

## Write-up on Bayesian Structural Time Series Analysis on Covid 19 Data

To understand the behavior of Covid-19 cases' time series data for different countries, Bayesian Structural Time Series models were applied to get some insight on forecasting behavior for different variants of the models.

The data on the number of confirmed cases, reported deaths and recovered cases are taken from the Johns Hopkins database. The data for country wise population size is taken from the World Bank database. The data is aggregated for different countries reported for the time period, from 22nd Jan 2020 to 30th March 2020, as of now.

The general structure of the Bayesian Structural Time Series model is defined by using two equations, namely,

$$y_t = Z_t^T \alpha_t + \epsilon_t$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t$$

Where  $y_t$  denotes the time series data,  $\alpha_t$  denotes the latent states of the model,  $\epsilon_t$  and  $\eta_t$  are Gaussian noise independent of all other quantities.  $Z_t$ ,  $T_t$  and  $R_t$  are structural parameters of the model.

For the time being, we are assuming that we do not have data on any covariates and we model the data using various latent state specifications. From existing literature we can see that there are several structures of the latent state variable,  $\alpha_t$  like the Local Trend, Local Linear Trend, Semi-Local Linear Trend etc. For a brief overview of the different structures and an overview on Bayesian Structural Time Series modeling, one may refer to : <http://www.unofficialgoogledatascience.com/2017/07/fitting-bayesian-structural-time-series.html>

