cox transformation is 0.6743. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

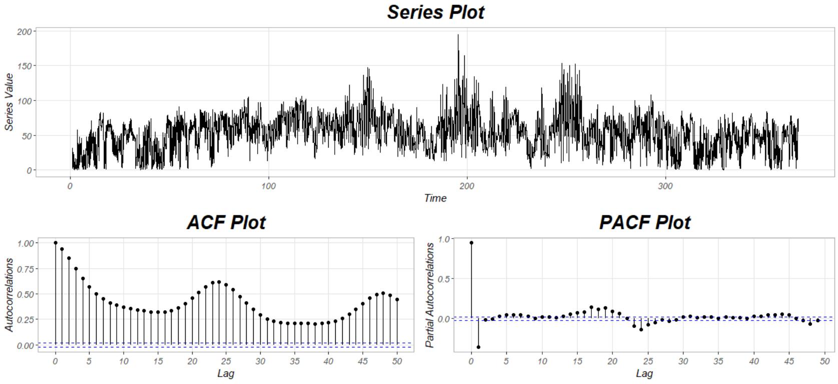


Figure 31: Restelo time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 32).

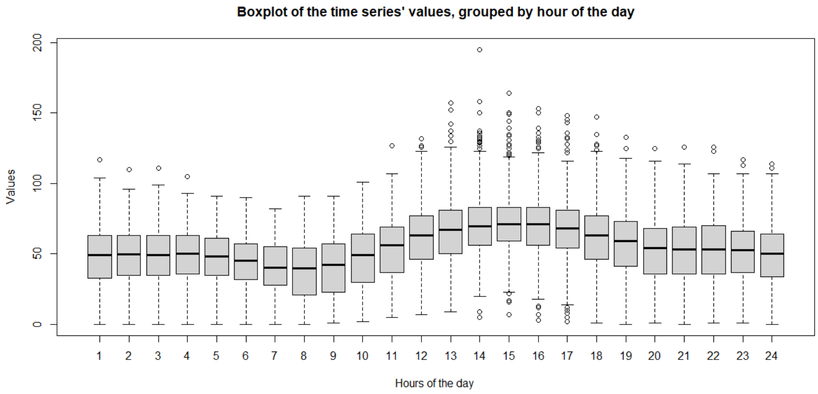


Figure 32: Restelo time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 33) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

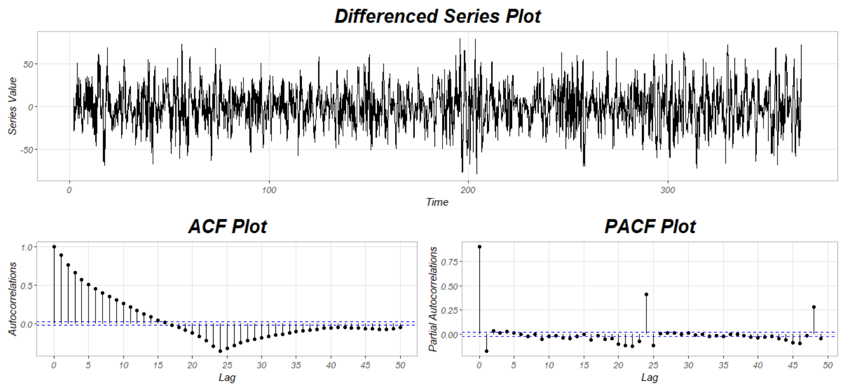


Figure 33: Differenced Time Series of Restelo with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 19 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	38012.33	38040.64	5.01	7.04
$(1,0,1)(1,1,1)_{24}$	37996.12	38031.51	5.01	7.04
$(1,0,1)(0,1,2)_{24}$	37996.3	38031.69	5.01	7.04
$(1,0,1)(1,1,2)_{24}$	37998.1	38040.57	5.01	7.04
$(2,0,0)(0,1,1)_{24}$	38031.24	38059.55	5.01	7.05
$(2,0,0)(1,1,1)_{24}$	38015.22	38050.61	5.01	7.04
$(2,0,0)(0,1,2)_{24}$	38015.37	38050.61	5.01	7.04

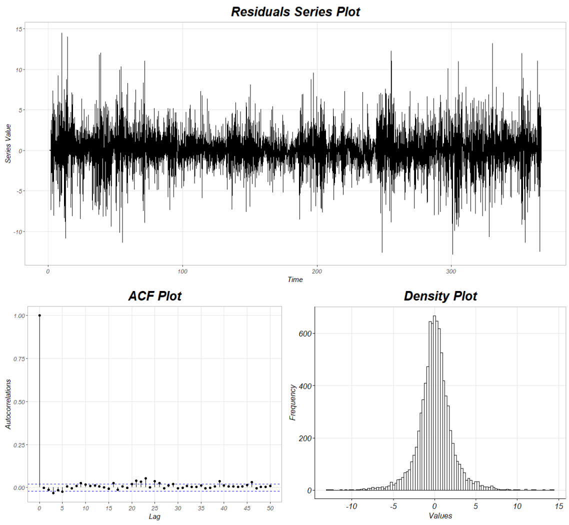
Table 19: Metrics of the best models for Restelo

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the models $i = 2$ and $i = 3$ have the minimum AIC and BIC values (see Table 20).

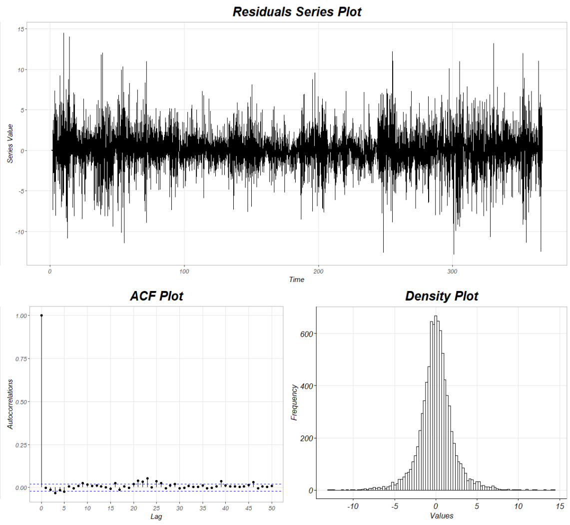
Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	16	9
i=2	$(1,0,1)(1,1,1)_{24}$	0	0
i=3	$(1,0,1)(0,1,2)_{24}$	0	0
i=4	$(1,0,1)(1,1,2)_{24}$	2	9
i=5	$(2,0,0)(0,1,1)_{24}$	35	28
i=6	$(2,0,0)(1,1,1)_{24}$	19	19
i=7	$(2,0,0)(0,1,2)_{24}$	19	19

Table 20: AIC and BIC comparisons for the models of Restelo

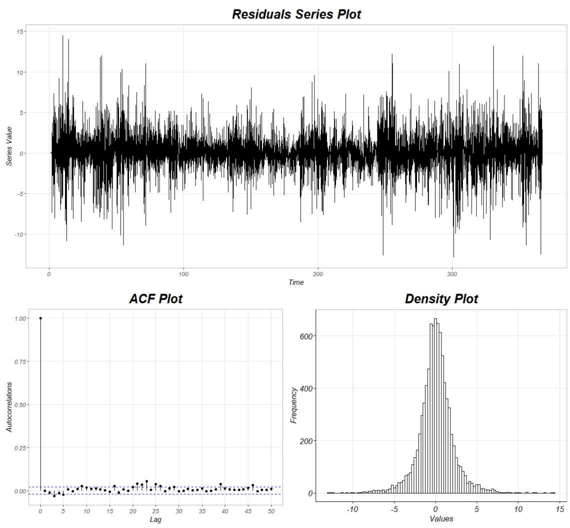
According the rules we defined above for comparing AIC and BIC values, only model $i = 4$ seems to be another possible option, while all the other models are significantly worse, so we will discard them. Figure 34 presents the residuals series, ACF and distribution plots of the two best models selected.



(a) $(1,0,1)(1,1,1)_{24}$ model



(b) $(1,0,1)(0,1,2)_{24}$ model



(c) $(1,0,1)(1,1,2)_{24}$ model

Figure 34: Residuals of the final models for Restelo

The plots from Figure 34 show that the residuals from both models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 19, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 2$, $i = 3$ and $i = 4$ were equal to, respectively, 0.0133, 0.0133 and 0.0152.

All the three models were viable for forecasting, but we decided to go ahead with model $i = 2$ since it's the one with the lowest AIC and BIC values (at a decimal place level).

The results seen in Figure 35 and Table 21 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

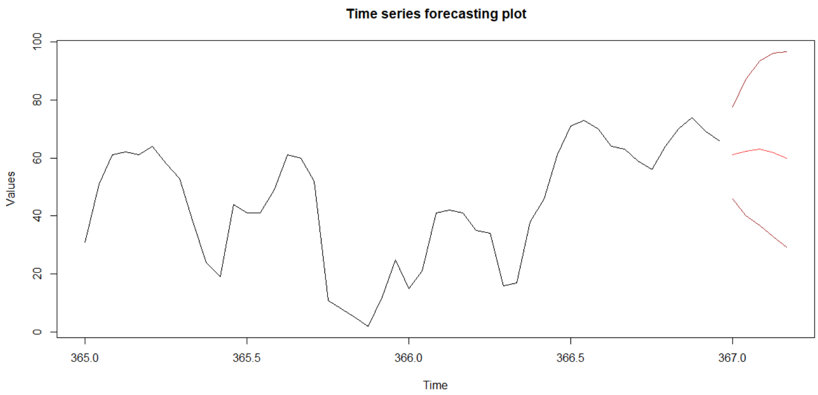


Figure 35: Time Series forecasting plot for Restelo

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	61.04	45.96	77.46
2nd hour	62.25	40.16	87.26
3rd hour	62.92	36.64	93.44
4th hour	61.82	32.93	96.06
5th hour	59.79	29.29	96.61

Table 21: Time Series forecasting values for Restelo

For a genereal expression of a SARIMA-type model found in [6], the values of the coefficients of this final model are: $\phi_1 = 0.9038$, $\theta_1 = 0.1976$, $\Phi_1 = 0.0490$ and $\Theta 1 = -0.9568$.

3.2.9 Sobreiras-Porto

Figure 36 shows the plot of the time series of Estarreja with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.7400. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

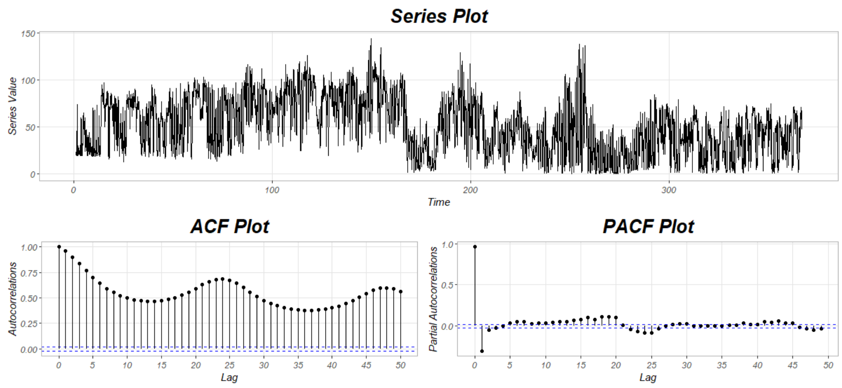


Figure 36: Sobreiras-Porto time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 37).

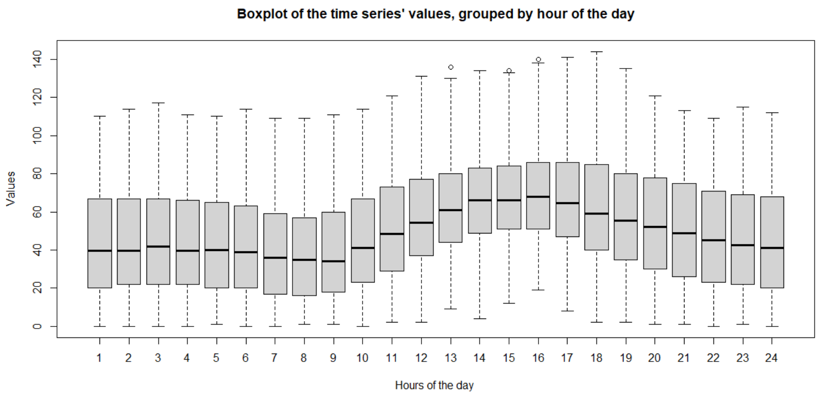


Figure 37: Sobreiras-Porto time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 38) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

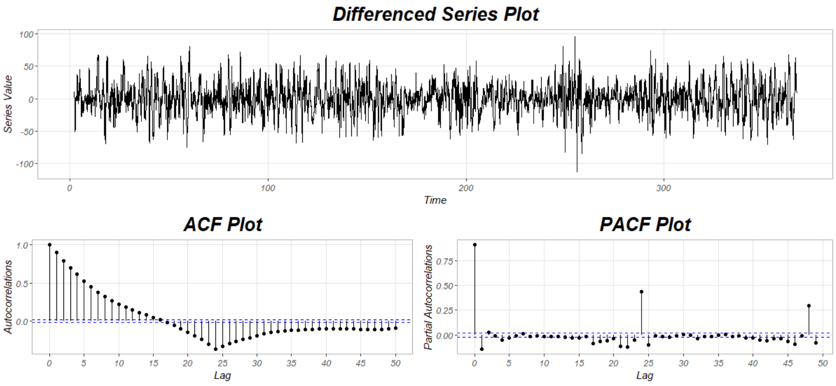


Figure 38: Differenced Time Series of Sobreiras-Porto with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 22 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	42080.77	42109.08	5.06	7.01
$(1,0,1)(1,1,1)_{24}$	42079.61	42115.00	5.06	7.01
$(1,0,1)(0,1,2)_{24}$	42079.64	42115.03	5.06	7.01
$(1,0,1)(1,1,2)_{24}$	42081.92	42124.39	5.06	7.01
$(2,0,0)(0,1,1)_{24}$	42087.73	42116.04	5.06	7.01
$(2,0,0)(1,1,1)_{24}$	42086.91	42122.30	5.06	7.01
$(2,0,0)(0,1,2)_{24}$	42086.93	42122.30	5.06	7.01

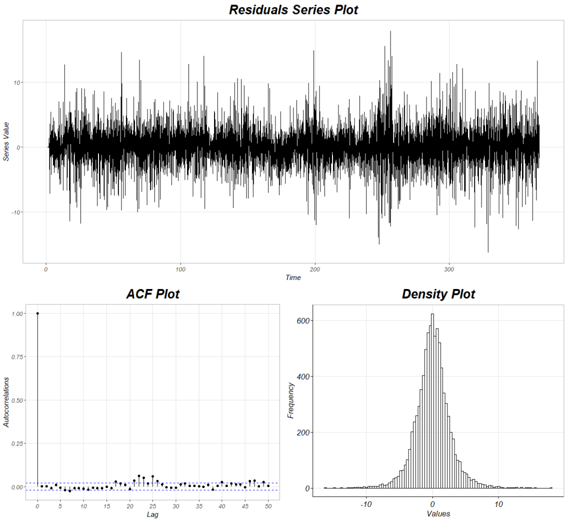
Table 22: Metrics of the best models for Sobreiras-Porto

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the models $i = 2$ and $i = 3$ have the minimum AIC and model $i = 1$ has the lowest BIC (see Table 23).

