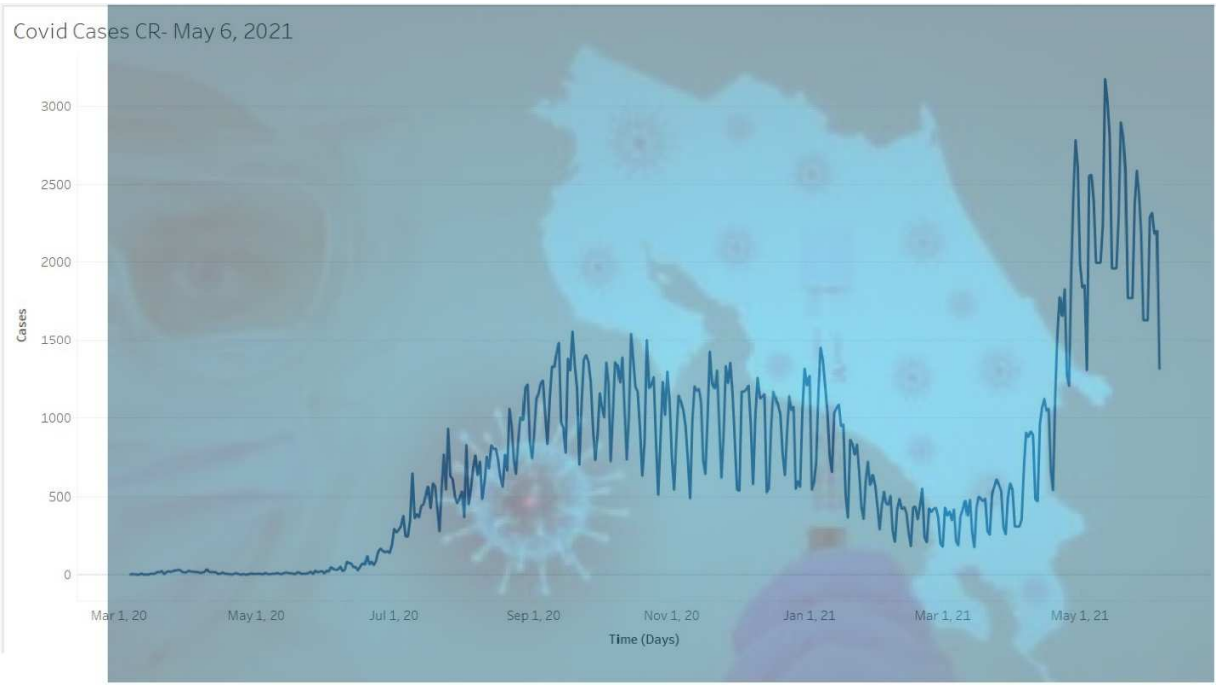
Analysis and Forecasts Covid-19 Third Wave in Costa Rica

An evaluation of media

June 2021

By: Juan Carlos Benavides B



Topics:

- Time series model
- Seasonal ARIMA
- Forecast
- Statistics

Tools:

- R

INTRODUCTION

The COVID-19 pandemic started in Costa Rica on March 5th when authorities detected the first case in the country.

Throughout 2020, the Costa Rican government managed the situation well, setting rules to avoid contact and increasing available ICU beds. When the curve peaked, between the beginning of September and the end of December, the capacity in hospitals never were overwhelmed. During this period there was a high volume of samples tested every day with the highest reported cases being 1,500 per day.

Then after a period of reporting fewer cases every day, the curve started to increase again announcing another peak was coming. In April the media and the authorities were expecting the maximum cases in May. Some part of the media were prudent with their reporting and did not announce forecasts but others were more negative (or alarmists) even publishing Costa Rica was going to reach the amount of 10,000 cases per day, overwhelming the capacity of hospitals and causing chaos in the health system and a catastrophe for the country.

For example the media CRHoy.com, an important digital media outlet in Costa Rica, said on May 4, 2021:

Pronóstico de 10 mil casos diarios por COVID-19 no se contendría con restricción vehicular

David Ulloa ✉ Mayo 4, 2021 7:12 am

(CRHoy.com) Las últimas semanas, Costa Rica ha regi **acelerado aumento en casos positivos** diarios por llegando a superar hasta los 2000 contagios varios días.

Sin embargo, esta no sería la peor etapa por la que atra país. Según datos del Centro Centroamericano de Pobla de la Universidad de Costa Rica (UCR), la tasa de reprodu COVID-19 pasó de **"R=1,22" a un alarmante "R=1,33"** en ta semana.

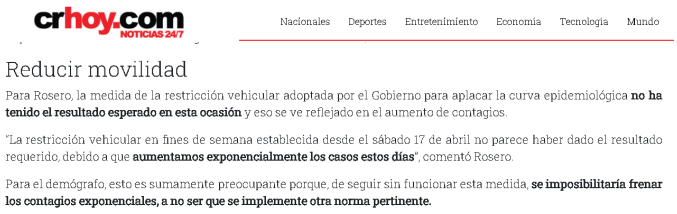
Con esta tasa, las proyecciones son catastróficas, ya que **que haya cerca de 10 mil contagios diarios dentro de un m**

este el escenario neutro.

Figure 1. Part of the article published by CRHoy.com.

Headline translation: Forecast of 10,000 cases per day will not be stopped with vehicle transit restriction.