In our case, the response data  $y_t$  denotes the number of deaths on day  $t$  per 100,000 size of the population. Since the response data is non-negative we use the log of number of deaths per 100,000 of the population size and fit a Bayesian Structural Time Series model using different state specifications for the trend. Since the data is spread across a few months only, Seasonality or cyclical patterns are not considered for this data. The time series fed into the model starts from the first instance of a confirmed case for a country till 30th March 2020.

Several state specifications were checked such as Local, Local Trend, Semi Local, AR and Sparse AR. The last 2 options are not performing well at all for any of the countries, using the criterion of MAPE (Mean Absolute Percentage Error) as compared to the other state specification models

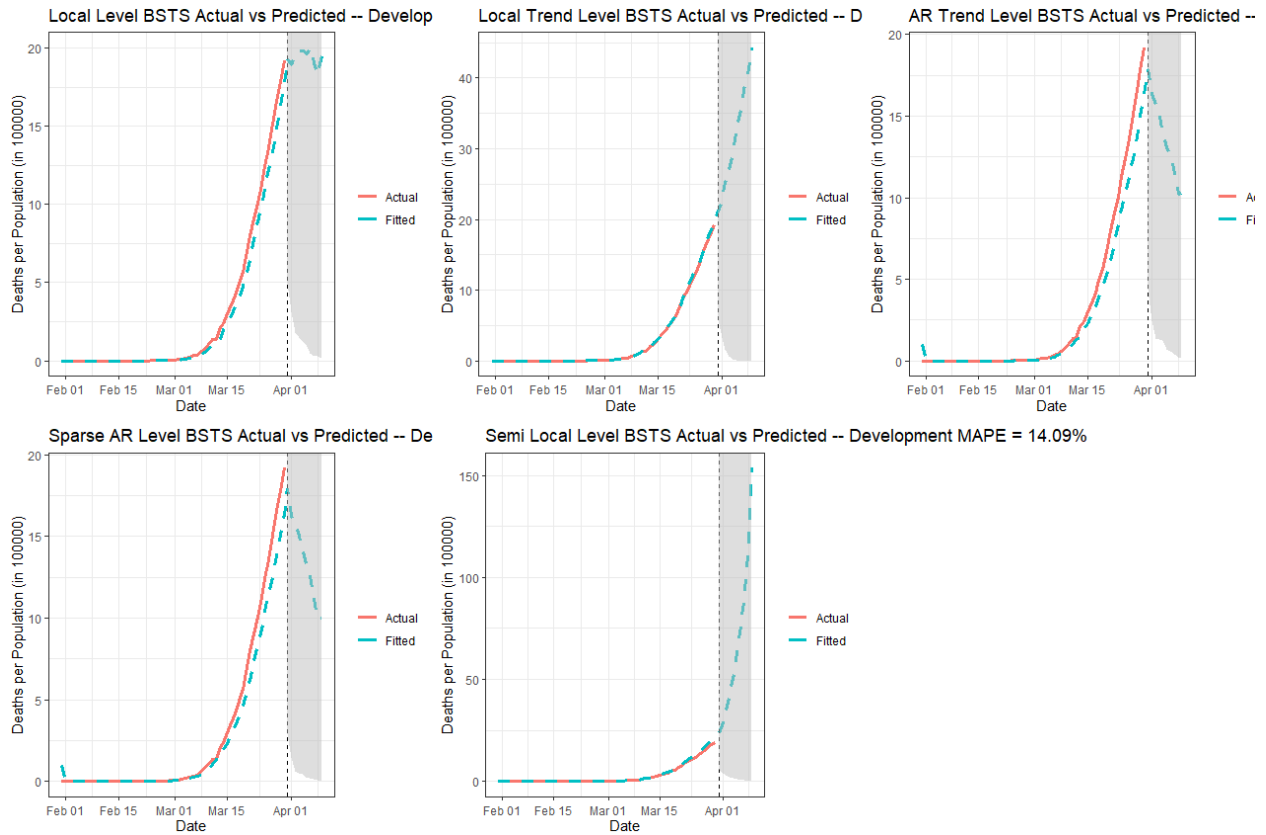
Once the best model is ascertained country wise using MAPE, we use the best-fitted model to predict the outcome (number of deaths per 100,000 size of the population) for the next 10 days. However, in selection of the best model even at a country wise level, there can be several other factors that are of interest such as nature of the predicted curve, over-fitting and interpretability.

