Modelling the Corona Spread

A simple model to predict the spread of the COVID-19 virus

Christoph Siebenbrunner[1]

Abstract

We present a simple model that estimates the ultimate spread of contagion of the COVID-19 virus in a country based on the latest available data.

Keywords: contagion model.

1. Introduction

This document explains the model behind the algorithm available at https://github.com/siebenbrunner/covid.

2. The Model

We posit a very simple contagion model that only considers the spread of the disease in an initial population starting with only one infected individual. In each time period every individual is exposed to a randomly drawn subset of the entire population, and every healthy individual that is exposed to an infected one transmits the disease with the same probability. All parameters are assumed to be constant in time. This leads to the initial value problem:

$$\frac{\mathrm{d}f}{\mathrm{d}t} = b(1-p)f \tag{1}$$

$$f(0) = 1$$

where f is the number of infected people in the population, b is a parameter and $p = \frac{f}{N}$ is the percentage of people that are infected. So N is the number of people that will eventually catch the disease, an unknown quantity that we will seek to learn from the data. The solution of the initial value problem in Eq. 1 is:

Email address: christoph.siebenbrunner@maths.ox.ac.uk (Christoph Siebenbrunner)

[1]Corresponding author. University of Oxford, Mathematical Institute, Woodstock Rd, Oxford OX2 6GG, United Kingdom. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Eurosystem or the OeNB. The authors declare that there is no conflict of interest.

$$f = \frac{e^{bt}N}{e^{bt} + N - 1} = \frac{N}{1 + (N-1)e^{-bt}} \tag{2}$$

taking $\frac{\frac{f}{N}}{1-\frac{f}{N}} = (N-1)e^{bt}$, we obtain the logit function:

$$\ln\left(\frac{\frac{f}{N}}{1-\frac{f}{N}}\right) = \ln(N-1) + bt \tag{3}$$

In order to find an appropriate N, we formulate a statistical model:

$$\ln\left(\frac{\frac{f_t}{N}}{1-\frac{f_t}{N}}\right) = c + b * t + \epsilon_t \tag{4}$$

where $f \in \mathbb{N}^T$ is a T-dimensional vector with the observed number of infections for all time periods $t \in 1 \dots T$ starting from the first period in which an infection was observed, c is an intercept and $\epsilon \in \mathbb{R}^T$ is a residual assumed to satisfy $\mathbb{E}(\epsilon^\intercal t) = 0$.

We now consider the model selection problem of finding N by maximizing the least squares coefficient of determination:

