Fitting SARIMA and GARCH Models
Time Series

Emanuel Frazão 76221 Margarida Campos 77039

June 2020

1 Introduction

This report was done in the context of the *Time Series* course, with the aim of fitting a SARIMA and a GARCH model respectively to two different time series.

1.1 Basic Concepts

In this section we define the main concepts required in analysing the model fitting, using the notation given in [1].

1.1.1 SARIMA Processes

Definition 1.1. Given a time series (X_t), we define three operators on X_t:

- *back shift operator* $B : BX_t = X_{t-1}$
- *difference operator* $\nabla : \nabla \equiv 1 - B$
- *seasonal difference operator* $\nabla_S : \nabla_S \equiv 1 - B^S$, for $S \in \mathbb{N}_1$.

Definition 1.2. A linear time series (X_t) is a *mixed seasonal* ARMA$(p, q) \times (P, Q)_S$ if

$$\Psi(B)\Phi_P(B^S)X_t = \theta(B)\Theta_Q(B^S)Z_t$$

where

- (Z_t) is a time series such that $Z_t \sim WN(0, \sigma_Z^2)$
- $\Psi(B) \equiv 1 - \sum_{i=1}^{p} \psi_i B^i$, with $\psi_i \in \mathbb{R}, \forall i \in \{1, ..., p\}$
- $\Phi_P(B) \equiv 1 - \sum_{i=1}^{P} \phi_i B^{iS}$, with $\phi_i \in \mathbb{R}, \forall i \in \{1, ..., P\}$
- $\theta(B) \equiv 1 + \sum_{i=1}^{q} \theta_i B^i$, with $\theta_i \in \mathbb{R}, \forall i \in \{1, ..., q\}$
- $\Theta_Q(B) \equiv 1 + \sum_{i=1}^{Q} \theta_i^* B^{iS}$, with $\theta_i^* \in \mathbb{R}, \forall i \in \{1, ..., Q\}$.

Definition 1.3. A mixed seasonal ARMA is *causal* if there exists (ω_i) with $\sum_{i=1}^{\infty} \omega_i < \infty$ such that

$$X_t = \sum_{i=1}^{\infty} \omega_i Z_{t-i}$$

Theorem 1.1. *A mixed seasonal ARMA is causal if and only if*

$$\Psi(z)\Phi_P(z^S) \neq 0, \forall_{z \in \mathbb{C}} : |z| \leq 1$$

Definition 1.4. A mixed seasonal ARMA is *invertible* if there exists if there exists (π_j) with $\sum_{j=1}^{\infty} \pi_j < \infty$ such that

$$Z_t = \sum_{j=1}^{\infty} \pi_j X_{t-j}$$

Theorem 1.2. *A mixed seasonal ARMA is invertible if and only if*

$$\theta(z)\Theta_Q(z^S) \neq 0, \forall_{z \in \mathbb{C}} : |z| \leq 1$$

.

Definition 1.5. A linear time series (Y_t) is a SARIMA$(p, d, q) \times (P, D, Q)_S$ if

$$X_t = \nabla^d \nabla_S^D Y_t$$

where (X_t) is a causal mixed seasonal ARMA$(p, q) \times (P, Q)_S$ and $d, D \in \mathbb{N}_0$.

1.1.2 GARCH Processes

Definition 1.6. A nonlinear time series (X_t) is a GARCH(p, q) if

$$\begin{cases} X_t = \mu + \sigma_t Z_t \\ \sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i (X_{t-i} - \mu)^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2 \end{cases} \tag{1}$$

where $a_0 > 0, a_i \geq 0, b_j \geq 0$, Z_t is a random variable with zero mean and variance one and μ is either constant or a random variable dependent on σ.

We use two generalizations of GARCH processes: APARCH, adapted from [1], and GJR-GARCH, adapted from [3].

Definition 1.7. A nonlinear time series (X_t) is an APARCH(p, q) if

$$\begin{cases} X_t = \mu + \sigma_t Z_t \\ \sigma_t^\delta = a_0 + \sum_{i=1}^{p} a_i (|X_{t-i} - \mu| - \gamma_i (X_{t-i} - \mu))^\delta + \sum_{j=1}^{q} b_j \sigma_{t-j}^\delta \end{cases} \tag{2}$$

where $a_0 > 0, a_i \geq 0, b_j \geq 0, \delta \geq 0, -1 < \gamma_i < 1$, Z_t is a random variable with zero mean and variance one and μ is either constant or a random variable dependent on σ.

Definition 1.8. A nonlinear time series (X_t) is a GJR-GARCH(p, q) if

$$\begin{cases} X_t = \mu + \sigma_t Z_t \\ \sigma_t^2 = a_0 + \sum_{i=1}^{p} (a_i + \gamma_i d_{t-1})(X_{t-i} - \mu)^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2 \end{cases} \tag{3}$$

$$\text{with dummy variable } d_{t-1} = \begin{cases} 1 & \text{if } X_{t-1} < 0 \\ 0 & \text{if } X_{t-1} \geq 0 \end{cases} \tag{4}$$

where $a_0 > 0, a_i \geq 0, b_j \geq 0, a_i + \gamma_i \geq 0$, Z_t is a random variable with zero mean and variance one and μ is either constant or a random variable dependent on σ.

Furthermore, we say that (X_t) is a *normal* GARCH (respectively GJR-GARCH or APARCH) if Z_t is normal and that (X_t) is a *ST* GARCH (respectively GJR-GARCH or APARCH) if Z_t follows a skewed T-student distribution. We assume μ is zero if not stated otherwise.

2 SARIMA Model

2.1 Data

The data consists of **43, 824** data points, corresponding to the hourly levels of $\mathbf{PM_{10}}(\mu g/m^3)$ particles in Avenida da Liberdade (Lisbon), collected between 2014 and 2018.
For the present analysis we are interested in the daily average levels. Computing the mean - ignoring missing values - we end up with **1, 826** points. The plot of the time series can be seen in Figure 1.

