A Time Series Approach to Forecast COVID 19 Reported Cases in Colombia

Jose Miguel Gil Valencia
jmgilv@eafit.edu.co

Juliana Lalinde Velasquez
jlalindev@eafit.edu.co

Nicolás Rengifo Campo
nrengifoc@eafit.edu.co

January 29, 2021

1 Introduction

The world is currently facing one of its biggest challenges: a global pandemic. What began as a few viral pneumonia cases in Wuhan, China, has now been recognized as a disease caused by a new coronavirus called SARS-CoV-2, more commonly known as COVID-19. There have been around 55,678,163 reported cases all around the world, with approximately 1,338,088 deaths. Since the start of the spread of COVID-19, countries have been faced with the difficult task of trying to understand the behaviour of the virus. In hopes of succeeding at this mission different approaches have been taken, studying economic, medical and social phenomenon. The total number of reported cases represents a key variable to analyze the emergency level a country is facing by allowing governments to estimate the number of deaths, risk of infection and determine containment measures, like mandatory quarantine. Taking this into account, predicting the total cases that will be reported in a specific country becomes a valuable tool for decision making processes.

A group of researchers [Maleki *et al.*, 2020] modeled the total number of confirmed and recovered COVID-19 cases around the world using an auto regressive time series model. In their case, in order to properly represent the asymmetrical distribution of the errors, a TP-SMN-AR model was implemented. This includes both the symmetric Gaussian and asymmetric heavy-tailed non-Gaussian models. For this specific study, data of total confirmed cases was used from January 22nd and data of total recovered cases was used from Feburary 2nd, both time series accumulating data until April 30th. Some suitable transformations were also applied to obtain stationary data (since the original was increasing and had signs of a trend) and, taking into account model selection criteria like Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Ljung-Box tests on the residuals, among others, the researchers conclude that a TP-SMN-AR model has better results than counterparts that assume a symmetric Gaussian behaviour for the errors.

The following work starts by defining some of the mathematical models and functions that are key to the problem. Then, some data analysis is shown in order to identify which are the best models and, in the fourth section, their results are shown. Finally, the forecast results are compared and the final conclusions stated. All in the hopes of analyzing how well the models studied through the semester fit and forecast over the cumulative daily reported cases of COVID-19 in Colombia

2 Mathematical Models

Before different models are tested to predict the cumulative reported cases in Colombia, some formal definitions are presented. The $ARIMA$ and $GARCH$ models are defined, assuming the reader has understanding of the followings basic models:

- AR(p)
- MA(q)

Then, the $Box - Cox$ transformation and the correlation functions are defined.

2.1 ARIMA(p,d,q) Model

The Auto Regressive Integrated Moving Average model with order p, d, q represented with lag operator is defined as follows.

$$\phi(L)(1-L)^d X_t = \theta(L)\varepsilon_t \tag{1}$$
$$(1-\phi_1 L - \cdots - \phi_p L^p)(1-L)^d X_t = (1-\theta_1 L - \cdots - \theta_q L^q)\varepsilon_t$$

Where

- p: order of $AR(p)$ model.
- d: number of differences required to make the time series stationary.
- q: order of $MA(q)$ model.

Then, the stationary process W_t with the d differences applied to X_t is an $ARMA(p,q)$ as follows.

$$\phi(L)W_t = \theta(L)\varepsilon_t \tag{2}$$

2.2 GARCH(p,q) Model

The Generalized Auto Regressive Conditional Heteroscedasticity model with order p, q is defined as follows.

$$U_t = \sigma_t \varepsilon_t \tag{3}$$

Where $\sigma_t^2 = E(U_t^2/\Omega_{t-1})$, then:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i U_{t-i}^2 + \sum_{j=1}^{p} \delta_j \sigma_{t-j}^2 = \alpha_0 + \alpha(L)U_t^2 + \delta(L)\sigma_t^2$$

