Statistical Methods for Data Science Project

Eros Fabrici, Doğan Can Demirbilek, Pietro Morichetti, Michele Rispoli

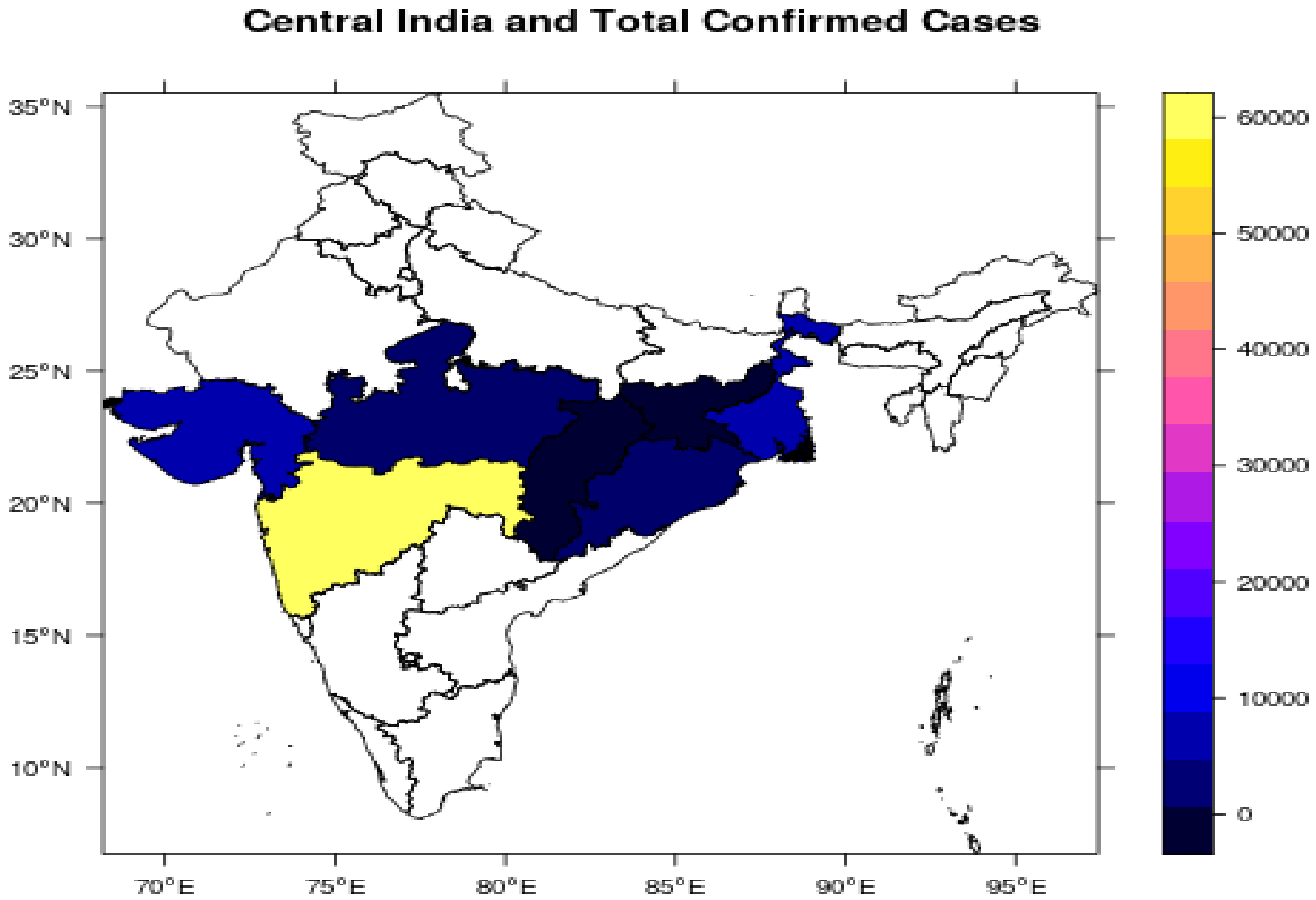
University of Trieste

2019/2020

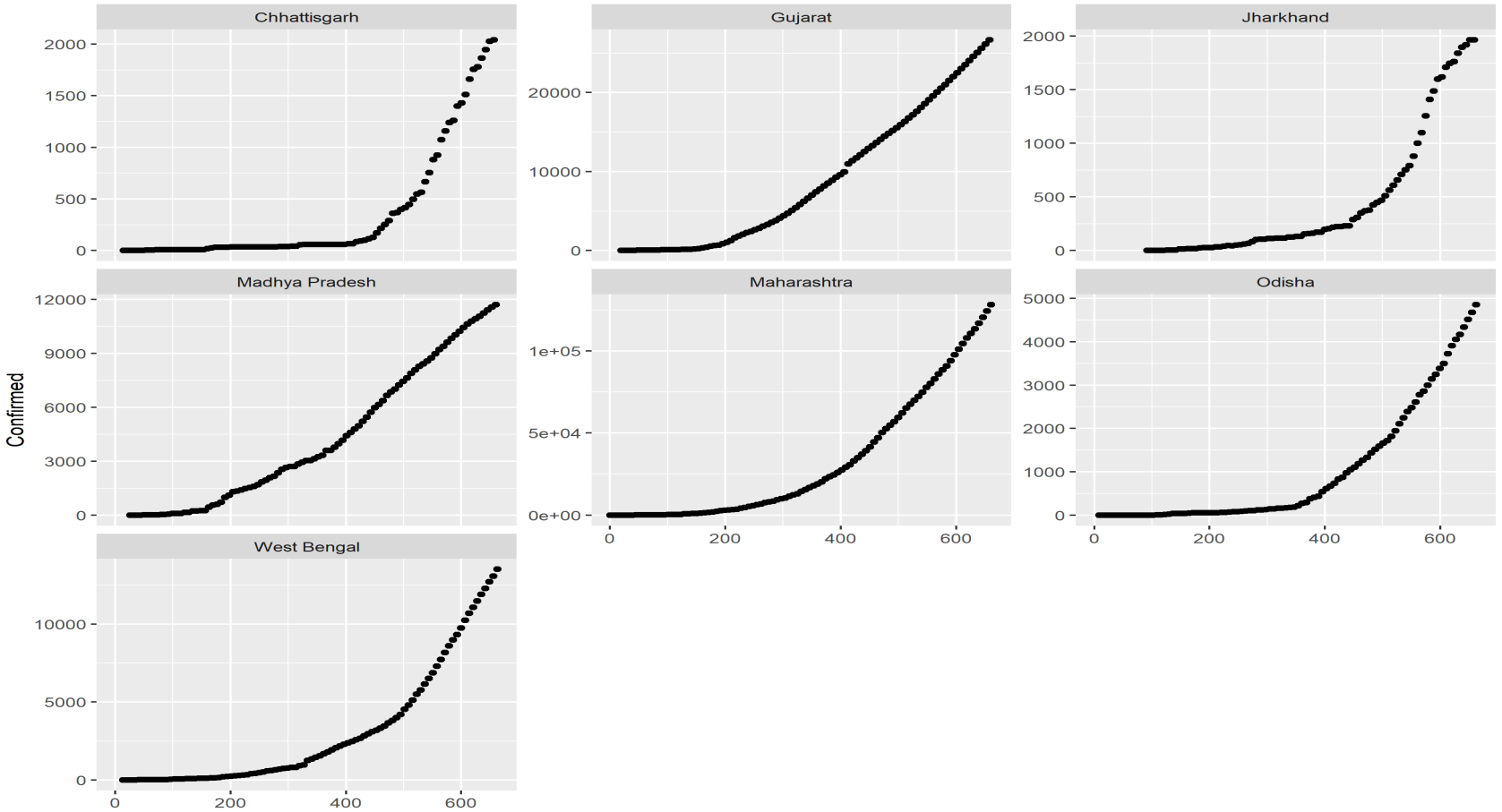
Table of Contents

State

India has been chosen for this project, in particular the following regions were picked: Gujarat, Maharashtra, Madhya Pradesh, Chhattisgarh, Jharkhand, Odisha, West Bengal.


Central India and Total Confirmed Cases
35°N
30°N
25°N
20°N
15°N
10°N
70°E
75°E
80°E
85°E
90°E
95°E
60000
50000
40000
30000
20000
10000
0

Forecasting confirmed cases



Models

After visualising the data, our initial approach is try to model the confirmed cases with a linear model.
We identified to possible models:

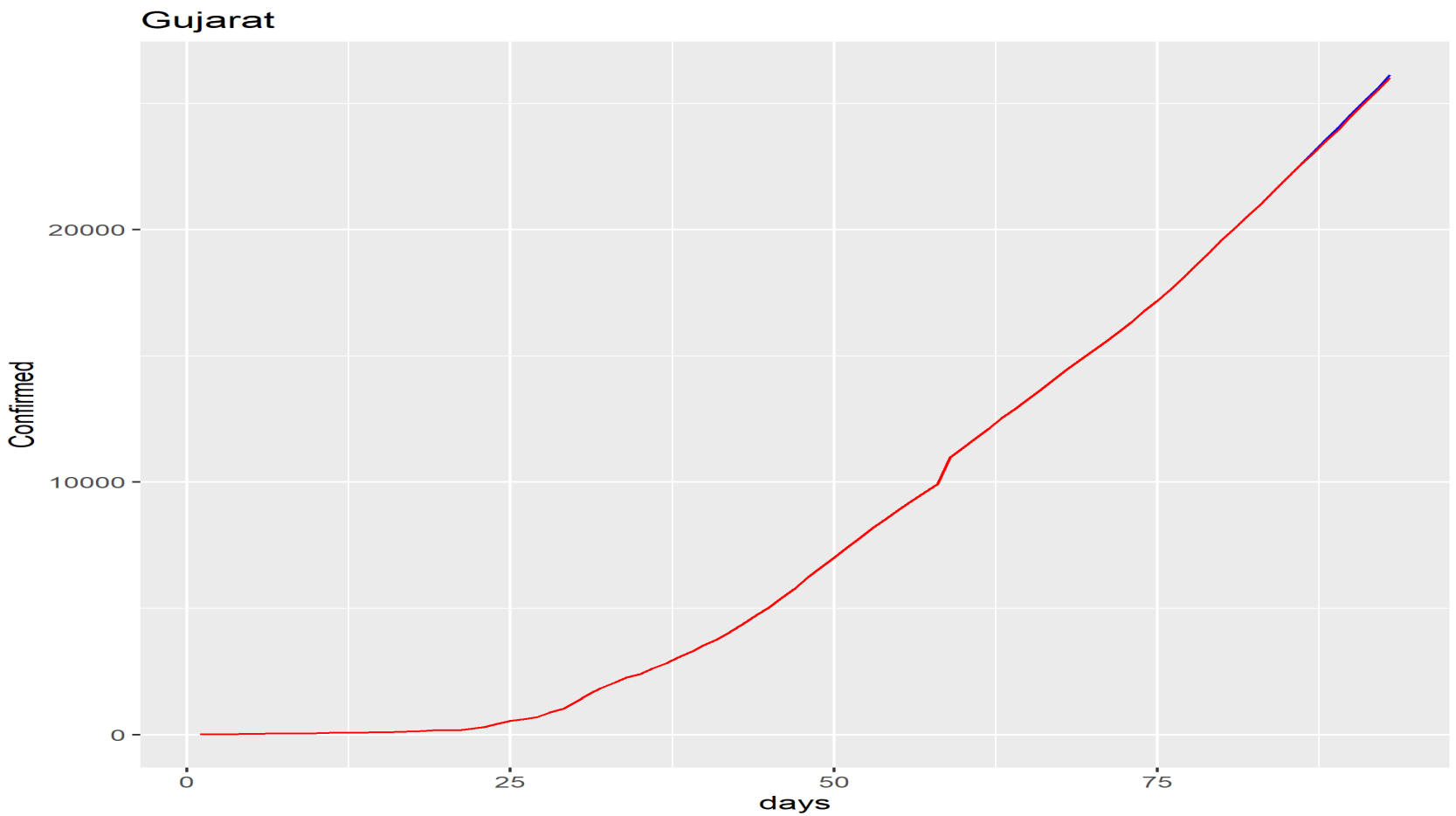
1. $Y_i = \beta_0 + \beta_1 * yesterday\_confirmed_i + \beta_2 * num\_day_i + \beta_3 * swabs_{i-1} + \epsilon_i$
2. $Y_i = \beta_0 + \beta_1 * yesterday\_confirmed_i + \beta_2 * num\_day_i + \beta_3 * swabs_{i-1} + \beta_4 * num\_day^2 + \epsilon_i$

where *num_day* is a counter representing the time. Now we proceed and asses the two models for each state.

Gujarat

Both models have $R^2 = 0.999$. By applying the F-test to the models the p-value is very high (0.92) therefore we cannot reject $H_0 : \beta_4 = 0$, thus the simpler model is preferable. Finally, we observed then that both the β_0 and β_2 have a very high p-value with the t-test. We can conclude then that the best model for this state in order to obtain short term predictions is
$Y_i = \beta_1 * yesterday\_confirmed_i + \beta_3 * swabs_{i-1} + \epsilon_i$

