

# Modelling the Spread of COVID-19 Cases

## TCD Final Year Project

Anthony Gibbons  
Project Supervisor: Athanasios Georgiadis

March 30, 2021

### Abstract

We construct various models of the Covid-19 pandemic for multiple countries in early 2021. We first construct simple model of the epidemic by using a recurrence equation. We also add a periodic complexity to these simpler models. These recurrence relation models are based on the paper *Mathematical riddles of COVID-19* by Grigorian [1], who studied the outbreak in the first half of 2020. We then use more statistical methods, modelling using time series forecasting methods such as Holt-Winters, ARIMA and Neural Network methods. All of this is with the aim of predicting the course of the pandemic within each chosen country. We implement both new and existing algorithms in R and visualise the results.

**Keywords**— ARIMA, Autoregressive model, COVID-19, Coronavirus, Data analysis, Forecasting, Holt-Winters, Mathematical model, Neural Network, NNAR, Pandemic, Parameter estimation, Prediction, SARS-CoV-2, Statistical model, Time series data.

## **Plagiarism Declaration**

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at <http://www.tcd.ie/calendar>.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at <http://tcd-ie.libguides.com/plagiarism/ready-steady-write>.

**Signed:** Anthony Gibbons

**Date:** 10/03/2021

# Contents

<b>1</b>	<b>Introduction</b>	<b>6</b>
1.1	Previous work on Covid-19 identification and modeling . . . . .	9
1.2	Key aims . . . . .	9
1.3	Limitations . . . . .	9
<b>2</b>	<b>Mathematical Models</b>	<b>11</b>
2.1	Base model . . . . .	11
2.1.1	Model Assumptions . . . . .	11
2.1.2	Notation . . . . .	11
2.1.3	How to select the best model . . . . .	12
2.1.4	Forecasting . . . . .	12
2.1.5	Implementation in R . . . . .	12
2.1.6	Plots . . . . .	12
2.2	Limiting curve $C_r^n$ . . . . .	13
2.3	Moving average . . . . .	14
2.4	Periodic model . . . . .	14
2.4.1	Implementation in R . . . . .	15
2.4.2	Plots . . . . .	16
2.5	Multi-phase model . . . . .	16
2.5.1	Plots . . . . .	17
<b>3</b>	<b>Statistical Models</b>	<b>20</b>
3.1	Holt-Winters' seasonal method . . . . .	20
3.1.1	Definitions and Theory . . . . .	20
3.1.2	Implementation in R . . . . .	20
3.1.3	Plots . . . . .	21
3.2	ARIMA models . . . . .	22
3.2.1	Definitions and Theory . . . . .	22
3.2.2	Implementation in R . . . . .	24
3.2.3	Plots . . . . .	24
3.3	Neural network models . . . . .	25
3.3.1	Implementation in R . . . . .	25
3.3.2	Plots . . . . .	26
<b>4</b>	<b>Model performance with Train/Test sets</b>	<b>27</b>
4.1	Summary of model parameters and results . . . . .	27
4.2	Base and Periodic models . . . . .	28
4.3	HoltWinters model . . . . .	29
4.4	ARIMA model . . . . .	30
4.5	NNAR model . . . . .	31
<b>A</b>	<b>Definitions and Theorems for Mathematical Models</b>	<b>32</b>
A.1	Recurrence equation . . . . .	32
A.2	Results . . . . .	32
<b>B</b>	<b>Source Code and Data Sources</b>	<b>39</b>
B.1	R packages . . . . .	39
B.2	Shapefiles and Datasets . . . . .	39
B.3	Source Code . . . . .	39

## List of Figures

1	Daily Cases Globally . . . . .	6
2	Cases per 1 million population, World . . . . .	7
3	Cases per 1 million population, Europe . . . . .	7
4	Situation by County, Ireland . . . . .	8
5	Individual Comparison of Ireland with various countries, daily cases per million of the population . . . . .	8
6	Comparison of Ireland with various countries, daily cases per million of the population . . . . .	9
7	Contours, Ireland . . . . .	12
8	Contours, Italy . . . . .	12
9	Contours, United States . . . . .	12
10	Basic model, Ireland . . . . .	12
11	Basic model, Italy . . . . .	13
12	Basic model, United States . . . . .	13
13	Limiting curve $Cr^n$ , Ireland . . . . .	13
14	Limiting curve $Cr^n$ , Italy . . . . .	13
15	Limiting curve $Cr^n$ , United States . . . . .	13
16	Limiting Curve Growing Exponentially, Ireland . . . . .	14
17	Moving average $x_n^*(3)$ , Ireland . . . . .	14
18	Moving average $x_n^*(3)$ , Italy . . . . .	14
19	Moving average $x_n^*(3)$ , United States . . . . .	14
20	Oscillating $a$ and $b$ parameters, Ireland, Italy and United States . . . . .	15
21	Periodic model, Ireland . . . . .	16
22	Periodic model, Italy . . . . .	16
23	Periodic model, United States . . . . .	16
24	Multi-phase model, Ireland . . . . .	17
25	Multi-phase model, Italy . . . . .	17
26	Multi-phase model, United States . . . . .	18
27	Multi-phase periodic model, Ireland . . . . .	18
28	Multi-phase periodic model, Italy . . . . .	18
29	Multi-phase periodic model, United States . . . . .	19
30	Normality checks, Ireland, Italy and United States . . . . .	20
31	Seasonal Holt-Winters' Additive Model Algorithm (denoted SHW <sub>+</sub> ) . . . . .	20
32	Comparison of HoltWinters multiplicative (left) and additive (right) algorithms . . . . .	21
33	HoltWinters model, Ireland . . . . .	21
34	HoltWinters model, Italy . . . . .	21
35	HoltWinters model, United States . . . . .	22
36	ARIMA model, Ireland . . . . .	24
37	ARIMA model, Italy . . . . .	24
38	ARIMA model, United States . . . . .	25
39	A linear regression model, or ARIMA( $p, 0, 0$ ) model. . . . .	25
40	A neural network with $p$ inputs and one hidden layer with $k$ hidden neurons. . . . .	25
41	Neural Network Autoregression model, Ireland . . . . .	26
42	Neural Network Autoregression model, Italy . . . . .	26
43	Neural Network Autoregression model, United States . . . . .	26
44	Periodic model with train/test split, Ireland . . . . .	28
45	Periodic model with train/test split, Italy . . . . .	28
46	Periodic model with train/test split, United States . . . . .	28
47	HoltWinters model with train/test split, Ireland . . . . .	29
48	HoltWinters model with train/test split, Italy . . . . .	29
49	HoltWinters model with train/test split, United States . . . . .	29
50	ARIMA model with train/test split, Ireland . . . . .	30
51	ARIMA model with train/test split, Italy . . . . .	30
52	ARIMA model with train/test split, United States . . . . .	30
53	Neural Network Autoregression model with train/test split, Ireland . . . . .	31
54	Neural Network Autoregression model with train/test split, Italy . . . . .	31
55	Neural Network Autoregression model with train/test split, United States . . . . .	31