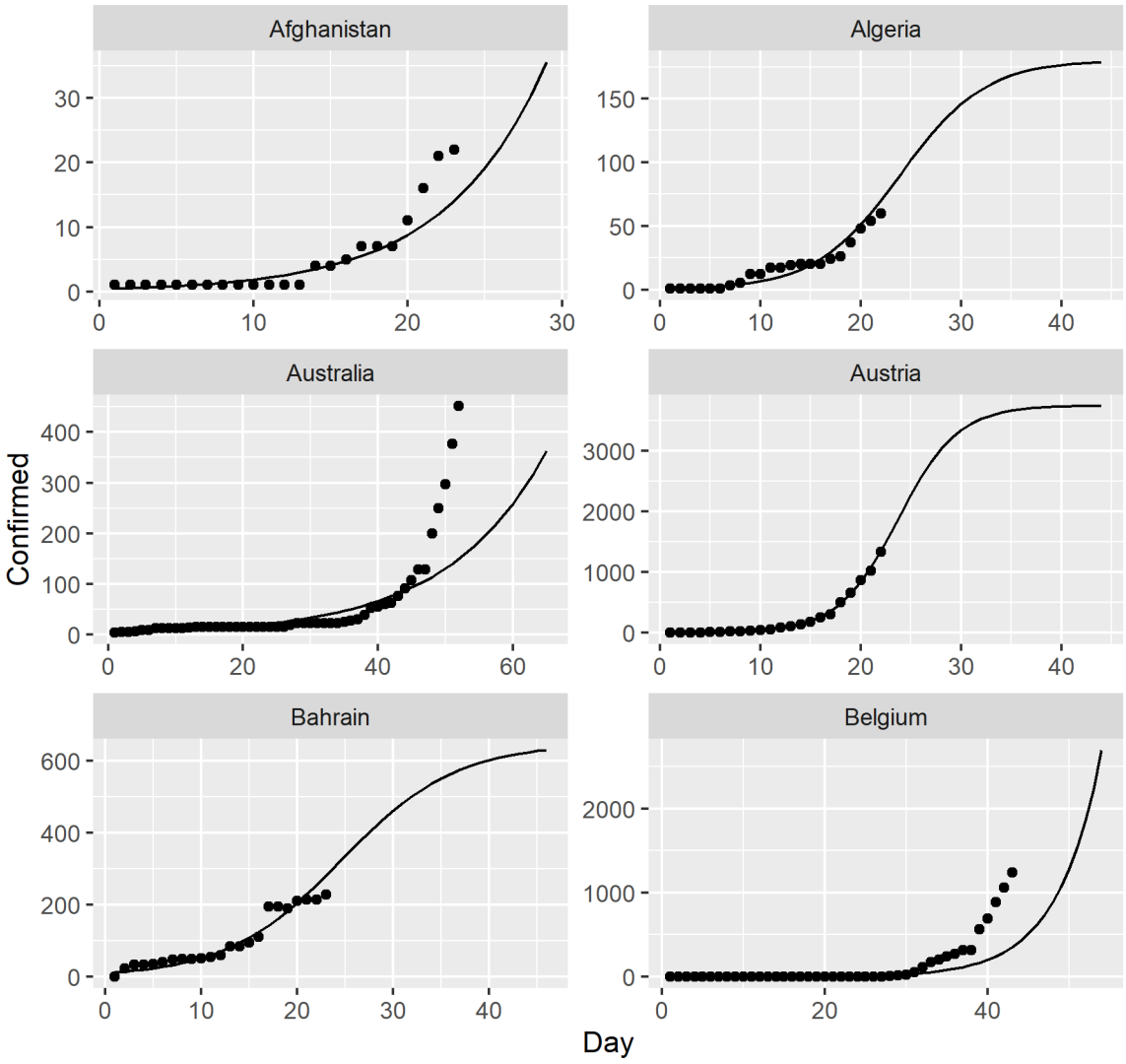
$$\begin{aligned} N = \underset{n \in (f_1 \dots P]}{\operatorname{argmax}} \quad & \frac{\sum_{i=1}^{T}(X_i^\intercal \hat{\beta}(n) - \bar{y}(n))^2)}{\sum_{i=1}^{T}(X_i^\intercal \hat{\beta}(n) - y_i(n))^2)} \\ \text{s.t.} \quad & \hat{\beta}(n) = (X^\intercal X)^{-1} X^\intercal y(n) \end{aligned} \tag{5}$$

where $y_t(n) = \ln\left(\frac{\frac{f_t}{n}}{1-\frac{f_t}{n}}\right)$, $\bar{y}(n)$ is its mean, $X_t = (1, t)^\intercal$ and P is the total population of the country in consideration. The current implementation on https://github.com/siebenbrunner/covid uses binary search to find N. A more robust method, e.g. combining grid search with binary search would be desirable (see discussion). If N is close to the total population of the country, we conclude that the country is currently still in the exponential growth phase.

3. Data

Our data are provided by the Whitening School of Engineering at John Hopkins university. Population numbers for countries are obtained from the World Bank database. [2]

