Time-series forecasting of economic outlier nations for population weighted COVID-19 case change

Austin Stevens

PSDS Capstone

Project purpose

The purpose of this project is to determine if LSTM or ARIMA models are effective in predicting COVID-19 rates for various countries. Initial analysis will be based on linear regression of year-over-year quarterly GDP change and population weighted COVID-19 rate change. The selected countries will be those identified as outliers from the regression model for case and death rates. This method is essentially a way to identify a subset of nations that for some reason or another exhibited a different economic response to COVID-19.

Project data

-COVID-19 Time Series data from Johns Hopkins

Hosted on Github: <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series>

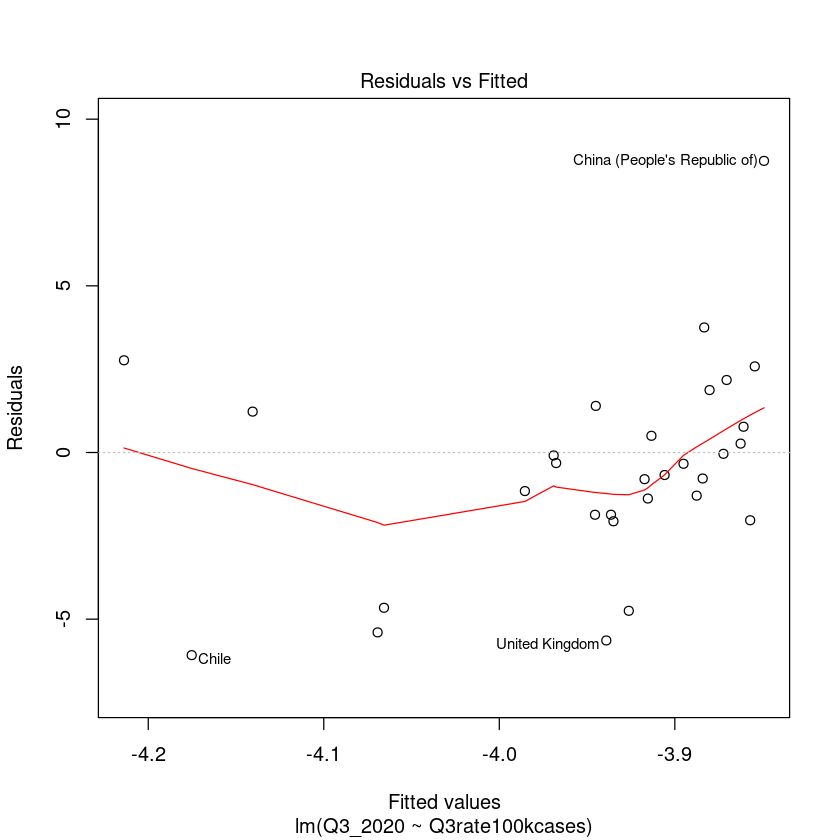
-Population data from <https://databank.worldbank.org/>

-OECD statistics on world GDP data – year over year quarterly data <https://stats.oecd.org/>

Economic results

Not all countries reported their Q3 2020 data at the initiation of this study, so I divided up the groups into Q3 and Q2 sets. Below is an example of the results of the lm() model for the Q3 GDP, population adjusted case change, and weighted for population density.

lm(Q3\_2020 ~ Q3rate100kcases, data=Q3case.sub, weights=Density2018)



The resulting list of countries for the time series analysis:

Argentina

Belgium

Chile

China (People's Republic of)

Colombia

India

Israel

Japan

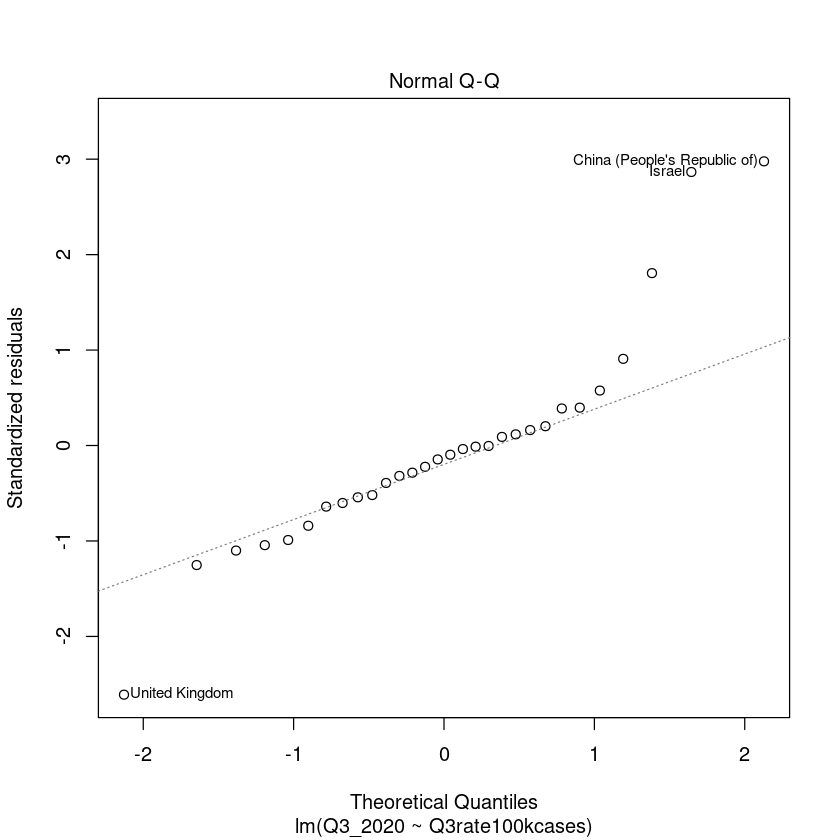
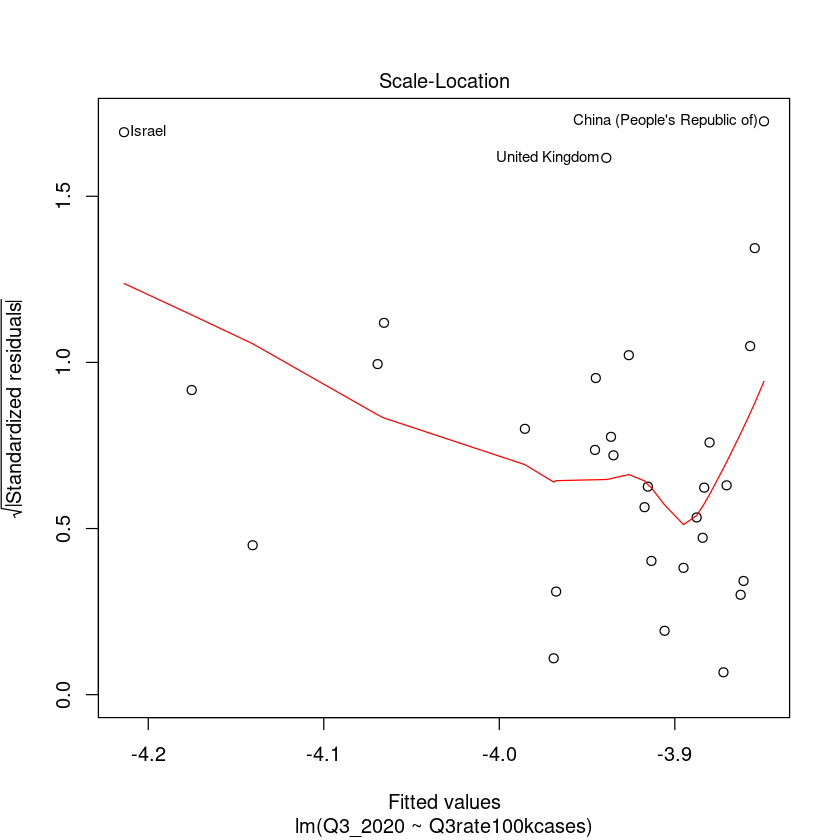
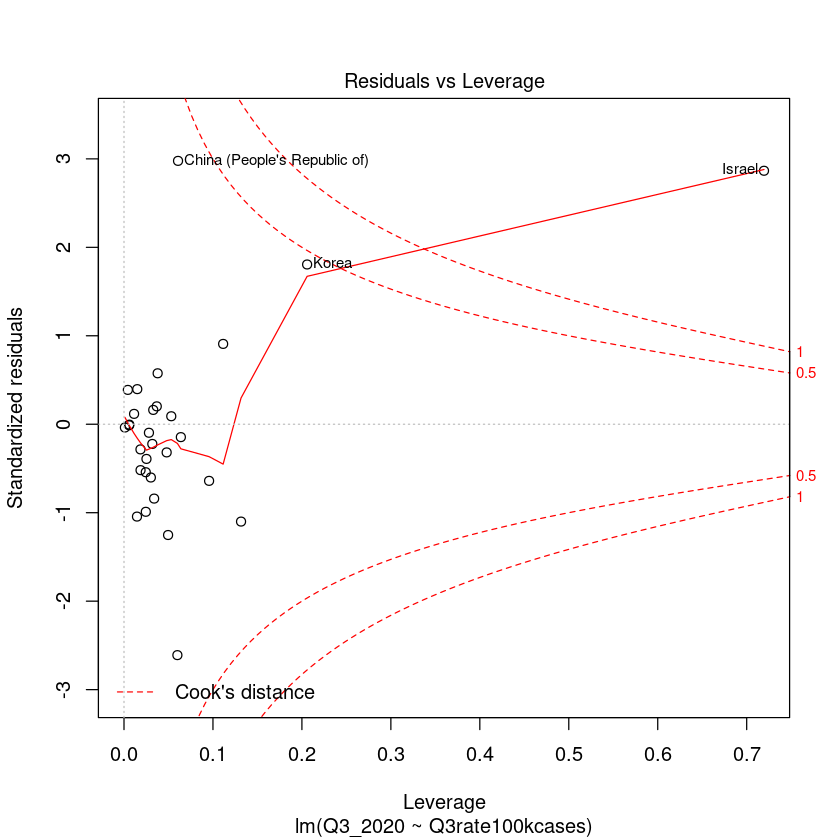
Korea