Analysis was performed for several countries checking the best latent state specification in terms of the lowest value of MAPE. Below are the results for some of the countries:

Country	BSTS Variant
Italy	Semi Local
Spain	Local Trend
United Kingdom	Local
Switzerland	Semi Local
Portugal	Semi Local
Iran	Semi Local
France	Local
India	Local
Pakistan	Semi Local
Bangladesh	Semi Local
Sri Lanka	Local Trend
US	Local
China	Local Trend
Australia	Local
Thailand	Local
United Arab Emirates	Local

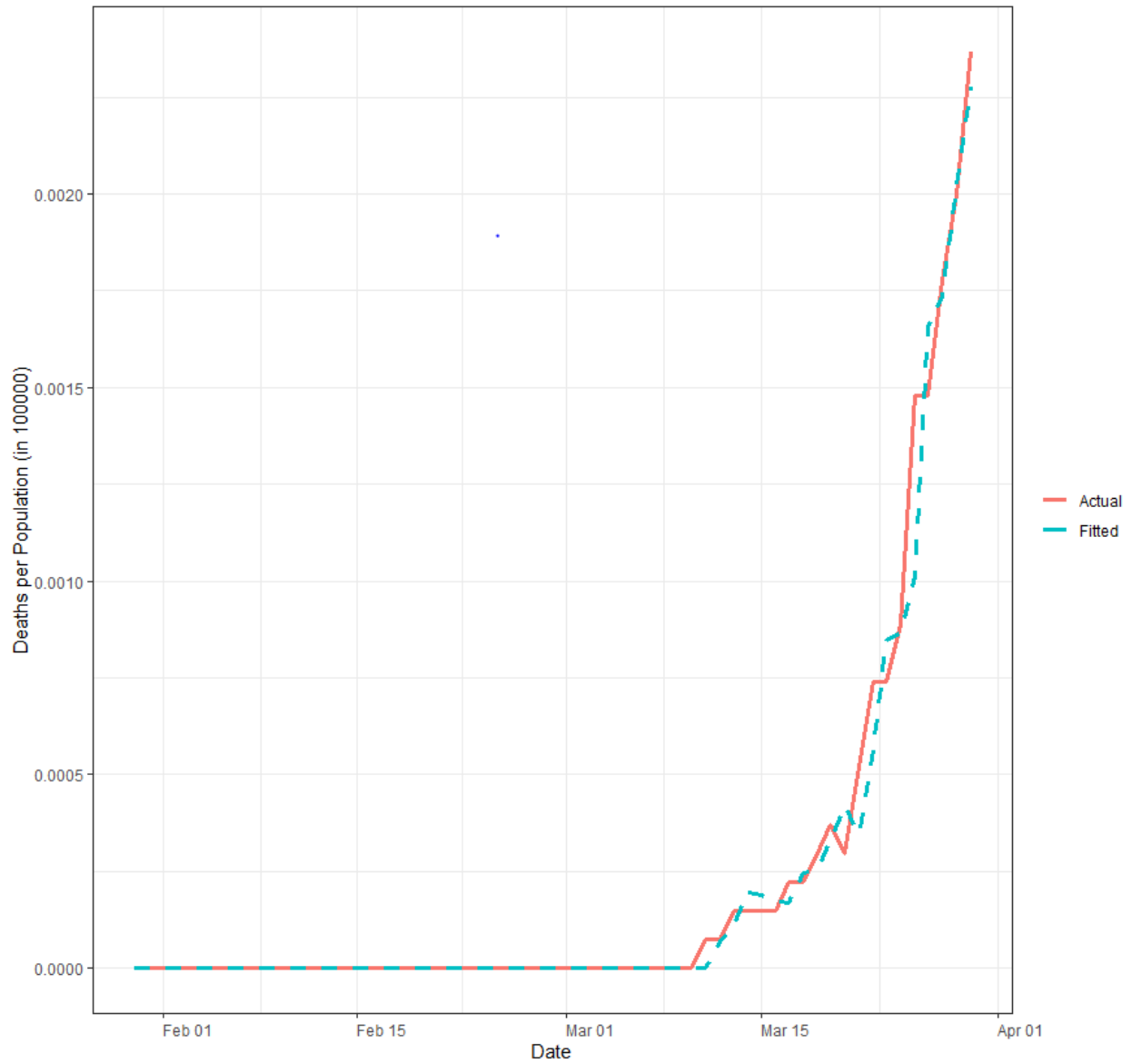
For some countries the best variant comes out to be the Local state specification, whereas for others the Semi Local seems to be working best. However, in the majority of the cases, there isn't a huge difference in terms of magnitude of MAPE for the top 2 variants. Thus, for final selection, the forecasted curve structure and any evidence of over-fitting might be crucial in identifying the best variant.

For each country, MAPE for the different state specification variants were observed along with forecasted graphs for the same. For example, the following graph is obtained for Italy :