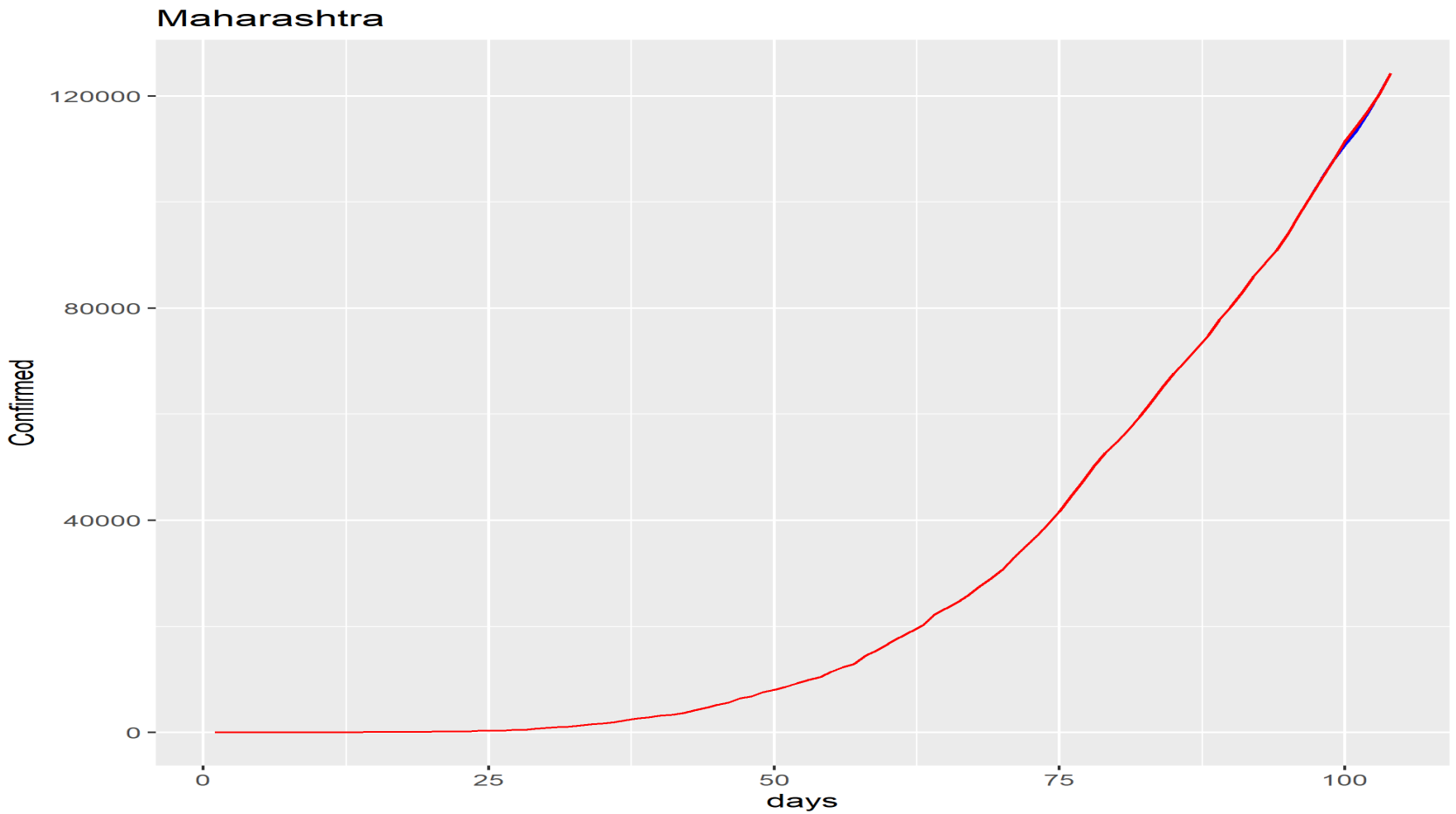
Gujarat prediction



Maharashtra

- Same $R^2 = 0.999$ for both models
- F-test: p-value $= 0.61$ we cannot reject H_0, e.g. we keep the simpler model
- t-test on the simpler model showed that only β_1 and β_3 have a significant p-value ($p \leq 0.05$, therefore we can get rid of the remaining covariates.
- Final model: $Y_i = \beta_1 * yesterday\_confirmed_i + \beta_3 * swabs_{i-1}$