Luxembourg

Spain

United Kingdom

United States



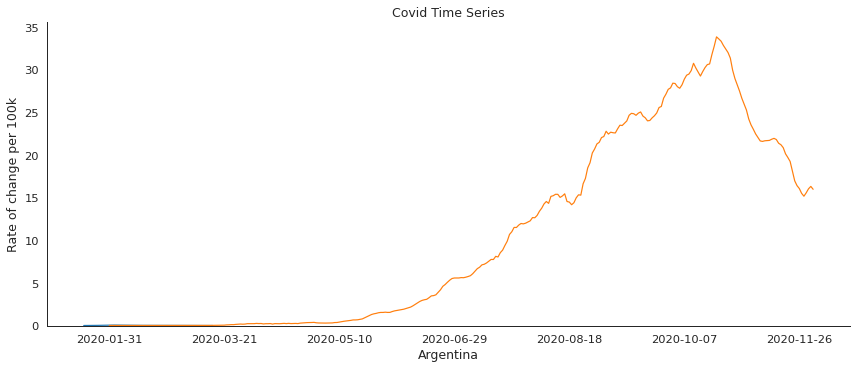
Time series forecasting

I evaluated LSTM and ARIMA models to predict values for this set of countries. For both, I decided to use roaming windows to judge effectiveness of each method of forecasting. The window size is 14 days, which is the upper range of generally accepted time in which COVID matures in the body. In addition, I tried both one direction and bi directional LSTM models.

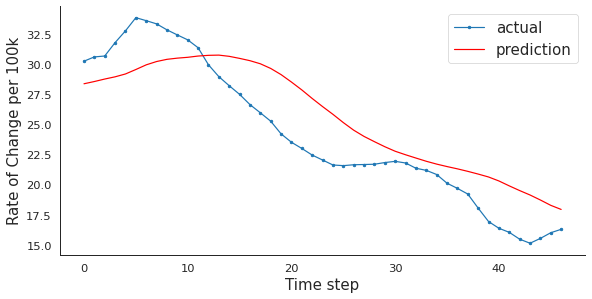
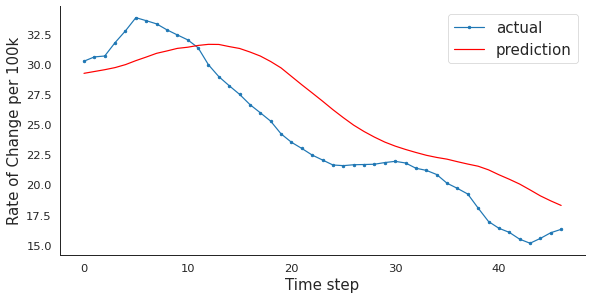
LSTM Results

|  |  |  |
| --- | --- | --- |
| Country | Test RMSE – LSTM Sequential | Test RMSE – LSTM Bidirectional |
| Argentina | 3.278 | 3.022 |
| Belgium | 15.338 | 13.160 |
| Chile | 0.796 | 0.526 |
| China, People’s Republic of | 0.001 | 0.001 |
| Colombia | 1.253 | 1.161 |
| India | 0.547 | 0.501 |
| Israel | 16.406 | 14.370 |
| Japan | 0.278 | 0.225 |
| Korea (South) | 0.060 | 0.044 |
| Luxembourg | 29.674 | 27.420 |
| Spain | 2.918 | 2.370 |
| United Kingdom | 7.038 | 5.921 |
| United States | 9.266 | 8.235 |

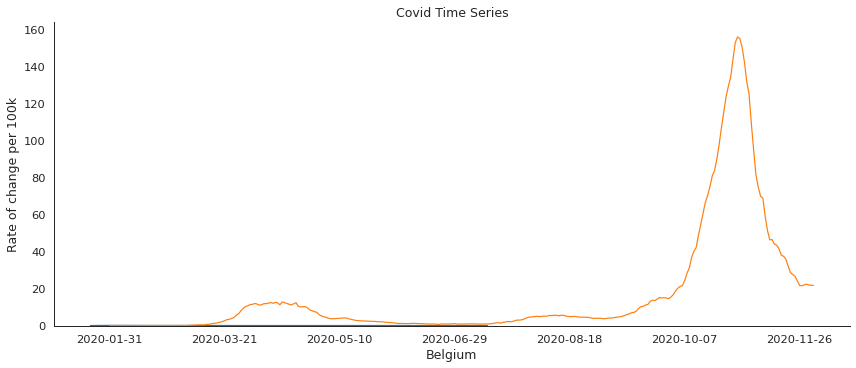
Argentina



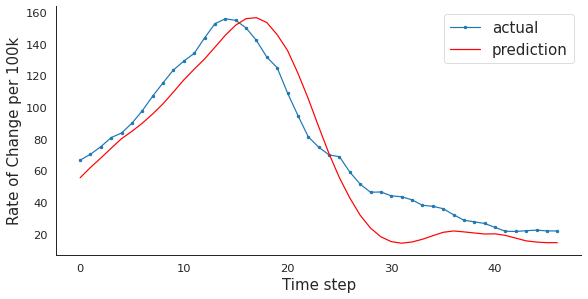
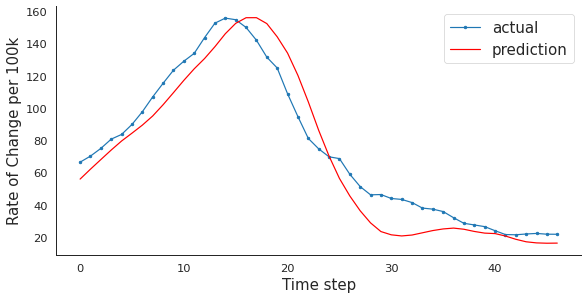
Forward Bidirectional



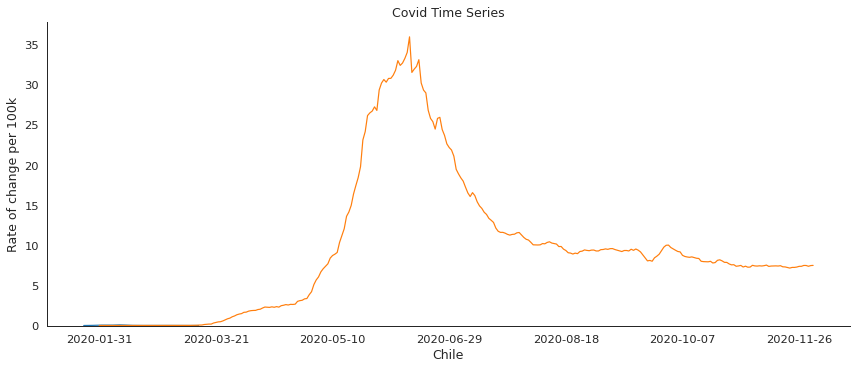
Belgium



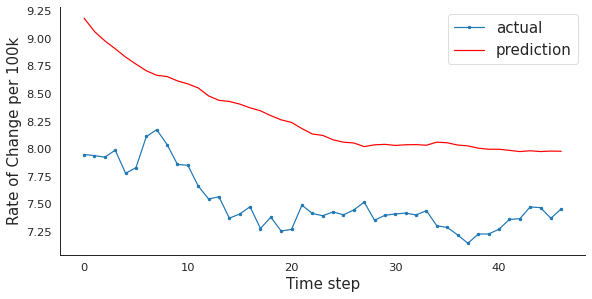
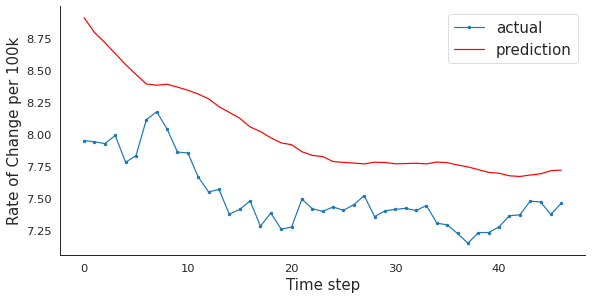
Forward Bidirectional

