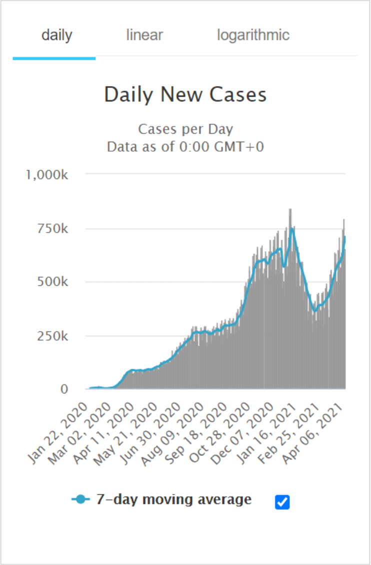
Review of Author’s selection of models

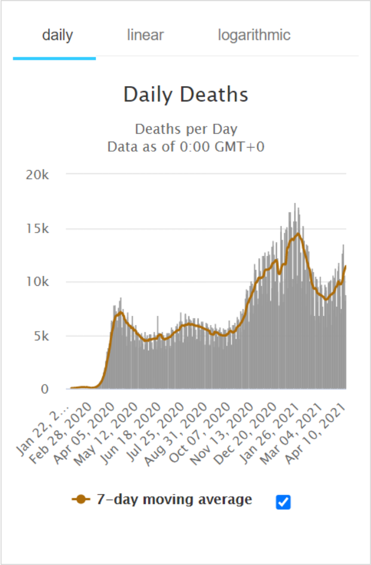
The authors justification for the selection of models is as follows:

1. LR: the authors explained that this was the standard and simplest predictive model to implement, and would serve as a good baseline to compare other models to
2. LASSO: LASSO was considered to account for potential multicollinearity in the dataset as daily data are not totally independent from each other. As the dataset used data across many days, the variable selection property of LASSO also helps to filter out variables that do not significantly improve the prediction results. As with regularization models, the shrinkage property of LASSO also helps to reduce the coefficients and prediction error
3. SVM Regression: On hindsight, the authors noted that SVM produced the least accurate results as the volatile case numbers data made it difficult to establish a clear hyperplane that best separates the data
4. ES: the authors explained that ES was a simple, powerful, and common time series method to do prediction for this time series dataset. Since it is known that many natural phenomena follow some form of natural logarithmic function, ES would be a good choice model to explore, a hypothesis later validated by the authors

Review of Exponential Smoothing

One of the key findings by Rustam, *et. al.* is that Exponential Smoothing was observed to be the best model to fit the COVID-19 case, death, and recovery numbers. As the paper used data from the first 66 days of the COVID-19 pandemic, one intuitive question that comes to mind is whether this claim still holds true in the long run. This is because unlike the first 66 days, the number of COVID-19 cases, and deaths in the long run have some form of wave shape, with increasing and decreasing trends over time, as pictured below [A].