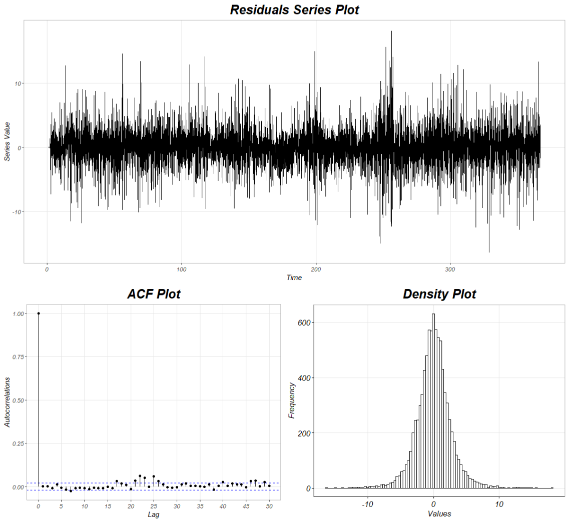
Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	1	0
i=2	$(1,0,1)(1,1,1)_{24}$	0	6
i=3	$(1,0,1)(0,1,2)_{24}$	0	6
i=4	$(1,0,1)(1,1,2)_{24}$	2	15
i=5	$(2,0,0)(0,1,1)_{24}$	8	7
i=6	$(2,0,0)(1,1,1)_{24}$	7	13
i=7	$(2,0,0)(0,1,2)_{24}$	7	13

Table 23: AIC and BIC comparisons for the models of Sobreiras-Porto

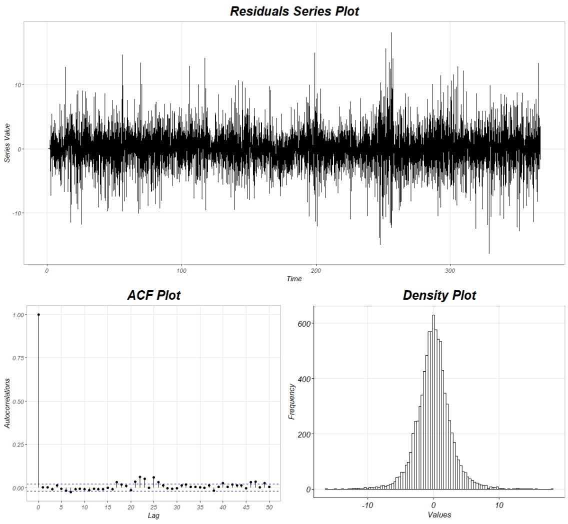
According the rules we defined above for comparing AIC and BIC values, we can see that the model $i = 1$ seems to be the best one. Models $i = 2$ and $i = 3$ have a BIC value slightly higher, but still a possible option. All the other models are significantly worse, so we will discard them. Figure 39 presents the residuals series, ACF and distribution plots of the theree best models selected.



(a) $(1,0,1)(0,1,1)_{24}$ model



(b) $(1,0,1)(1,1,1)_{24}$ model



(c) $(1,0,1)(0,1,2)_{24}$ model

Figure 39: Residuals of the final models for Sobreiras-Porto

The plots from Figure 39 show that the residuals from the three models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 20, we don't reject the null hypothesis of the residuals being uncorrelated. The p-values for models $i = 1$, $i = 2$ and $i = 3$ were, respectively: 0.0308, 0.0318 and 0.0317.

The three models were viable for forecasting, but we decided to go ahead with model $i = 1$ since it's the one with the minimizes AIC and BIC values at the same time.

The results seen in Figure 40 and Table 24 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

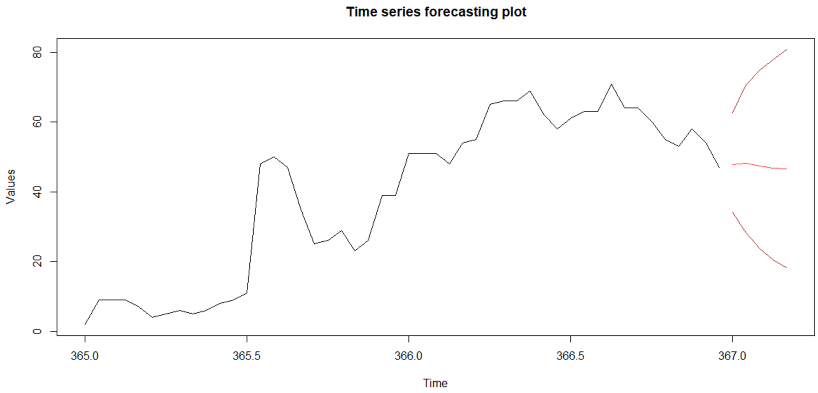


Figure 40: Time Series forecasting plot for Sobreiras-Porto

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	47.90	34.21	62.70
2nd hour	48.25	28.26	70.71
3rd hour	47.46	23.71	74.91
4th hour	46.75	20.38	77.88
5th hour	46.69	18.28	80.79

Table 24: Time Series forecasting values for Sobreiras-Porto

For a genereal expression of a SARIMA-type model found in [6], the values of the coefficients of this final model are: $\phi_1 = 0.9276$, $\theta_1 = 0.1783$ and $\Theta 1 = -0.9568$.

3.2.10 V.N. Telha - Maia

Figure 41 shows the plot of the time series of V.N. Telha - Maia with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.6455. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

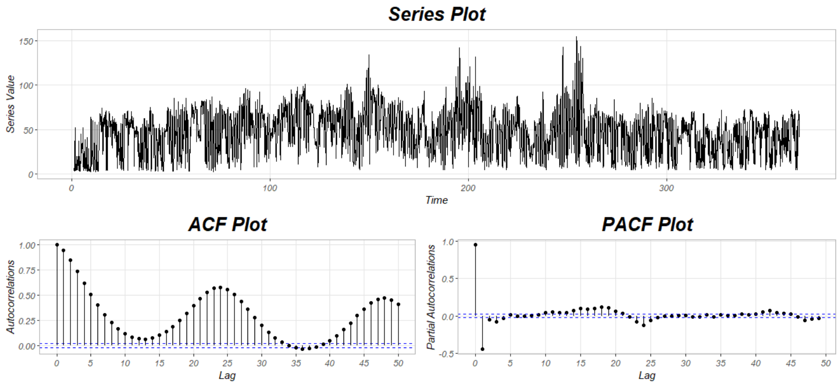


Figure 41: V.N. Telha - Maia time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 42).

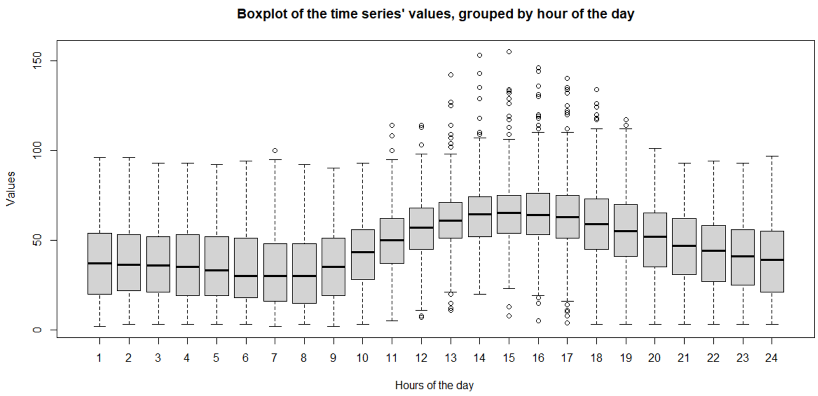


Figure 42: V.N. Telha - Maia time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 43) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

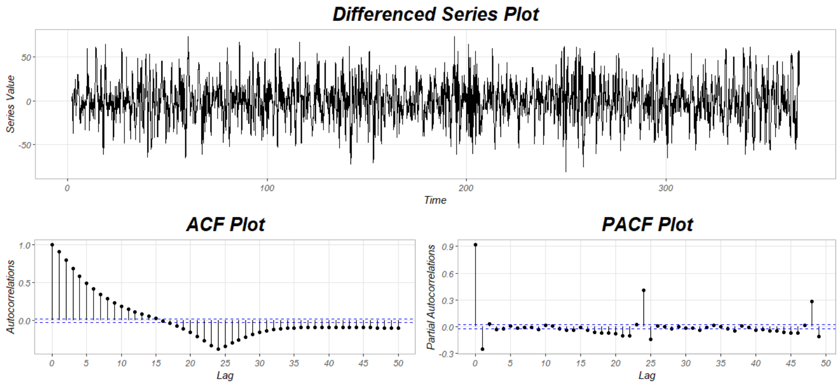


Figure 43: Differenced Time Series of V.N. Telha - Maia with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 25 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	34714.89	34743.2	4.53	6.31
$(1,0,1)(1,1,1)_{24}$	34699.96	34735.35	4.53	6.30
$(1,0,1)(0,1,2)_{24}$	34699.43	34734.82	4.53	6.30
$(1,0,1)(1,1,2)_{24}$	34699.27	34741.74	4.53	6.30
$(2,0,0)(0,1,1)_{24}$	34715.76	34744.07	4.52	6.31
$(2,0,0)(1,1,1)_{24}$	34700.7	34736.09	4.52	6.30
$(2,0,0)(0,1,2)_{24}$	34700.19	34736.09	4.52	6.30

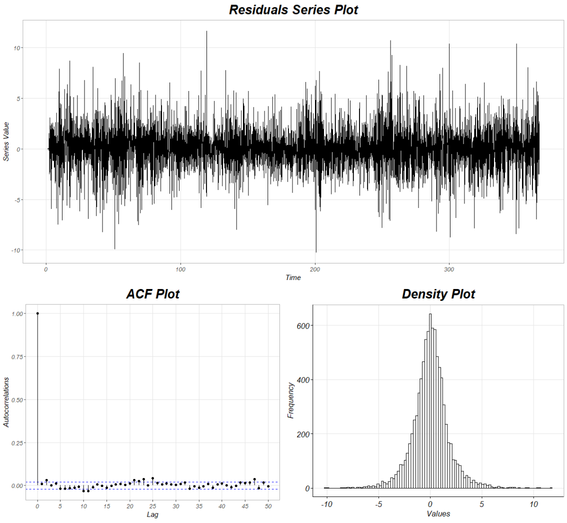
Table 25: Metrics of the best models for V.N. Telha - Maia

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the model i=3 has the minimum AIC and BIC (see Table 26).

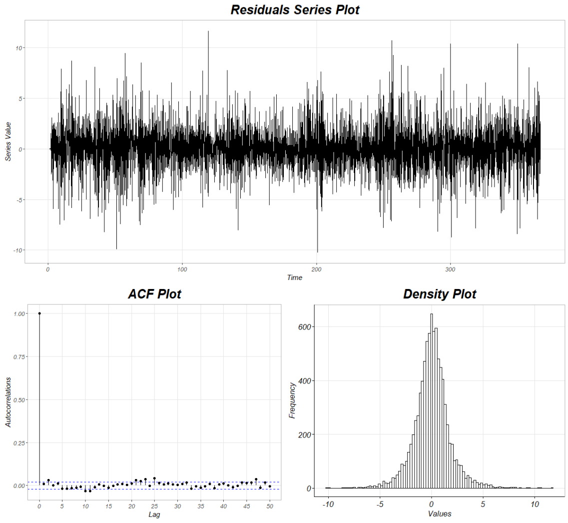
Models		Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	16	8
i=2	$(1,0,1)(1,1,1)_{24}$	1	1
i=3	$(1,0,1)(1,1,1)_{24}$	0	0
i=4	$(1,0,1)(0,1,2)_{24}$	0	7
i=5	$(2,0,0)(0,1,1)_{24}$	16	9
i=6	$(2,0,0)(1,1,1)_{24}$	1	1
i=7	$(2,0,0)(0,1,2)_{24}$	1	1

Table 26: AIC and BIC comparisons for the models of V.N. Telha - Maia

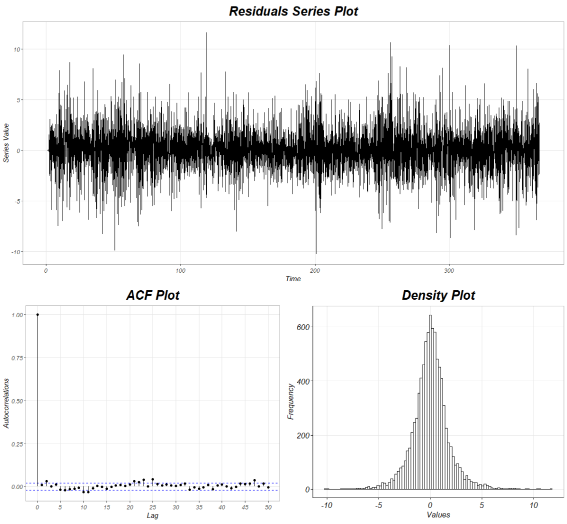
According the rules we defined above for comparing AIC and BIC values, we conclude that the models $i = 2$, $i = 3$, $i = 6$ and $i = 7$ are practically identical. The model $i = 4$ will also be considered, although it has a higher BIC value. All the other models are significantly worse, so we will discard them. Figure 44 presents the residuals series, ACF and distribution plots of all the these models selected.