Further in the article we can read:



crhoy.com NOTICIAS 24/7 Nacionales Deportes Entretenimiento Economía Tecnología Mundo

Reducir movilidad

Para Rosero, la medida de la restricción vehicular adoptada por el Gobierno para aplacar la curva epidemiológica **no ha tenido el resultado esperado en esta ocasión** y eso se ve reflejado en el aumento de contagios.

"La restricción vehicular en fines de semana establecida desde el sábado 17 de abril no parece haber dado el resultado requerido, debido a que **aumentamos exponencialmente los casos estos días**", comentó Rosero.

Para el demógrafo, esto es sumamente preocupante porque, de seguir sin funcionar esta medida, **se imposibilitaría frenar los contagios exponenciales, a no ser que se implemente otra norma pertinente.**

Figure 2. Part of the article published by CRHoy.com.

Decreasing of movability: According to Romero (a demographer) the measures of transit restriction imposed by the government since April are not working and in May Costa Rica health system capacity will collapse.

The whole article can be read in the next link:

https://www.crhoy.com/nacionales/pronostico-de-10-mil-casos-diarios-por-covid-19-no-se-contendria-con-restriccion-vehicular/

According to the experts consulted by CRHoy in the article, more restrictive measurements will be necessary.

But there are several questions we should pose after reading the article:

- Based on the size sample of Costa Rica authority's test, is it possible to reach 10 000 cases per day?
- Are the rules set by the government sufficient? Or do Costa Ricans need more and stricter rules?
- Is the article realistic or alarmist?

A model on time series performed weekly on May 2021 provides insights to determine if the article makes an accurate prediction and to answer these questions.

This report contains the results and the conclusions from the model.

METHODOLOGY

Data has been taken from the daily official report of the government of Costa Rica.

This analysis is based on the report on positive cases and it does not consider the report of the death count. A comparison between the two-time series: daily cases and the daily deaths report could be a better approach for a better prediction, but for the purpose enumerated in the introduction, using time series based solely on cases is enough.

It is important to explain that the government of Costa Rica (since September 2020) has been issuing reports of weekend data for COVID-19 every Monday, it means every Monday cases detected from Friday, Saturday, and Monday are reported. To avoid distortions in the data analysis for weekends the data has been considered as the mean of the three days.

The statistical method chose for the analysis is Seasonal ARIMA (SARIMA).

The stochastic analysis will be performed in R.

The performing of the forecast has been done as follows:

Estimated Date	**Prediction**	
	Initial date	Final date
5-May-2021	6-May-2021	17-May-2021
20-May-2021	21-May-2021	1-Jun-2021
28-May-2021	29-May-2021	9-Jun-2021
9-Jun-2021	10-Jun-2021	21-Jun-2021

The complete development of analysis was made for the estimate calculated on May 5, 2021, then for the following week's forecast. The analysis is exactly the same because the behavior is the same as well as the model. To avoid repetition only the main results for the forecasts performed on the dates: May 20th, May 28th, and June 9th are shown.

ANALYSIS

Exploring initial data

Figure 3 shows the evolution of Covid-19 in Costa Rica, since the beginning of March 5, 2020 until May 5, 2021. As the graph shows, there is different behavior in the first 100 days, this follows the strict measures the authorities imposed to avoid the spread. But after some months in order to keep in balance the impact on the economy and the management of Covid-19, the authorities decided to relax the measures, which is remains similar to current controls.

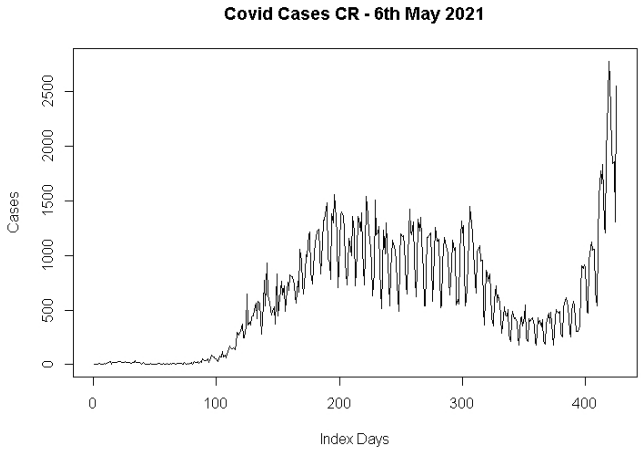


Figure 3. Total cases from March 2020 until May 6, 2021

The plot of the autoregressive correlation function (ACF) suggests there are correlations among the data with some periodicity. Figure 4 corresponds to ACF and figure 5 to partial autocorrelation function (PACF).

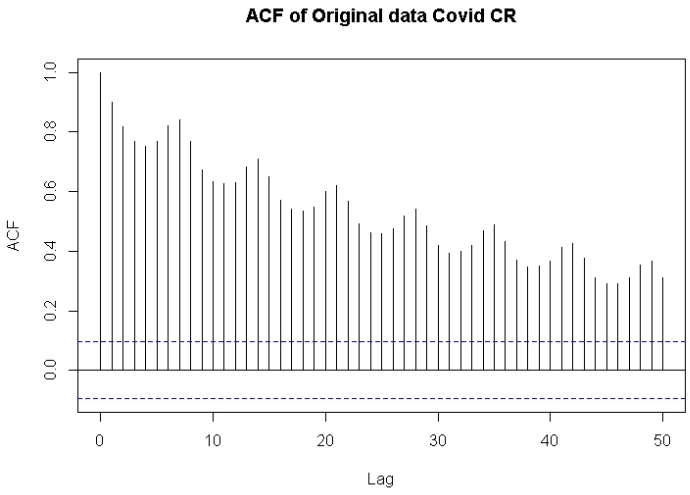


Figure 4. Autocorrelation function.