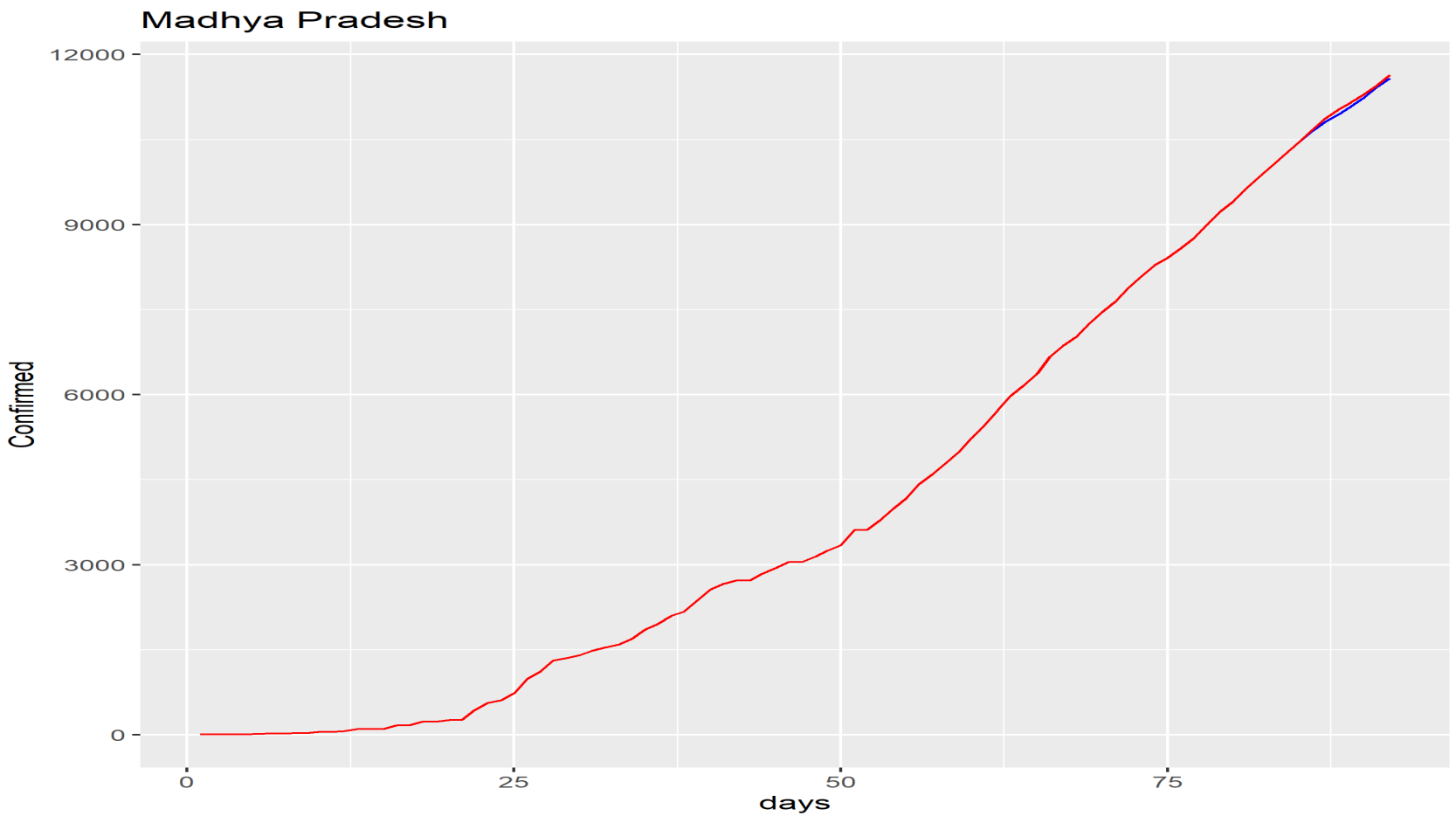
Maharashtra prediction



Madhya Pradesh

- Same $R^2 = 0.997$ for both models
- F-test: p-value $= 0.09$ we reject H_0, e.g. we keep the more sophisticated model
- t-test on the selected model showed that all covariates are significant ($p \leq 0.05$) except for the num_day^2 which has a p-value$= 0.09$, but has we decided to keep after the F-test we continue to maintain it in the model
- Final model: $Y_i = \beta_0 + \beta_1 * yesterday\_confirmed_i + \beta_2 * num\_day_i + \beta_3 * swabs_{i\_1} + \beta_4 * num\_day^2 + \epsilon_i$