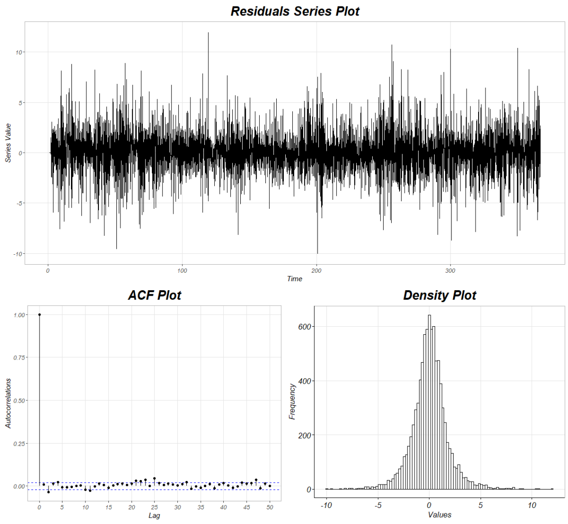
(a) $(1,0,1)(1,1,1)_{24}$ model



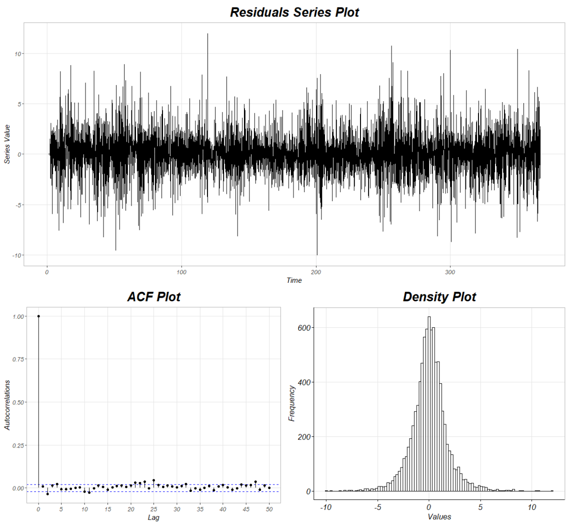
(b) $(1,0,1)(0,1,2)_{24}$ model



(c) $(1,0,1)(1,1,2)_{24}$ model



(d) $(2,0,0)(1,1,1)_{24}$ model



(e) $(2,0,0)(0,1,2)_{24}$ model

Figure 44: Residuals of the final models for V.N. Telha - Maia

The plots from Figure 44 show that the residuals from the five models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that up to lag 6, we don't reject the null hypothesis of the residuals being uncorrelated for models $i = 2$, $i = 3$ and $i = 4$. Apparently, for the remaining two models the test is rejected, even for smaller lags. Correlations may still be meaningful in these two models, so they will be discarded. The p-values for models $i = 2$, $i = 3$ and $i = 4$ were, respectively: 0.0110, 0.0110 and 0.0101.

These three last models were viable for forecasting, but we decided to go ahead with model $i = 3$ since it's the one with the lowest AIC and BIC values.

The results seen in Figure 45 and Table 27 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

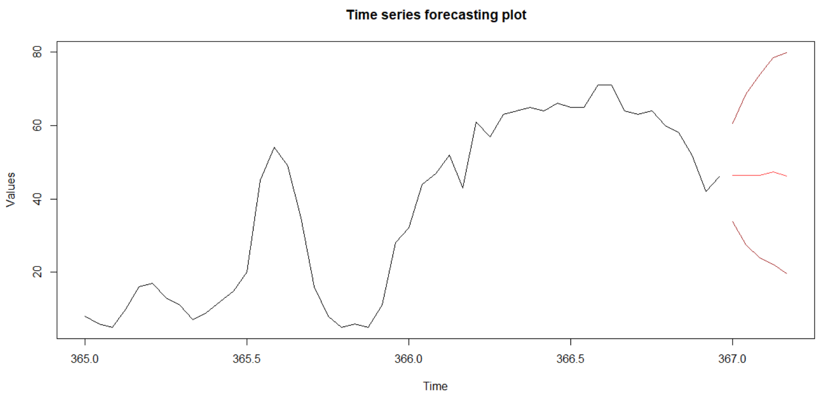


Figure 45: Time Series forecasting plot for V.N. Telha - Maia

	Predicted Value	**Lower Bound**	**Upper Bound**
1st hour	46.50	33.84	60.51
2nd hour	46.39	27.46	68.58
3rd hour	46.49	23.97	73.84
4th hour	47.29	22.19	78.45
5th hour	46.18	19.64	79.85

Table 27: Time Series forecasting values for V.N. Telha - Maia

For a genereal expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 0.9014$, $\theta_1 = 0.2780$, $\Phi_1 = -0.9047$ and $\Theta_1 = -0.0454$

3.2.11 Ilhavo

Figure 46 shows the plot of the time series of Ilhavo with its ACF and PACF. In particular, the ACF shows a very slow decay, typical of a non-stationary time-series

The optimal lambda parameter for the Box-Cox transformation is: 0.6453. We won't plot the series with this transformation applied for the sake of simplicity, but this transformation was applied to the initial series to stabilize the variance.

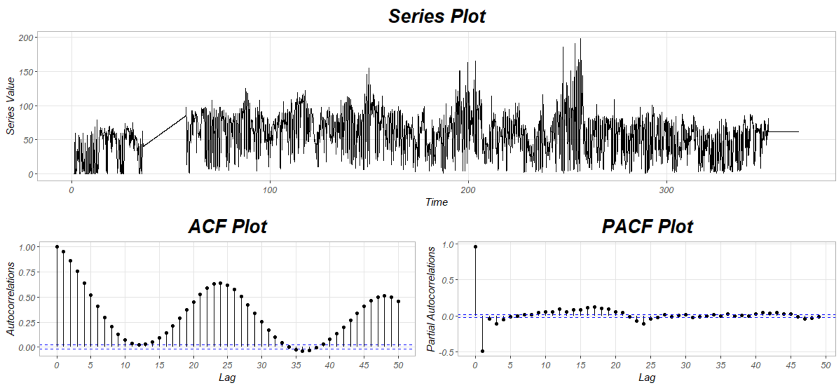


Figure 46: Ilhavo time series, ACF and PACF

Grouping the observations per hour of the day, we can see a pattern that repeats every day (see Fig. 47).

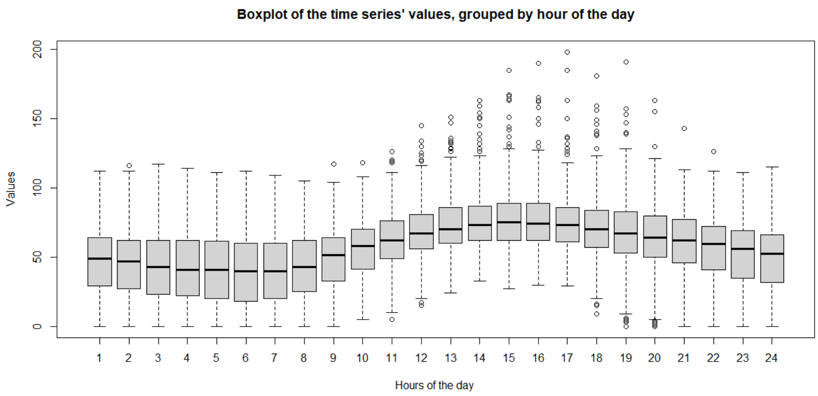


Figure 47: Ilhavo time series observations grouped by hour of the day

Applying the seasonal difference operator allows to obtain a stationary Time-Series. Now the ACF decays exponentially (see Fig. 48) and now we are in conditions of fitting the autoregressive and moving average terms of a SARIMA-type model.

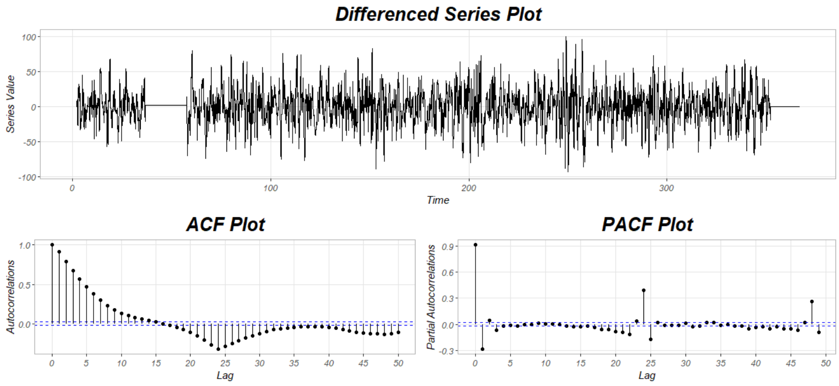


Figure 48: Differenced Time Series of Ilhavo with ACF and PACF

Now we apply our step-by-step approach (as defined above) in order to find models that capture all the important correlations in the ACF and PACF. The models presented in Table 28 have residuals whose ACF and PACF values are very low.

Models	**AIC**	**BIC**	**MAE**	**RMSE**
$(1,0,1)(0,1,1)_{24}$	34714.89	34743.2	4.53	6.31
$(1,0,1)(1,1,1)_{24}$	34699.96	34735.35	4.53	6.30
$(1,0,1)(0,1,2)_{24}$	34699.43	34734.82	4.53	6.30
$(1,0,1)(1,1,2)_{24}$	34699.27	34741.74	4.53	6.30
$(2,0,0)(0,1,1)_{24}$	34715.76	34744.07	4.52	6.31
$(2,0,0)(1,1,1)_{24}$	34700.7	34736.09	4.52	6.30
$(2,0,0)(0,1,2)_{24}$	34700.19	34736.09	4.52	6.30

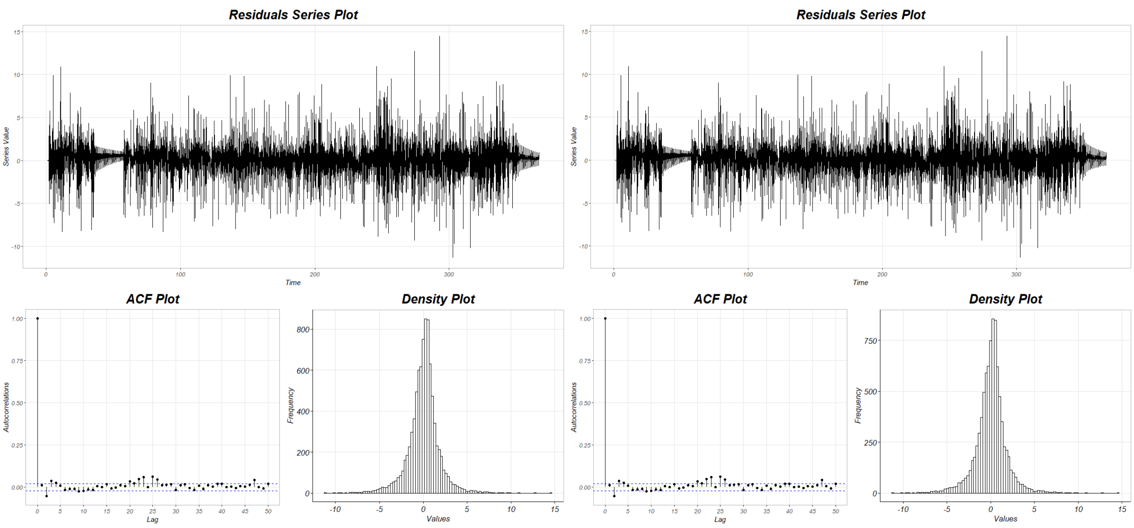
Table 28: Metrics of the best models for Ilhavo

The differences we observe in the MAE and RMSE are negligible, since the time series measurements are integer numbers, and the error differences appear in the second decimal place. Let's compare the AIC and BIC values, taking into consideration that the models $i = 6$ and $i = 7$ have the minimum AIC and BIC (see Table 29).

	Models	Δ_i^{AIC}	Δ_i^{BIC}
i=1	$(1,0,1)(0,1,1)_{24}$	73	65
i=2	$(1,0,1)(1,1,1)_{24}$	9	9
i=3	$(1,0,1)(1,1,1)_{24}$	9	9
i=4	$(1,0,1)(0,1,2)_{24}$	11	18
i=5	$(2,0,0)(0,1,1)_{24}$	57	50
i=6	$(2,0,0)(1,1,1)_{24}$	0	0
i=7	$(2,0,0)(0,1,2)_{24}$	0	0

Table 29: AIC and BIC comparisons for the models of Ilhavo

According the rules we defined above for comparing AIC and BIC values, the models $i = 6$ and $i = 7$ are the only viable options. All the other models are significantly worse, so we will discard them. Figure 49 presents the residuals series, ACF and distribution plots of the selected models.



(a) $(2,0,0)(1,1,1)_{24}$ model (b) $(2,0,0)(0,1,2)_{24}$ model

Figure 49: Residuals of the final models for Ilhavo

The plots from Figure 49 show that the residuals from the two models are very low correlated, with approximately normal distributions with zero mean. In order to confirm the non-correlation of the residuals we conducted the Ljung-box test (with a 1% significance level) and found that these models still have significant correlations. The p-values for any lag superior to 1 had p-values extremely low, we rejected the hypothesis of the residuals being uncorrelated, for both odels

Even though we should only forecast with models that have uncorrelated residuals, we will still forecast for the best model we have. There are no differences between $i = 6$ and $i = 7$, so we just picked randomly i=6.

The results seen in Figure 50 and Table 30 show, respectively, the plot and values of the point forecasts, with lower and upper bounds, for a 95% confidence interval.

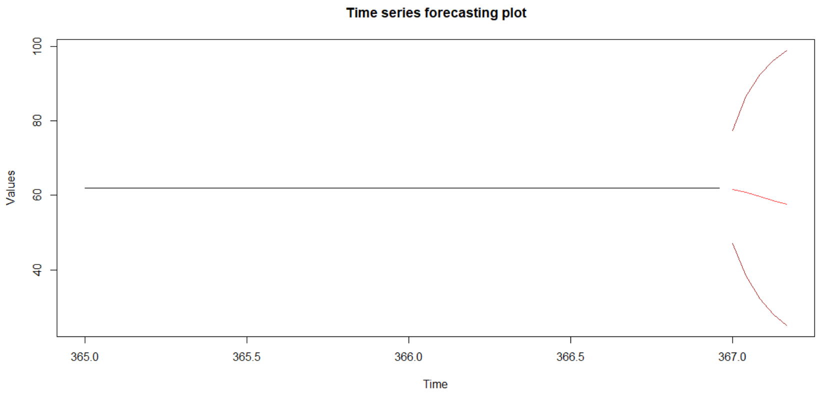


Figure 50: Time Series forecasting plot for Ilhavo

	Predicted Value	**Lower Bound**	**Upper Bound**
1^{st} hour	61.66	47.15	77.50
2^{nd} hour	60.79	38.40	86.59
3^{rd} hour	59.65	32.29	92.49
4^{th} hour	58.53	28.00	96.27
5^{th} hour	57.70	25.08	98.91

Table 30: Time Series forecasting values for Ilhavo

For a general expression of a SARIMA-type model found in [6], the value of the coefficients of this final model are: $\phi_1 = 1.2460$, $\phi_2 = -0.3295$, $\Phi_1 = 0.0895$ and $\Theta_1 = -0.9502$

3.3 Conclusions and Future work

We conclude our analysis for the hourly-ground-levels of ozone series with satisfactory results. We were able to find models that have low errors and information criteria values on the fitting set, while having a small set of parameters. We were then able to generate forecasts for five periods into the future. An important aspect to mention is that some models were very similar to each other. For instance, the SARIMA-type model $(1, 0, 1)(1, 1, 1)_{24}$ was used 5 times, and the coefficients did not vary a lot between different series. This is likely a consequence of the similarity of the ozone level behaviour between different regions in Portugal, something that was visualized, for example, in the daily seasonal pattern. For future work, one suggestion we leave for the researchers is to evaluate errors in one additional set of data not used for training. This can be particularly important if the goal such project is more oriented for finding the best models for prediction purposes.