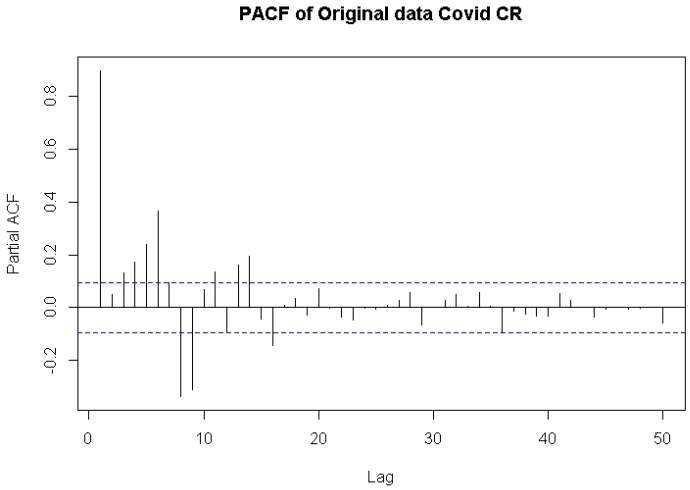


Figure 5. Partial autocorrelation function.

Data Transformation

In order to obtain the best transformation over the data that provides white noise behavior, and to start the stochastic analysis, it is convenient to cut the first 100 days and to analyze the remaining data set. As previously stated, it is easy to see the behavior at the beginning of Covid-19 is completely different due to the changes in measures imposed by authorities.

After some trial and errors (to see the code for more detail) the transformation obtained (the one that offers white noise or most similar) has the form:

$$transformed\_data = diff(diff(\ln(data), 7))$$

This implies there is a seasonal period of seven days, plotting the transformed data the following figure is observed:

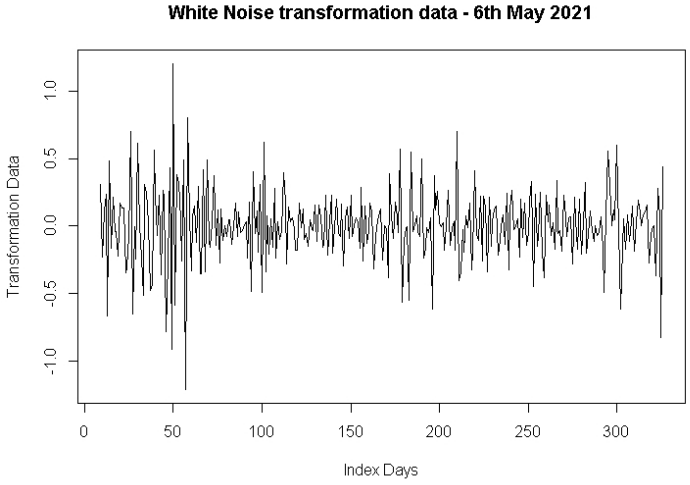


Figure 6. Plot after transformation over data.

Figure 6 shows there are some outliers, especially close to day 50, but Ljung – Box test gives a p-value of almost zero, according to theory the outliers will not impact the analysis.

Building the SARIMA model

The plots for ACF and PACF over the transformed data will allow to determine the degrees in the autoregressive and moving averages components.

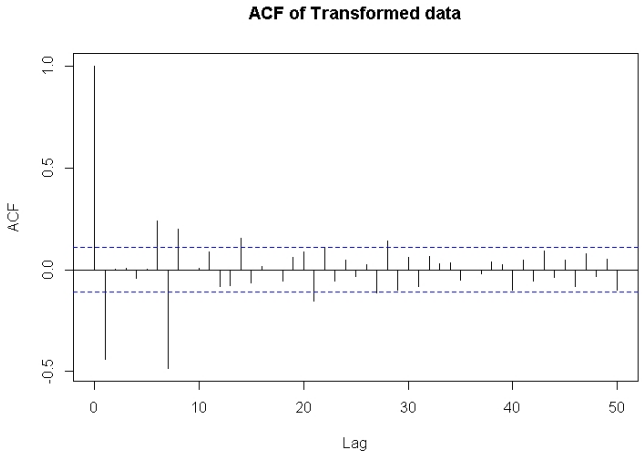


Figure 7. Graph of ACF over transformed data.

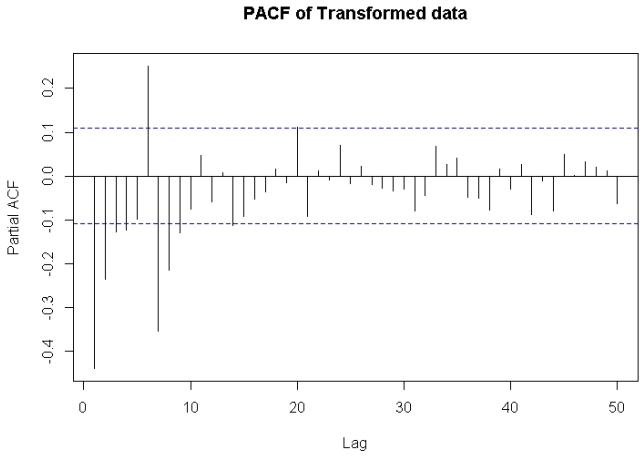


Figure 8. Graph of PACF over transformed data.