2.3 Box-Cox transformation

This transformation of the dependent variable ($Box-Cox$ [Box & Cox, 1964]) is a transformation from a non lineal to normal shape. The value of λ varies in $[-5, 5]$, this parameter is the optimal value -for the specific data- to approach a normal distribution. The transformation follow the form of equation (4).

$$X_t^\lambda = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ Log(X_t) & \text{if } \lambda = 0 \end{cases} \tag{4}$$

The equation (4) only works for positive data ($X_t > 0$), for negative values ($X_t > -\lambda_2$) equation (5) is used.

$$X_t^\lambda = \begin{cases} \frac{(X+\lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ Log(X_t + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases} \tag{5}$$

2.4 Auto Correlation Function

The Auto Correlation Function, known as ACF, gives the auto-correlation values of the series. It describe how well the current value of the data is related with the past values. The plot of the ACF represents the values of the auto-correlations -between the data- and the significance level, to identify statistically significant values. It also shows trends, seasonal behaviour, cyclic behaviour, etc. The ACF values are calculated by the following equation.

$$\hat{\rho_j} = \frac{\hat{COV}(X_t, X_{t-j})}{\hat{VAR}(X_t, X_{t-j})} \tag{6}$$

2.5 Partial Auto Correlation Function

The Partial Auto Correlation Function or $PACF$ is similar to ACF (section 2.4), but it finds the correlation of the residuals with the next lag value. The plot of the $PACF$ represents the values of the partial auto-correlations -between the residuals of the data- and the significance level, to identify statistically significant values. The plot shows hidden information in the residuals.

3 Data Analysis

An analysis to identify which models best represent the cumulative daily reported cases in Colombia was executed. The data set considered is the official report of the Colombian Ministry of Health through the National Institute of Health [Colombian Ministry of Health, 2020]. An initial exploratory analysis is shown in this section. March third of 2020 is used as day one in all of the graphics. This project is developed in the statistical software R.

Original data

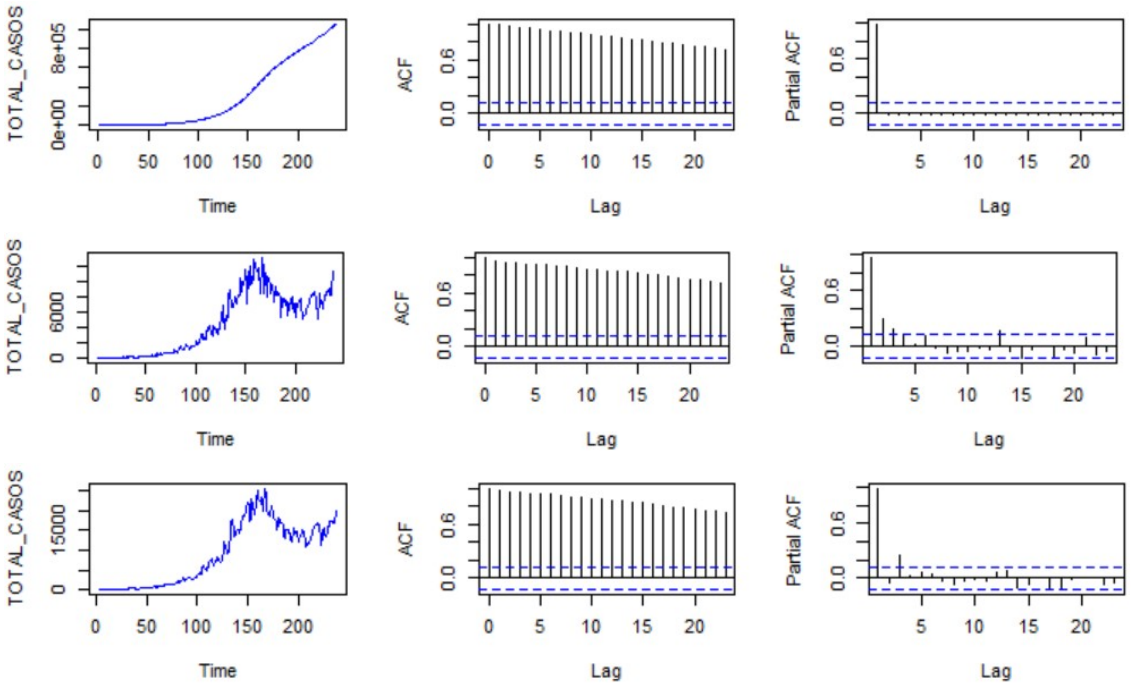

Figure 1: Level Variable. First and second difference, acf and pacf.