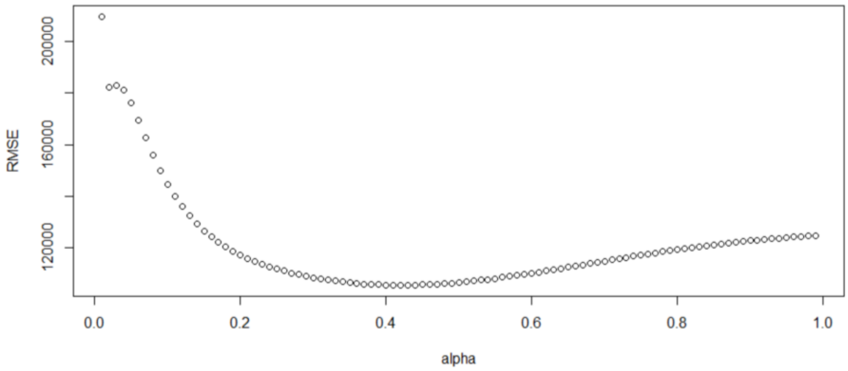




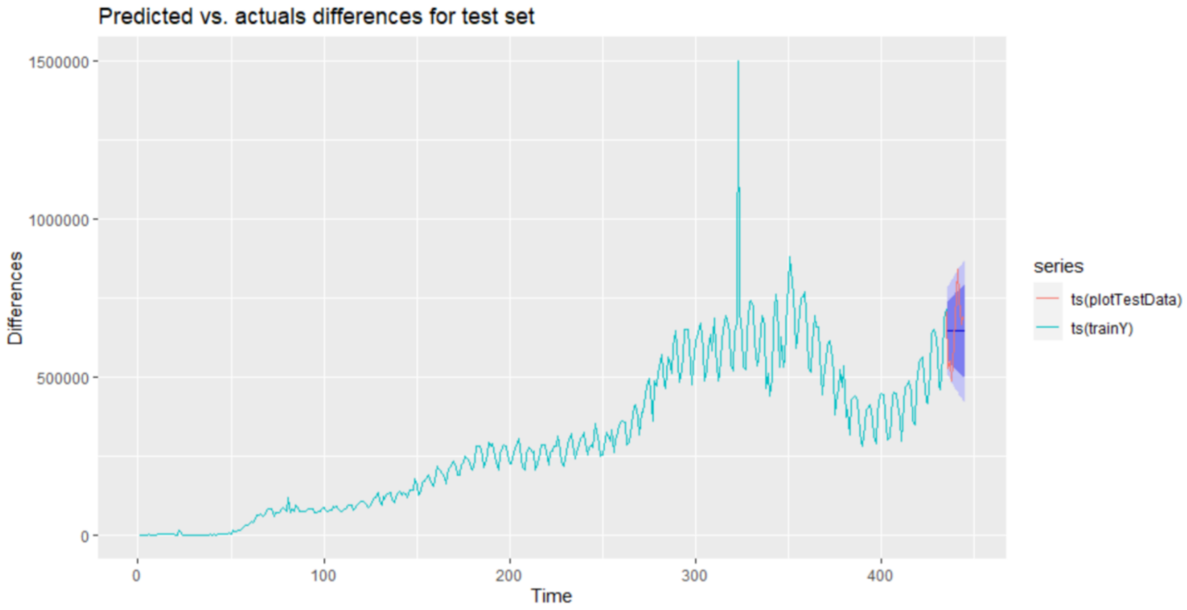
This may pose a problem for the exponential smoothing model developed by the authors, since it has only been trained on predicting case numbers that followed a generally increasing trend in the early stage of the pandemic. As the pandemic progresses over time, it is important to have a model that can predict the case numbers robustly regardless of the trend pattern. Indeed, it is therefore no wonder that this was one of the areas for future research proposed by the authors themselves.

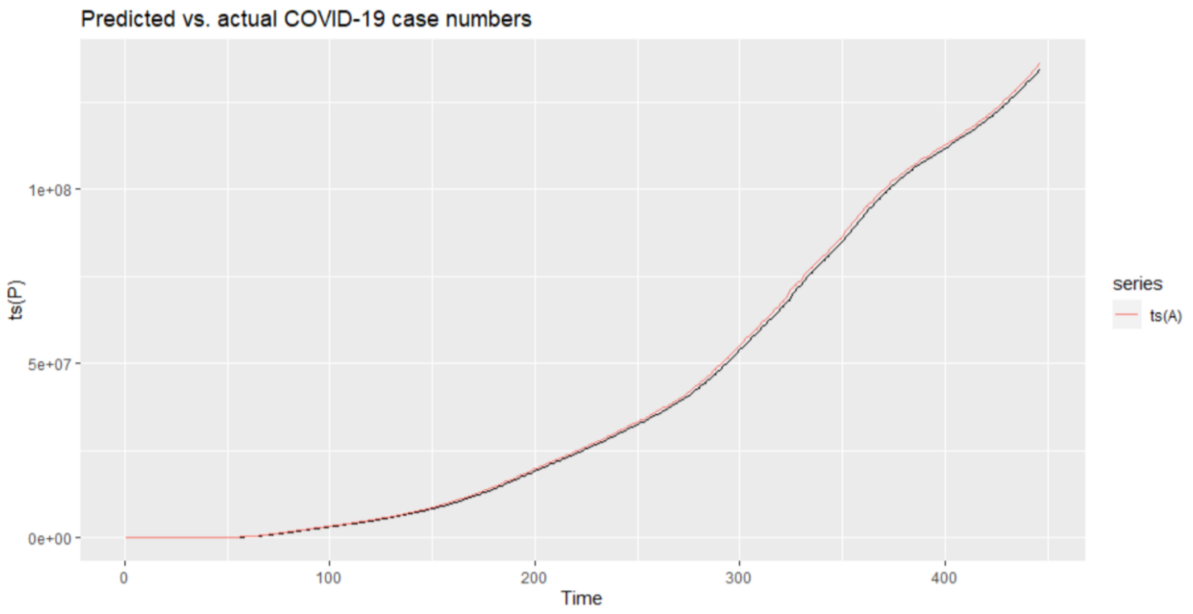
To test whether the models still work well with the updated case numbers, we adopted the same algorithm (Simple Exponential Smoothing (SES)) used by the authors, and tested it on an updated data set with 466 days of COVID-19 data. For brevity, this discussion shall focus only on the new COVID-19 case numbers and not the death and recovery numbers. As the authors intended the SES algorithm to perform a local prediction (more specifically, to forecast the case numbers for the next 10 days), we first evaluated the model's performance in predicting the case numbers for next 10 days. As the ratio of training to test data would be very disproportional, we also explored if if SES could be used to make more long-term predictions.