The lags in figure 7 show the possibilities for the order of moving average component are zero or one, and the possibilities in the order of seasonal moving averages are zero or one.

Figure 8 is harder to interpret, but in a similar way, it shows the possibilities in the order of autoregressive terms could be p=0,1,2,3,4 and for seasonal autoregressive terms the possibilities are zero or one.

Since what is looked at is the form for SARIMA Model:

$$SARIMA\ (p,d,q,P,D,Q)_s$$

Where:

p: order of autoregressive terms.

d: first difference over transformed data.

q: order of moving average.

P: order in seasonal autoregressive terms.

D: second difference over transformed data.

Q: seasonal moving average terms.

s: seasonal period.

As stated earlier, the seasonal period in the transformation is equal to seven, and because it was applied difference twice then the value of d is equal one and finally D is equal one.

Then testing the different values obtained from ACF and PACF plots: p,q,P,Q in SARIMA models it is possible to choose the best model according to AIC, SSE, and p-values.

From the possibilities for: SARIMA (p,1,q,P,1,Q,)$_7$ the following models are calculated:

```
0 1 0 0 1 0 7 AIC= 81.24087  SSE= 23.8877  p-VALUE= 5.266898e-12
0 1 0 0 1 1 7 AIC= -34.75092  SSE= 16.19725  p-VALUE= 2.49454e-10
0 1 0 1 1 0 7 AIC= -9.193627  SSE= 17.74003  p-VALUE= 2.244327e-11
0 1 0 1 1 1 7 AIC= -32.7527  SSE= 16.19562  p-VALUE= 2.48884e-10
0 1 1 0 1 0 7 AIC= -17.46211  SSE= 17.37184  p-VALUE= 0.8842284
0 1 1 0 1 1 7 AIC= -123.0357  SSE= 12.22412  p-VALUE= 0.9298975
0 1 1 1 1 0 7 AIC= -103.6821  SSE= 13.07869  p-VALUE= 0.626899
0 1 1 1 1 1 7 AIC= -121.0736  SSE= 12.22634  p-VALUE= 0.9184689
1 1 0 0 1 0 7 AIC= 13.77894  SSE= 19.18709  p-VALUE= 0.0001937882
1 1 0 0 1 1 7 AIC= -91.60211  SSE= 13.50117  p-VALUE= 0.001023915
1 1 0 1 1 0 7 AIC= -72.67303  SSE= 14.43504  p-VALUE= 0.002585914
1 1 0 1 1 1 7 AIC= -89.60289  SSE= 13.50034  p-VALUE= 0.001024398
1 1 1 0 1 0 7 AIC= -22.72265  SSE= 16.95656  p-VALUE= 0.5122809
1 1 1 0 1 1 7 AIC= -121.8621  SSE= 12.19295  p-VALUE= 0.9930834
1 1 1 1 1 0 7 AIC= -103.7328  SSE= 12.99068  p-VALUE= 0.9089214
1 1 1 1 1 1 7 AIC= -119.9499  SSE= 12.19488  p-VALUE= 0.9911647
2 1 0 0 1 0 7 AIC= -2.042277  SSE= 18.13501  p-VALUE= 0.08911109
2 1 0 0 1 1 7 AIC= -110.627  SSE= 12.6171  p-VALUE= 0.07615764
2 1 0 1 1 0 7 AIC= -86.60796  SSE= 13.72723  p-VALUE= 0.03761595
2 1 0 1 1 1 7 AIC= -108.7161  SSE= 12.60715  p-VALUE= 0.08517612
2 1 1 0 1 0 7 AIC= -26.90385  SSE= 16.6282  p-VALUE= 0.8465922
2 1 1 0 1 1 7 AIC= -119.873  SSE= 12.19171  p-VALUE= 0.9910522
2 1 1 1 1 0 7 AIC= -103.0251  SSE= 12.93696  p-VALUE= 0.9985437
3 1 0 0 1 0 7 AIC= -5.082793  SSE= 17.84713  p-VALUE= 0.57896
3 1 0 0 1 1 7 AIC= -113.395  SSE= 12.43062  p-VALUE= 0.2819304
3 1 0 1 1 0 7 AIC= -90.39803  SSE= 13.47622  p-VALUE= 0.1181439
3 1 1 0 1 0 7 AIC= -26.02486  SSE= 16.56992  p-VALUE= 0.9795179
4 1 0 0 1 0 7 AIC= -8.116688  SSE= 17.56327  p-VALUE= 0.7318844
```

Figure 9. Outcomes of SARIMA models knowing d=1, D=1, s=7 and for different combinations of p,q,P,Q.