- The first row represent the original data. The first plot shows the increasing behaviour of the cases in time. The ACF plot of the data shows a small decrease between lags, indicating that there seems to be significant correlations for a large number of lags. The PACF show only one significant correlation, which indicates that the high number of correlations observed in the ACF plot may be due to the propagation of the autocorrelation in that lag. This indicates that a AR model can represent the data.
- The second row represents the first difference of the original data. The ACF shows the small decrease and the PACF shows some significant values, that means, as explained above, an AR model with some lags.
- The final row represents second difference of the original data. The ACF shows small decrease and the PACF some significant lags, that means, as explained above, an AR model with some lags.

These interpretations shows that the AR model is a good option to use, based on ACF and PACF plots. The interpretation of the ACF and PACF plots gives a suggestion for the values of the parameters, that needs to be verified using a hypothesis test.

Natural logarithm of the variable

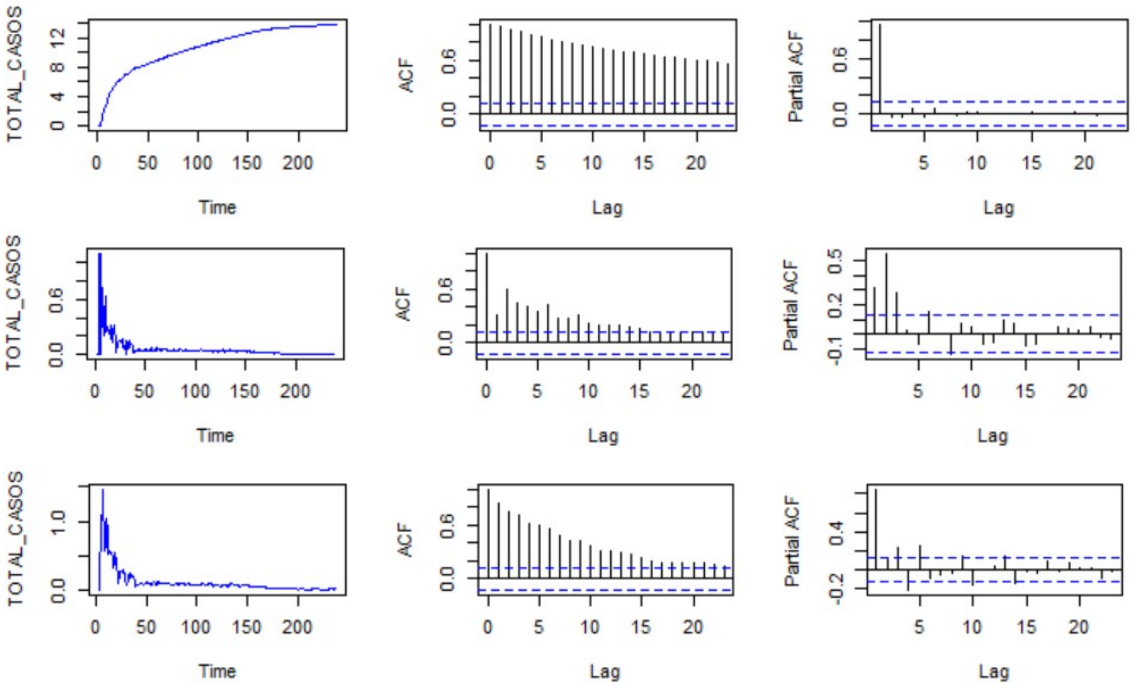

Figure 2: Natural logarithm of Variable. First and second difference, acf and pacf.