For the 10 days prediction, the SES model was observed to fit the data well. We first selected the best alpha that the minimised the RMSE, as seen in the figure below.

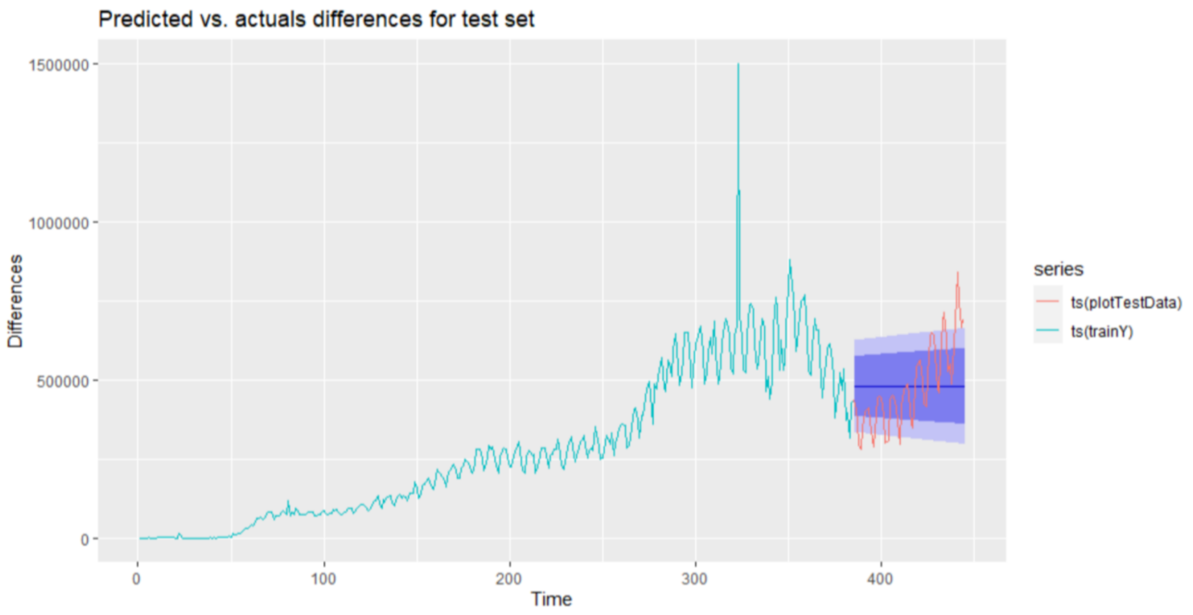


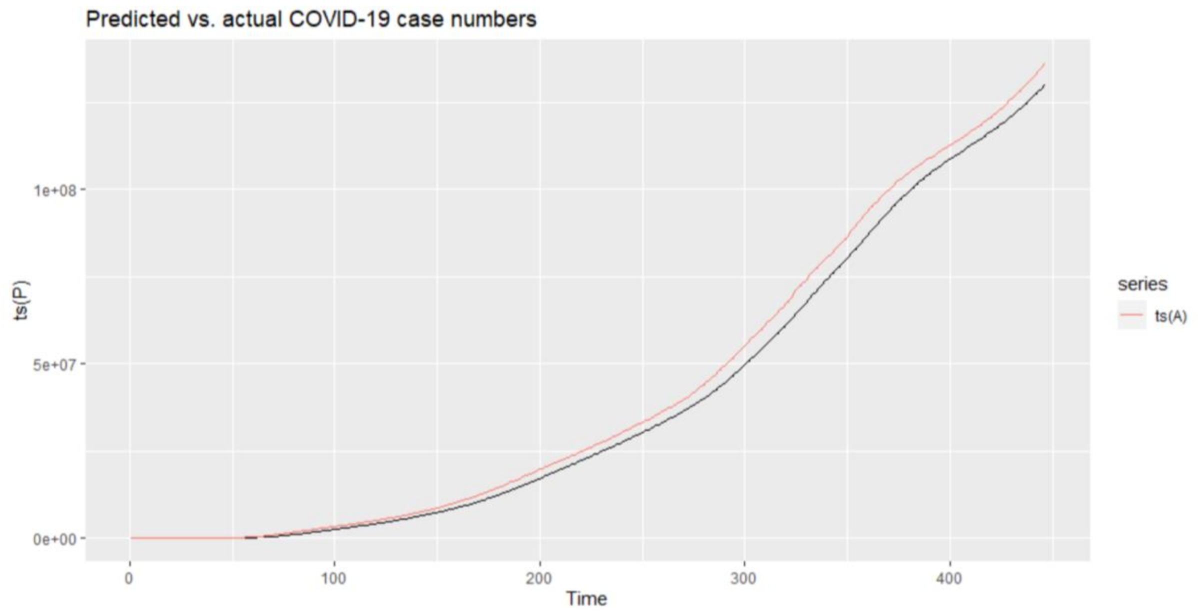
With the optimized alpha, the time series differences plot and predicted case numbers are shown below. Details of the test parameters like the RMSE can be found in the Appendix, along with the code to analyse the data. As seen from the differences plot in Figure 3 below, the SES does a good job at predicting the 10 days differences as most the predictions (red line) lie within the 95% and 80% confidence interval of the prediction (blue regions). As seen from the predicted case numbers plot in Figure 4, the predicted numbers (black line) closely the actual case numbers (orange line). Hence, SES can make good 10 day predictions throughout the duration of COVID-19.





When exploring the time interval which SES could be applied to make a prediction, SES found to be reasonably accurate when making predictions of around 30 days. However, when making predictions for the case numbers 60 days ahead, the predicted case numbers began to show significant deviations from the actual case numbers, as seen in the differences plot and predicted case numbers below. Even with a re-optimised alpha, the forecasted differences were shown to be outside the 80% confidence band. In the predicted case numbers plot, the predicted case numbers (black) were also noticeably different from the actual case numbers (orange). This indicates that SES cannot be used to make predictions up to two months ahead.





This finding paints one of the author's claims (that SES works well even when the training data set was small) in a slightly different light. This claim was made by the authors based on the observation that a reasonably good 10 day prediction could be made by the SES model with 15 days of training data. However, as noted previously, the data set used by the authors had a generally increasing trend because it was the beginning of the pandemic. Once fluctuations appear in the case numbers, one can argue that the SES model requires the training to test data ratio to be significantly more than the 80:20 conventional split, requiring even more than a 90:10 split. Therefore, the author's claims about the size of the training dataset required might only hold true in the initial phase and cannot be claimed to be a generic property of the SES model.