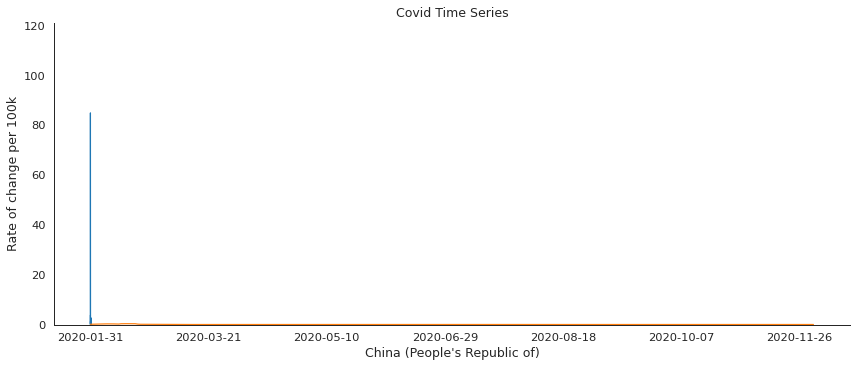
Chile



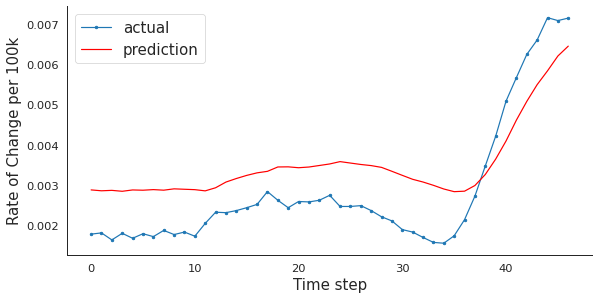
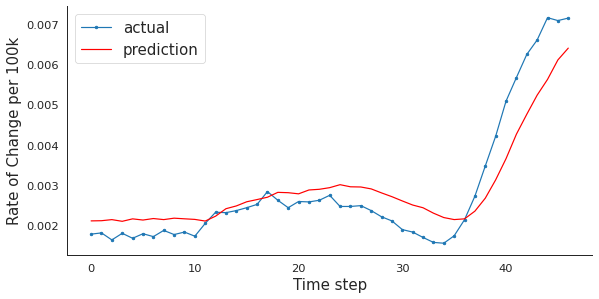
Forward Bidirectional

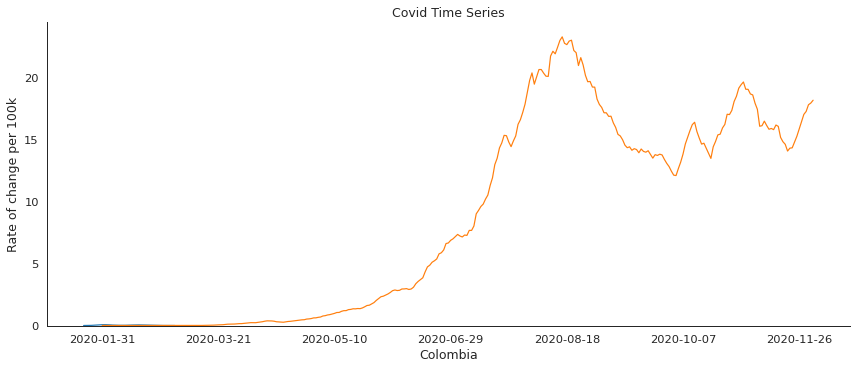
China



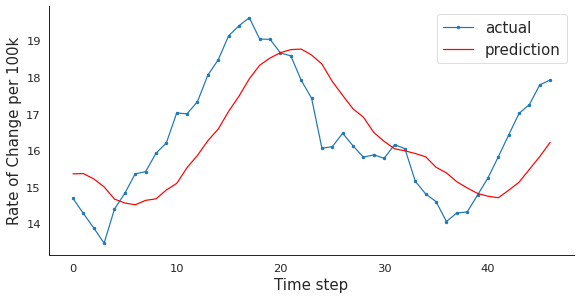
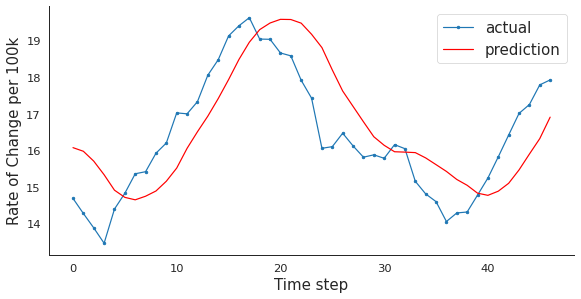
Forward Bidirectional

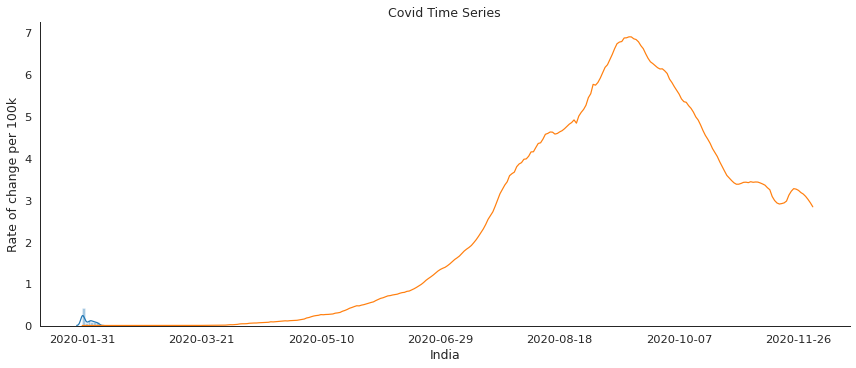
Colombia



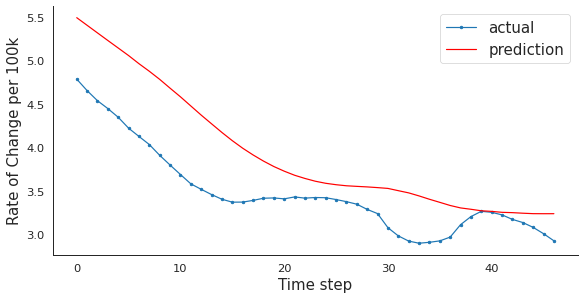
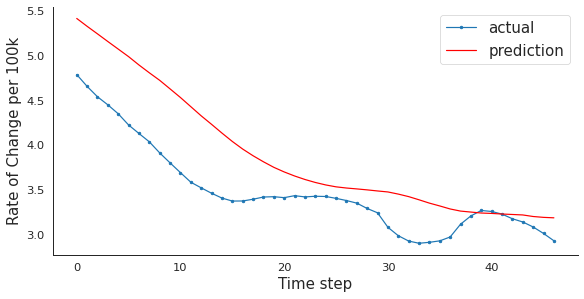
Forward Bidirectional