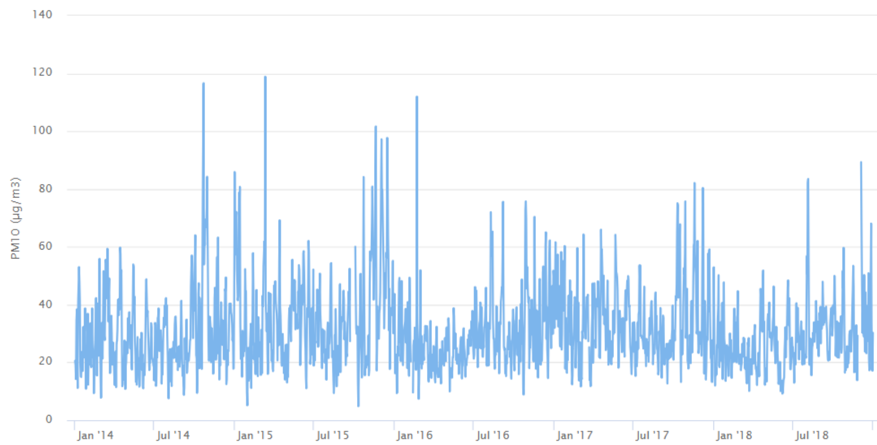


Figure 1: Daily average values of PM_{10} particles

2.1.1 Missing Values

There are **26** days without records. For imputation we used linear interpolation.

2.1.2 Stationarity

To assess if the series is stationary we applied the following tests:

- **ADF**(Augmented Dickey-Fuller): the null hypothesis is the presence of a unit root in the time series; the alternative hypothesis is stationarity
- **KPSS**(Kwiatkowski-Phillips-Schmidt-Shin) - null hypothesis is stationarity againgst the alternative presence of unit root

Test	**Test-Stat**	**P-Value**
ADF	0.1927	>0.1
KPSS	-8.7223	<0.01

Table 1: Stationarity test results

Based on the results in Table 1, it is fair to assume the series is stationary.

2.2 Method

To analyze the forecasting capabilities of the models we leave out the last 21(3 weeks) points of the series for testing purposes. The models are therefore tune on 1805 points.
We start by applying a Box-Cox transformation on the data, in order to smooth the variance. Afterwards, since both the data and its transformation are stationary, we first try to fit it directly with mixed seasonal ARMA processes (models M_1, M_2 and M_3). We then use the

R function `auto.arima` to fit the remaining parameters of a SARIMA process whilst forcing differencing with $d = 1$ (model M_4) and with $D = 1$ (model M_5).

2.2.1 Transformations

In order to attenuate variance in the data, we explored a Box-Cox class of power transformations for $\lambda \in \mathbb{R}$, bearing in mind the trade-off in the ability to forecast.
Using the criteria given in [2] for deciding which λ to use, we got $\lambda = -0.08 \simeq 0$ for the data series at hand, hence we applied a log transformation. The log transformation of the data can be seen in Figure 2

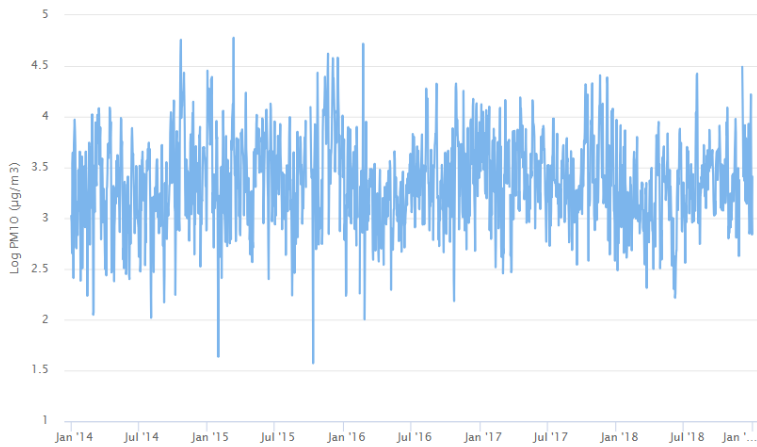


Figure 2: Time series after Log transformation

2.2.2 Order Dependence

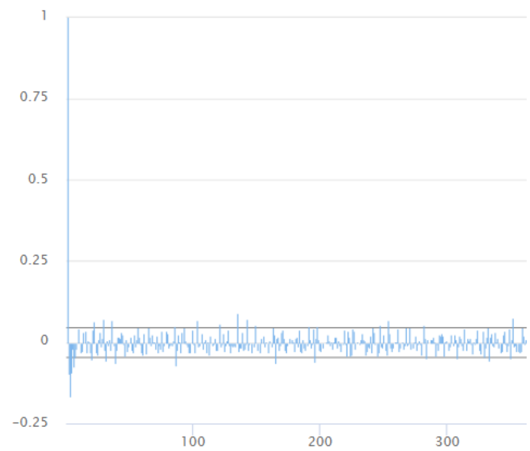
Considering the time series measures the pollution in a Lisbon avenue, we observe that if the model is to have a seasonal component of parameter S (which considering the phenomena, it certainly should), then the best value of S to take into consideration is 7, since the seasonal component should depend mostly on the day of the week.
When analysing the sample ACF and the sample PACF in Figures 3 and 4, we notice that both measures tail off. Still, when trying to fit an ARMA$(p, q) \times (P, Q)_7$, we can use ACF to estimate the parameters q and Q and PACF to estimate p and P.
From Figure 3 (right) we estimate that $q = 2$ (models M_1 and M_2) or $q = 1$ (M_3), since these are the values of higher ACF, and that $Q = 0$, since there doesn't seem to exist any weekly correlation (we tried models with $Q > 0$, and they grossly missfitted the data).
From Figure 4 (right) we estimate that $p = 2$ and $P > 0$, since there is clear correlation between same week days: we set $P = 1$ for model M_1 and $P = 2$ for M_2 and M_3.