- The first row represent the original data with a logarithmic transformation. The first plot shows the logarithmic behaviour of the cases in time after the transformation was applied. The ACF plot of the data shows a small decrease between lags, indicating that there seems to be significant correlations for a large number of lags. The PACF show only one significant correlation, which indicates that the high number of correlations observed in the ACF plot may be due to the propagation of the autocorrelation in that lag. This indicates that a AR model with one lag can represent the data.
- The second row represents the first difference of the original data with a logarithmic transformation. The ACF of the first difference shows small decrease but show increase at the first part, and the PACF show some significant values at the start but non consecutive. that suggest a ARIMA model, both effects at the same time
- The final row represents second difference of the original data with a logarithmic transformation. The ACF shows a exponential decrease and the PACF show a very small decrease, that suggest a AR model with multiple lags.

Its important to remember that these results need to be verified with a hypothesis test before implementing them.

Box-Cox transformation of the variable

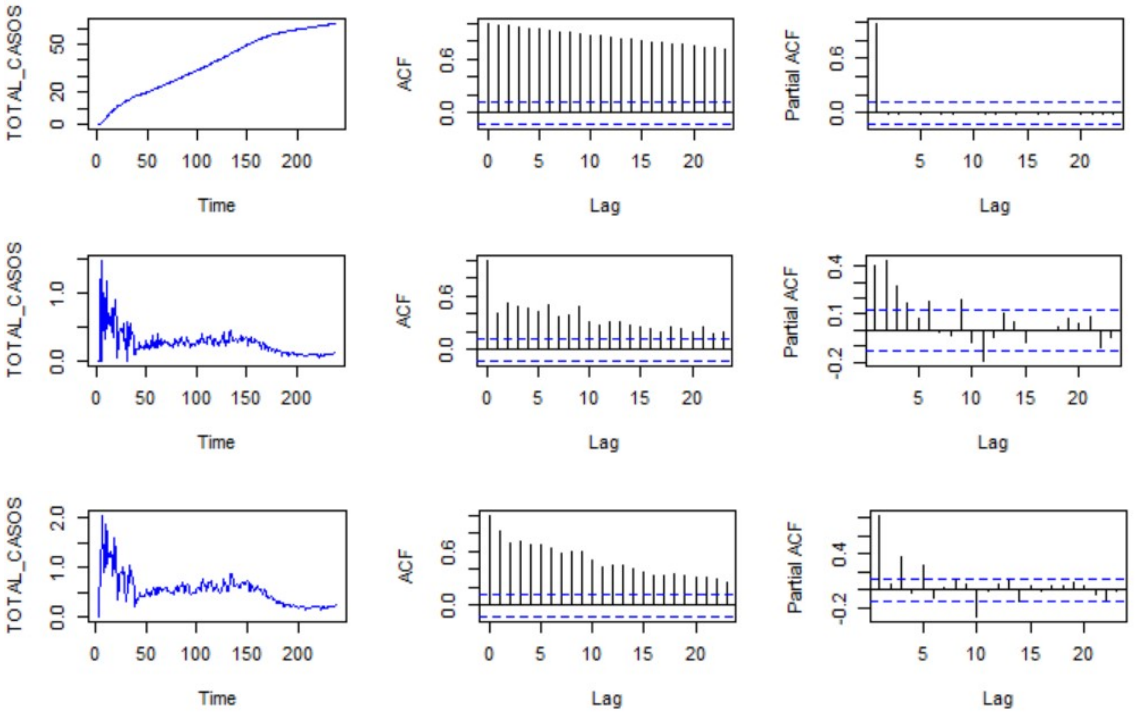

Figure 3: Variable with Boxcox transformation. First and second difference, acf and pacf.