[2] https://github.com/CSSEGISandData/COVID-19



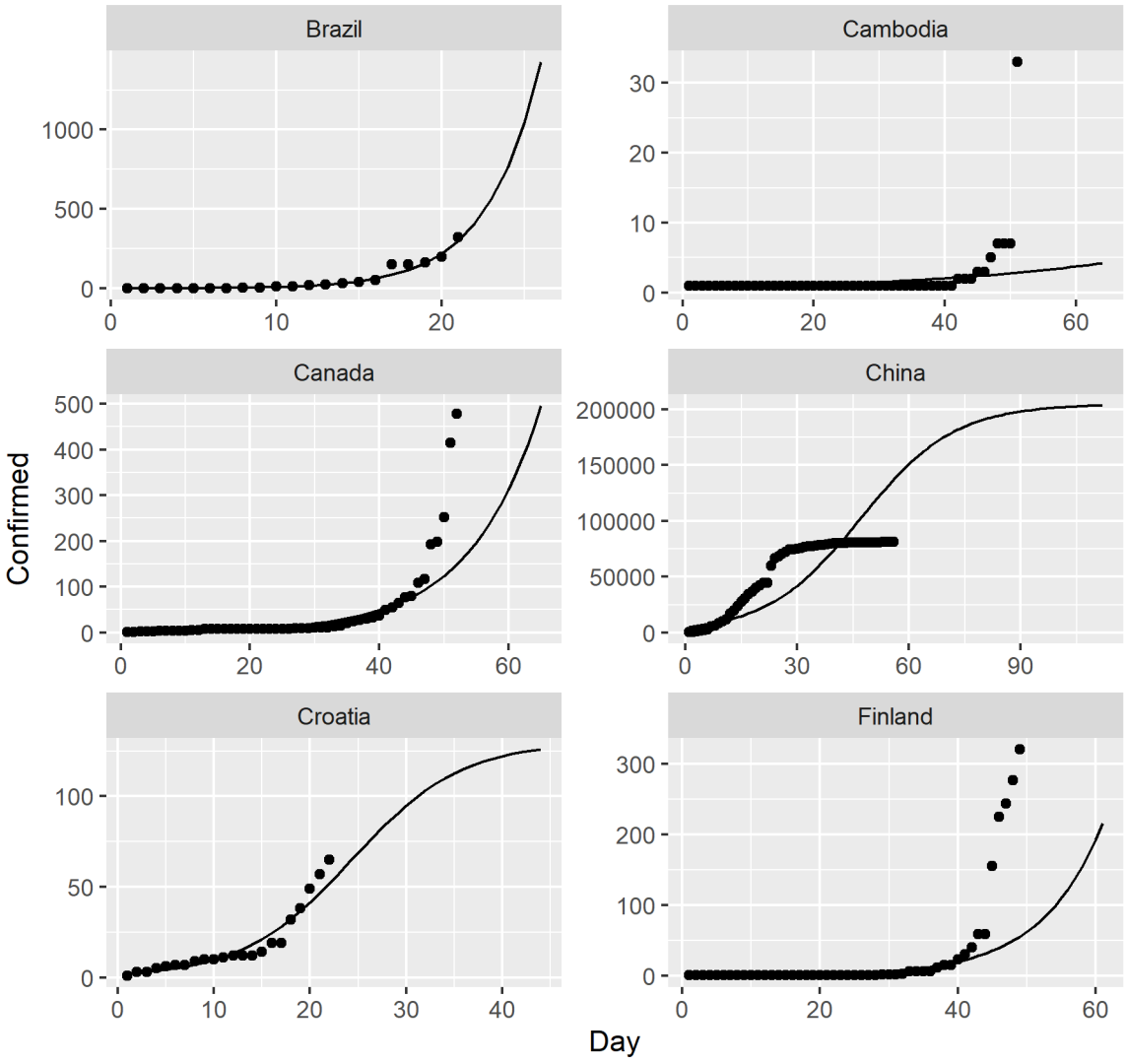
References

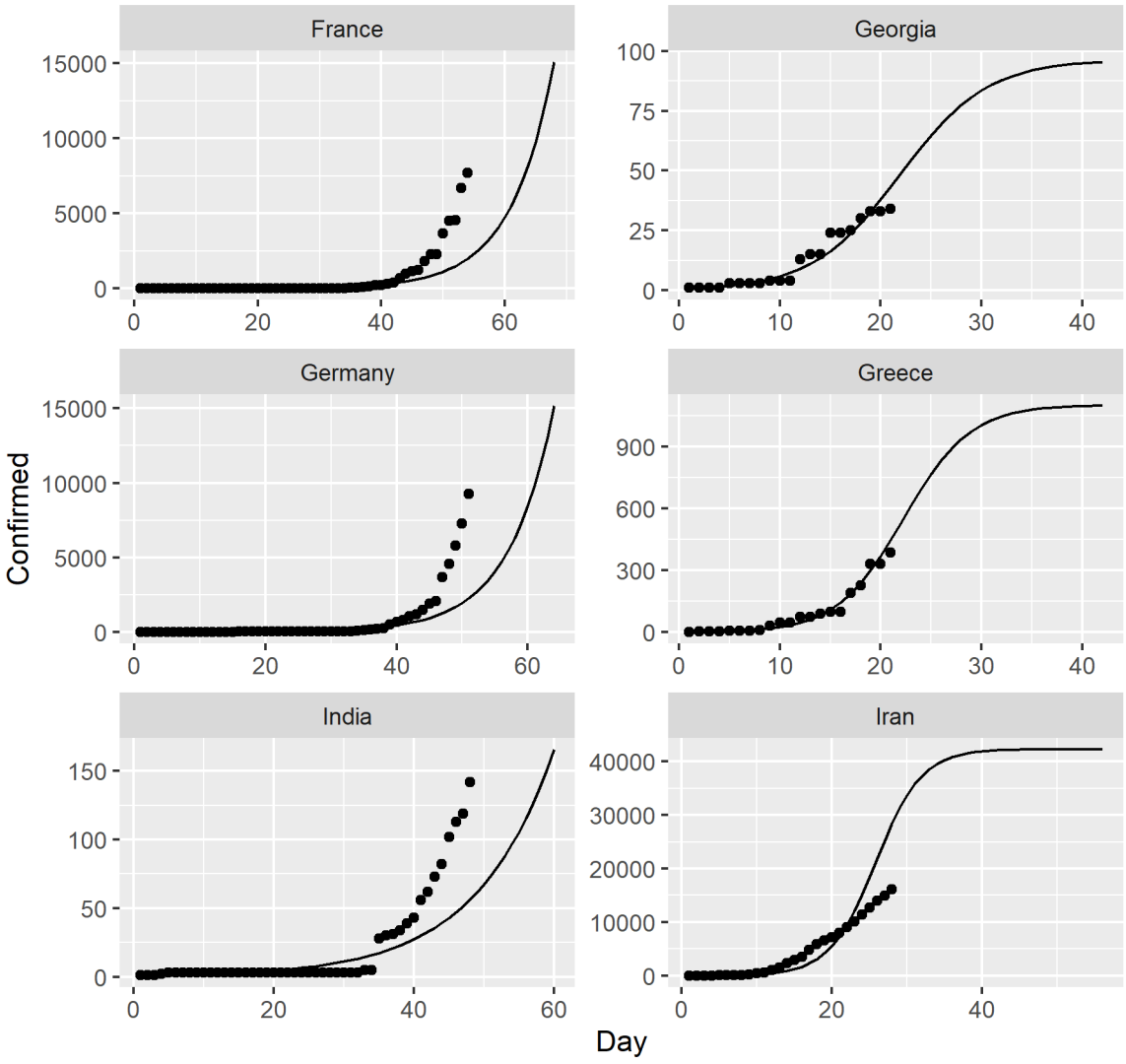
4. Results

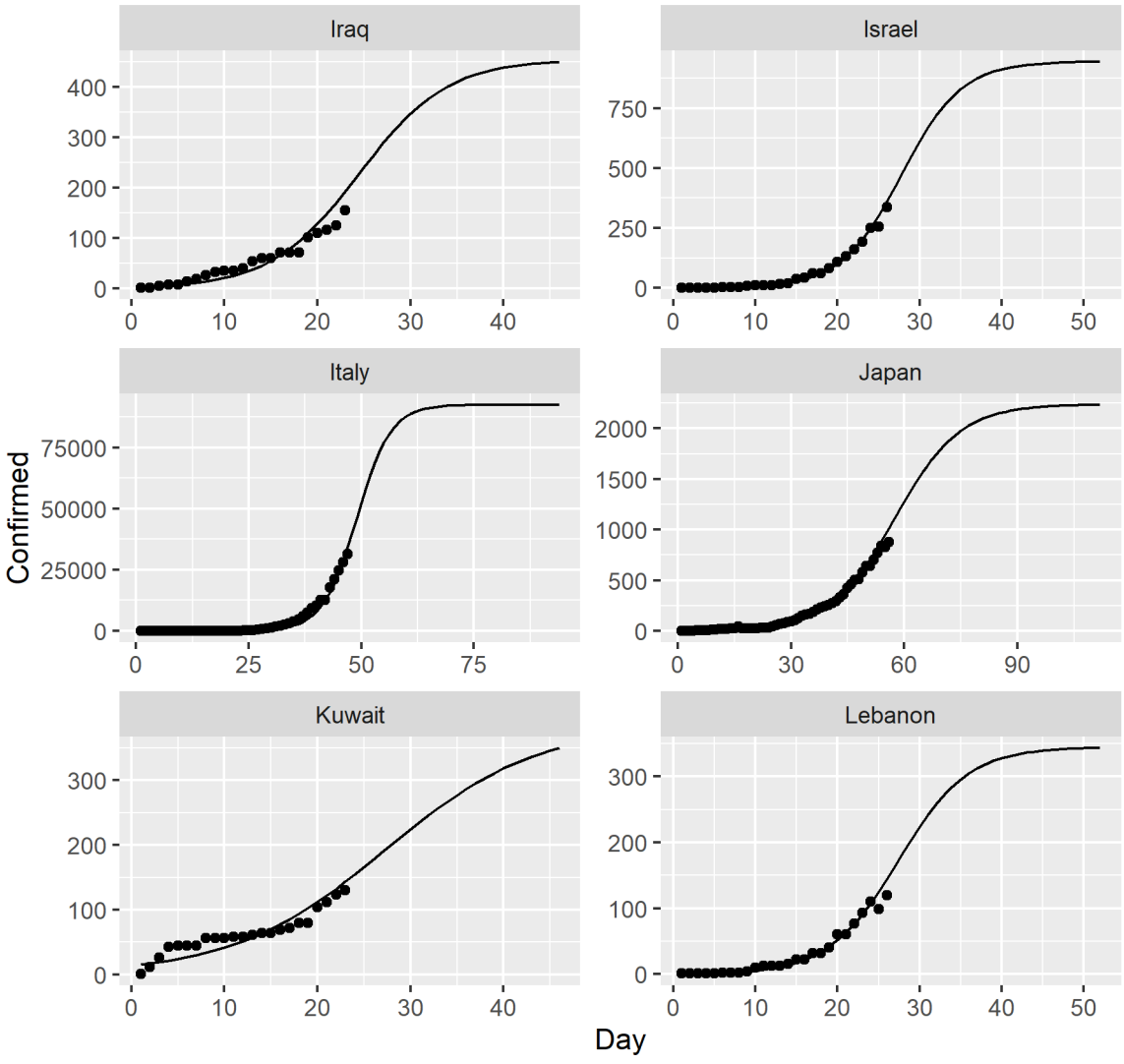
In the figures we plot the logistic growth function for twice the time period of available data, to show a forecast. For those countries that we find to be in the exponential growth phase, we instead show an exponential growth model for a forecast horizon corresponding to 25% of the available data (to ensure that the charts are legible).