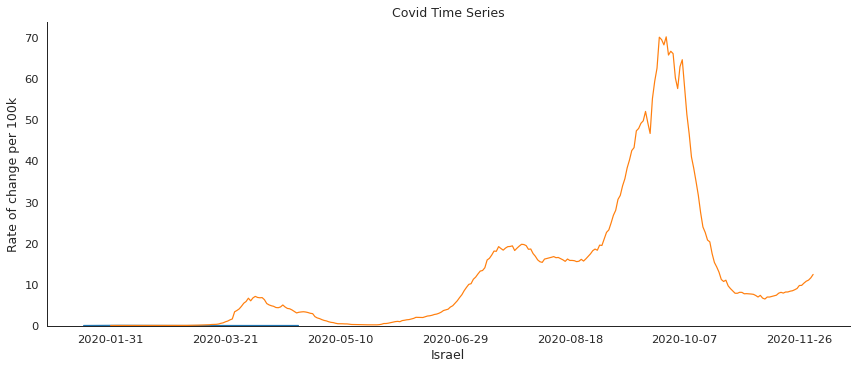
India



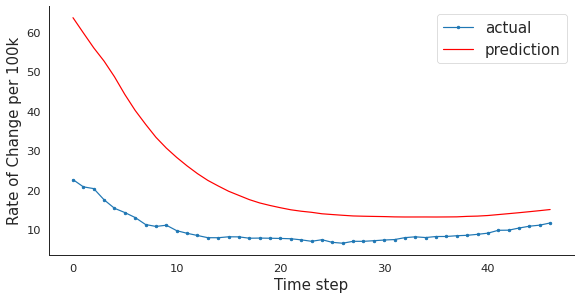
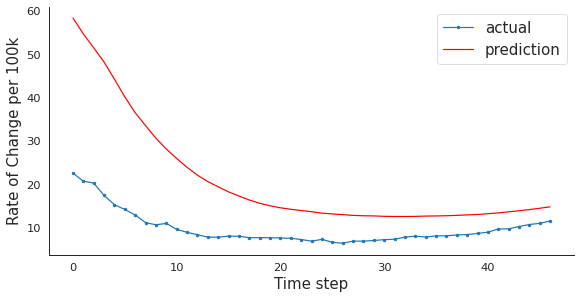
Forward Bidirectional

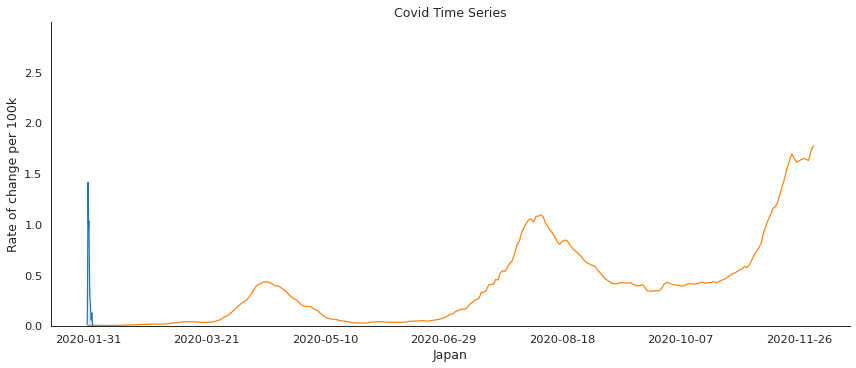
Israel



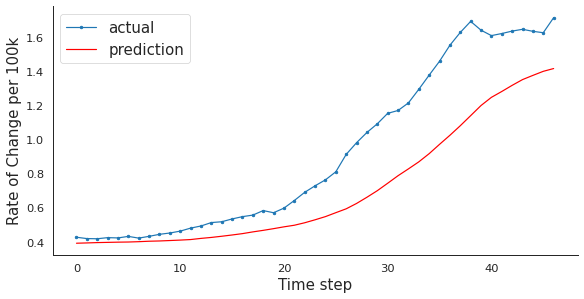
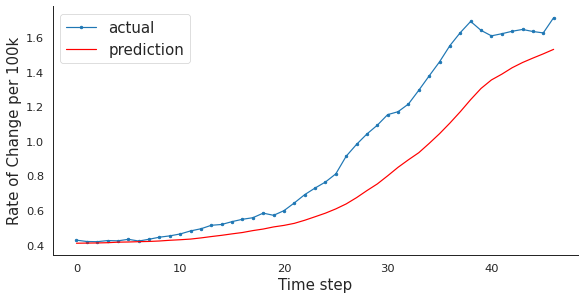
Forward Bidirectional

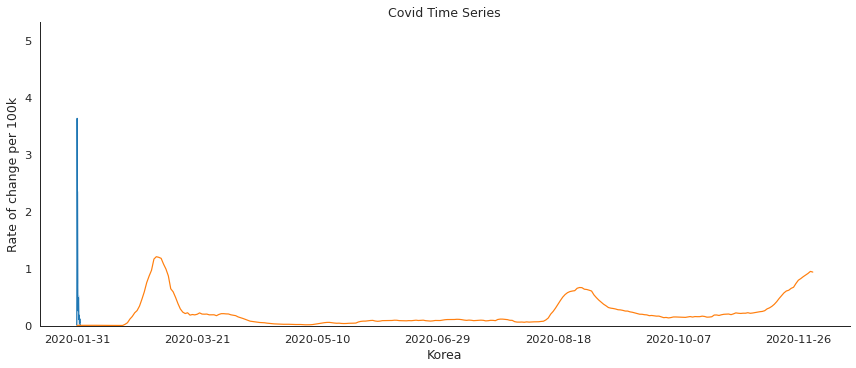
Japan



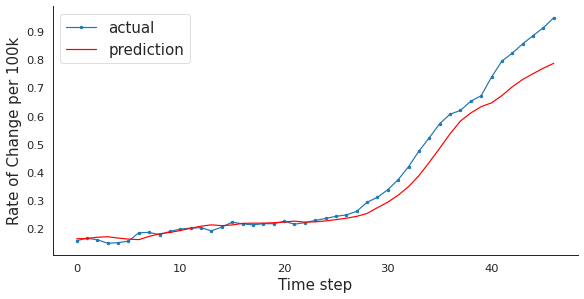
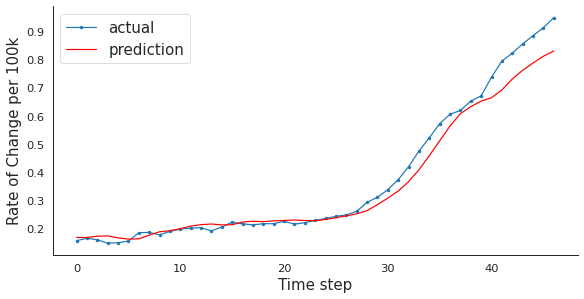
Forward Bidirectional

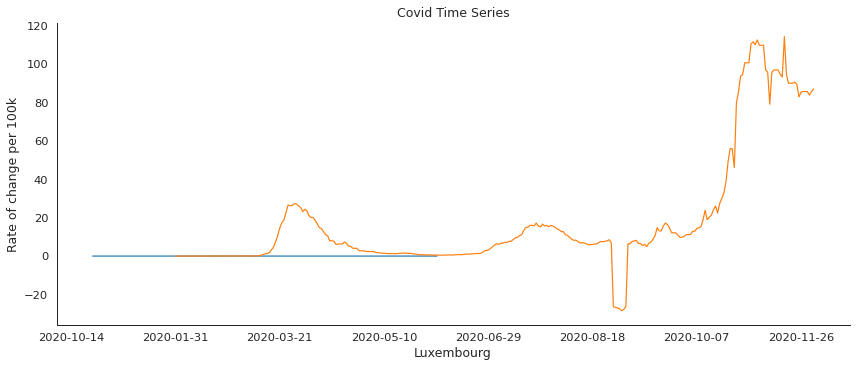
Korea



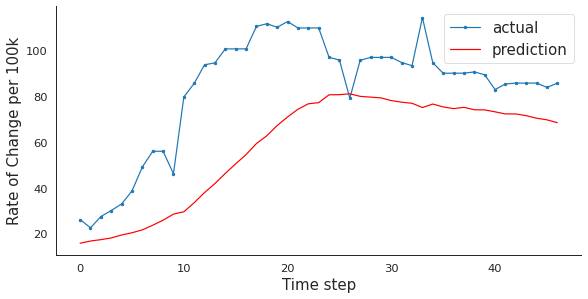
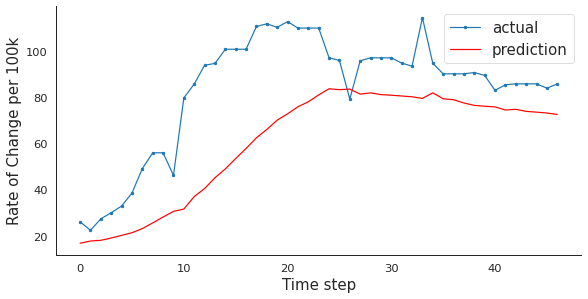
Forward Bidirectional