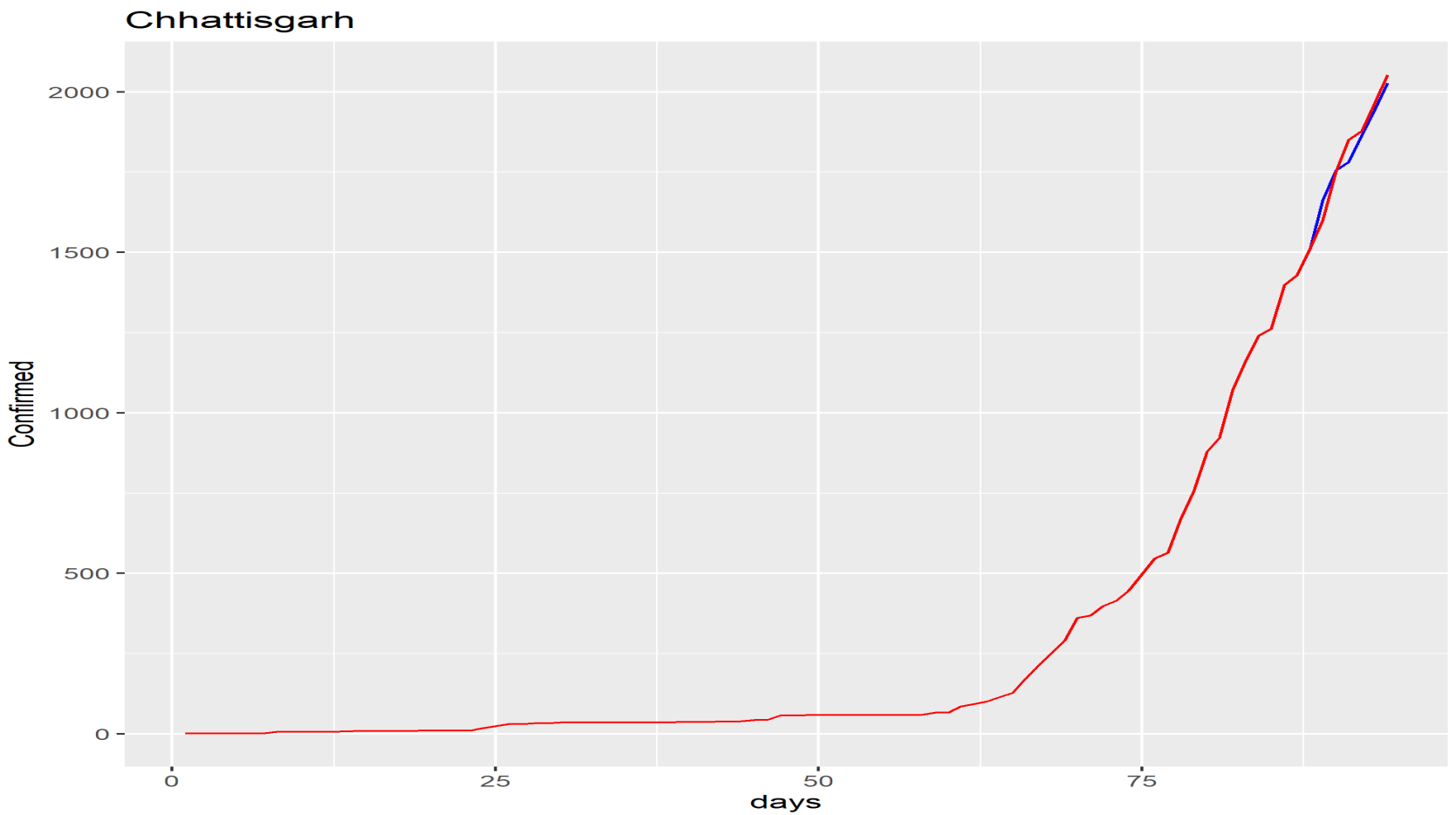
Madhya Pradesh prediction



Chhattisgarh

- $R^2 = 0.9972$ for the first model and $R^2 = 0.9974$ for the second model.
- F-test: p-value $= 0.005$ we reject H_0, e.g. we keep the more sophisticated model.
- t-test on the selected model showed that β_0's p-value is 0.92 and β_2's p-value is equal to 0.25, while for the remaining ones the p-values are very significant.
- Final model: $Y_i = \beta_1 * yesterday\_confirmed_i + \beta_3 * swabs_i\_1 + \beta_4 * num\_day^2 + \epsilon_i$