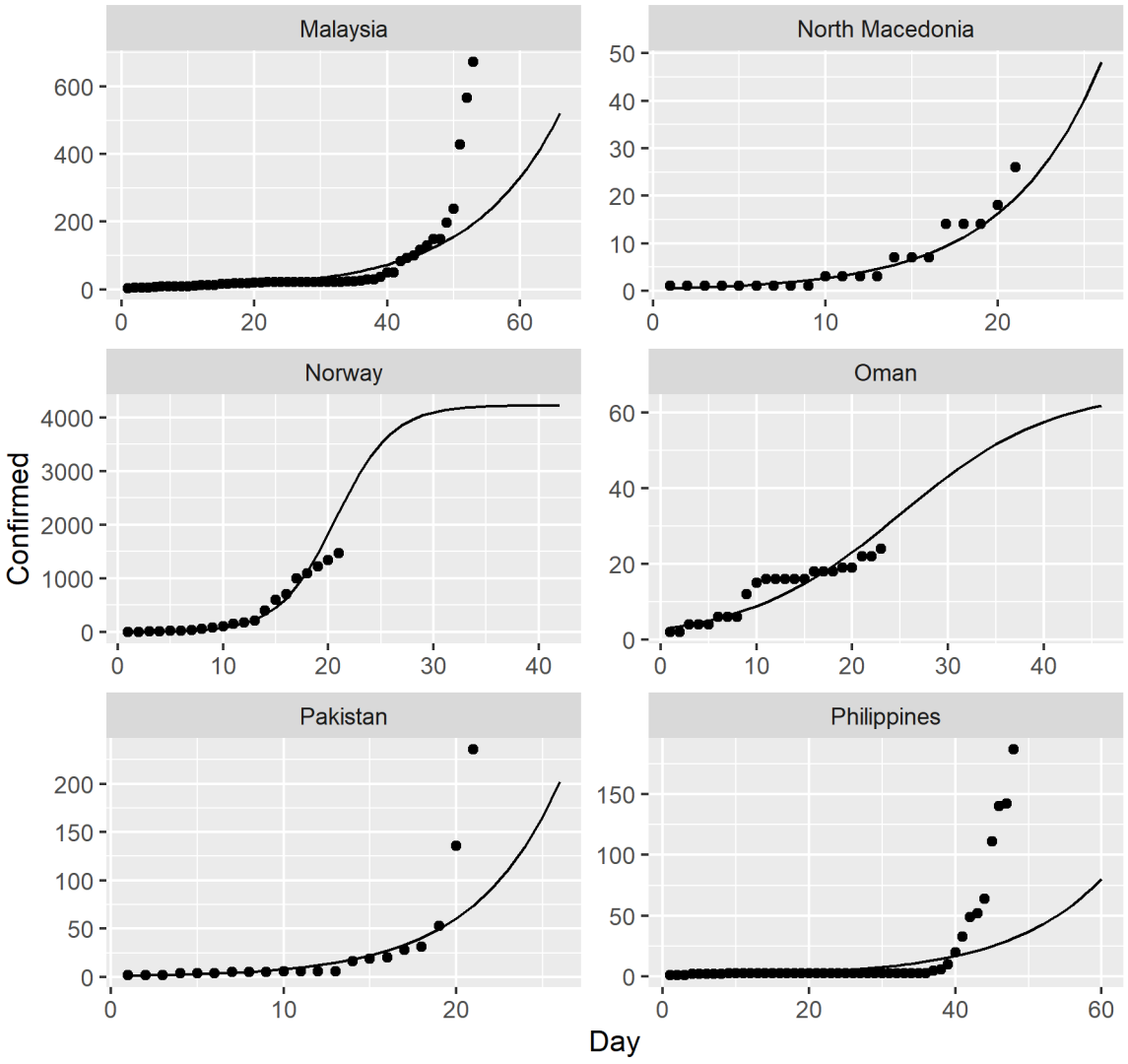
5. Discussion

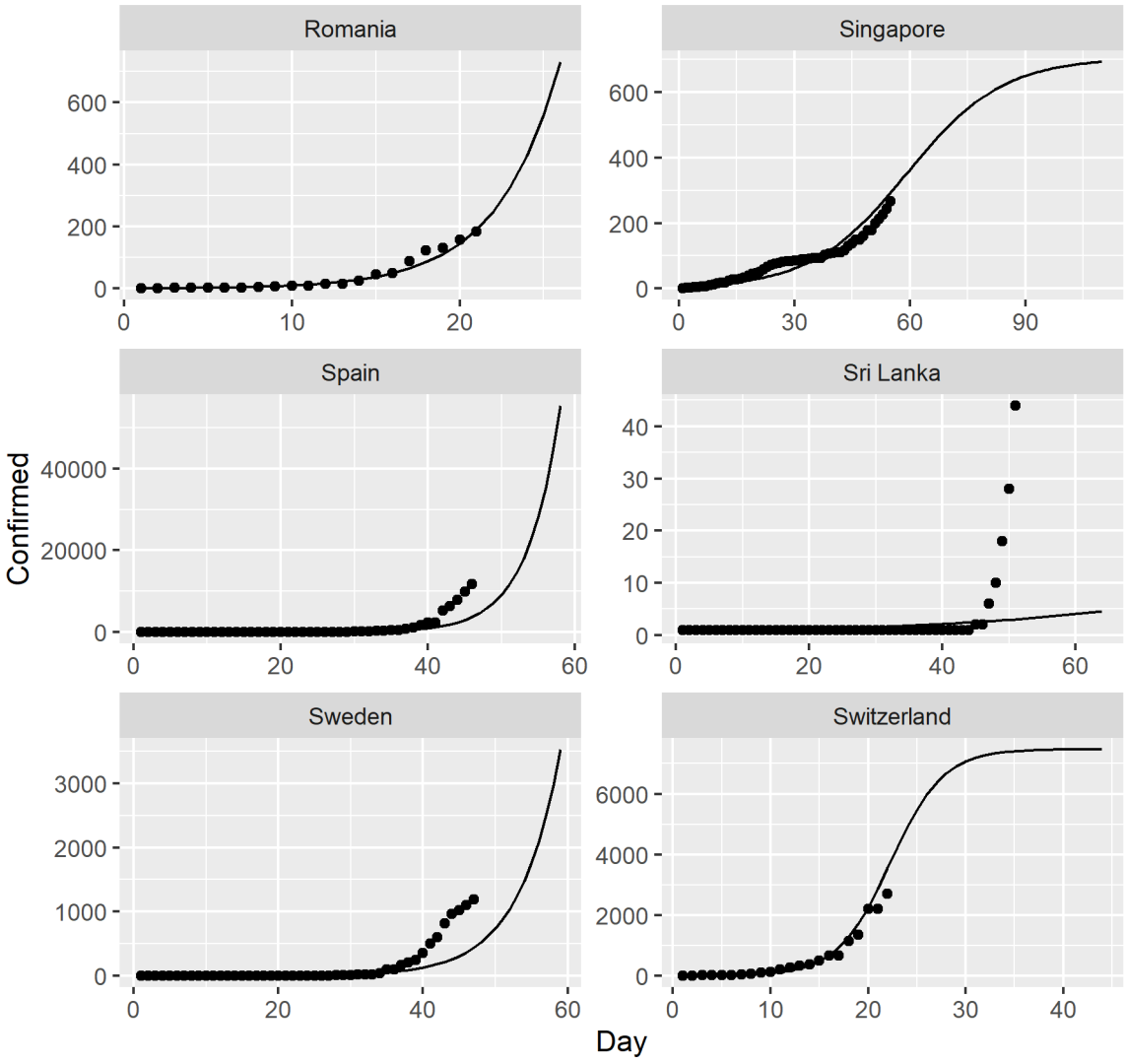
The model suggests that several countries, including China and Italy, are already on track to containment, while others such as the US or the UK are still in the plain exponential growth phase.