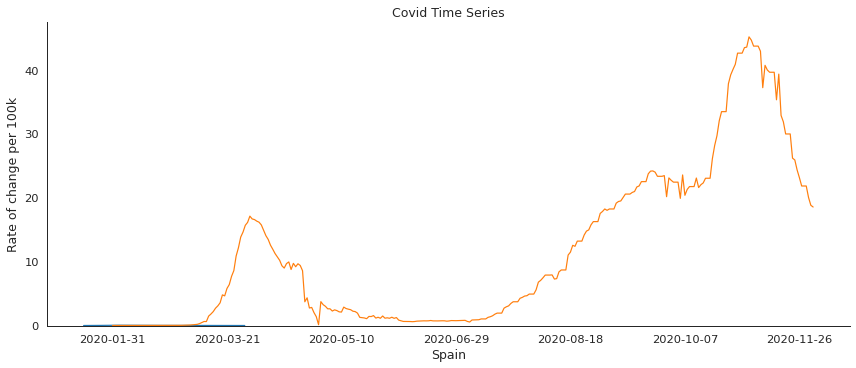
Luxembourg



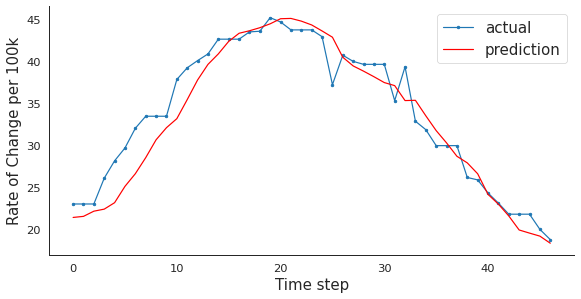
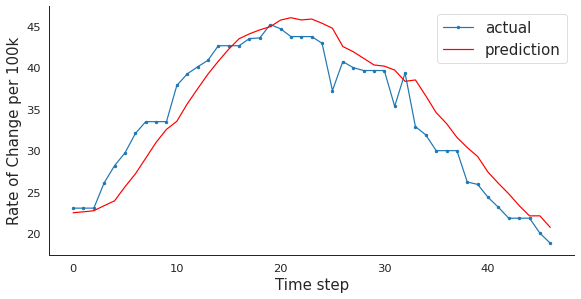
Forward Bidirectional

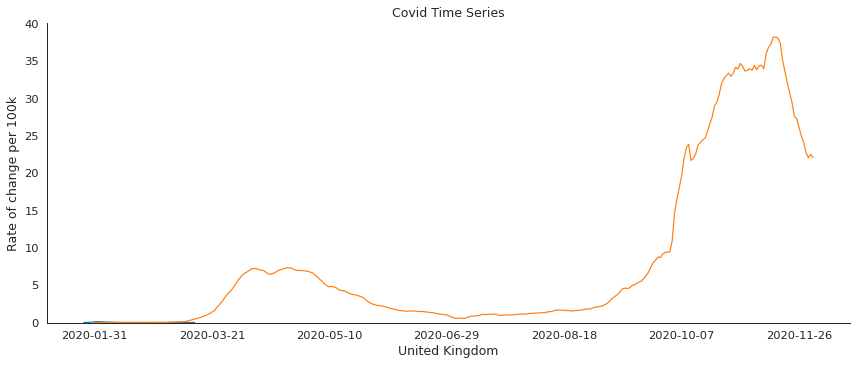
Spain



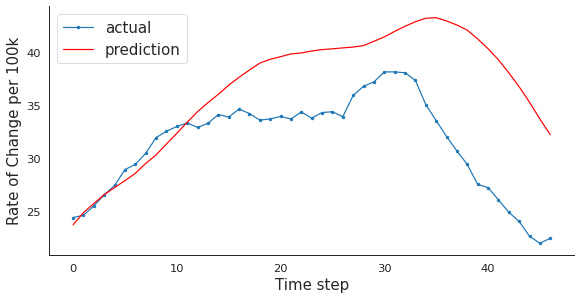
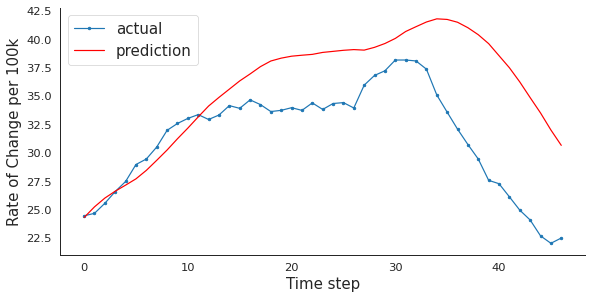
Forward Bidirectional



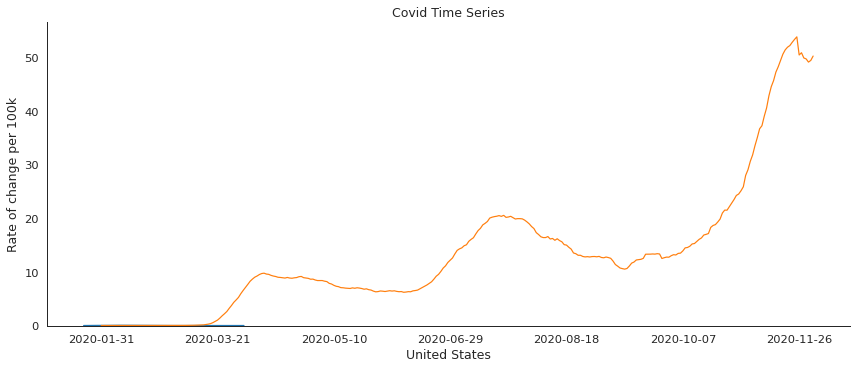
United Kingdom



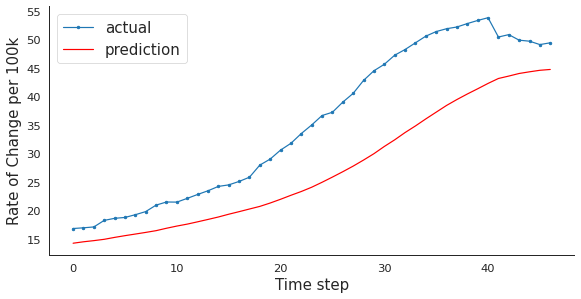
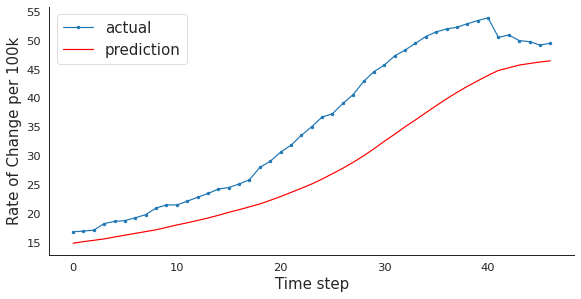
Forward Bidirectional