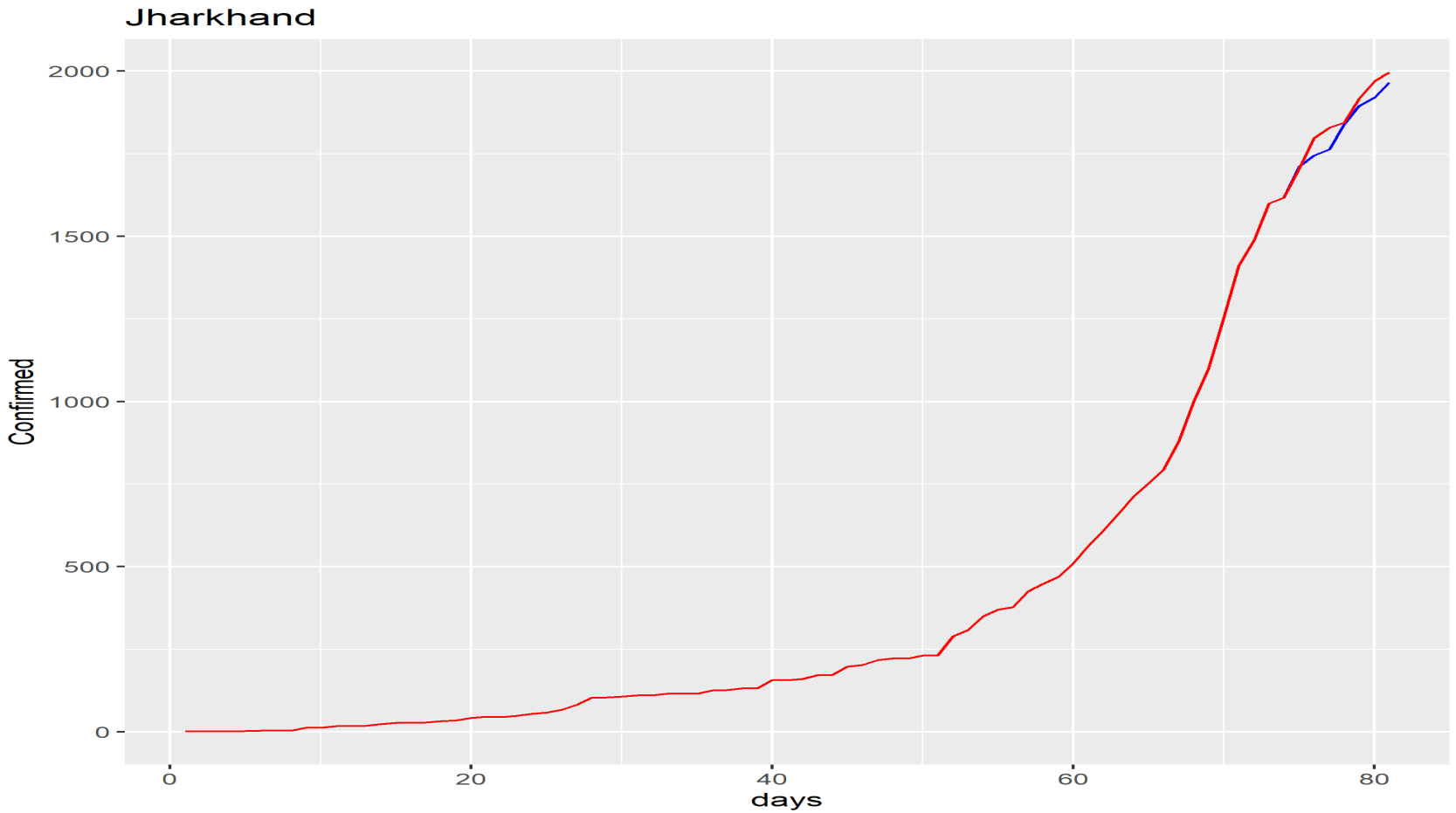
Chhattisgarh prediction



Jharkhand

- $R^2 = 0.997$ for the first model and $R^2 = 0.998$ for the second model.
- F-test: p-value $= 0.001$; we reject H_0, e.g. we keep the more sophisticated model.
- t-test on the selected model showed that β_0's p-value is 0.28, while for the remaining coefficients the p-values are ≤ 0.05, therefore we can remove β_0.
- Final model: $Y_i = \beta_1 * yesterday\_confirmed_i + \beta_2 * num\_day_i + \beta_3 * swabs_{i-1} + \beta_4 * num\_day^2 + \epsilon_i$