Now for the criteria of AIC, minimum errors and high p-values (according to theory) the chosen model is:

$$SARIMA\ (0,1,1,0,1,1)_7$$

The evaluation of residuals analysis shows:

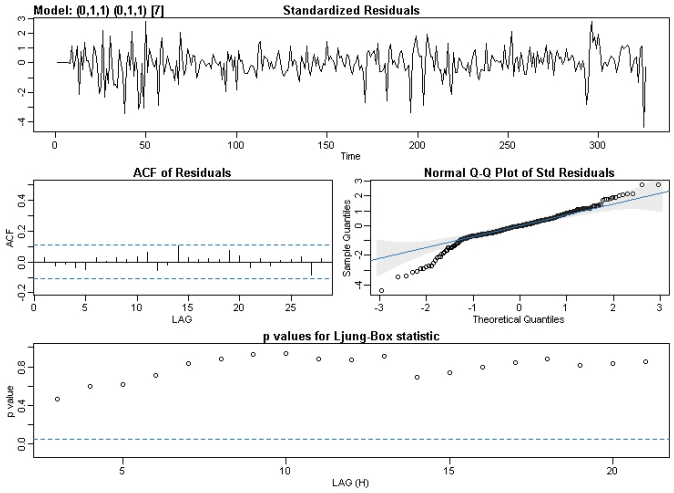


Figure 10. Residuals analysis plots.

Figure 10 shows:

- There is no pattern in the residuals plot.
- Residuals do not have autocorrelation.
- Normal Q-Q plot is almost a line.
- P-values according to Ljung-Box test are high.

In summary, all of this is what is expected in the residuals analysis according to theory, and it can be concluded the model SARIMA (0,1,1,0,1,1,)$_7$ fits the data and the forecast can be performed.

From the analysis following coefficients are obtained:

$$ma_1 = -0.5995$$

$$sma_1 = -0.6433$$

And the correspondent p-values to these coefficients (tend to be zero) tell they are significant.

Also we can see:

$$\sigma^2 = 0.03723$$

Then according to theory, the polynomial that governs the time series can be constructed using the general expression:

$$\Phi_P(B^s)\varphi_p(B)(1-B^s)^D(1-B)^d Y_t = \Theta_Q(B^s)\theta_q(B)Z_t$$

Where:

B: backward shift operator.

$Y_t = \ln(X_t)$

X_t: cases Covid-19.

Z_t: stochastic perturbation.

Θ: sma_1 from analysis.

θ: ma_1 from analysis.

Using the values obtained from analysis the expression becomes:

$$Y_t = Y_{t-1} + Y_{t-7} - Y_{t-8} + Z_t + \theta Z_{t-1} + \Theta Z_{t-7} + \theta\Theta Z_{t-8}$$

And finally:

$$Y_t = Y_{t-1} + Y_{t-7} - Y_{t-8} + Z_t - 0.5995 Z_{t-1} - 0.6433 Z_{t-7} + 0.3857 Z_{t-8}$$

Where:

$Y_t = \ln(cases)$

$Z_t = N{\sim}(0, 0.03723)$

Forecast

Now the model is complete and it is possible to perform a forecast.

On May 5, 2021, the forecast was released for the next 12 days (from May 6th until May 17th), figure 11 shows there is a fluctuation over the days of forecasting but in general maintains a constant and consistent number of cases reported.

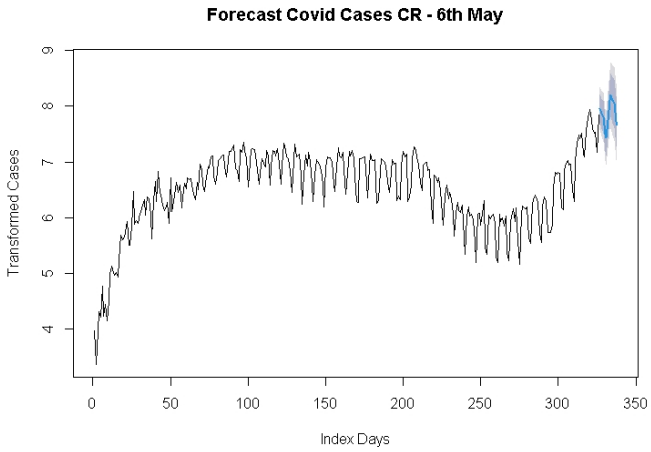


Figure 11. Total curve of cases (transformed data) plus forecast for dates between May 6th and May 17th.

The region of prediction is hard to read from figure 11, figure 12 provides a more detailed view of the forecast as well a comparison with the official report of Covid-19 cases according to the authorities.

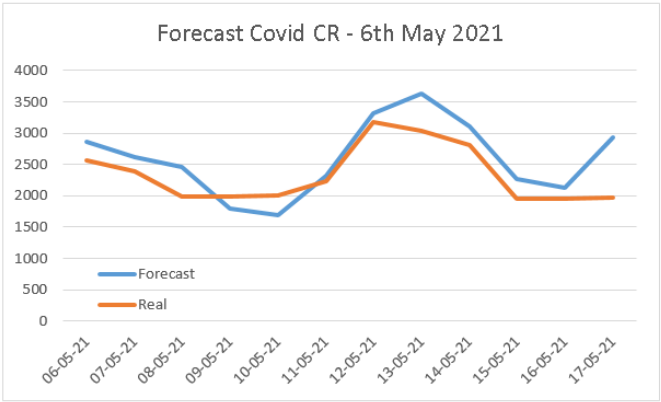


Figure 12. Forecast (blue) and real number of cases (orange) for dates between May 6th and May 17th.

The data of the forecast and real cases from figure 12 can be seen in figure 13. The update of real cases given by the government allows the calculation of Mean Squared Error (MSE) in the forecast.