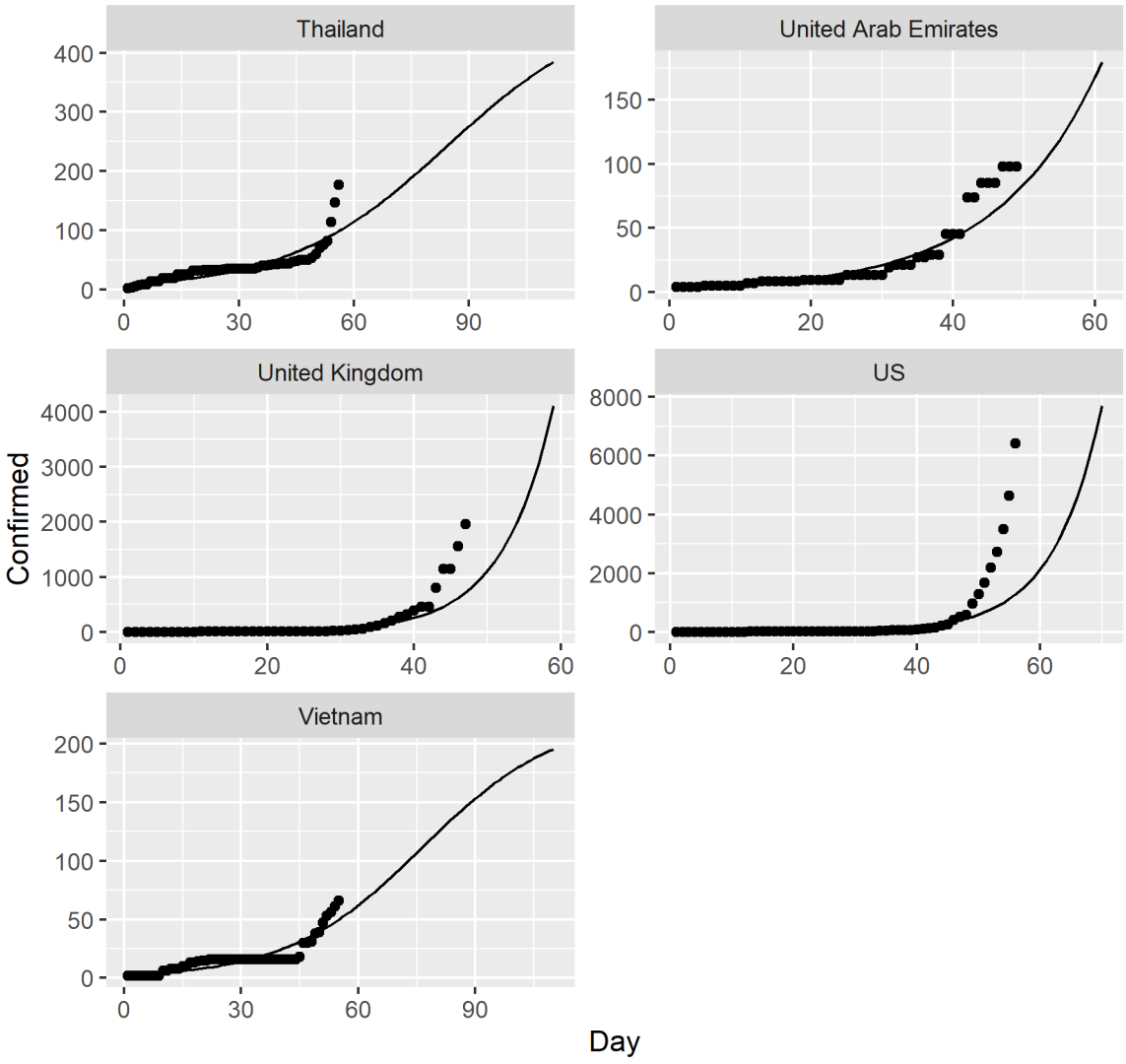

Brazil
Cambodia
Canada
China
Croatia
Finland
Confirmed
Day
0
500
1000
10
20
30
40
60
90
50000
100000
150000
200000
100
200
300
400
50


France
Georgia
Germany
Greece
India
Iran
Confirmed
Day
15000
10000
5000
0
100
75
50
25
900
600
300
150
40000
30000
20000
10000
0
20
40
60
10
30


Iraq
Israel
Italy
Japan
Kuwait
Lebanon
Confirmed
Day


Malaysia
North Macedonia
Norway
Oman
Pakistan
Philippines
Confirmed
Day


Romania
Singapore
Spain
Sri Lanka
Sweden
Switzerland
Confirmed
Day


Thailand
United Arab Emirates
United Kingdom
US
Vietnam
Confirmed
Day

The model could be extended in various dimensions, e.g. a richer statistical model could be considered that includes policy measures such as lockdowns, travel restrictions and others. The theoretical model could be extended to include contagion between different countries, account for recoveries, potential immunity or re-infections of recovered individuals. The results in the current implementation also appear to be sensitive to a hidden hyperparameter choice, namely the lower bound for N taken in the initialization of the binary search. A more robust search method would be desirable here.