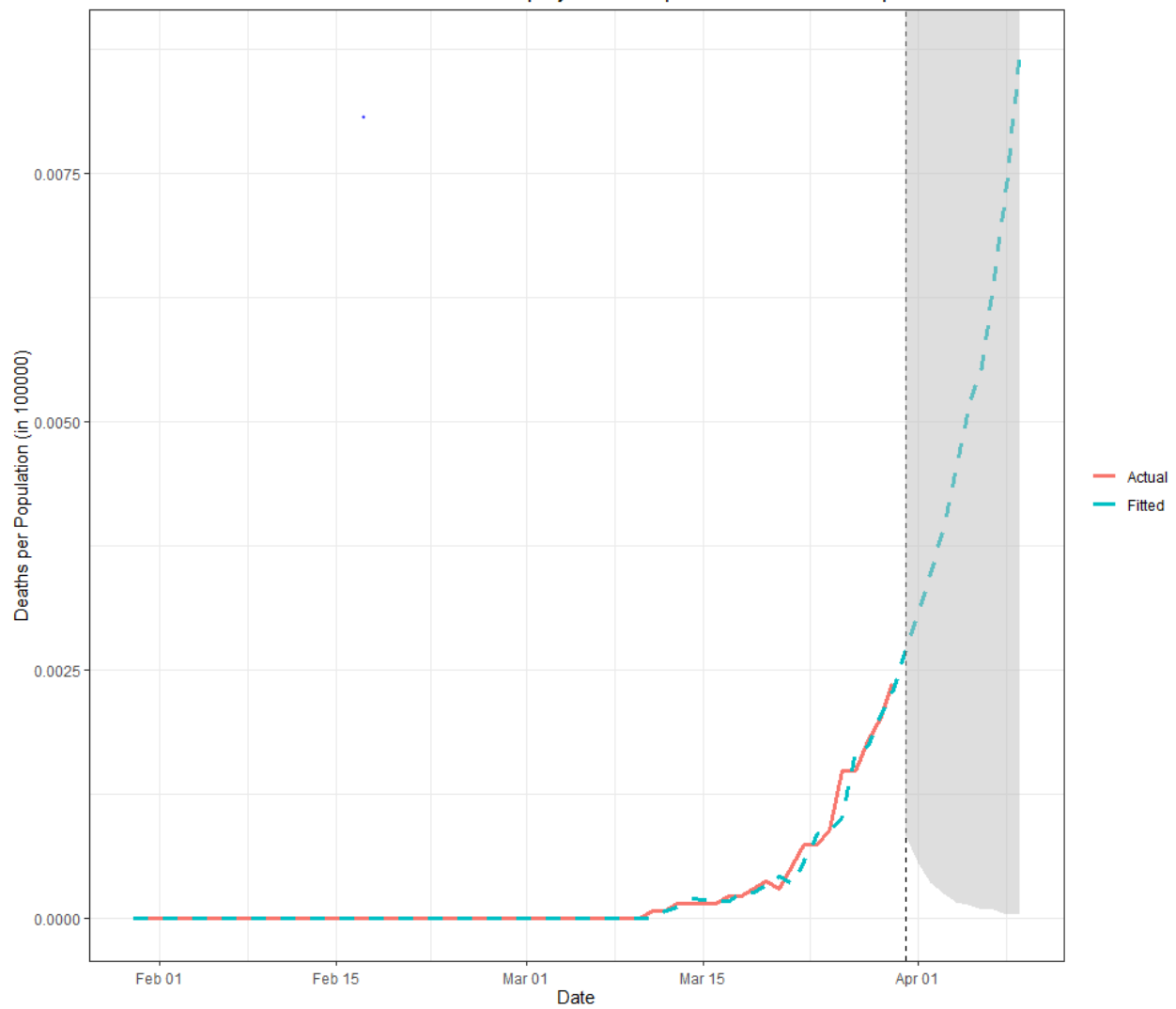
To give an example of the fitted values obtained, considering the case of India, below is a plot of the actual vs fitted values (using Local Level BSTS model as it has the lowest MAPE for India among other variants) across the period for which data is available:

INDIA Local Level BSTS - Actual vs Predicted till 30th March 2020-- Development MAPE = 8.48%



Next, forecasts for the next 10 days are obtained using the same model. The mean of the posterior distribution are plotted for the forecasted range, and the 95% credible intervals are shaded as grey. Based on preliminary analysis, it seems that modeling the data only on latent state specifications and not covariates are resulting in quite wide credible intervals. Below is a plot for the same:

INDIA Local Level BSTS Actual vs Predicted projected till April 9th, 2020-- Development MAPE = 8.48%



The next phase is to identify any one state specification variant which can be used for all the countries and add on to that by using regressors or covariates, which can make the forecasting more robust.