- The first row represent the original data with a Boxcox transformation. The first plot shows the increasing behaviour of the cases in time after the transformation was applied. The ACF plot of the data shows a small decrease between lags, indicating that there seems to be significant correlations for a large number of lags. The PACF show only one significant correlation, which indicates that the high number of correlations observed in the ACF plot may be due to the propagation of the autocorrelation in that lag. This indicates that a AR model with one lag can represent the data.
- The second row represents the first difference of the original data with a Boxcox transformation. The ACF of the first difference shows small decrease but non uniformly, and the PACF shows some significant values at beginning and some significant values non consecutive. That suggest a ARIMA model, some effects of both models.
- The final row represents the second difference of the original data with a Boxcox transformation. The ACF shows decrease and the PACF some significant values but non consecutive. That surges a AR model or ARIMA model, because the ACF's decrease in non uniformly.

Its important to remember that these results need to be verified with a hypothesis test before implementing them. The AR or ARIMA model may adjust well to the data, but the tests need to be used first.

Now we are going to perform a unit root test for the data that we are using for estimate our models. Specifically, we will perform Augmented Dickey-Fuller test, Phillips-Perron test and KPSS test for the second difference of natural logarithm and Box-Cox transformation of our variable. The results can be seen in the following table.

Stationarity			
	Aug. Dickey-Fuller	Phillips-Perron	KPSS
2 difference LN	6,96744E-06	0,01	0,01
2 difference BC	0,01948941	0,01	0,01

Table 1: Unit root test for stationary

*reject if $values < 0.05$

- First row: reject all the test. For D-F the data is stationary, for P-P the data is stationary and for KPSS the data is not stationary.
- Second row: reject all the test. For D-F the data is stationary, for P-P the data is stationary and for KPSS the data is not stationary.

In conclusion, the KPSS test is ignored because the others prove stationary behaviour and the plots of the data, with visual test, show the same type of behaviour. So the data in both cases is stationary.

4 Results

4.1 Chosen Models

We validate the incoming models by testing on their residuals Ljung-Box and White Neural-net test.
For all models, the LjungBox test (test independence) is rejected, so the error data are not independently distributed; they exhibit serial correlation.
In White test where use it to validate that the respective model use enough parameters to represent the data generating process (DGP), for model 1 and model 2 don't reject linearity in mean of the residuals (error), and the model 3 and 4 reject linearity in mean.

Validation		
	LjungBox	Neuralnet White
Mdl1	2,63E-10	0,07047465
Mdl2	2,06E-12	0,07047465
Mdl3	2,18E-09	0,01952116
Mdl4	1,00E-13	0,03129175
Mdl5	3,81E-03	

Table 2: Validation

4.1.1 Natural logarithm, arima(6,2,0)

The decision to choose this model comes from the need to transform the original data. We know that as was explained in the section 3, the variable that we are studying has an exponential behaviour so naturally the first transformation that comes in mind is to take the natural logarithm of the serie. The coefficients of the arima model associated are chosen in order to create an auto regressive process in the second difference given the acf and pacf analyzed in the past section. Another reason to decide a sixth auto regressive order is because it is estimated that six days is the mean time that it takes to show symptoms of the COVID-19 [Lauer *et al.*, 2020].

4.1.2 Natural logarithm, arima(14,2,0)

Under the same criteria to chose the past model, we decide to include another arima model in the second difference of the natural logarithm of the variable. The difference is that in this model we consider a higher auto regressive order trying to capture as much information as possible. The model was validated as was explained in the table at the beginning of this section. A more specific reason to chose the parameters of this model is that fourteen days are the mean time of recuperation for COVID-19 [Ganyani *et al.*, 2020].