2.2.3 Degree of Differencing

We use differencing (a SARIMA process instead of a mixed seasonal ARMA) in order to obtain stationarity. Since the data was already stationary, we started by trying to fit an ARMA model to the data. In order to use differencing and fit it with the SARIMA model, we resorted to the function `auto.arima` in R to search for the best parameters p, d, q, P, D, Q whilst forcing it to use differencing, thus resulting in models M_4 and M_5.

2.3 Results

In this section we start by discussing the results for the best models we found: the test of whiteness with *Ljung-Box-Pierce* Q-statistic, *Akaike Information Criterion* corrected (AICc) and *Bayesian Information Criterion* (BIC). We then make an analysis of the residuals of the best model, and give its expression.



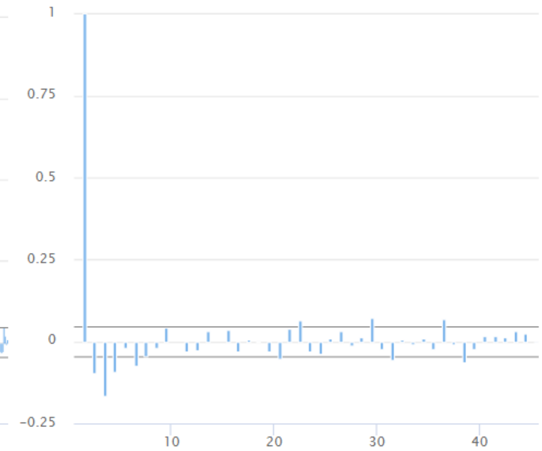
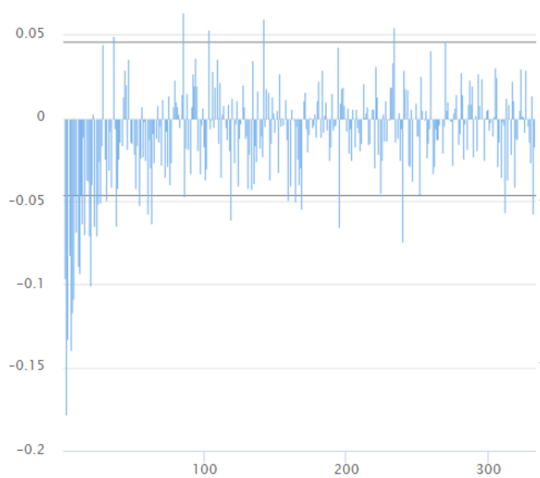


Figure 3: ACF



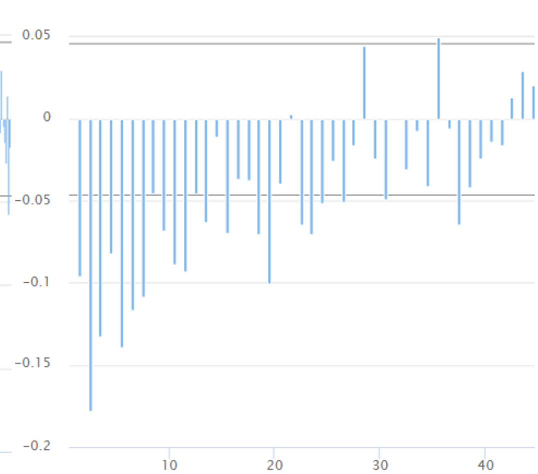


Figure 4: PACF

2.3.1 Residual Diagnostics and Model Choice

Table 2 shows the performance metrics for each model explored. We chose the model $\mathbf{M_3}$: $(Y_t) \sim SARIMA(2,0,1)(2,0,0)_7$ as it minimizes both AICc and BIC, whilst passing the Ljung-Box test for the randomness of the residuals and adequacy of the model.

SARIMA Model	$(p,d,q) \times (P,D,Q)_7$	**AICc**	**BIC**	**Ljung-Box p-value**
M_1	$(2,0,2)(1,0,0)_7$	1012.06	1050.48	0.0942
M_2	$(2,0,2)(2,0,0)_7$	1009.74	1053.65	0.1687
$\mathbf{M_3}$	$\mathbf{(2,0,1)(2,0,0)_7}$	**1008.98**	**1047.41**	**0.1486**
M_4	$(4,1,3)(2,0,0)_7$	1009.34	1064.19	0.0982
M_5	$(2,0,2)(1,1,0)_7$	1645.66	1684.05	$< 10^{-16}$

Table 2: Performance metrics for some of the tested models.

In Figure 5 we can see the plot of these residuals. Looking at the ACF plot we can see that, even though the autocorrelation passes the desired threshold in three of the lags, it lies mostly within the bounds. Looking at the histogram it is also clear that the residuals approximate a normal distribution, as it is also confirmed by the Q-Q Plot on Figure 6.