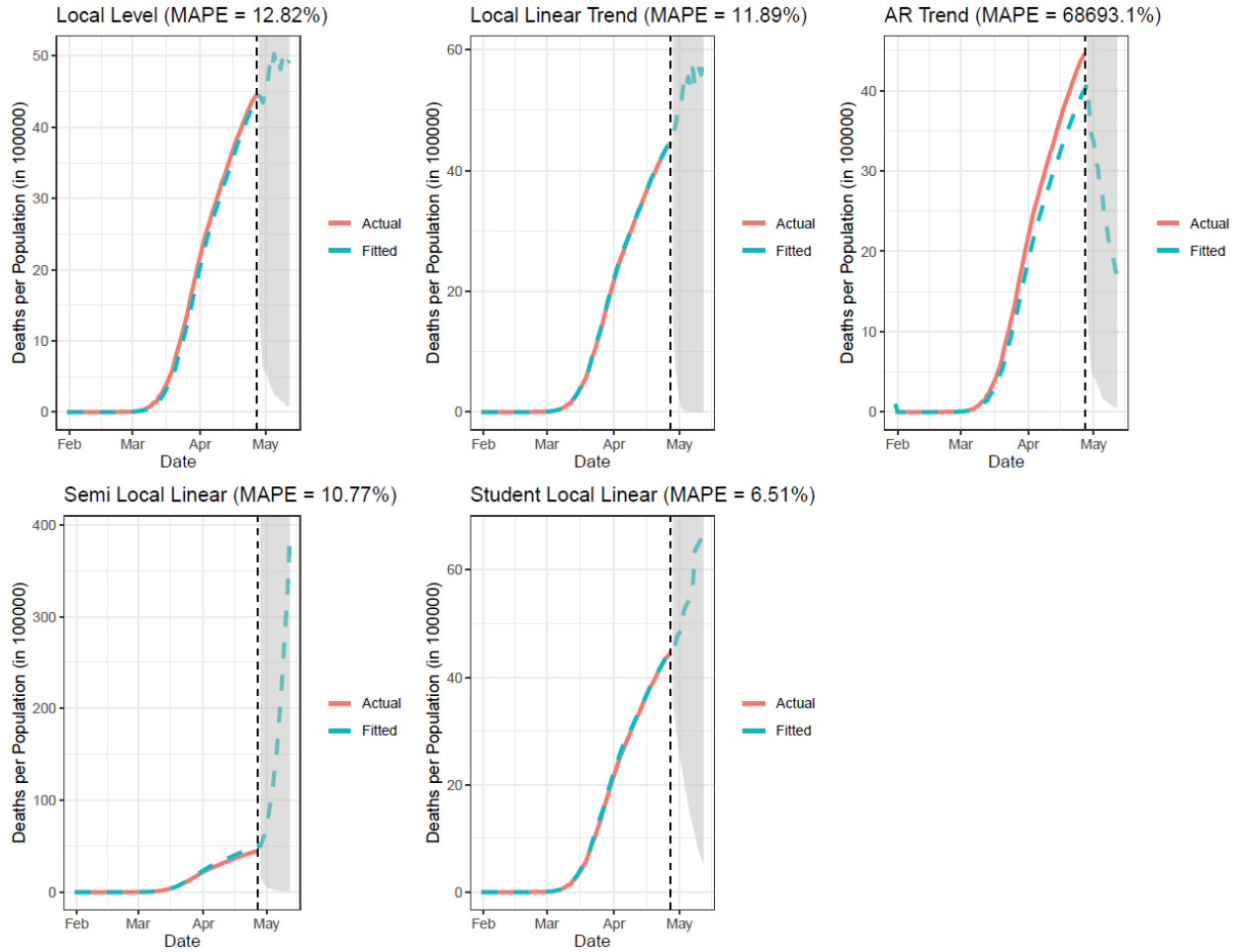
## UPDATE TO THE PREVIOUS VERSION

Recently the *Student Local Linear* specification was also considered as a robust alternative to the *Local Linear trend* specification. The data is aggregated for different countries reported for the time period, from 22nd Jan 2020 to 28th April 2020, as of now

Analysis was performed for countries in the previous table by adding the *Student Local Linear* trend specification and checking the best latent state specification in terms of the lowest value of MAPE. Below are the results for some of the countries:

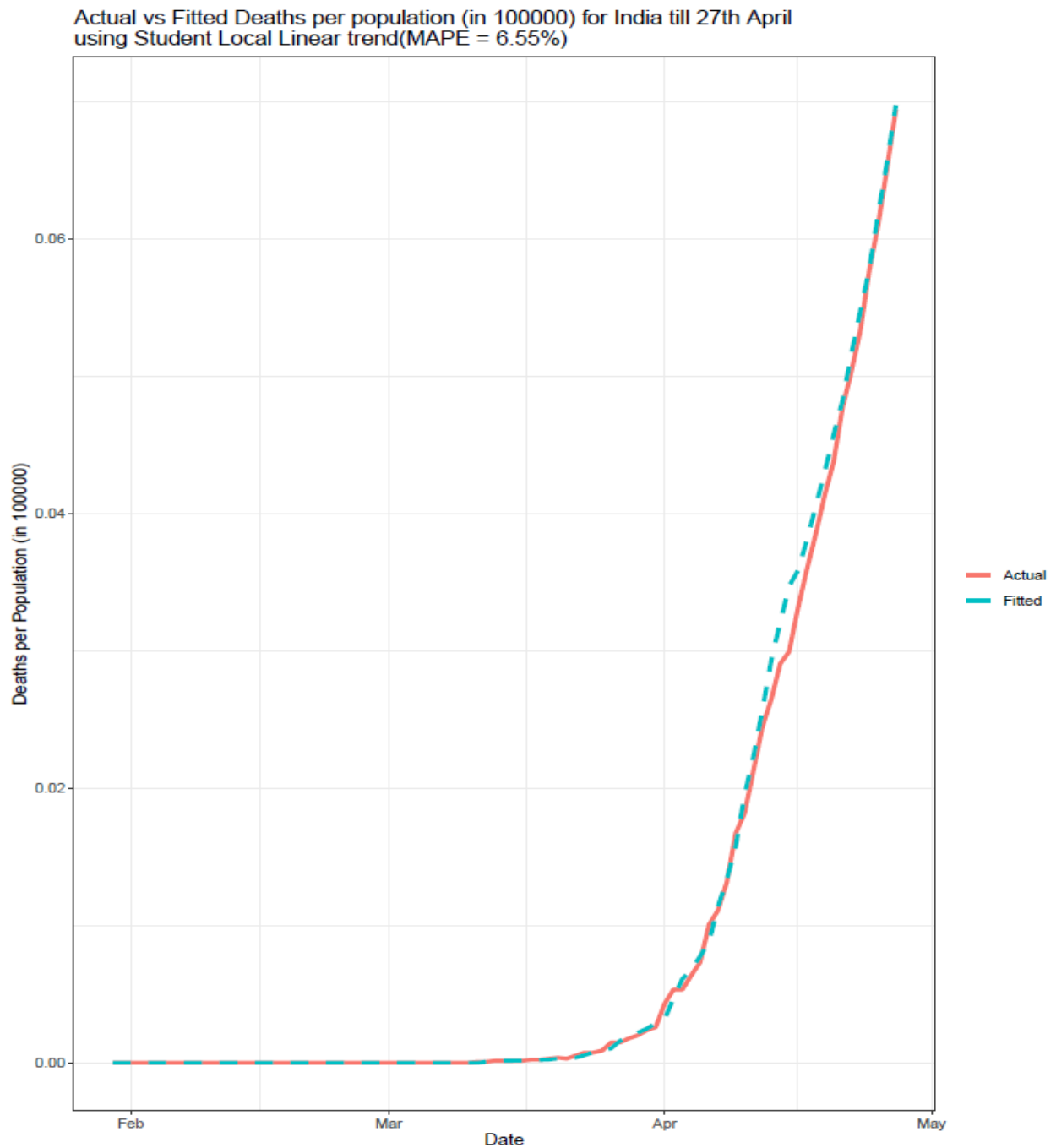
Country	BSTS Variant
Italy	Student Local Linear Trend
Spain	Student Local Linear Trend
United Kingdom	Student Local Linear Trend
Switzerland	Student Local Linear Trend
Portugal	Student Local Linear Trend
Iran	Local Linear Trend
France	Student Local Linear Trend
India	Student Local Linear Trend
Pakistan	Student Local Linear Trend
Bangladesh	Student Local Linear Trend
Sri Lanka	Student Local Linear Trend
US	Student Local Linear Trend
China	Student Local Linear Trend
Australia	Student Local Linear Trend
Thailand	Student Local Linear Trend
United Arab Emirates	Student Local Linear Trend

For almost all countries the best variant comes out to be the *Student Local Linear* state specification, whereas for Iran the *Local Linear Trend* seems to be working best. The plot for Italy is shown below.