4.1.3 Boxcox transformation, arima(6,2,2)

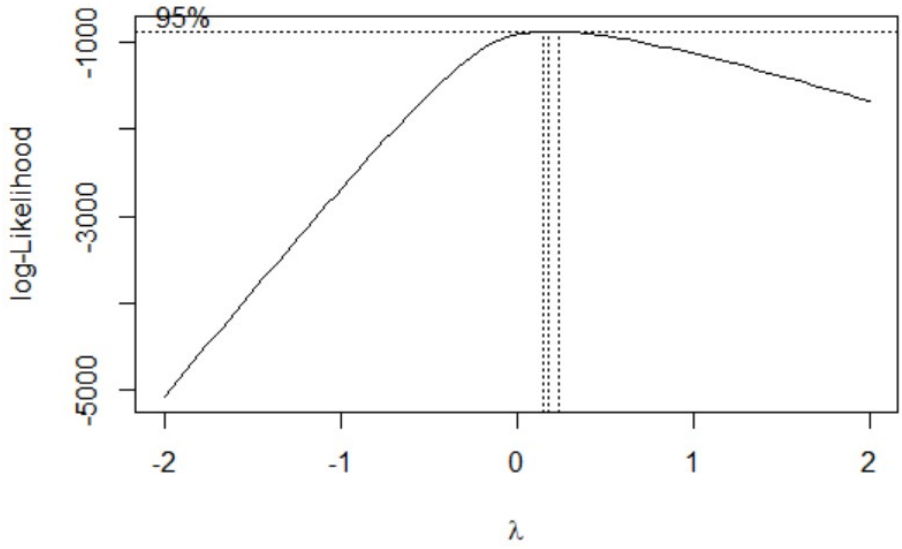

Figure 4: Boxcox transformation parameter Log-Likelihood.

Another idea that we wanted to consider was to include a model that depends of the boxcox transformation. As can be observed in the figure 4, according to the idea that this transformation proposes, a transformation needs to be done to reach the weak stationary condition. Also we can see the parameter of the transformation that is required to also maximize the log-likelihood. Finally the parameters of the arima process are chosen thanks to both acf and pacf of the second difference of the boxcox transformation exposed in figure 3.

4.1.4 Natural logarithm, autoarima(3,2,0)

The auto arima command in R can offers us a good idea of the model to choose, because it finds the one that minimizes information criteria.
Decide to use it to compare our models (based in tests, ACF and PACF information) with this one, to know if our decisions were correct.
This model show that the biggest component of the model is a AR model, and our models too, so that means our decisions were appropriated. Second, the *autoarima* use the second difference to become the data stationary, our models use the same difference. Finally, our models don't have MA effects, the value of q is zero ($MA(q)$); same happen in the auto arima model.
So, our theoretical arima decisions were a good approach, compare between our models and this one.

4.1.5 First difference, Garch(1,1)

Finally our last idea was based in using a Garch methodology trying to fit the first difference of our level variable, this would be the daily confirmed cases. The decision is taken since this serie looks strongly heterokedastic. The Garch (1,1) is chosen by its desirable properties, as explained by [Nelson, 1990].

4.2 Forecast

We decided to do a forecast with a time horizon of ten days, using our proposed models we can see the graphic results in figure 5.