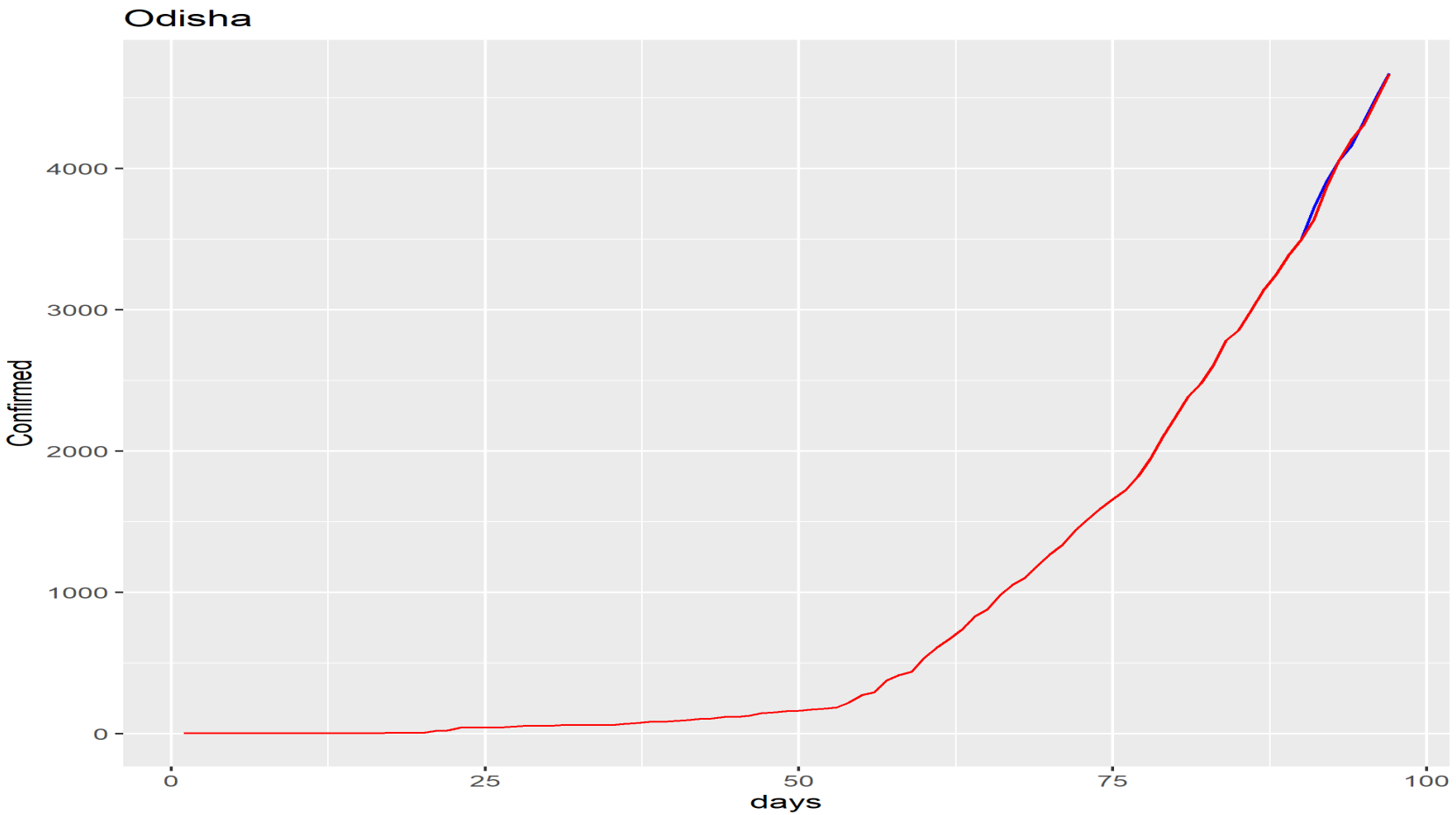
Jharkhand prediction



Odisha

- $R^2 = 0.997$ for both models.
- F-test: p-value $= 0.51$; we cannot reject H_0, e.g. we keep the simpler model.
- t-test on the selected model's coefficients showed that β_0's p-value is 0.32 and β_2's p-value$= 0.115$, while for the remaining coefficients the p-values are ≤ 0.05, therefore we can remove β_0 and $\beta_2 * num\_day$.
- Final model:
  $Y_i = \beta_1 * yesterday\_confirmed_i + \beta_3 * swabs_{i-1} + \epsilon_i$

Odisha prediction



West Bengal

- $R^2 = 0.9998$ for both models.
- F-test: p-value $= 0.67$; we cannot reject H_0, e.g. we keep the simpler model.
- t-test on the selected model's coefficients showed that β_0's p-value is 0.88 and β_2's p-value$= 0.09$, while for the remaining coefficients the p-values are ≤ 0.05, therefore we can remove β_0 and $\beta_2 * num\_day$.
- Final model:
  $Y_i = \beta_1 * yesterday\_confirmed_i + \beta_3 * swabs_{i-1} + \epsilon_i$

West Bengal prediction