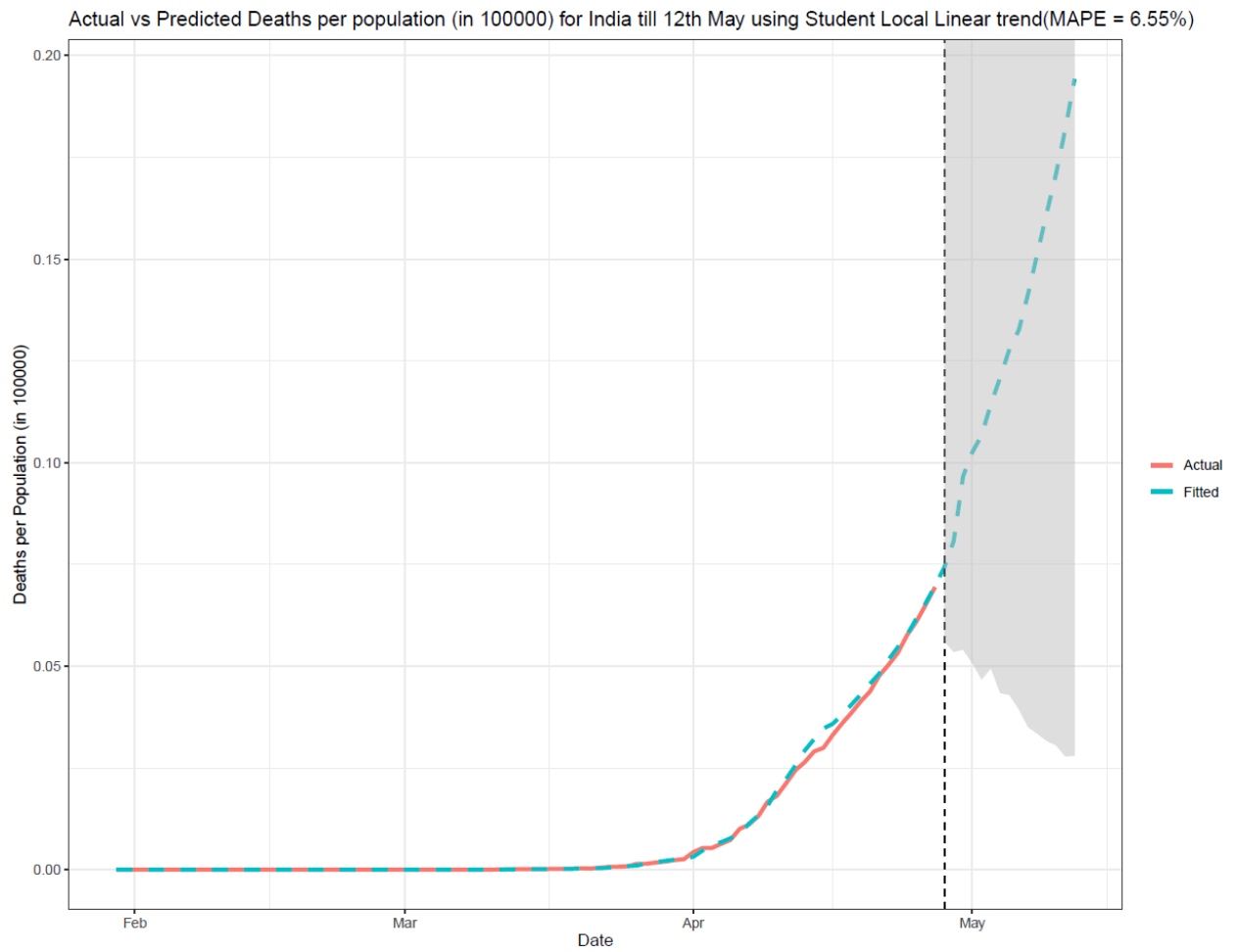
From the above plot we can see that the Student local linear trend performs the best as it has the least MAPE and there is a considerable difference between the top 2 choices of trend specifications.

For India, we fit the Bayesian Structural Time Series model using the Student Local Linear Trend specification as it has the lowest MAPE.



The Student local linear trend specification performs very well as we can see from the above plot. Next, we look at the predicted number of deaths for India using the Student Local Linear trend for the next 15 days.





Although an improvement over the previously considered model, the credible intervals are still quite wide.