United States



Forward Bidirectional

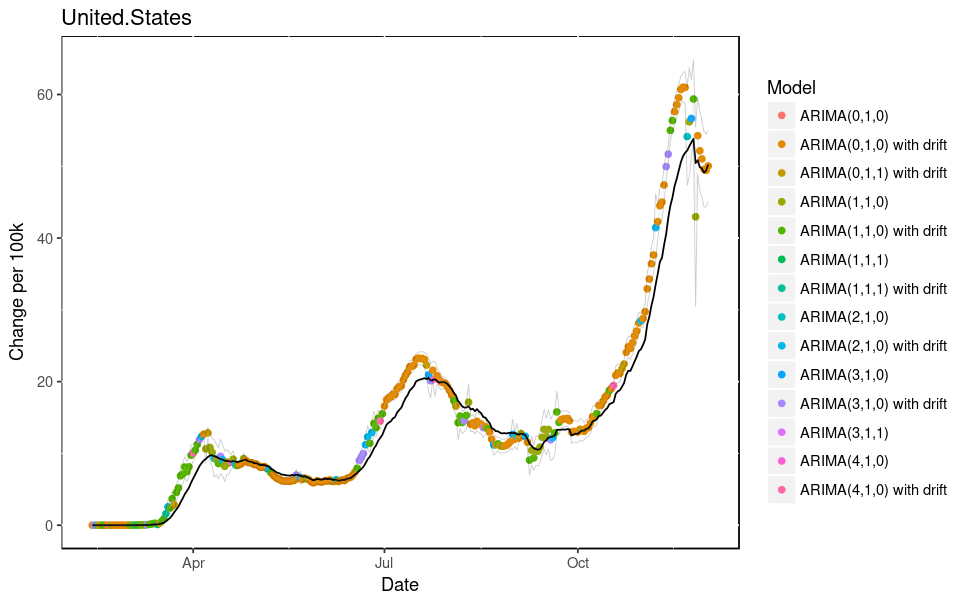
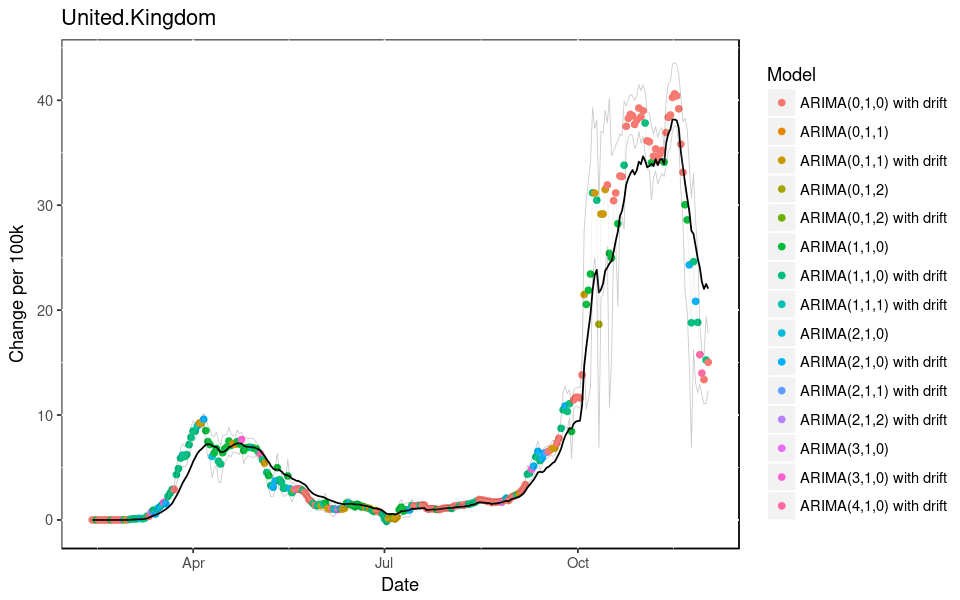
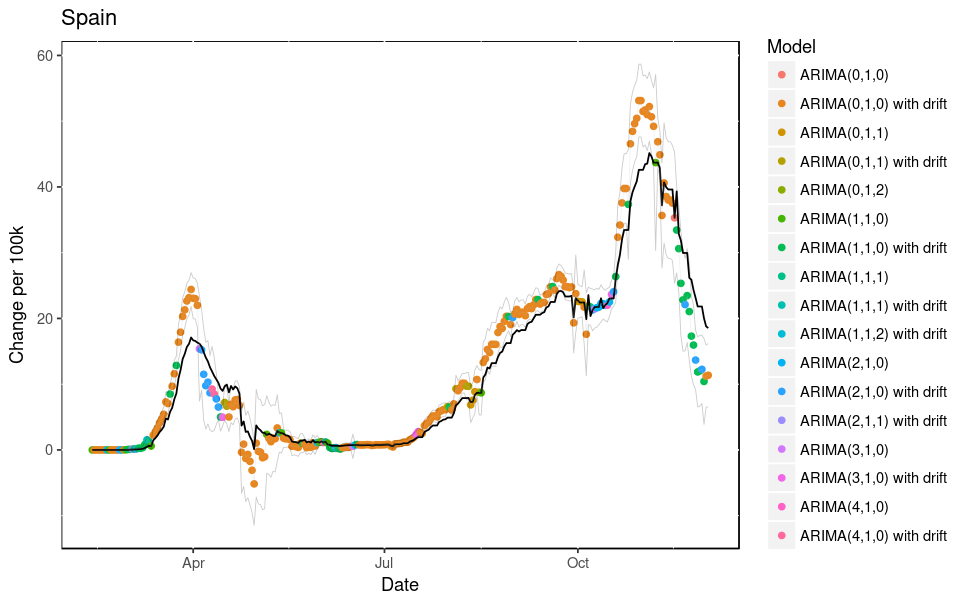
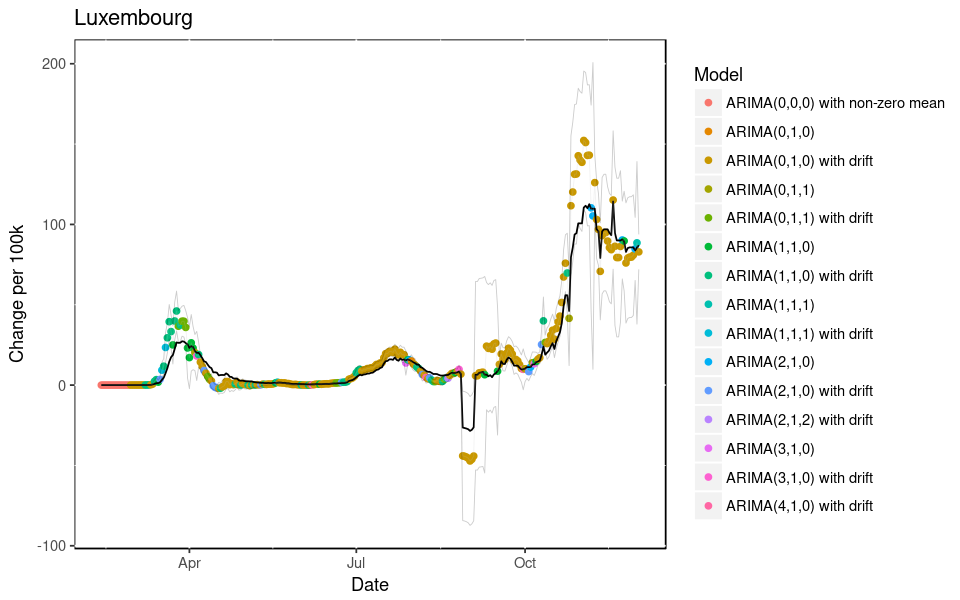
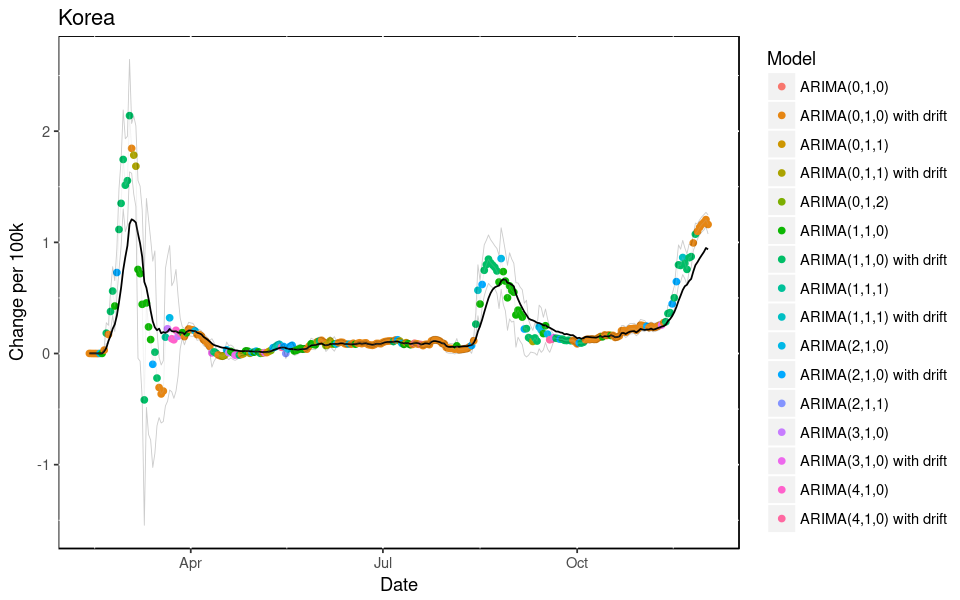
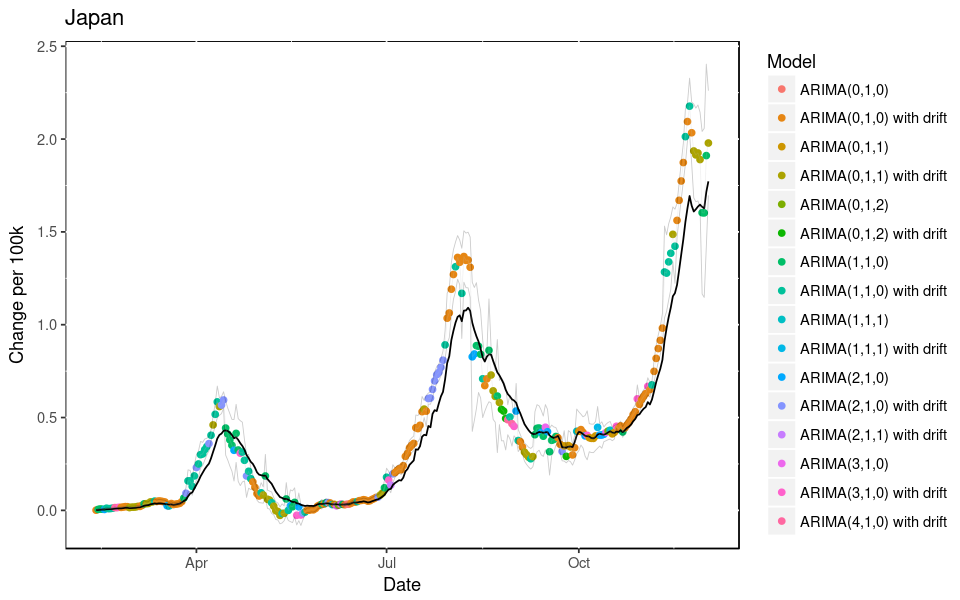
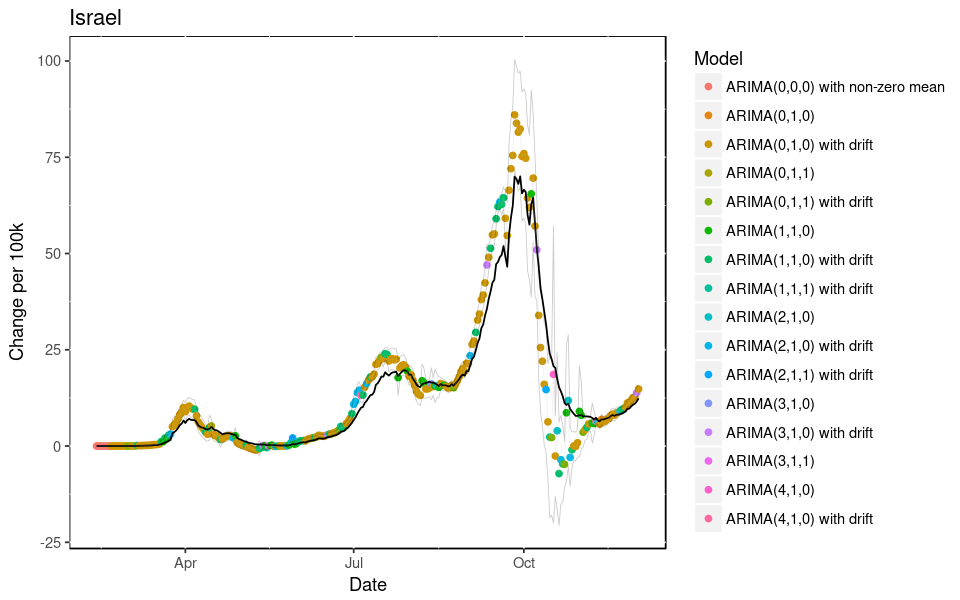
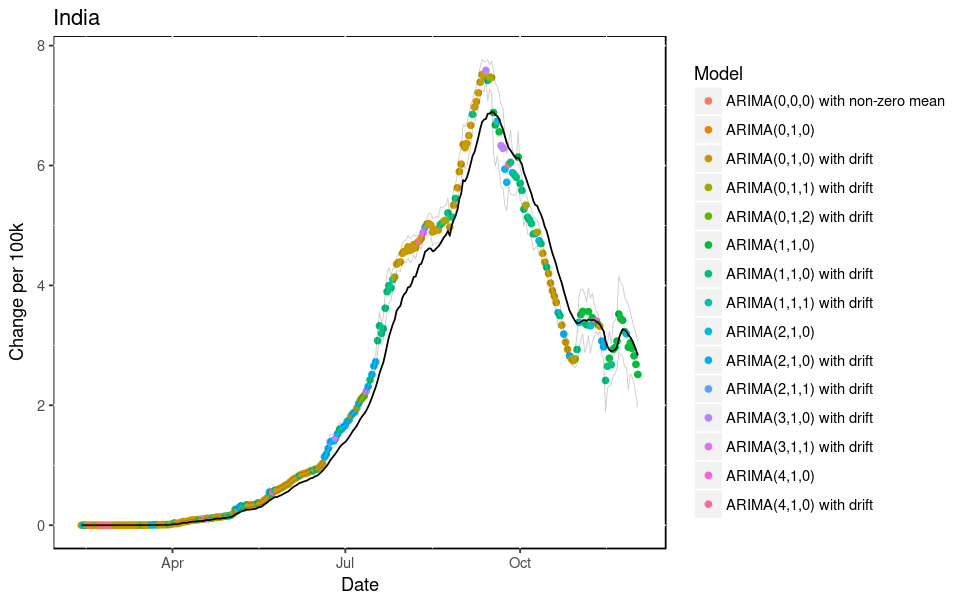