Index Day	Date	Forecast	Real
327	06-05-21	2863	2559
328	07-05-21	2618	2392
329	08-05-21	2460	1997
330	09-05-21	1795	1997
331	10-05-21	1687	1998
332	11-05-21	2314	2236
333	12-05-21	3314	3173
334	13-05-21	3625	3039
335	14-05-21	3115	2815
336	15-05-21	2273	1961
337	16-05-21	2136	1961
338	17-05-21	2930	1963
MSE	2.82 %		

Figure 13. Table of forecast and real number of cases for dates between May 6th and May 17th, and MSE for this forecast period.

By using the same model (to see details in the R code) it is possible to make a prediction on the dates: May 20th, May 28th, and Jun 9th.

The results are shown in the following figures:

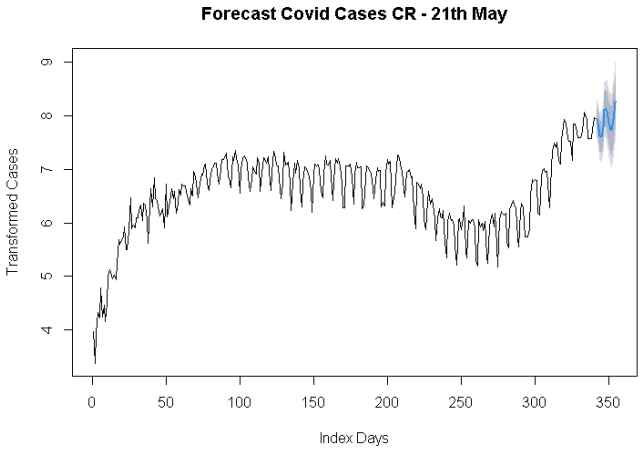


Figure 14. Total curve of cases (transformed data) plus forecast for dates between May 21st and June 1st.

The forecast window in figure 14 can be observed with a better detail:

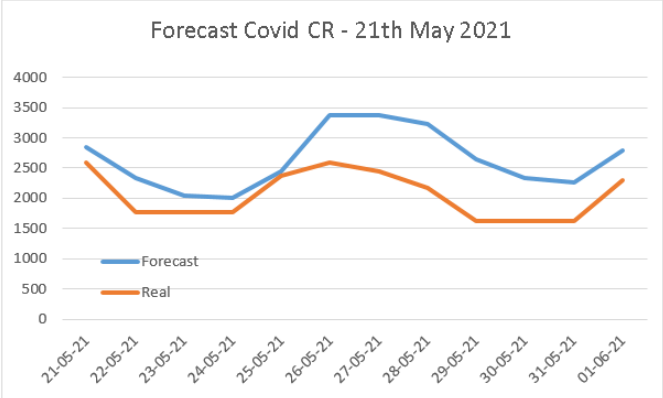


Figure 15. Forecast (blue) and real number of cases (orange) for dates between May 21st and June 1st.

Index Day	Date	Forecast	Real
342	21-05-21	2838	2587
343	22-05-21	2334	1770
344	23-05-21	2050	1770
345	24-05-21	2000	1770
346	25-05-21	2450	2370
347	26-05-21	3373	2587
348	27-05-21	3382	2436
349	28-05-21	3222	2180
350	29-05-21	2649	1628
351	30-05-21	2327	1628
352	31-05-21	2271	1628
353	01-06-21	2782	2293
MSE	8.05 %		

Figure 16. Table of forecast and real number of cases for dates between May 21st and June 1st, and MSE for this forecast period.

Similar to the past forecast, data shows a fluctuation in the period of study but it does not show an increase in the cases per day.

The next forecast was performed on May 28th, predicting the window between May 29th and June 9th. The results are shown in the following figures:

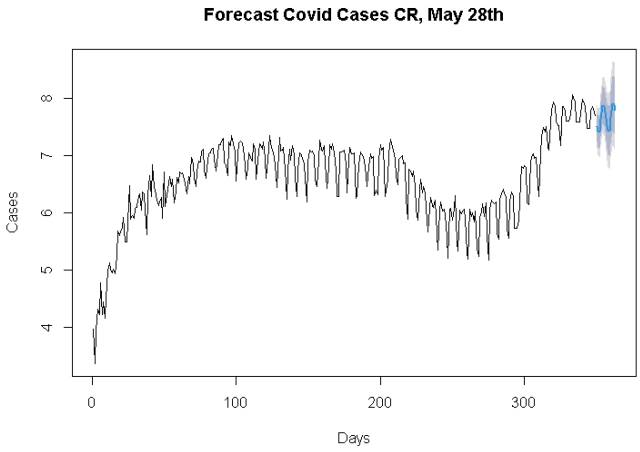


Figure 17. Total curve of cases (transformed data) plus forecast for dates between May 28th and June 9th.

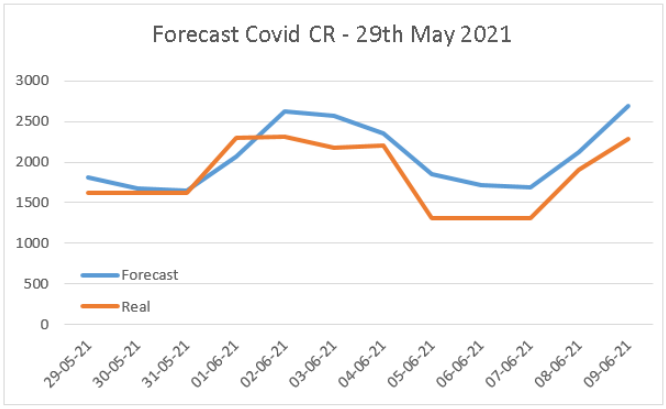


Figure 18. Forecast (blue) and real number of cases (orange) for dates between May 28th and June 9th.