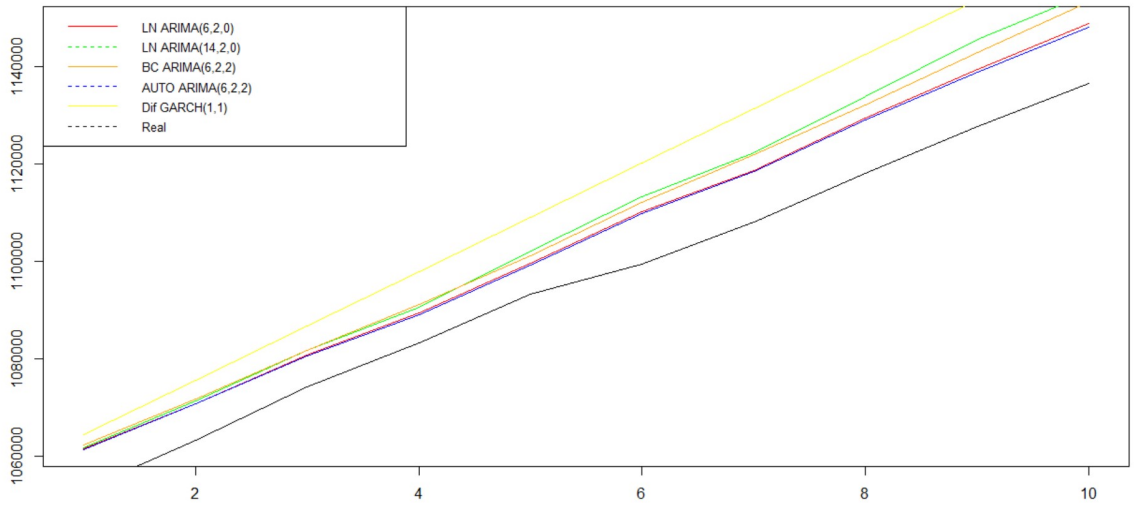

Figure 5: Forecast performance of proposed models.

Now we are going to see the performance of our models. We consider the percentage error defined as:

$$\frac{abs(z - \hat{z})}{z}$$

Where $\hat{z}$ is an estimation of the real observed value in the time series Z.

	Percentual Daily Error				
	Mdl1	Mdl2	Mdl3	Mdl4	Mdl5
Obs 1	0,007888	0,008059	0,008706	0,007787	0,010618
Obs 2	0,007179	0,00773	0,007958	0,007179	0,011598
Obs 3	0,006043	0,007046	0,006958	0,005913	0,011608
Obs 4	0,005524	0,006652	0,007264	0,005238	0,013384
Obs 5	0,005818	0,008031	0,007186	0,005414	0,014384
Obs 6	0,009841	0,012618	0,011451	0,009394	0,018871
Obs 7	0,009481	0,012789	0,01238	0,00925	0,020943
Obs 8	0,010236	0,014224	0,012671	0,009845	0,021879
Obs 9	0,010243	0,015816	0,013361	0,00975	0,022918
Obs 10	0,010797	0,016308	0,014755	0,010194	0,024875

Table 3: Forecast error

We must be carefully analyzing these results, the might look like a little error, but it is because of the data scale. Just looking at the graphic of the forecast we can see that there is a great difference between the real data and the predicted one. We had validated our models by some different tests, yet still, as can be observed, that does not guarantees that we will have a good fit in the forecast contrasting.
But, if compare all models without the real data, you can see that all those are very close, which mean that different models can be used for represent the same data. The models with BoxCox transformation don't fit very well, same with the GARCH model.
The best approaches are the autoarima model and the ARIMA(6,2,0), both with the natural logarithm transformation over the data.

5 Conclusions

The chosen models and the transformations used might not be the proper to represent the interest variable. The variable considered has a exponential nature, so a better approach could be to use exponential family models. We did some test to validate our models and the data that feeds them,

nonetheless the fit in the forecast is not as well as expected. A significant model does not guarantee a good performance.
The assumptions that the models that we propose require might be too strong for this variable, for example the stationary conditions.

References

[Box & Cox, 1964] Box, G. E. P., & Cox, D. R. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**(2), 211–252.

[Ganyani *et al.*, 2020] Ganyani, Tapiwa, Kremer, Cécile, Chen, Dongxuan, Torneri, Andrea, Faes, Christel, Wallinga, Jacco, & Hens, Niel. 2020. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, **25**(17), 2000257.

[Lauer *et al.*, 2020] Lauer, Stephen A, Grantz, Kyra H, Bi, Qifang, Jones, Forrest K, Zheng, Qulu, Meredith, Hannah R, Azman, Andrew S, Reich, Nicholas G, & Lessler, Justin. 2020. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, **172**(9), 577–582.

[Maleki *et al.*, 2020] Maleki, Mohsen, Mahmoudi, Mohammad Reza, Wraith, Darren, & Pho, Kim-Hung. 2020. Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, 101742.

[Nelson, 1990] Nelson, Daniel B. 1990. Stationarity and persistence in the GARCH (1, 1) model. *Econometric theory*, 318–334.