Since the roots from both the autoregressive polynomial - $\Psi(z)\Phi_2(z^7)$ - and the moving average polynomial - $\theta(z)$ - are inside the unit circle (Figure 7), M_3 is neither causal nor invertible, which means we can neither write X_t as a linear combination of (Z_t) nor Z_t as a

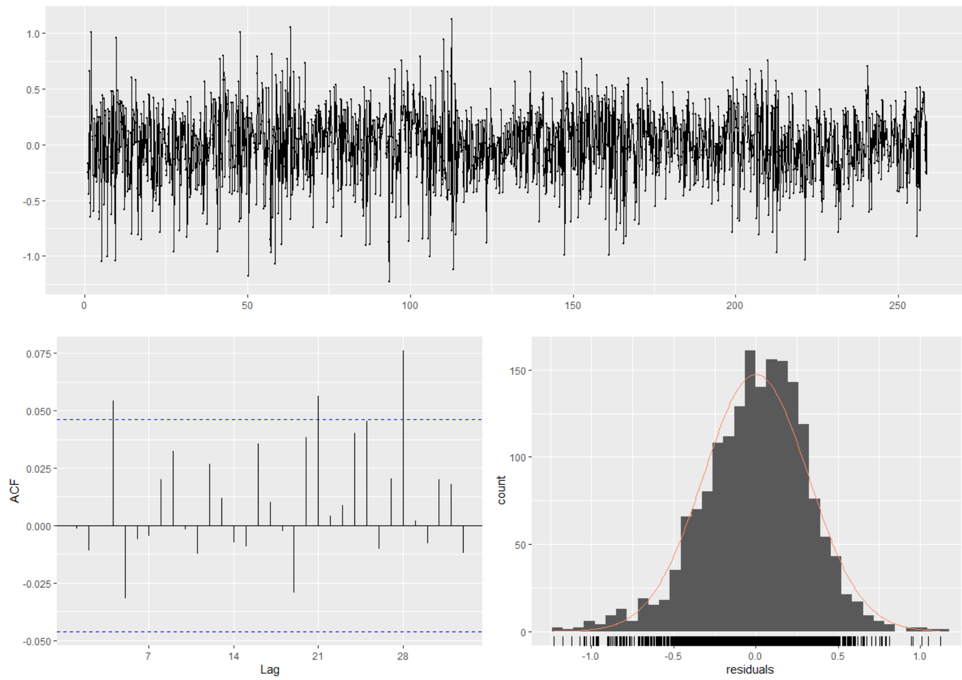


Figure 5: Residuals' details for final model M_3.

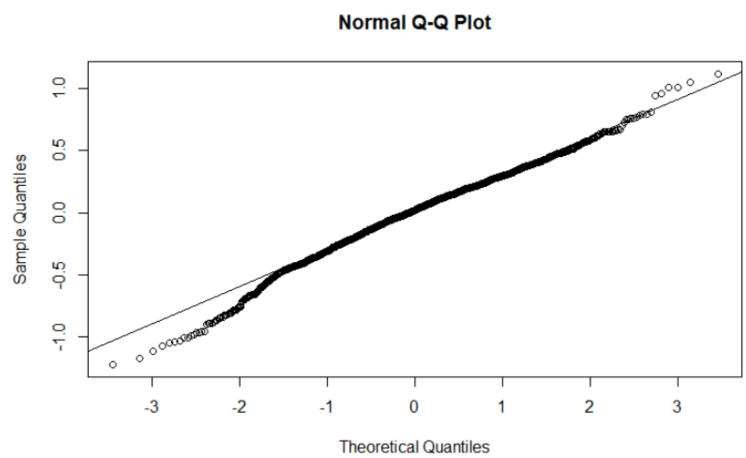


Figure 6: Normal Q-Q Plot for M_3 residuals.

linear combination of (X_t). Still, their roots are far from the unit circle, which adds on the evidence that the process is stationary.

In Figure 8 we can see the comparison of the actual transformed values against the fitted values of the model. The model is able to capture the oscillations of the real data, but it clearly has a lower variance and therefore never matches the amplitude of the spikes.

2.3.2 Parameters Estimation

We estimate the parameters $\phi_1, \phi_2, \psi_1, \psi_2, \theta_1$ and σ_Z^2 for model $M_3 : (Y_t) \sim \text{SARIMA}(2,0,1)(2,0,0)_7$ resorting to the `arima` function in `R`, and give its expression.
$M_3 : Y_t = \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \phi_1 Y_{t-7} + \phi_1\psi_1 Y_{t-8} + \phi_1\psi_2 Y_{t-9} + \phi_2 Y_{t-14} + \phi_2\psi_1 Y_{t-15} + \phi_2\psi_2 Y_{t-16} + Z_t + \theta_1 Z_{t-1}$, with $Z_t \sim WN(0, 0.1019)$ and $\phi_1, \phi_2, \psi_1, \psi_2, \theta_1$ given in Table 3.

2.3.3 Forecasting

We forecasted the log-values of the pollution in the forward five days of our time series, given in Figure 9 and Table 4, by a 95% confidence interval. The actual prediction of the values

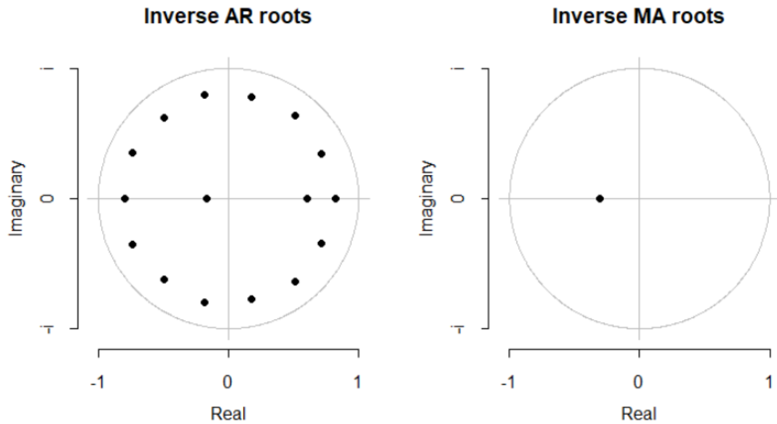


Figure 7: Autoregressive polynomial roots (left) and moving average polynomial roots (right) for model M_3.