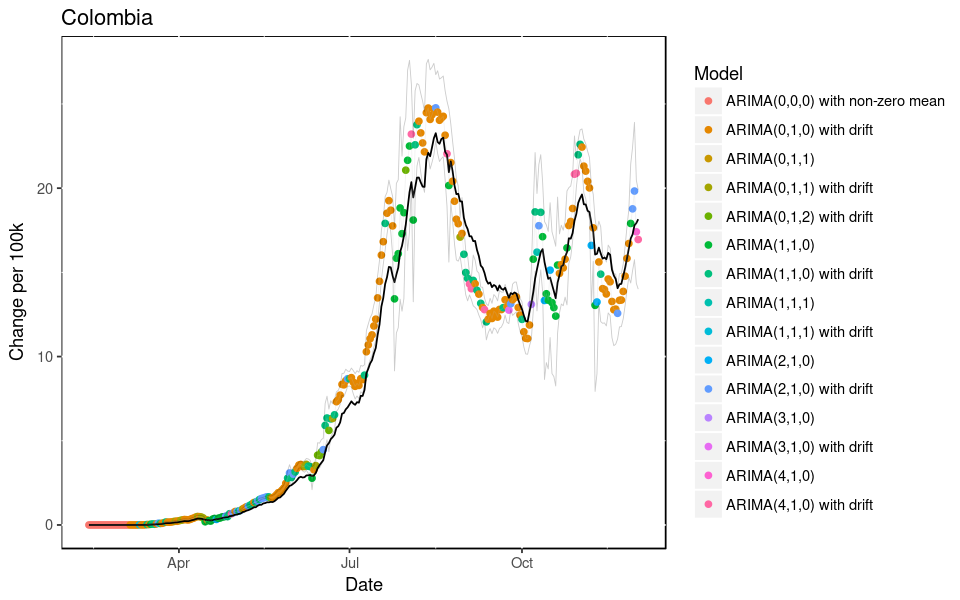
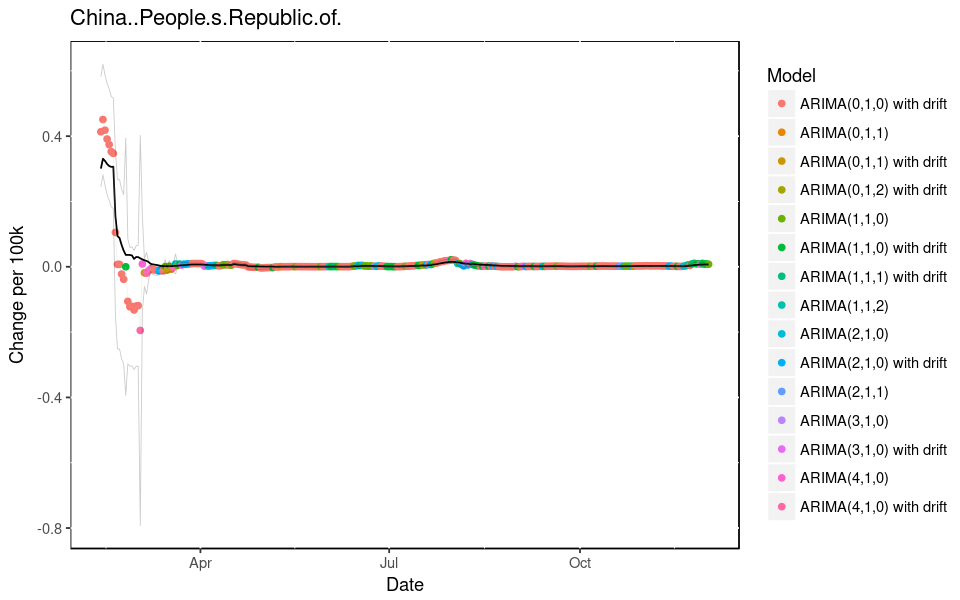
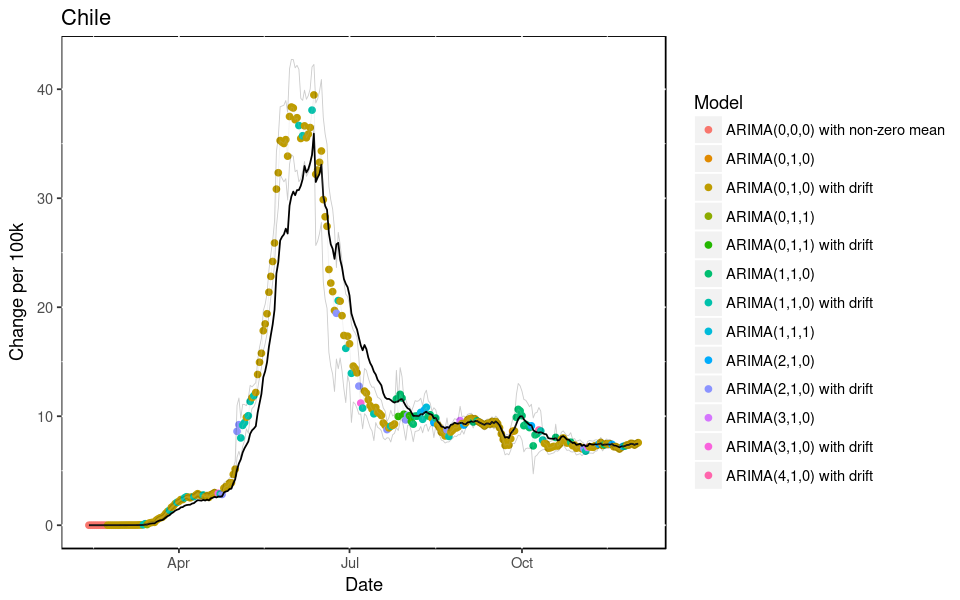
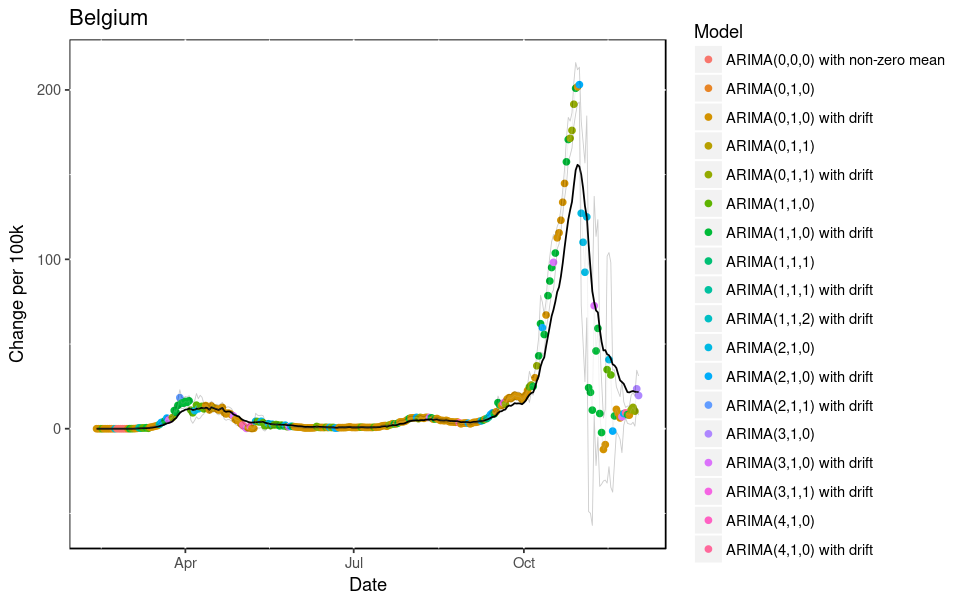
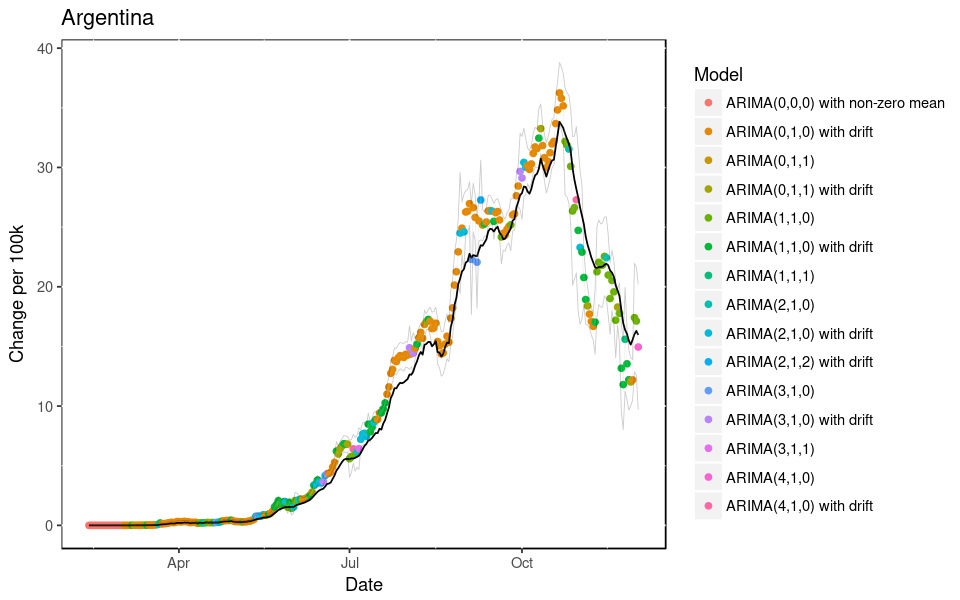
The LSTM models had varying levels of accuracy. After performing the ARIMA forecasting, it had overall better results. I used a roaming 14 days window predicting 7 days out for the ARIMA process.