Index Day	Date	Forecast	Real
350	29-05-21	1809	1628
351	30-05-21	1670	1628
352	31-05-21	1645	1628
353	01-06-21	2073	2293
354	02-06-21	2628	2319
355	03-06-21	2575	2181
356	04-06-21	2350	2200
357	05-06-21	1856	1314
358	06-06-21	1713	1314
359	07-06-21	1688	1313
360	08-06-21	2126	1900
361	09-06-21	2695	2287
MSE	3.01 %		

Figure 19. Table of forecast and real number of cases for dates between May 28th and June 9th, and MSE for this forecast period.

Similar to the past forecasts, data shows a fluctuation in the period of study but it does not show an increment in the cases per day.

The final forecast for this study was performed on June 9th, predicting the window between June 10th and June 21th.The results are shown in the following figures:

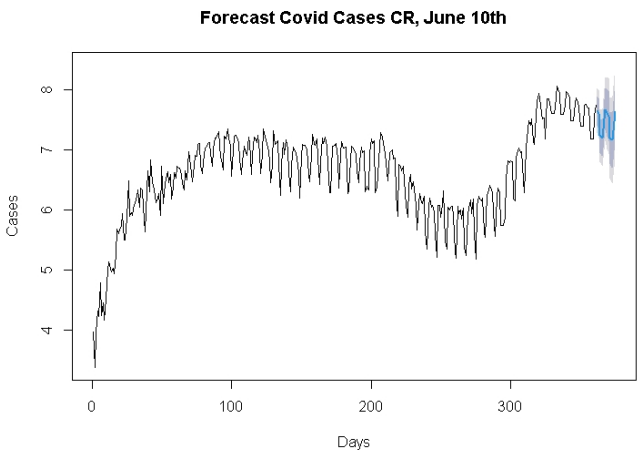


Figure 20. Total curve of cases (transformed data) plus forecast for dates between June 10th and June 21st.

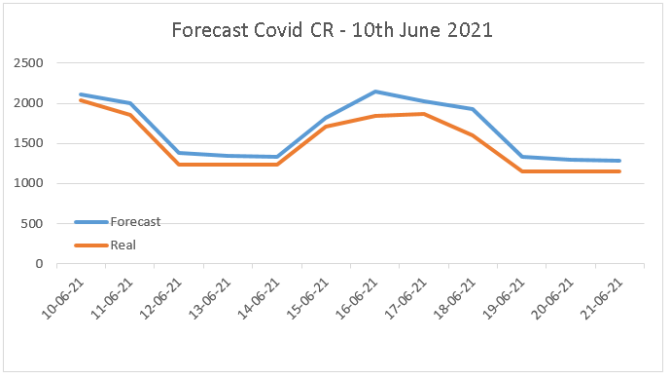


Figure 21. Forecast (blue) and real number of cases (orange) for dates between June 10th and June 21st.

Index Day	Date	Forecast	Real
362	10-06-21	2107	2041
363	11-06-21	1995	1852
364	12-06-21	1384	1235
365	13-06-21	1344	1235
366	14-06-21	1337	1234
367	15-06-21	1816	1708
368	16-06-21	2141	1845
369	17-06-21	2030	1869
370	18-06-21	1922	1604
371	19-06-21	1334	1155
372	20-06-21	1295	1155
373	21-06-21	1288	1155
MSE		1.20 %	

Figure 22. Table of forecast and real number of cases for dates between June 10th and June 21st, and MSE for this forecast period.

Similar to the past forecasts, data shows a fluctuation in the period of study but in average there is a decrease in the reported cases per day.

The forecast is supported by the actualized data of real cases reported throughout the month of May. Cases did not reach the number of 10 000 per day, as the article told. Instead the highest number of cases reported in a day was 3173.

This agrees with some simple analysis that can be done comparing the statistics of second waves in countries with reliable data and strong health systems such as Italy, Spain, United Kingdom or countries with similar population size such as Panamá and Ireland. What is observed in these analogs are that the peak of second waves was two or three times the peak of the first wave. This would mean that Costa Rica could expect a maximum report of 3000 cases of 4500 per day following this simple approximation. A peak of 10 000 cases per day would be catastrophic for the country.

While it is not part of this study to continue the forecasting for June and so on, it is interesting to note that the official reports definitively demonstrate a decreasing curve.

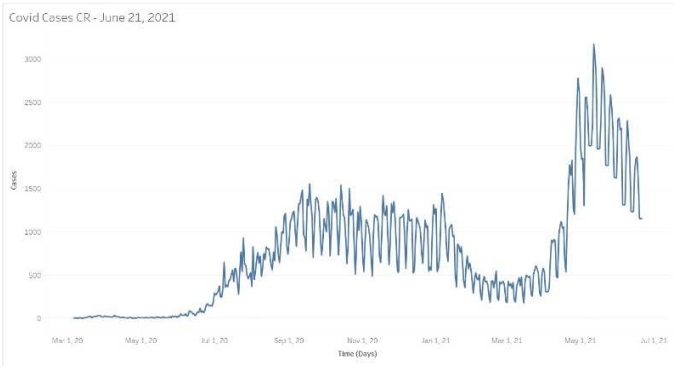


Figure 23. Total curve of Covid-19 cases in Costa Rica, last update: June 21, 2021.