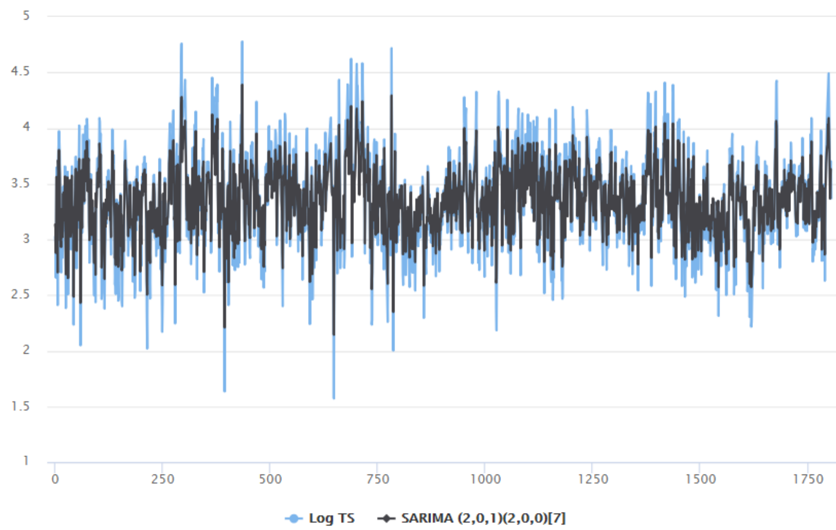


Figure 8

Parameter	ψ_1	ψ_2	θ_1	ϕ_1	ϕ_2
Value	0.434	0.101	0.142	0.024	0.024

Table 3: Final model (M_3) parameters

is given in Table 4 as a point estimated, since the confidence intervals do not preserve their properties through transformations.

Time	Pred. Log	Lower 95%	Upper 95%	Pred. Actual
$T+1$	3.440	2.812	4.068	31.183
$T+2$	3.440	2.661	4.219	31.188
$T+3$	3.435	2.614	4.257	31.041
$T+4$	3.409	2.572	4.245	30.229
$T+5$	3.339	2.497	4.181	28.186

Table 4: Forecast of 5 future points

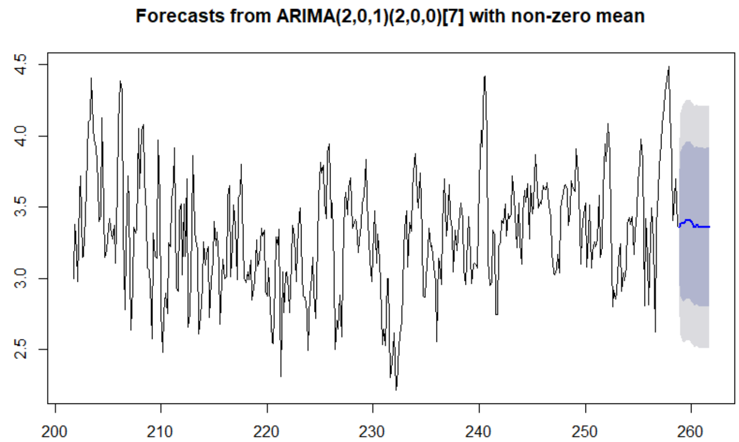


Figure 9: Forecasting of test points with M_3

2.4 Discussion

Since the data considered was stationary, we first applied directly a model with no differencing, which turned out to be the best fit to the data. The model chosen $M_3 : (Y_t) \sim SARIMA(2,0,1) \times (2,0,0)_7$ gives us insight into the past dependencies of the data: $P = 2$ represents the fact that the pollution on a given day is dependent on the pollution in the previous two weeks, where $p = 2$ (together with $P = 2$) represents the dependence on the previous two days (*e.g.*, the pollution in a given monday is dependent on the pollution of the previous weekend, as well as of the two previous mondays and respective weekends); $q = 1$ can be seen as the existence of cumulative effect of randomness (due to unexplained factors) from one day to the next.

We see the first limitation of the model regarding the fitted values, as it consistently underestimates the volatility of the phenomena (Figure 8). In a further analysis, one could explore different processes that would get around this factor. The second limitation observed regards the forecasting uncertainty (Figure 9), which is due to high volatility in the data and the fact that the process itself has a significant random component - given it is a real-world complex phenomena, it has no inherent regularity: one specific day's pollution might be affected by unusual traffic caused by a big accident or by lack of accidents, movement in nearby streets, a riot that blocks the avenue, a holiday, etc.

3 GARCH Model

The family of GARCH processes can be used to model financial time series of asset prices through their log-returns. Considering (P_t) the original time series of asset prices, we obtain the series of *log-returns* (X_t) using:

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

Usually, in time series of this sort, both the sample mean is close to zero and the sample variance lower or similar than 10^{-4}. GARCH processes are parametric models for both the sample mean and the sample variance (if one assumes constant mean, they model the sample variance alone). These models also have the feature of accommodating the heavy-tails present in most financial time series and the usual asymmetric responses in the volatility: it rises higher in response to negative shocks than to positive shocks. In order to accommodate the latter fact we resort to two generalizations of GARCH's: GJR-GARCH and APARCH.

3.1 Data

For the study of GARCH models we will use the log-returns of the closing values for the Nasqad Composite Index between the years of 2015 and 2019. The data consists of **1.257** observations as there are approximately 252 trading days in a year. The time series of closing values can be seen in Figure 10 and their log-returns in Figure 11.