4 Time Series for Financial Data

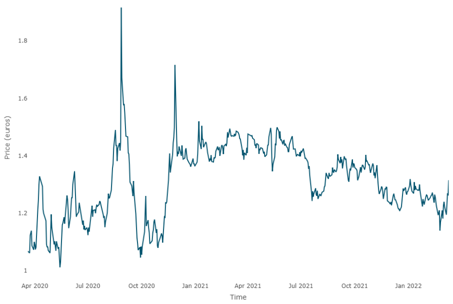
In the second assignement it was asked to study five daily share related-prices from five different companies (EDP, GALP, MOTAENGIL, NOS and NOVABASE) listed on Euronext Lisbon, and to make a financial study about them.

4.1 Data Analysis

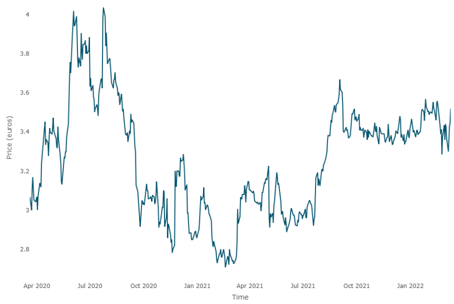
It should be noticed that the data related with NOVABASE had a missing value relative to the day 18-01-2022. In order to complete the dataset, and since the original database did not had the value aswell, we applied interpolation (average). Figure 51 shows the closing values by day for all of this companies:

(a) Closing Values - EDP

(b) Closing Values - GALP



(c) Closing Values - MOTAENGIL



(d) Closing Values - NOS

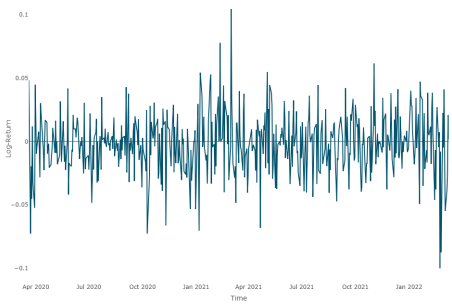
(e) Closing Values - NOVABASE

Figure 51: Closing Values by day.

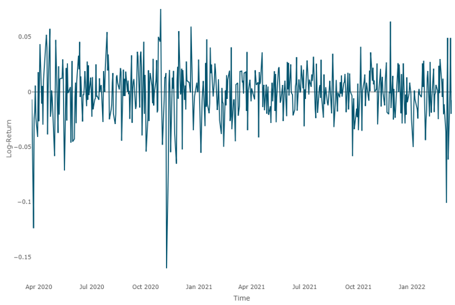
As one would expect, the prices of a stock tend to be non-stationary since economic growth and inflation take effect during time. Therefore, when one studies financial data, asset equity **returns** are typically the main variable of interest, rather than prices. Their **easier interpretability** and **statistical properties** (such as stationarity) are two of the reasons that makes us want to work with them. The first step of this assignment consisted in computing the **log-returns** associated to the daily closing values, given by:

$$X_t = \log(P_t) - \log(P_{t-1}) = \log\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right)$$

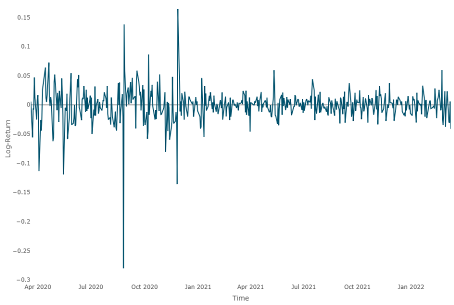
where P_t is the closing price at time index t. The distribution of the log-returns through time for each company are:



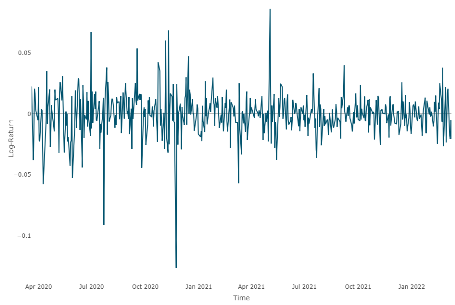
(a) Log-Returns - EDP



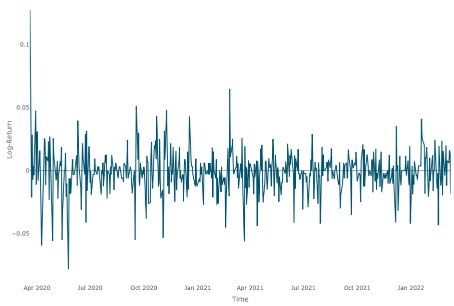
(b) Log-Returns - GALP



(c) Log-Returns - MOTAENGIL



(d) Log-Returns - NOS



(e) Log-Returns - NOVABASE

Figure 52: Log-Returns by day.

Values of the **sample mean close to zero and low variance** (in the order of 10^{-4} or less) are some of the stylized facts on log returns. These metrics were computed and Table 31 resumes the values obtained, supporting our initial beliefs.

	EDP	GALP	MOTAENGIL	NOS	NOVABASE
Sample Mean	-0.00175	-0.00061	-0.00041	-0.00027	-0.00086
Sample Variance	0.00056	0.00059	0.00079	0.00031	0.00031
Kurtosis	4.565	8.162	26.253	11.097	10.002

Table 31: Sample mean, variance and *kurtosis* for the 5 companies log-returns

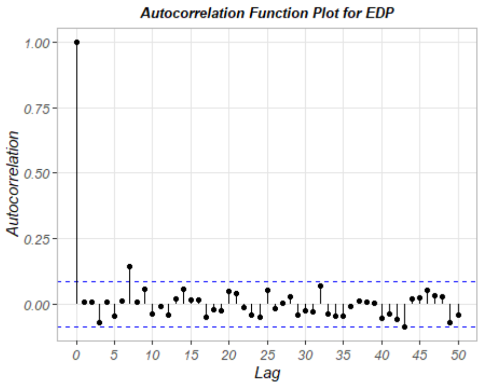
It is also expected that the returns X_t **do not show a great deal of autocorrelation**, and therefore, be negligible at all lags. On the other hand, it is anticipated that **squared returns or returns in absolute value show some kind of correlation (non-linearity)**. Figures 53, 54 and 55 plot the sample ACF for this type of time series values.

The inspection of both the daily log-returns plot and the autocorrelograms suggests that returns appear to have weak or no serial dependence, since the majority of the values fall in the 95% confidence intervals. Absolute and squared returns hint us that there is weak to moderated dependence for many lags (for all time series), as one would expect, but this is not as clear as one would want it to be.

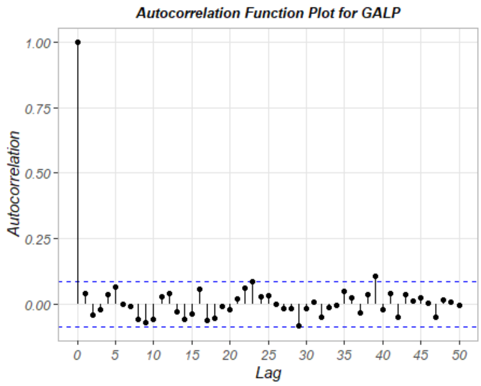
Due to the market volatility, it is more likely that an investment will move beyond the usual standard deviations, that is, it will happen an investment which may be larger than otherwise anticipated. Therefore, financial time series present **fat-tailed distribution of returns** that the ones of Gaussian distribution. This can be assessed with the aid of the QQ-plot, where it is expected that the normality comparison starts to fail for extreme values, and the estimation of the *kurtosis*, where it is expected that it exceeds the value of 3, i.e, that the distribution is leptokurtic. As one can observe in figure 56 and in table 31, the *kurtosis* values are way higher than 3 all of the companies data and the qqplot starts to fail for extreme values, evidencing the presence of fat tails, uncharacteristic of Gaussianity.

Not only that, these **exceedances tend to occur in clusters**, that is, large returns (in absolute value) tend to be followed by large returns (in absolute value), and vice versa. Furthermore, there is the presence of **leverage effects**, i.e, negative returns tend to increase volatility by a larger amount than positive returns. Figures 51 and 52 clearly evidence this kind of behaviour, since decreasing price implies higher volatility.

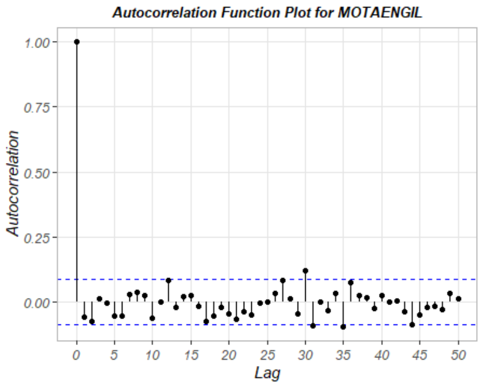
The strong evidence of dependence in absolute and square returns suggest that the scale of returns changes in time, i.e, the variance of the process is time varying, alongside other untypical attributes. To capture all of them, we need to use appropriate time series models that are able to model this type of data.



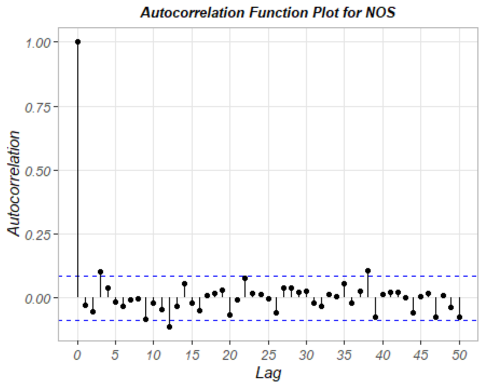
(a) ACF - EDP



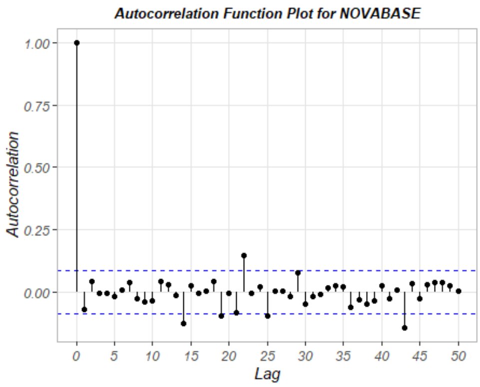
(b) ACF - GALP



(c) ACF - MOTAENGIL

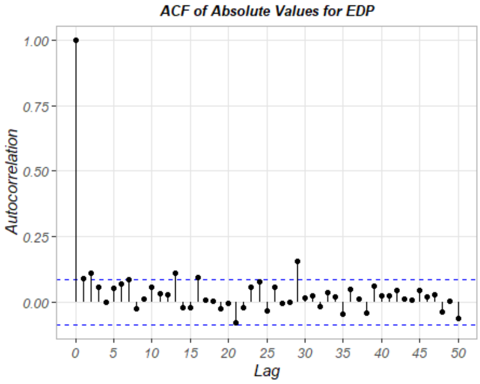


(d) ACF - NOS

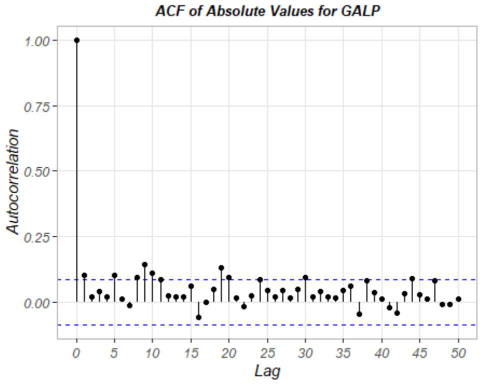


(e) ACF - NOVABASE

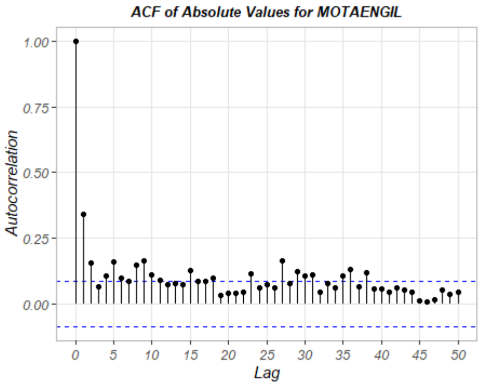
Figure 53: Autocorrelation Function Plot for the Values.