From Exponential Smoothing to Splines

To explore a more accurate method of making long-term predictions, one might be inclined to explore a spline method using SES splines. Looking at the plot of daily new COVID-19 cases, one can discern that there are roughly four waves of COVID-19 cases. If each wave can be "localised" and separated from the others, implementing an SES spline unique to that wave might give a more accurate forecast throughout the duration of the wave. Such an approach is particularly suited for the COVID-19 pandemic as each wave of the pandemic is appearing to be increasingly volatile with more dramatic spikes and drops occurring more recently. Therefore, theoretically, the SES spline method will be a good candidate to model the current wave as splines from the previous wave will have minimal impact on the case numbers of the current wave, and the model will be trained using relevant and current data.

In practice, one drawback of this spline method would be that it requires manual intervention to determine when the wave starts and ends. Therefore, the implementation will not be as straightforward and elegant as the ES approach.

Appendix

1. ES forecast for 10 days

```
Forecast method: Simple exponential smoothing

Model Information: Simple exponential smoothing
```

```
Call: ses(y = trainY, h = 10, alpha = bestalpha)

  Smoothing parameters:     alpha = 0.42

  Initial states:     l = 459.24

  sigma:  70848.34

     AIC     AICc      BIC 
12361.19 12361.22 12369.34 

Error measures:
                   ME     RMSE      MAE       MPE
Training set 3527.548 70685.28 39081.56 -3.340086
                 MAPE     MASE      ACF1
Training set 19.24691 1.135243 0.2268081

Forecasts:
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
436       644942.2 554146.4 735738.0 506082.0 783802.4
437       644942.2 546463.2 743421.1 494331.7 795552.7
438       644942.2 539337.6 750546.7 483434.0 806450.4
439       644942.2 532663.3 757221.0 473226.5 816657.8
440       644942.2 526364.1 763520.2 463592.7 826291.7
441       644942.2 520383.0 769501.3 454445.4 835438.9
442       644942.2 514676.3 775208.0 445717.7 844166.6
443       644942.2 509209.3 780675.1 437356.6 852527.7
444       644942.2 503954.1 785930.3 429319.5 860564.8
445       644942.2 498887.9 790996.5 421571.4 868312.9

                     ME      RMSE      MAE       MPE
Training set 3527.54751  70685.28 39081.56 -3.340086
Test set       35.38482 105497.01 89469.98 -2.711762
                 MAPE     MASE      ACF1
Training set 19.24691 1.135243 0.2268081
Test set     14.33896 2.598928        NA
```

2. ES Forecast for 60 days

```
Forecast method: Simple exponential smoothing

Model Information: Simple exponential smoothing

Call: ses(y = trainY, h = 60, alpha = bestalpha)

  Smoothing parameters:    alpha = 0.1

  Initial states:    l = 1951.77

  sigma:  73785.66

     AIC     AICc      BIC 
10924.86 10924.89 10932.77 

Error measures:
                   ME     RMSE      MAE       MPE
```

```
Training set 12425.38 73593.76 40076.08 -7.456873
                 MAPE     MASE      ACF1
Training set 32.65069 1.226588 0.4205849

Forecasts:
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
386       480328.9 385768.8 574889.0 335711.7 624946.2
387       480328.9 385297.2 575360.7 334990.4 625667.4
388       480328.9 384827.9 575830.0 334272.7 626385.2
389       480328.9 384360.9 576297.0 333558.4 627099.4
390       480328.9 383896.1 576761.7 332847.7 627810.1
...
441       480328.9 362602.5 598055.3 300281.9 660375.9
442       480328.9 362223.3 598434.5 299702.0 660955.8
443       480328.9 361845.4 598812.4 299124.0 661533.8
444       480328.9 361468.7 599189.2 298547.9 662110.0
445       480328.9 361093.1 599564.7 297973.5 662684.3

                   ME      RMSE       MAE       MPE
Training set 12425.38  73593.76  40076.08 -7.456873
Test set     -2366.78 129186.06 104781.63 -7.855710
                 MAPE     MASE      ACF1
Training set 32.65069 1.226588 0.4205849
Test set     23.36365 3.206997        NA
```

References

[A] Worldometer COVID-19 Coronavirus Pandemic Data. *Worldometer*. Accessed: 15 April 2021. [Online]. Available: https://www.worldometers.info/coronavirus/