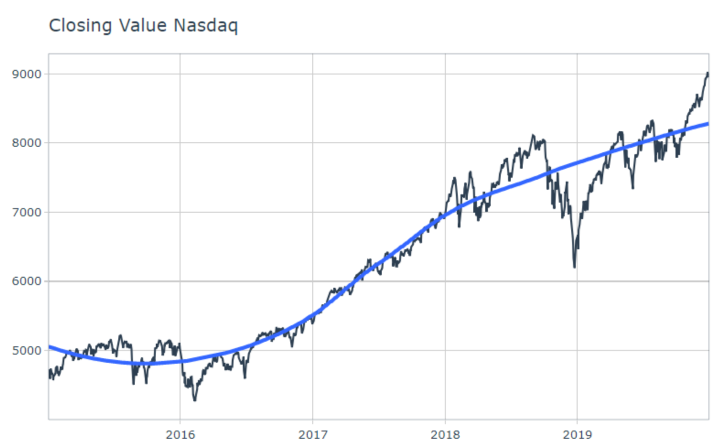


Figure 10: Closing values of Nasqad Index.

3.1.1 Parameters & Volatility

As expected, the log-returns series has a close to zero sample mean ($\mu \approx 0.0005$) and a low sample variance ($\sigma^2 \approx 0.0001$). We can see the volatility of the stock changes over time, *e.g*, the data from 2017 has periods of low volatility. In Figure 12 we can see the monthly volatility as well as the series absolute deviance from the mean making the mentioned variations more clear.

Another interesting measure of a stock's log-return is the *annualized volatility*, which gives us a measure of the overall volatility and is defined as:

$$\sigma_{annualized} = \sigma_{daily} \times \sqrt{252}$$

The Nasqad log-returns (2015-2019) has an annualized volatility of 16.23%.

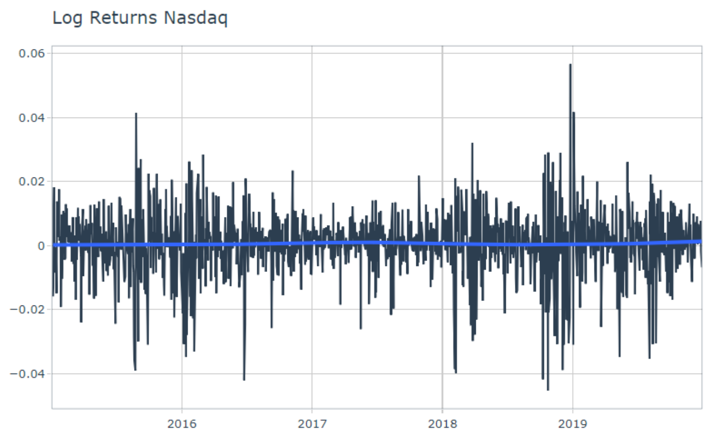


Figure 11: Log-returns of Nasqad Index.

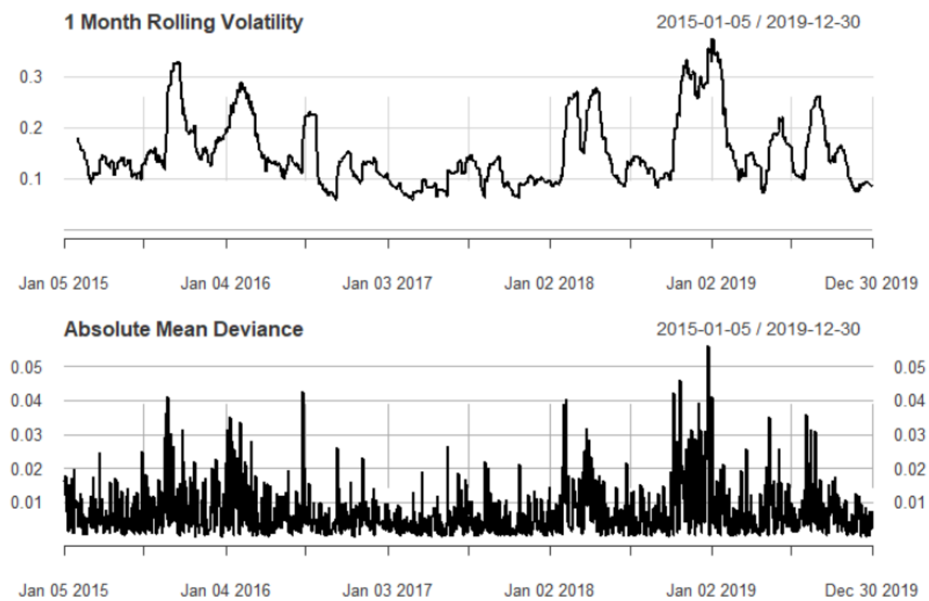


Figure 12: Monthly volatility (up) and absolute mean deviance of log-returns (down).

3.1.2 Distribution

To apply GARCH models we need to assume a distribution (through Z_t) for the data, therefore it is essential to analyze its distribution. The simpler approach is to assume normality but more often than not this is not the best assumption. We can see in Figure 13 the histogram of the log-returns against a Normal distribution. From this comparison two differences are clear: the data has a higher and narrower peak around the mean and it is not symmetric - it is negatively skewed (as it is often the case with financial time series). In order to handle this fact we will consider a **Skewed T Distribution**:

$$Z_t \sim ST(\mu, \sigma, \nu, \xi) \text{ with } \mu = 0, \sigma = 1$$

where the additional parameters (with respect to the normal distribution) ν and ξ represent respectively the number of degrees of freedom - which controls the scale factor and the fatness of tails - and the skewness parameter, where:

- $|\xi| = 1$ yields a symmetric distribution
- $\xi < 1$ yields a negatively skewed distribution
- $\xi > 1$ yields a positively skewed distribution