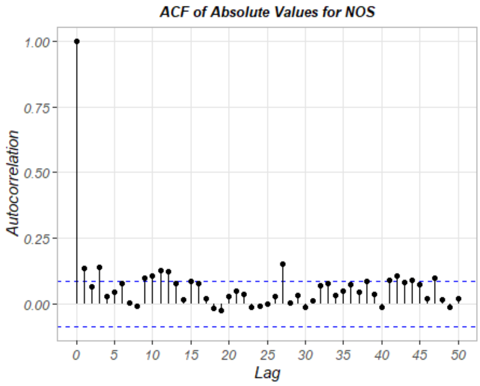
(a) ACF abs - EDP



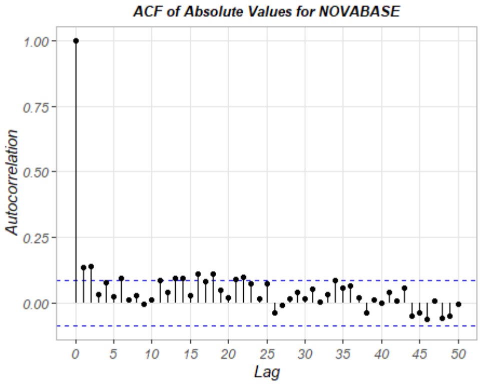
(b) ACF abs - GALP



(c) ACF abs - MOTAENGIL

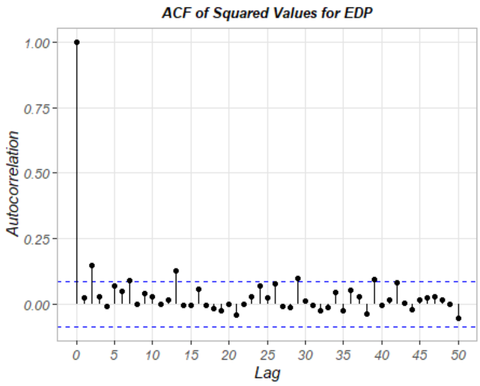


(d) ACF abs - NOS

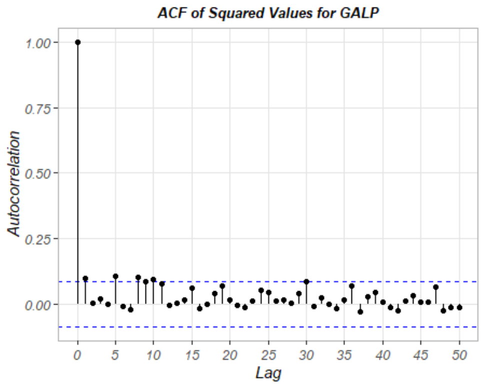


(e) ACF abs - NOVABASE

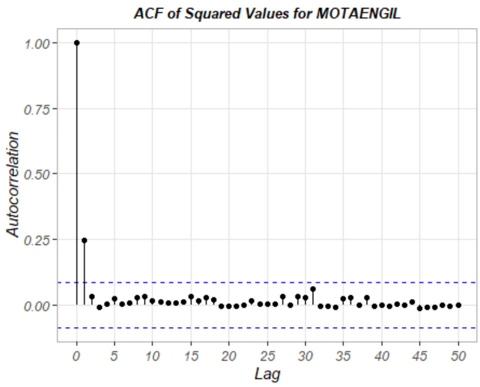
Figure 54: Autocorrelation Function Plot for the Absolute Values



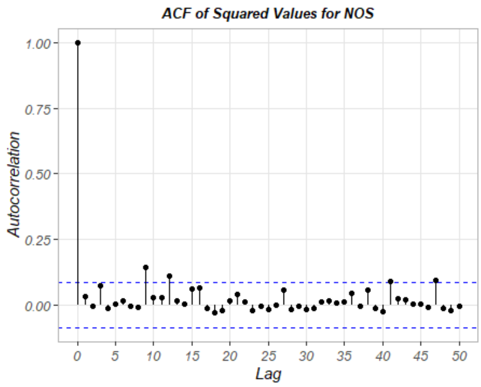
(a) ACF squared - EDP



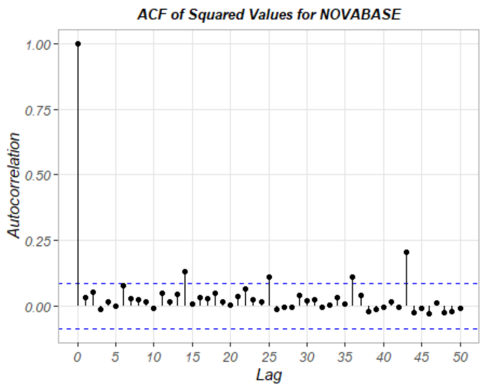
(b) ACF squared - GALP



(c) ACF squared - MOTAENGIL

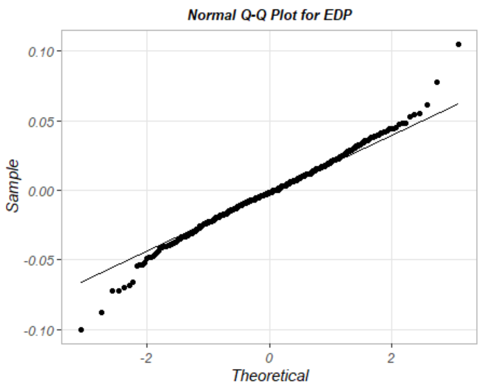


(d) ACF squared - NOS

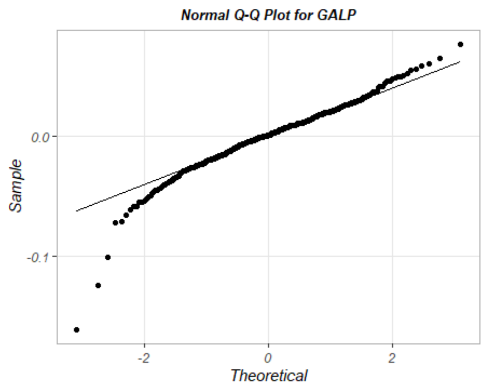


(e) ACF squared - NOVABASE

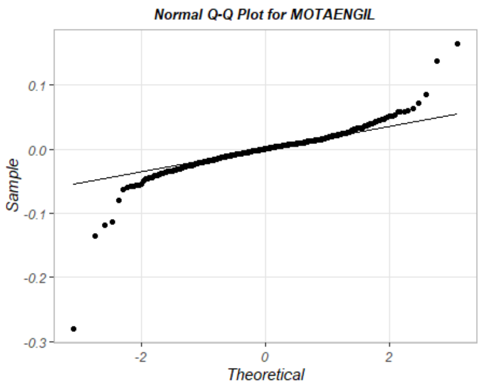
Figure 55: Autocorrelation Function Plot for the Squared Values



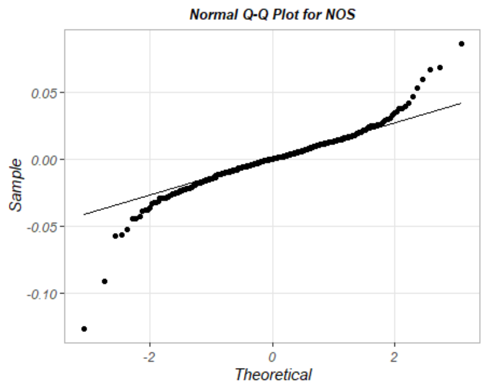
(a) QQ-Plot - EDP



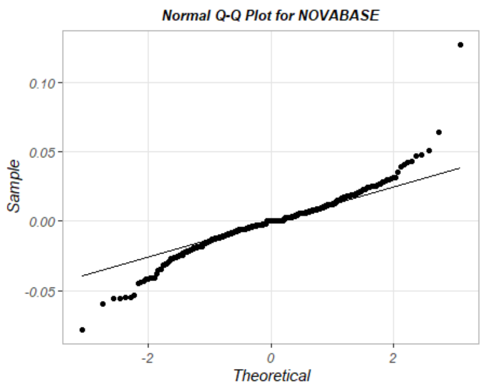
(b) QQ-Plot - GALP



(c) QQ-Plot - MOTAENGIL



(d) QQ-Plot - NOS



(e) QQ-Plot - NOVABASE

Figure 56: QQ-Plots for the log-returns

4.2 Models

The second assignment consisted in the application of non-linear multiplicative models, more specifically the **GARCH-type models**, to our time series data. These models are used whenever there's a strong belief that the variance error is serially autocorrelated.

We decided that the first procedure should be to model the data according to 5 different methods given in class: GARCH(p, q), IGARCH(p, q), FIGARCH(p, q), APARCH(p, q) and $\mathrm{GARCH-M(p, q)}$, **assuming conditional Normal distribution**. The choice for the values of p and q, respectively, the number of lag variances and the number of lag residual errors to include in the model, are trickier to justify than for the SARIMA-type model. Even though one can retrieve some information about the initial analysis made, there is no fully reliable analysis one can do in order to assess which (most likely) are the best values. Therefore, we decided to estimate all possible subset models with $p, q \in \{0, \cdots, 5\}$ (whenever 0 in p or q is possible). The reasoning behind this is that we wanted to have a large number of methods for comparison but still having low parameters to estimate, so that the computations would still be feasible. Then, in order to choose the best, we decided to use the ***AIC* and *BIC* criterion**, since both of them punish models with more parameters, that also may lead to overfit.

Since we have to analyze 5 different companies, we will present in the next subsection a detailed analysis for EDP, and resume the main results for the other 4. We used R packages *fGarch* and *rugarch* to fit our log return series to the various GARCH-type models.

4.3 EDP Results

4.3.1 Model Fitting

As a first step, we decided to select the best p and q for each of the models previously stated, based on AIC and BIC. As an example, table 32 shows the values the best values obtained (AIC) for the GARCH(p,q) methods [1]:

The best model, both in the Bayesian and the Akaike Information Criterion is the GARCH(1,1) model. Not only this model has few parameters, but also it is known to work well with real data, being a simple model that encompasses the problem of volatility. Table 33 presents the best model for all methods considered, and the respective parameters.

Here, we can see that for both criteria, the Integrated GARCH with $p = q = 1$ is the best model. We notice aswell that for the majority of the GARCH-type models, this were the best parameters. Clearly the model does not gain much by adding a lot of information about the past, but just the necessary to adapt

[1]It was space spending and unfeasible to present all 35, so the top 10 were chosen.

Model	AIC	BIC
GARCH(1,1)	-4.676449	-4.643138
GARCH(1,2)	-4.670858	-4.629220
GARCH(2,1)	-4.670858	-4.629220
GARCH(1,3)	-4.668405	-4.618439
GARCH(3,1)	-4.667923	-4.617957
GARCH(2,2)	-4.666921	-4.616955
GARCH(4,3)	-4.666168	-4.591219
GARCH(1,4)	-4.665802	-4.607508
GARCH(2,3)	-4.664550	-4.606256
GARCH(1,5)	-4.664483	-4.597861

Table 32: 10 best models for the GARCH(p,q) with conditional Normal Distribution.

GARCH-type	$(\mathbf{p}, \mathbf{q})$	AIC	BIC
GARCH	(1,1)	-4.670858	-4.629220
IGARCH	(1,1)	-4.673386	-4.648402
GARCH-M	(1,1)	-4.672512	-4.630873
APARCH	(2,3)	-4.671095	-4.587818
FIGARCH	(1,1)	-4.667401	-4.625762

Table 33: Best results for each GARCH-type model applied assuming conditional Normal Distribution

to this kind of problem. Nonetheless, there are ways that one could improve this kind of models. One of the most glaring assumptions that we have done and is most likely wrong is that the conditional variance follow a Gaussian distribution, since we have already seen in the initial data analysis that this financial time series presented fat-tailed distribution of returns. Therefore, the next step is to test if by changing the conditional distribution of the variance to one with a more fat-tail, the results improved or worsened. The distributions considered were the skewed normal distribution, generalized error (normal) distribution and student-t distribution. The results are resumed in table 34 [2].

Taking a deeper look at the table, there is a clear pattern - the best models usually are those whose p and q are 1. Even though the information criteria have similar values for every model, the best one is the Student-t GARCH(1,1) model.

4.3.2 Diagnostics Check

To validate our model, we must perform a diagnostic of both the residuals and the square of the residuals in order to investigate the presence of conditional heteroscedasticity (volatility clustering). In other words, it is necessary to check if

[2] In this table and in others, some values are filled with *NA* due to convergence problems of the *R* functions used.

GARCH-Type	Conditional Distribution	(p,q)	AIC	BIC
GARCH	*Normal*	(1,1)	-4.670858	-4.629220
	Skewed Normal	(1,1)	-4.673100	-4.631462
	GED	(1,1)	-4.700493	-4.658854
	Student-t	(1,1)	-4.702752	-4.661114
IGARCH	*Normal*	(1,1)	-4.673386	-4.648402
	Skewed Normal	(1,1)	-4.670219	-4.636908
	GED	(1,1)	-4.69803	-4.664719
	Student-t	(1,1)	-4.698744	-4.665433
GARCH-M	*Normal*	(1,1)	-4.672512	-4.630873
	Skewed Normal	(1,1)	-4.669163	-4.619197
	GED	(1,1)	-4.696643	-4.646677
	Student-t	(1,1)	-4.699009	-4.649043
APARCH	*Normal*	(2,3)	-4.671095	-4.587818
	Skewed Normal	(2,3)	-4.668199	-4.576594
	GED	(2,3)	-4.693618	-4.635324
	Student-t	*NA*	*NA*	*NA*
FIGARCH	*Normal*	(1,1)	-4.667401	-4.625762
	Skewed Normal	(1,1)	-4.664172	-4.614206
	GED	(1,1)	-4.692455	-4.642489
	Student-t	(1,1)	-4.69382	-4.643854

Table 34: Best models implemented with the 4 distributions - EDP

they are **independent, identically distributed and driven by a normal distribution**.

To evaluate the mutual independence of the residuals, we plotted the ACF of the standardized residuals (57). As one can see, almost all plotted lags are inside the 95% confidence interval, leading us to believe that the residuals are mutually independent. We also have in the same image the ACF plot for the squared values of the residuals, and we can make the same conclusion. We also applied the *Wald–Wolfowitz runs test* on the standardized residuals to confirm this fact. The null hypothesis of this test is that the residuals are mutually independent and the p-value obtained is 0.859. Therefore, we do not reject the null hypothesis for the usual levels of significance that the standardized residuals are mutually independent.

We also applied the **Jarque-Bera** and **Shapiro-Wilk tests** to check normality of the residuals. Both of them have low p-values (put val here) that leads us to reject the normality assumption. As can be seen in the QQ-Plot (58), the fact that extreme values are more than usual and tend to fluctuate a lot lead to this rejection, since this tests are very sensitive to this extreme points and therefore, easily manipulated.