Auto.ARIMA Results

|  |  |
| --- | --- |
| Country | Test RMSE |
| Argentina | 1.598 |
| Belgium | 15.198 |
| Chile | 2.509 |
| China, People’s Republic of | 0.031 |
| Colombia | 1.412 |
| India | 0.358 |
| Israel | 5.703 |
| Japan | 0.128 |
| Korea (South) | 0.188 |
| Luxembourg | 9.603 |
| Spain | 3.346 |
| United Kingdom | 2.347 |
| United States | 2.817 |

ARIMA Model summary for all study countries and counts

|  |  |
| --- | --- |
| ARIMA(0,1,0) with drift | 1971 |
| ARIMA(1,1,0) with drift | 567 |
| ARIMA(1,1,0) | 377 |
| ARIMA(2,1,0) with drift | 233 |
| ARIMA(0,1,1) with drift | 173 |
| ARIMA(2,1,0) | 100 |
| ARIMA(0,0,0) with non-zero mean | 89 |
| ARIMA(3,1,0) with drift | 76 |
| ARIMA(4,1,0) with drift | 52 |
| ARIMA(0,1,1) | 39 |
| ARIMA(3,1,0) | 27 |
| ARIMA(1,1,1) with drift | 27 |
| ARIMA(1,1,1) | 20 |
| ARIMA(4,1,0) | 16 |
| ARIMA(0,1,2) with drift | 16 |
| ARIMA(0,1,0) | 12 |
| ARIMA(2,1,1) with drift | 8 |
| ARIMA(0,1,2) | 6 |
| ARIMA(2,1,2) with drift | 3 |
| ARIMA(3,1,1) | 3 |
| ARIMA(1,1,2) with drift | 2 |
| ARIMA(3,1,1) with drift | 2 |
| ARIMA(2,1,1) | 2 |
| ARIMA(1,1,2) | 1 |



Something interesting to note is the top ARIMA model is a random walk model, which means the future really can’t be predicted, even though it appears to fit in the past. However, this is an issue with both the LSTM and ARIMA approaches, because we are using the past to predict the future. In addition, the prominence of the random walk model for each of these 14 day windows indicates there is no seasonality, though that is unlikely to occur in such a short timeframe. When we use an auto.arima model for the entire time series, we do get different models:

|  |  |
| --- | --- |
| **Argentina** | ARIMA(3,2,3) |
| **Belgium** | ARIMA(2,1,1) |
| **Chile** | ARIMA(0,2,3) |
| **China..People.s.Republic.of.** | ARIMA(2,1,2) |
| **Colombia** | ARIMA(2,1,3) |
| **India** | ARIMA(0,2,2) |
| **Israel** | ARIMA(3,1,3) |
| **Japan** | ARIMA(1,2,2) |
| **Korea** | ARIMA(2,1,3) |
| **Luxembourg** | ARIMA(0,1,0) with drift |
| **Spain** | ARIMA(2,1,3) |
| **United.Kingdom** | ARIMA(1,1,1) |
| **United.States** | ARIMA(1,2,1) |

With this approach, only Luxembourg presents as a random walk. Still, as time and the pandemic progress, these models can change, which tells me to revert back to the rolling window with an understanding that out of trend changes can occur with a random walk model and the forecast will be prone to errors. With that understanding, the rolling window auto.arima approach appears to work the best.