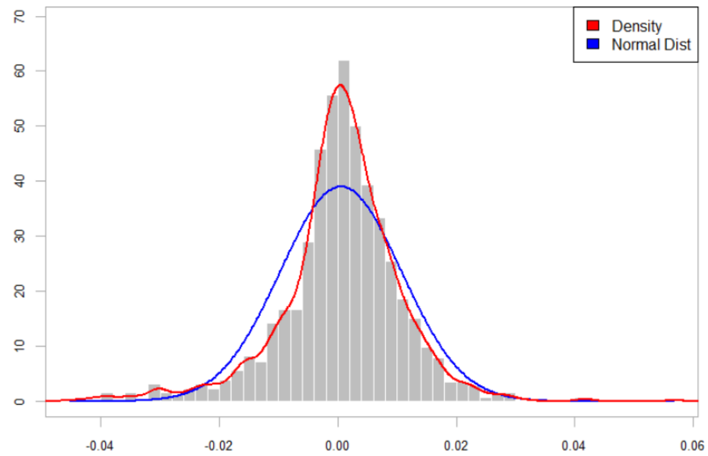


Figure 13: Log-returns distribution against Normal distribution

3.1.3 ACF Analysis

From Figure 14 we can see (as expected for series of this sort) that the sample ACF of the log-returns is negligible for almost all lags but the first, and that the sample ACF of their absolute value is approximately constant and positive for large lags.

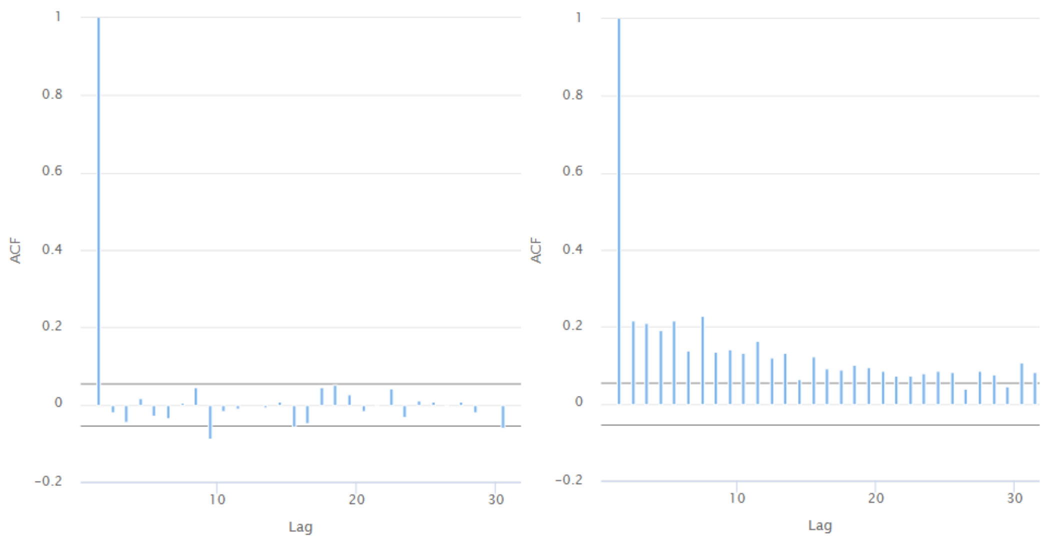


Figure 14: Sample ACF for log-returns (left) and absolute log-returns (right)

3.2 Method

To allow comparisons across different types of models we present only models that have significant specification differences. In Table 5 we can see a summary of the structure of the different models. Considering the distribution of the data we consider mostly models assuming a Skewed T-distribution, only using the normal distribution with the simple GARCH base model to analyze differences.

For the mean modelling, we considered both constant and varying means. The modelling of the mean is done in two different ways:

- **ARMA(1,1)** - assuming mean varies with the past observations of deviance from the mean return
- **ARCH-M** (as given in [4]) - assuming mean varies with the previous variance values

Model Name	$\mathbf{Z_t}$	**Mean Model**	**Variance Model**
G_1	Normal	constant	GARCH(1,1)
G_2	ST	constant	GARCH(1,0)
G_3	ST	constant	GARCH(1,1)
G_4	ST	ARMA(1,1)	GARCH(1,1)
G_5	ST	ARCH-M	GJR-GARCH(1,1)
$\mathbf{G_6}$	**ST**	**ARMA(1,1)**	**APARCH(1,1)**

Table 5: Models' specifications

It is important to note that models including mean modelling have a higher number of coefficients to estimate and therefore are more complex, which is taken into account when choosing the final model through the BIC.

3.3 Results

In Table 6 we can see the summary of the most relevant performance results for the different models.

Model Name	**AIC**	**BIC**	**Likelihood**	**Ljunj-Box p-value**
G_1	-6.542	-6.527	4112.9	0.192
G_2	-6.522	-6.502	4101.0	0.047
G_3	-6.625	-6.601	4166.6	0.237
G_4	-6.634,	-6.607	4177.7	0.079
G_5	-6.667	-6.634	4194.6	0.233
$\mathbf{G_6}$	**-6.694**	**-6.653**	**4213.8**	**0.288**

Table 6: Performance metrics for GARCH models

As expected, the ARCH model (G_2) and the normal GARCH perform significantly worse than other models.
It is interesting to see that, in fact, the best performing models are those who account for the asymmetry of the variance with regard to the sign of the deviance as this is aligned to the mentioned expected asymmetric behavior of stock markets.

3.3.1 Model Choice

The chosen model was the APARCH with ARMA mean - G_6 - highlighted in the table of results. The choice was simple, as it not only maximizes the likelihood of the model generating the actual output, but also minimizes the AIC and BIC scores that take into account model complexity.
It is important, to avoid unnecessary complexity, to analyze if any of the parameters are dispensable, *i.e.*, if there is statistical evidence that the coefficients are not null. For this purpose we use the T-statistic for all parameters. In Table 7 we can see the estimated values for the coefficients as well as the results of the significance tests.