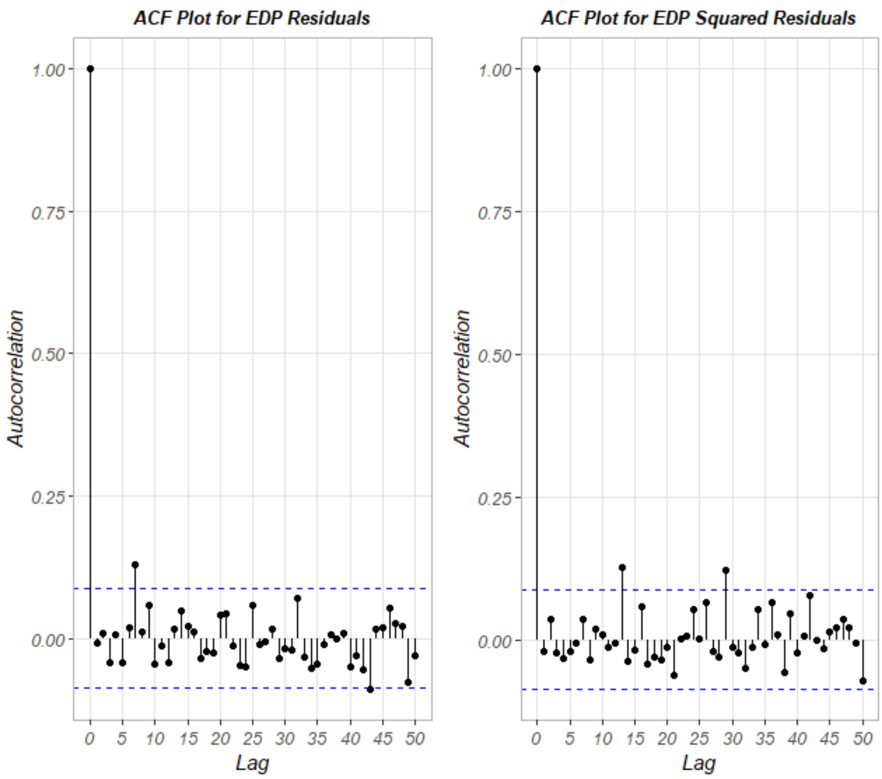


Figure 57: ACF and PACF for Residuals - EDP

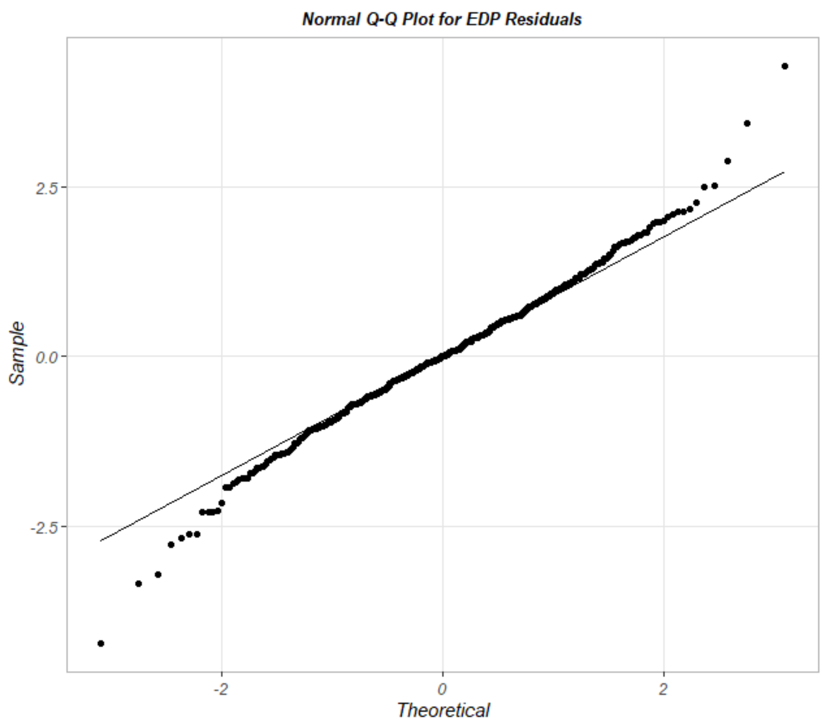


Figure 58: QQ-Plot for the Residuals - EDP

4.3.3 Final Model

Even though the residuals fail the normality tests due to the extreme values, it is not inappropriate to consider that this model fits well the data. So, the final model is a GARCH(1,1) model where the conditional variance follows a student-t distribution:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{cases}$$

where Z_t is White Noise with zero-mean and finite variance. The estimators for each coefficient were:

$$\begin{cases} \hat{a_0} = 0.000039 \\ \hat{a_1} = 0.072642 \\ \hat{b_1} = 0.857017 \end{cases}$$

4.4 GALP Results

In table 35 are compiled the best results for each model. This time, the best model was an IGARCH(1,5) model with generalized error distribution for the variance. This model was chosen as the best since it had the best AIC. Even though the BIC was higher for the GARCH(1,2) with student-t distribution, we opted for this more complex model, since it was computationally feasible to use it.

GARCH-Type	Conditional Distribution	(p,q)	AIC	BIC
GARCH	*Normal*	(1,4)	-4.665798	-4.607504
	Skewed Normal	(1,2)	-4.681184	-4.631218
	GED	(1,2)	-4.724441	-4.674475
	Student-t	(1,2)	-4.741072	-4.691106
IGARCH	*Normal*	(2,5)	-4.695918	-4.629296
	Skewed Normal	(2,5)	-4.716199	-4.641249
	GED	(1,5)	-4.754249	-4.687628
	Student-t	(2,5)	-4.750067	-4.675118
GARCH-M	*Normal*	(1,4)	-4.690216	-4.623594
	Skewed Normal	(1,4)	-4.700658	-4.625709
	GED	(1,4)	-4.731193	-4.656244
	Student-t	(1,2)	-4.745409	-4.687115
APARCH	*Normal*	(1,1)	-4.691917	-4.641950
	Skewed Normal	(1,1)	-4.703192	-4.644898
	GED	(1,1)	-4.734469	-4.676175
	Student-t	*NA*	*NA*	*NA*
FIGARCH	*Normal*	(1,3)	-4.654428	-4.596134
	Skewed Normal	(2,1)	-4.669741	-4.611447
	GED	(5,5)	-4.720163	-4.603575
	Student-t	(1,1)	-4.734411	-4.684445

Table 35: Best models implemented with the 4 distributions - GALP

From the residuals ACF and Q-Q plot (figures 59 and 60) we get to the same conclusions as for the EDP data. The p-value for the Wald-Wolfowitz runs test is 0.1553 and for the Jarque-Bera and Shapiro-Wilk are approximately 0 and $2.854 \cdot 10^{-08}$, respectively. So, the same conclusions about the residuals are drawn.

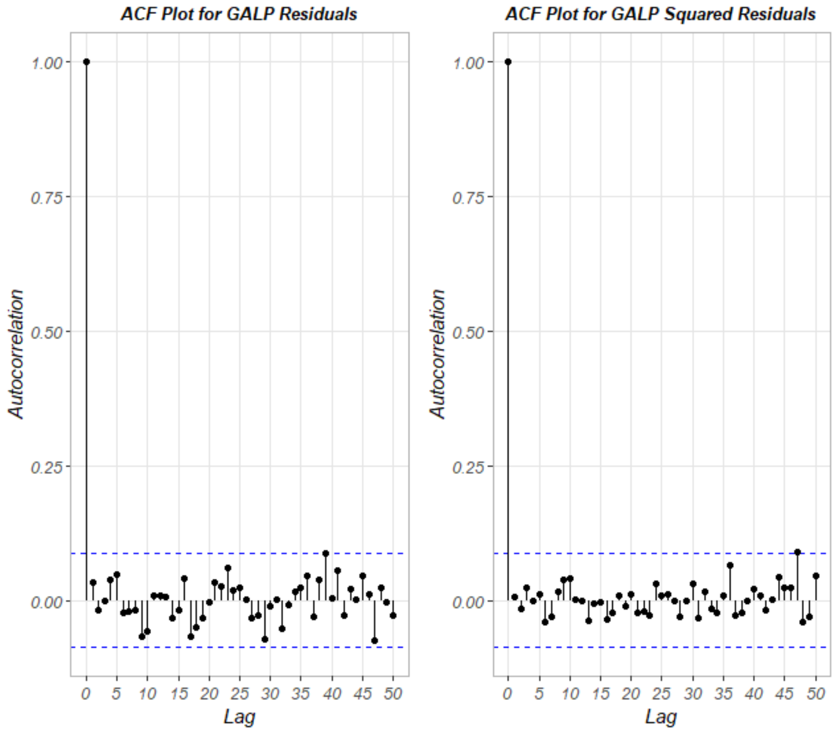


Figure 59: ACF and PACF for Residuals - GALP

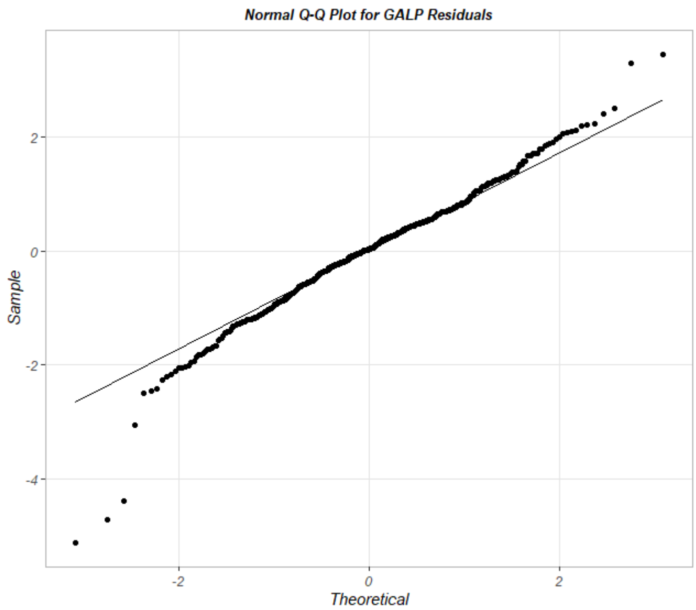


Figure 60: QQ-Plot for the Residuals - GALP

So, the final model is the IGARCH(1,5) model where the conditional variance follows a generalized error distribution:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + a_1 X_{t-1}^2 + \sum_{i=1}^{5} b_i \sigma_{t-i}^2 \end{cases}$$

where Z_t is White Noise with zero-mean and finite variance and $b_5 = 1 - (a_1 + b_1 + b_2 + b_3 + b_4)$. The estimators for each coefficient were:

$$\begin{cases} \hat{a_0} = 0.000011 \\ \hat{a_1} = 0.112328 \\ \hat{b_1} = 0.242460 \\ \hat{b_2} = 0.158429 \\ \hat{b_3} = 0.214836 \\ \hat{b_4} = 0.792265 \\ \hat{b_5} = -0.520318 \end{cases}$$

4.5 MOTAENGIL Results

In table 36 are compiled the best results for each model. This time, the best model was a GARCH(3,1) model with student-t distribution for the variance. This model was chosen as the best BIC and one of the best AIC's. Even though the AIC was higher for the FIGARCH(5,4) with student-t distribution and 2 other models, we opted for this much more simpler model since it has less coefficients to estimate.

From the residuals ACF and Q-Q plot (figures 61 and 62) we get to the same conclusions as for the EDP data. The p-value for the Wald-Wolfowitz runs test is 0.3285 and for the Jarque-Bera and Shapiro-Wilk are both approximately 0. So, the same conclusions about the residuals are drawn for this company.

So, the final model is the GARCH(3,1) model where the conditional variance follows a student-t distribution:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + \sum_{i=1}^{3} a_i X_{t-i}^2 + b_1 \sigma_{t-1}^2 \end{cases}$$

where Z_t is White Noise with zero-mean and finite variance. The estimators for each coefficient were:

$$\begin{cases} \hat{a_0} = 0.000061 \\ \hat{a_1} = 0.374969 \\ \hat{a_2} = 1.237 \cdot 10^{-9} \\ \hat{a_3} = 0.004488 \\ \hat{b_1} = 0.411730 \end{cases}$$

GARCH-Type	Conditional Distribution	(p,q)	AIC	BIC
GARCH	*Normal*	(4,5)	-4.665018	-4.573413
	Skewed Normal	(2,2)	-4.68516	-4.626866
	GED	(3,1)	-4.809155	-4.750861
	Student-t	(3,1)	-4.858492	-4.800198
IGARCH	*Normal*	(5,5)	-4.646545	-4.55494
	Skewed Normal	(1,5)	-4.686373	-4.619752
	GED	(1,5)	-4.809718	-4.743096
	Student-t	(1,5)	-4.860695	-4.794073
GARCH-M	*Normal*	(5,4)	-4.645368	-4.545436
	Skewed Normal	(5,4)	-4.674392	-4.566131
	GED	(1,4)	-4.815761	-4.740812
	Student-t	(1,4)	-4.860170	-4.785220
APARCH	*Normal*	(5,5)	-4.731803	-4.581904
	Skewed Normal	(5,1)	-4.736795	-4.61188
	GED	(1,2)	-4.824247	-4.757625
	Student-t	(1,2)	-4.861259	-4.794637
FIGARCH	*Normal*	(5,1)	-4.683925	-4.608975
	Skewed Normal	(5,1)	-4.699715	-4.616438
	GED	(5,1)	-4.821254	-4.737976
	Student-t	(5,4)	-4.866900	-4.758640

Table 36: Best models implemented with the 4 distributions - MOTAENGIL

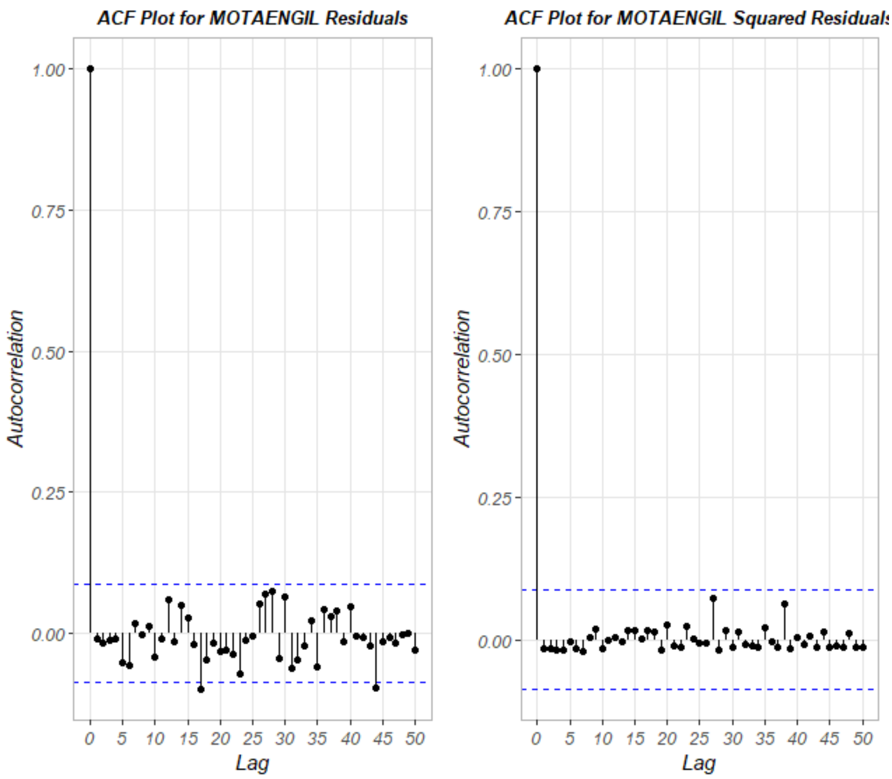


Figure 61: ACF and PACF for Residuals - MOTAENGIL

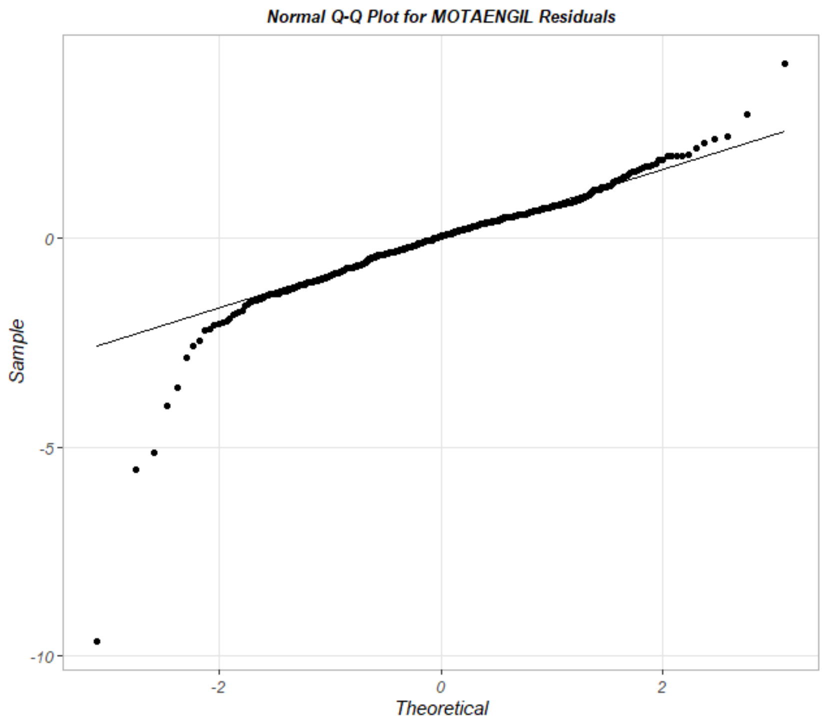


Figure 62: QQ-Plot for the Residuals - MOTAENGIL

4.6 NOS Results

In table 37 are compiled the best results for each model. This time, the best model was the IGARCH(1,2) model with student-t distribution for the variance. This model was chosen since it had the best BIC and one of the best AIC, with low parameters.