Figure 23 shows the real cases per day according to authorities, updated until June 21st, it is easy to see that in May Costa Rica reached the maximum of the wave and throughout June the cases were decreasing.

It is important to say that the government of Costa Rica did not make big changes in its measures to avoid the spread of Covid-19 during the months of this wave; it would seem that vehicular movability restrictions were enough to avoid the health system collapse despite what to the article led its readers to believe.

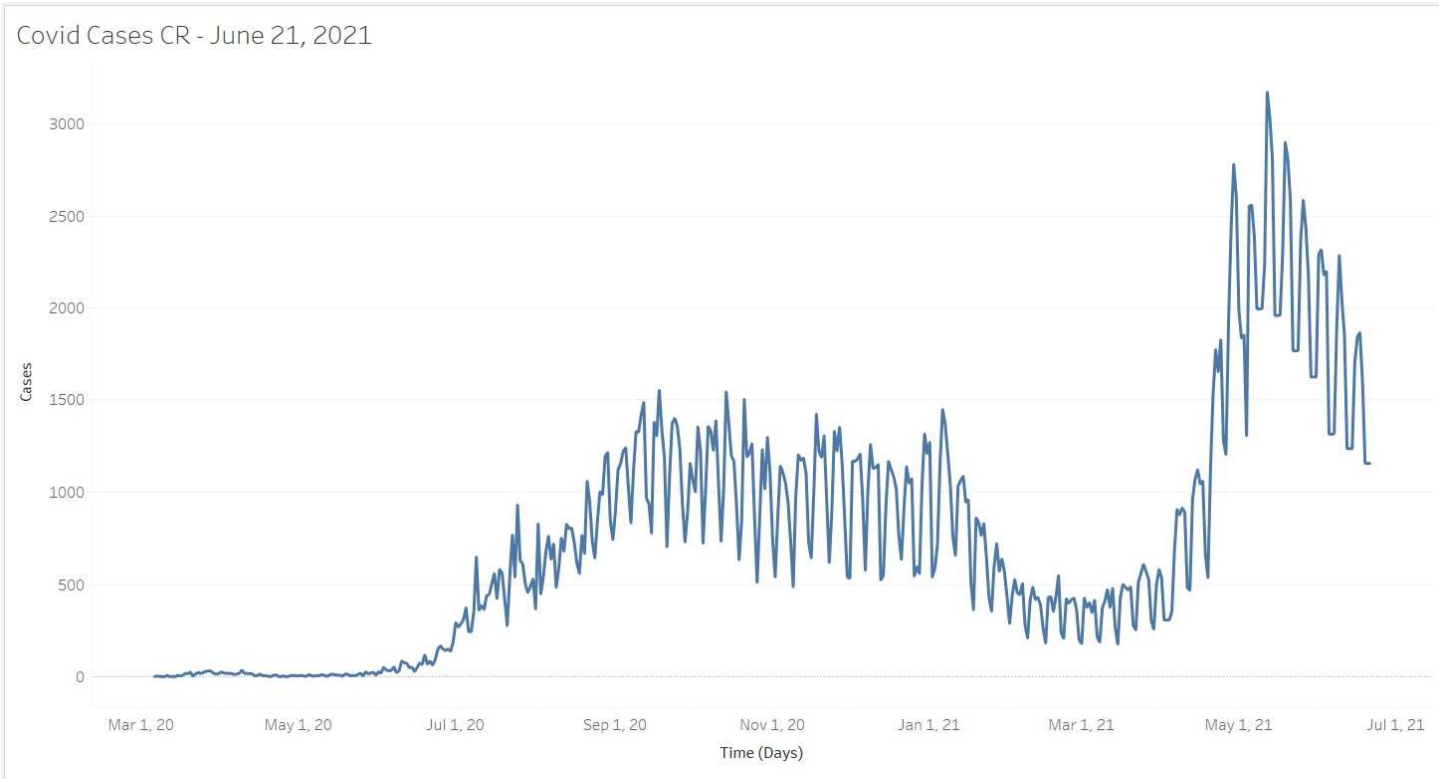


Figure 23. Total curve of Covid-19 cases in Costa Rica, last update: June 21,2021.

Conclusions

The wave studied in this analysis extended from April 10th to July 25th, and peaked in May 2021.

The model found in the analysis: SARIMA (0,1,1,0,1,1,)7, fits with sufficient accuracy and confidence for the behavior of cases in CR (in the window of time analyzed).

For the periods of forecasting with the model SARIMA (0,1,1,0,1,1,)7, the MSE calculated are: 2.82%,8.05% 3.01% and 1.20%.

Answering introduction questions:

Costa Rica did not reach 10,000 cases per day throughout May 2021.

It was not necessary for Costa Rica to issue stricter measures throughout May 2021, as the article suggested.

Due to the evidence and Covid-19 experience in Costa Rica, a forecast of 10,000 cases per day was not realistic.

Final Recommendation

The media has the function to keep people informed every day, and to educate people.

Reliable information is beneficial for everyone; as such the media has a responsibility about the information it reports and should avoid publishing fake data or alarmist news, especially in the middle of a pandemic plagued by misinformation, conspiracy beliefs, and ignorance all of which has resulted in low rates of vaccination in several countries.