Several interesting insights can be drawn from the table:

- all parameters are important - looking at the tests' *p-values*[1]
- skew parameter $\xi \approx 0.8$ is less than 1, revealing the already observed negative skewness
- scale parameter $\nu \approx 7.8$ is large due to the peak around the mean of the distribution
- $\gamma_1 \approx 1$ is significantly larger than zero, showing the aforementioned different impacts on variance between negative and positive returns (See Figure 15)

[1] ω, albeit not passing the significance test, is not dispensable as it is the intercept coefficient

Parameter	Estimate	Std. Error	t value	Pr(> |t|)
μ	0.001	0.000	11.412	0.00
ψ_1	0.959	0.008	114.024	0.00
θ_1	-0.980	0.000	-4119.905	0.00
a_0	0.002	0.001	1.475	0.14
a_1	0.120	0.015	8.173	0.00
b_1	0.844	0.024	34.763	0.00
γ_1	1.000	0.000	2083.529	0.00
δ	0.780	0.130	5.993	0.00
ξ	0.802	0.034	23.317	0.00
ν	7.797	1.647	4.734	0.00

Table 7: Final Model (G_6) parameters and T-statistic summary

- autoregressive coefficient $\psi_1 \approx 1$ is positive, indicating that higher than average returns are followed by higher than average returns - there is momentum in returns - this can be due to a market's underreaction to news or, considering the value is smaller than 1, that the deviances are merely transitory

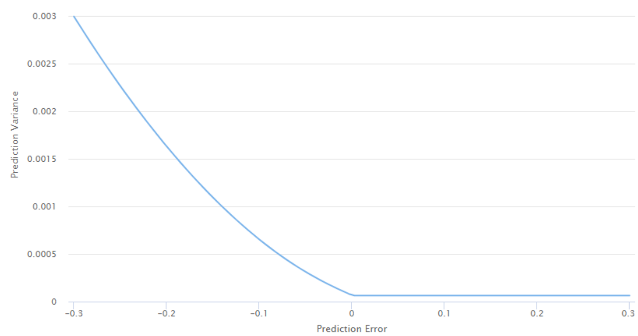


Figure 15: Predicted variance variation with predicted error

To check model validity we also need to look at the residuals, to make sure there is no autocorrelation. We can see in Figure 16 that the residuals' autocorrelation has the expected behavior, being below the threshold for all lag values.

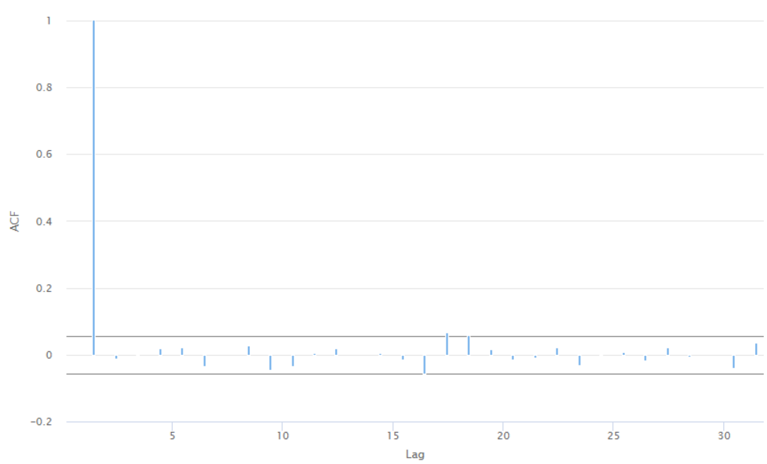


Figure 16: ACF of G_6 residuals

Finally, we can see in Figure 17 the fitted values of the final model for the returns' variance. Notice that the behavior is very similar to the original log-returns volatility, with high and low volatility periods falling within the same range of days.

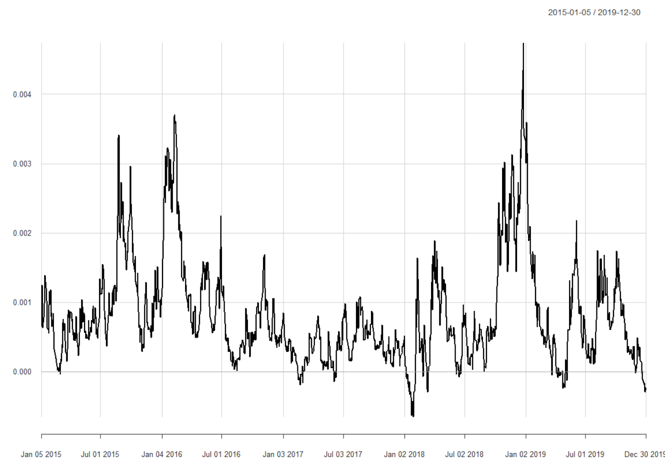


Figure 17: Fitted values of final model (G_6)

3.4 Discussion

Looking at the results we can see the validation of a few theoretical assumptions: the GARCH model are an improvement on ARCH models; APARCH and GJR-GARCH are better than GARCH at dealling with asymetric volatility and heavy tails; assuming a normal distribution on log-returns is not always the best option; and there are improvements in modelling the mean when compared to assuming it is constant.

4 Conclusion

This report allows us to see both the power and limitations of applying SARIMA and GARCH models to real world data.
SARIMA models offer a bigger challenge when dealing data without strong obvious seasonality and with forecasting when non linear transformations are applied to the original data.
GARCH models offer a very efficient and insightful way to model financial data as can be seen by the results.

References

[1] Manuel G. Scotto *Time Series Lecture Notes* IST Lisboa, 2020

[2] Victor M. Guerrero *Time-series analysis supported by Power Transformations* ITAM México, 1993.

[3] Eleftherios I. Thalassinos, Erginbay Ugurlu, Yusuf Muratoglu *Comparison of Forecasting Volatility in the Czech Republic Stock Market* UP Greece, FEAS Turkey, 2014.

[4] Tim Bollerslev, Ray Y. Chou, Kenneth F. Kroner *ARCH Modeling in Finance: A review of the theory and empirical evidence* Journal of Econometrics, 1992.