From the residuals ACF and Q-Q plot (figures 63 and 64) we get to the same conclusions as for the other companies. The p-value for the Wald-Wolfowitz runs test is 0.7899 and for the Jarque-Bera and Shapiro-Wilk are approximately 0 and $1.117 \cdot 10^{-11}$. So, mutually independence is very likely however normality is rejected.

So, the final model is the IGARCH(1,2) model where the conditional variance follows a student-t distribution:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 + b_2 \end{cases}$$

where Z_t is White Noise with zero-mean and finite variance and $b_2 = 1 - (a_1 + b_1)$. The estimators for each coefficient were:

$$\begin{cases} \hat{a_0} = 0.000015 \\ \hat{a_1} = 0.186256 \\ \hat{b_1} = 0.000004 \\ \hat{b_2} = 0.813740 \end{cases}$$

GARCH-Type	Conditional Distribution	(p,q)	AIC	BIC
GARCH	*Normal*	(2,2)	-5.339265	-5.289299
	Skewed Normal	(2,2)	-5.336814	-5.278520
	GED	(1,2)	-5.474422	-5.424456
	Student-t	(1,2)	-5.496421	-5.446455
IGARCH	*Normal*	(3,3)	-5.344537	-5.286243
	Skewed Normal	(3,3)	-5.341361	-5.274739
	GED	(1,2)	-5.474499	-5.432860
	Student-t	(1,2)	-5.495466	-5.453828
GARCH-M	*Normal*	(3,3)	-5.339217	-5.264268
	Skewed Normal	(3,3)	-5.336034	-5.252757
	GED	(1,2)	-5.470662	-5.412368
	Student-t	(1,2)	-5.494721	-5.436427
APARCH	*Normal*	(3,4)	-5.350944	-5.242684
	Skewed Normal	(3,5)	-5.345998	-5.221082
	GED	(1,2)	-5.467924	-5.401302
	Student-t	(1,2)	-5.489998	-5.423376
FIGARCH	*Normal*	(3,1)	-5.345912	-5.287618
	Skewed Normal	(3,3)	-5.349683	-5.266406
	GED	(3,3)	-5.480332	-5.397055
	Student-t	(2,3)	-5.521199	-5.446249

Table 37: Best models implemented with the 4 distributions - NOS

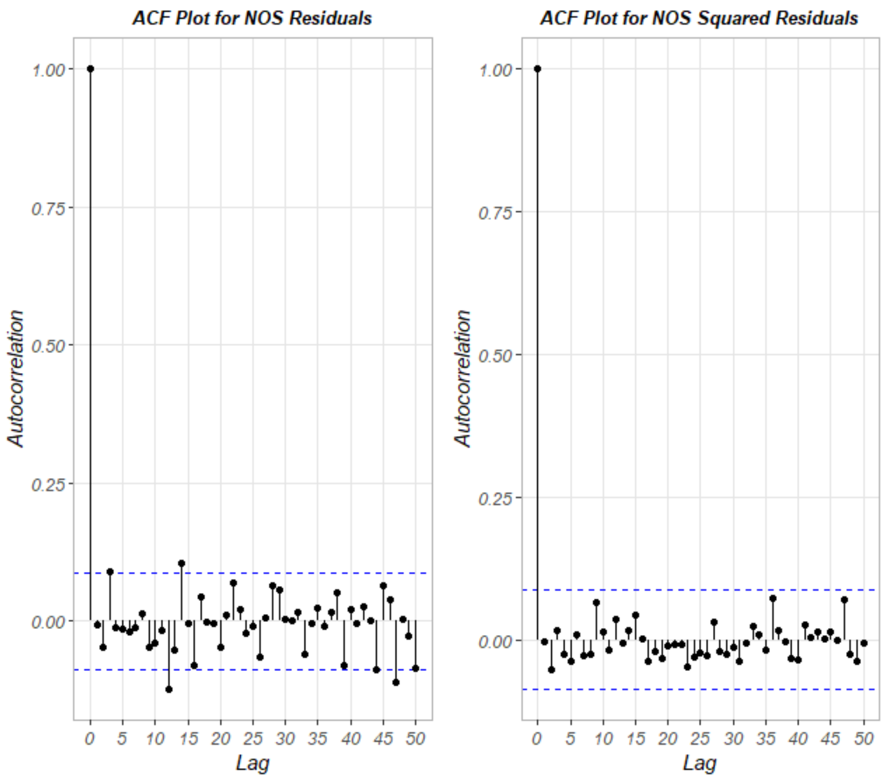


Figure 63: ACF and PACF for Residuals - NOS

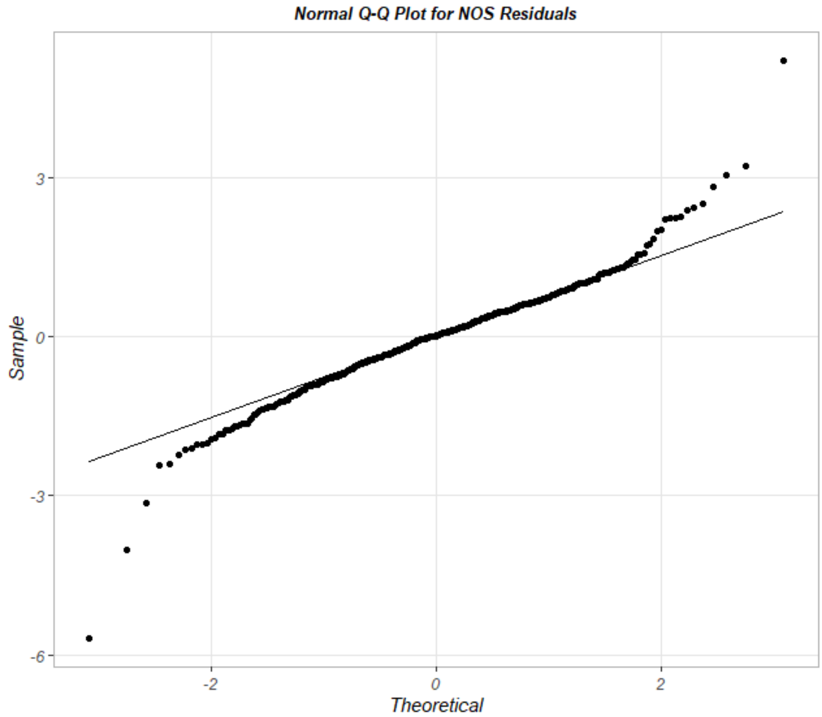


Figure 64: QQ-Plot for the Residuals - NOS

4.7 NOVABASE Results

In table 38 are compiled the best results for each model. This time, the best model was an IGARCH(1,1) model with student-t distribution for the variance. This model was chosen as the best BIC and one of the best AIC's, with a few number of parameters to estimate.

From the residuals ACF and Q-Q plot (figures 65 and 66) we get to the same conclusions as for all the other companies. The p-value for the Wald-Wolfowitz runs test is 0.2482 and for the Jarque-Bera and Shapiro-Wilk are all approximately 0. Yet again, the same conclusions are drawn for this data, rejecting the normality of the residuals.

So, the final model is the IGARCH(1,1) model where the conditional variance follows a student-t distribution:

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = a_0 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{cases}$$

where Z_t is White Noise with zero-mean and finite variance and $b_1 = 1 - a_1$. The estimators for each coefficient are:

$$\begin{cases} \hat{a_0} = 0.000007 \\ \hat{a_1} = 0.082131 \\ \hat{b_1} = 0.917869 \end{cases}$$

GARCH-Type	Conditional Distribution	(p,q)	AIC	BIC
GARCH	*Normal*	(1,1)	-5.301689	-5.268378
	Skewed Normal	(1,1)	-5.298072	-5.256433
	GED	(1,1)	-5.481711	-5.440072
	Student-t	(1,1)	-5.480118	-5.438479
IGARCH	*Normal*	(1,3)	-5.281732	-5.240094
	Skewed Normal	(1,4)	-5.29504	-5.236746
	GED	(1,1)	-5.479213	-5.445902
	Student-t	(1,1)	-5.482147	-5.448836
GARCH-M	*Normal*	*NA*	*NA*	*NA*
	Skewed Normal	*NA*	*NA*	*NA*
	GED	*NA*	*NA*	*NA*
	Student-t	*NA*	*NA*	*NA*
APARCH	*Normal*	(2,3)	-5.360521	-5.277244
	Skewed Normal	(2,3)	-5.364451	-5.272846
	GED	(1,1)	-5.483932	-5.425638
	Student-t	(1,1)	-5.476690	-5.418396
FIGARCH	*Normal*	(3,5)	-5.304627	-5.213022
	Skewed Normal	(3,5)	-5.300754	-5.200821
	GED	(3,5)	-5.493413	-5.393481
	Student-t	(1,5)	-5.483805	-5.400528

Table 38: Best models implemented with the 4 distributions - NOVABASE

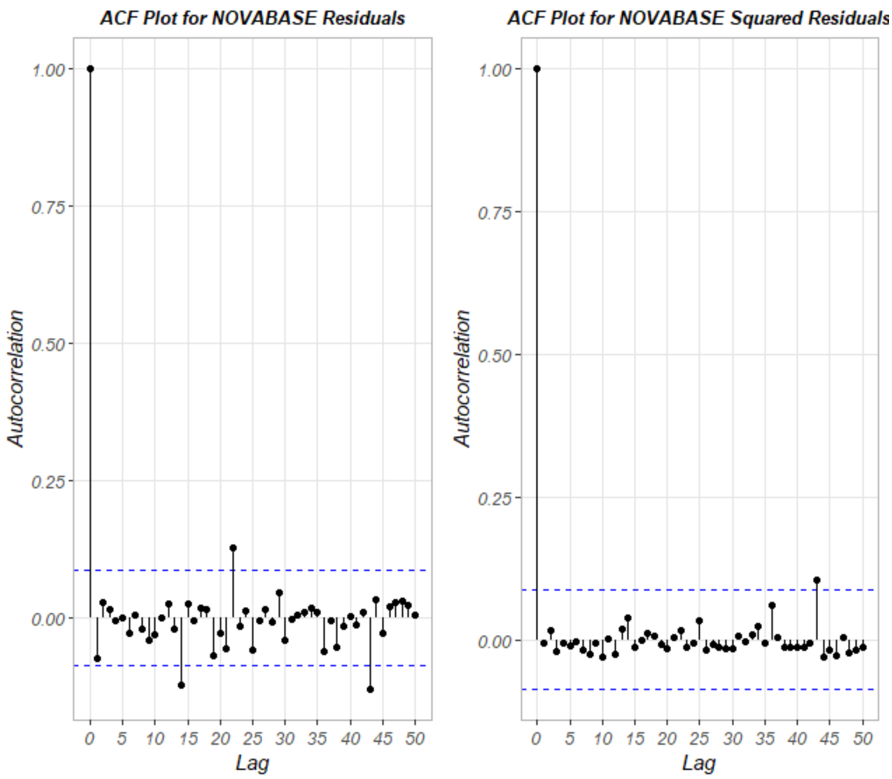


Figure 65: ACF and PACF for Residuals - NOVABASE

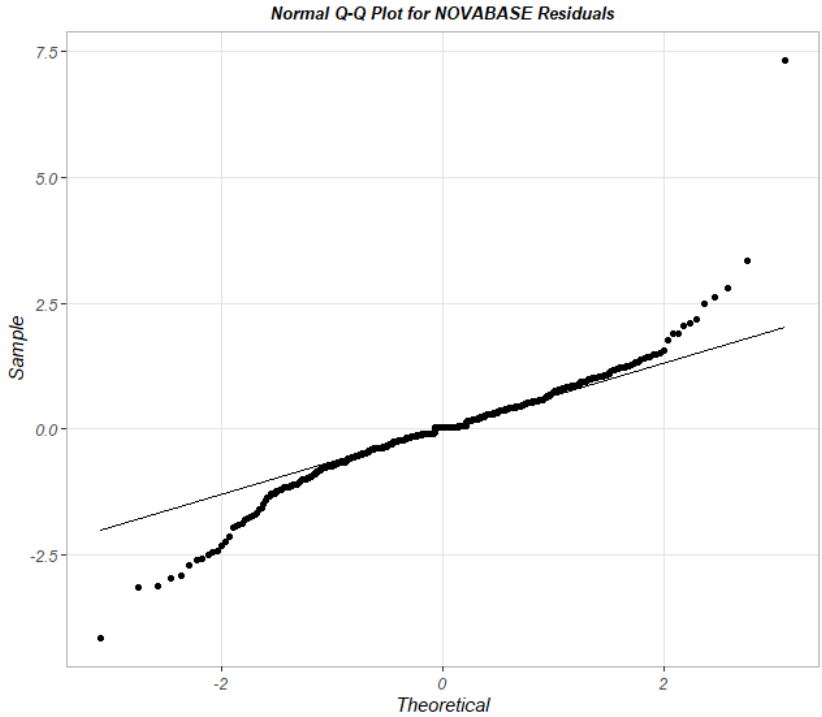


Figure 66: QQ-Plot for the Residuals - NOVABASE

4.8 Conclusion

In conclusion, we have seen that even though different types of GARCH models bring something new to the equation, they almost predict the same way the data at hand, and for the sake of complexity we always will choose those who are computationally more efficient. During each time series fit we chose the best model, usually being an IGARCH, reinforcing the idea that the coefficients sum to one in financial time series. Although this model can probably be further improved, we obtained several good methods that do not violate too many assumptions of the residuals.

This assignment also allowed us to better understand how to model the log-returns of a very popular stock market index although the results weren't great we were still able to have a general idea of how to derive such models and what things one should consider when dealing with this kind of analysis.

For future work a deeper analysis should be done. In particular, we only tested four distributions, but an extensive study can be made here and therefore more could be used. We also only focused only on the metrics AIC and BIC to choose the best model, which should but other measures exist and other types of comparisons should be executed.

Bibliography

[1] *ARIMA Model Building*. http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/xegbohtmlnode39.htMl. Accessed: 2022-06-15.

[2] Shawn Bauldry. "Structural Equation Modeling". In: *International Encyclopedia of the Social & Behavioral Sciences* 47 (2015), pp. 615–620. DOI: https://doi.org/10.1016/b978-0-08-097086-8.44055-9.

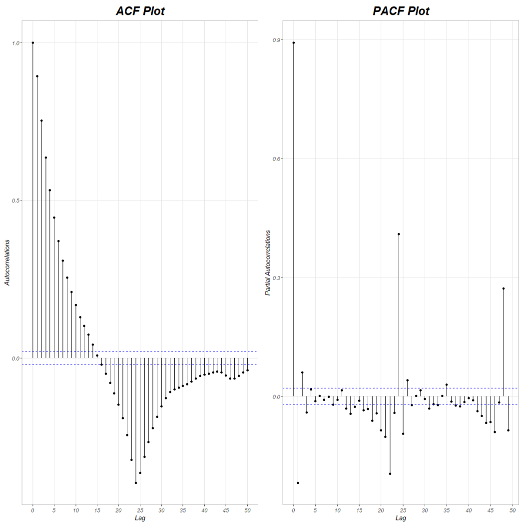
[3] Kenneth P. Burnham and David R. Anderson. "Multimodel Inference: Understanding AIC and BIC in Model Selection". In: *Sociological Methods & Research* (2004), p. 159.

[4] *Concentrações de ozono superficial em Portugal: Avaliação dos padrões temporais e dos contrastes espaciais em estações de fundo.* https://bibliotecadigital.ipb.pt/bitstream/10198/11789/1/Afonso_Paulo.pdf. Accessed: 2022-06-15.

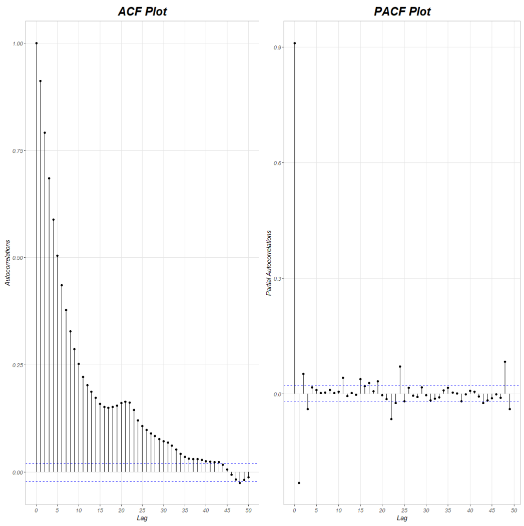
[5] *MAE and RMSE — Which Metric is Better?* https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d. Accessed: 2022-06-15.

[6] Robert H. Shumway and David S. Stoffer. "Time Series Analysis and Its Applications With R Examples". In: (2004), pp. 1–12.

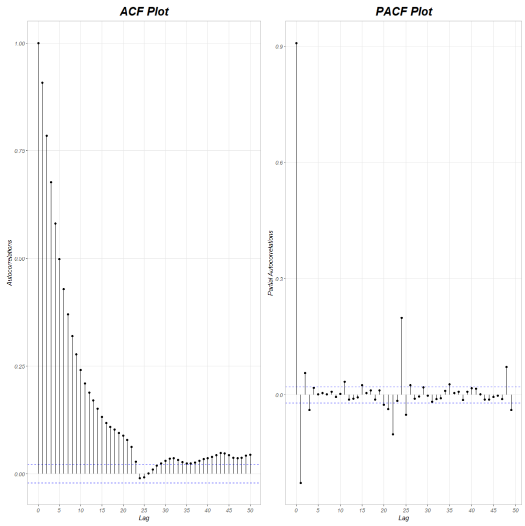
Appendix Images and Tables



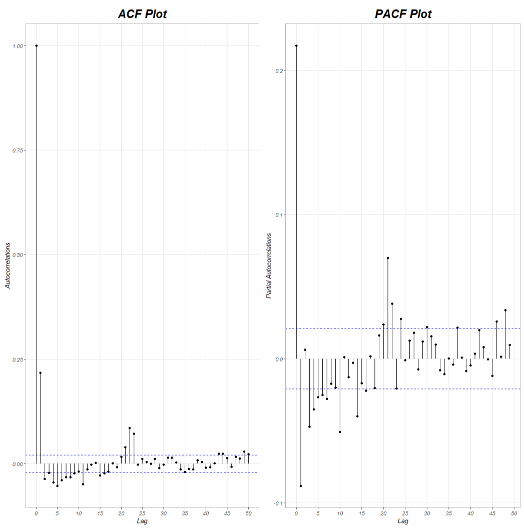
(a) Spike in ACF at lag h = 24, so Q = 1



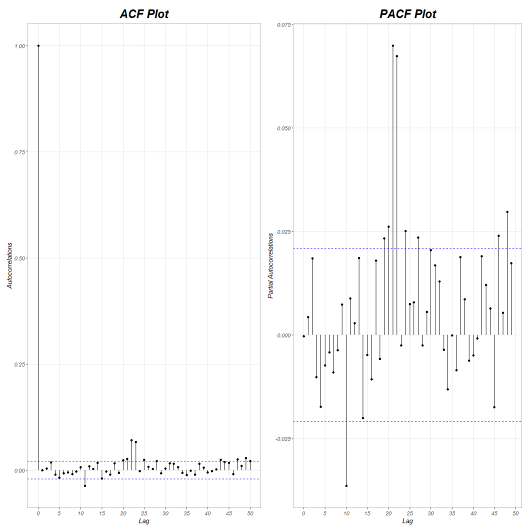
(b) Spike in PACF at lag h = 24, so P = 1



(c) Spike in PACF at lag h = 1, so p = 1

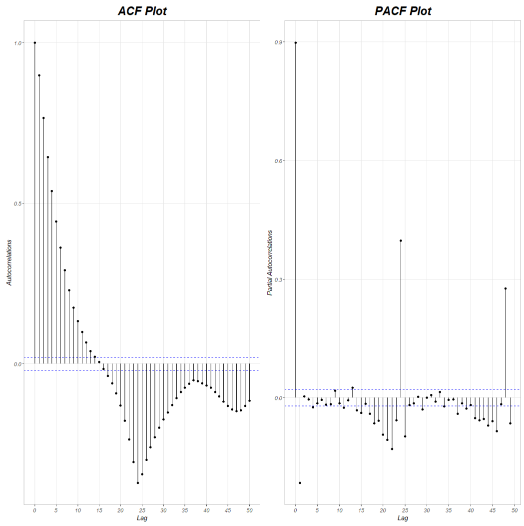


(d) Spike in ACF at lag h = 1, so q = 1

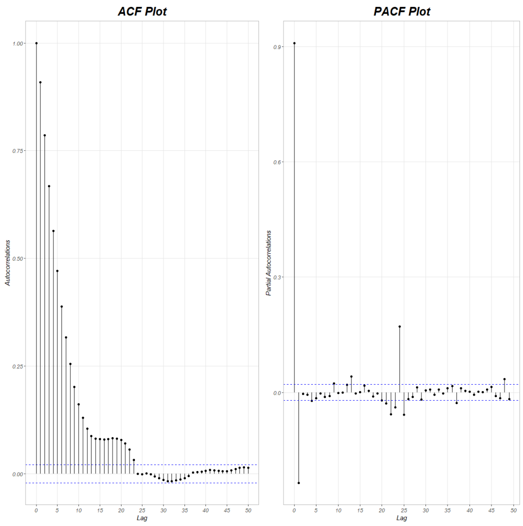


(e) ACF and PACF values are negligible, white noise achieved

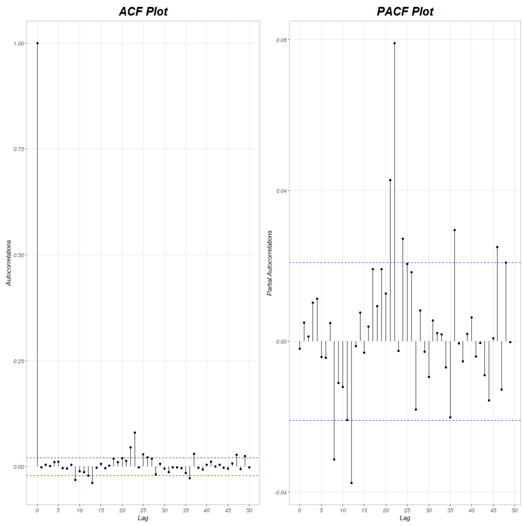
Figure 67: Proceedure for finding the best model for Antas-Espinho



(a) Spikes in ACF at lags h = 24 and 48, so Q = 2

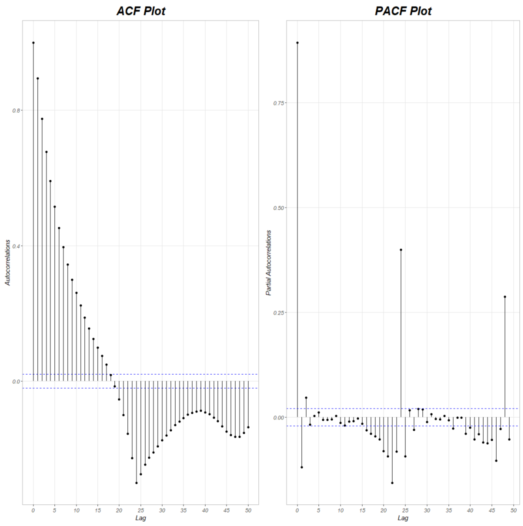


(b) Spikes in PACF at lags h = 1 and 2, so p = 2

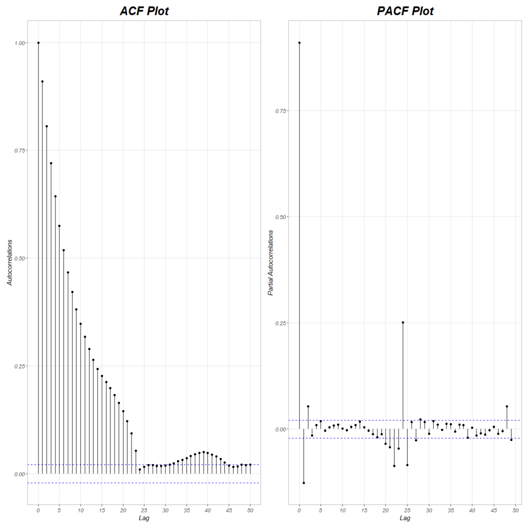


(c) ACF and PACF values are negligible, white noise achieved

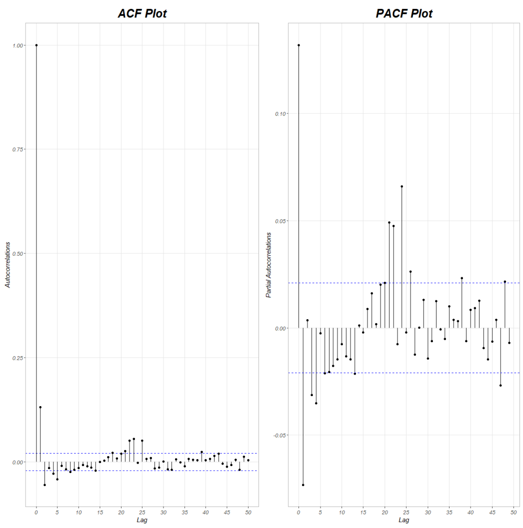
Figure 68: Proceedure for finding the best model for Estarreja



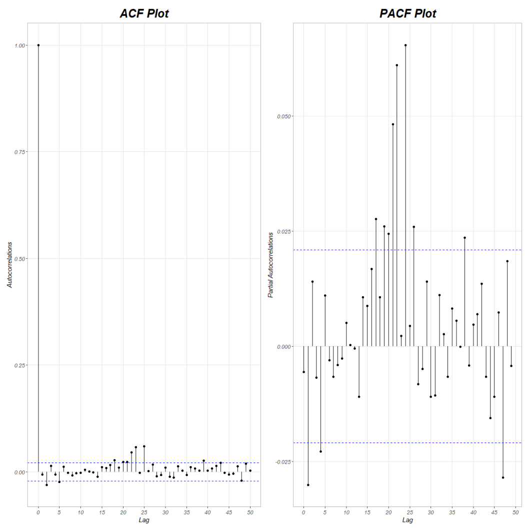
(a) Spike in ACF at lags h = 24 and 48, so Q = 2



(b) Spike in PACF at lag h = 1, so p = 1



(c) Spike in ACF at lag h = 1, so q = 1



(d) ACF and PACF values are negligible, white noise achieved

Figure 69: Proceedure for finding the best model for Laranjeiro-Almada

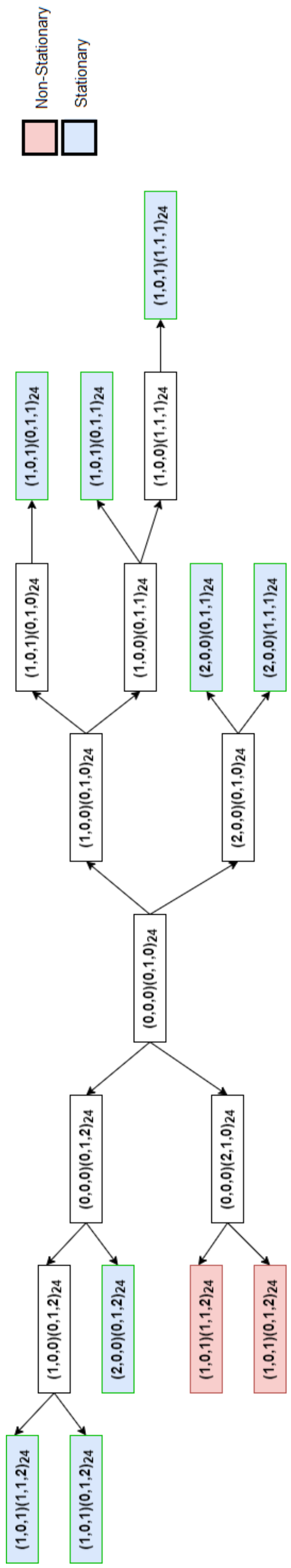


Figure 70: Possible paths for achieving White Noise