## List of Code

1	Algorithm for Base Model . . . . .	12
2	Algorithm for Limiting Curve . . . . .	13
3	Algorithm for Periodic Model . . . . .	15
4	Algorithm for HoltWinters Model . . . . .	20
5	Algorithm for ARIMA Model . . . . .	24
6	Algorithm for NNAR Model . . . . .	25
7	Model helper functions . . . . .	39
8	Main algorithm . . . . .	40
9	Multi-phase models . . . . .	45

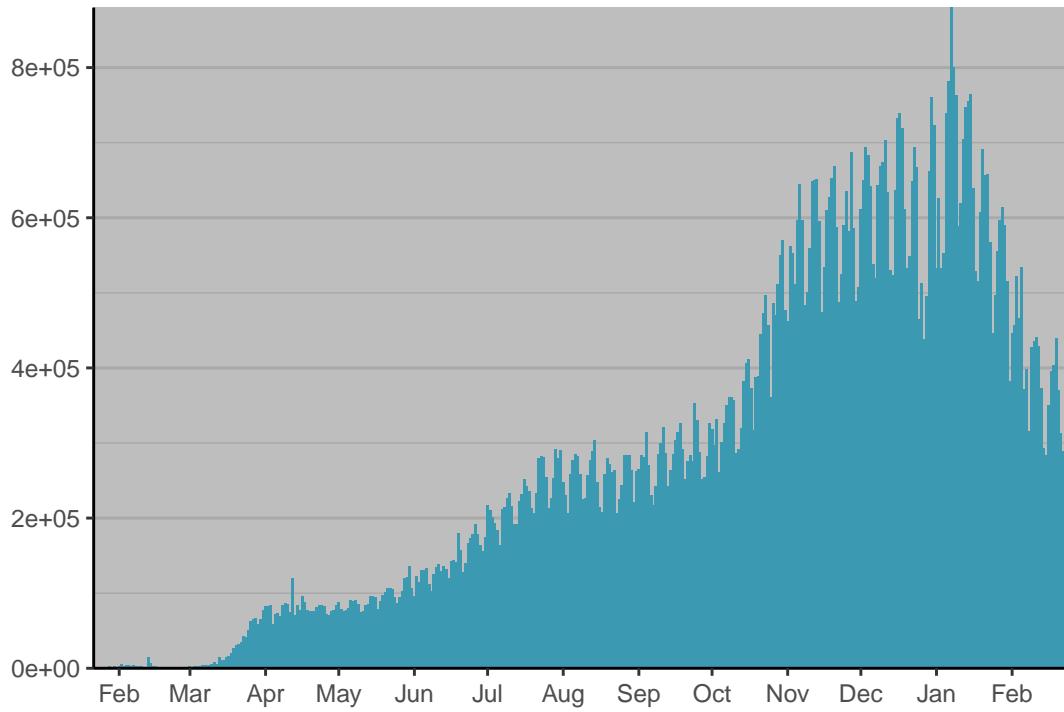
## List of Tables

1	Ireland: mathematical models parameters . . . . .	27
2	Italy: mathematical models parameters . . . . .	27
3	United States: mathematical models parameters . . . . .	27
4	Ireland, 2021-01-12 to 2021-02-16 . . . . .	27
5	Italy, 2021-01-02 to 2021-02-16 . . . . .	27
6	United States, 2021-01-06 to 2021-02-16 . . . . .	27

# 1 Introduction

The Coronavirus disease (COVID-19) was first characterized by the World Health Organisation as pandemic on 11th March 2020 [2]. The outbreak has affected almost every aspect of human life throughout 2020, and is expected to continue for much of 2021.

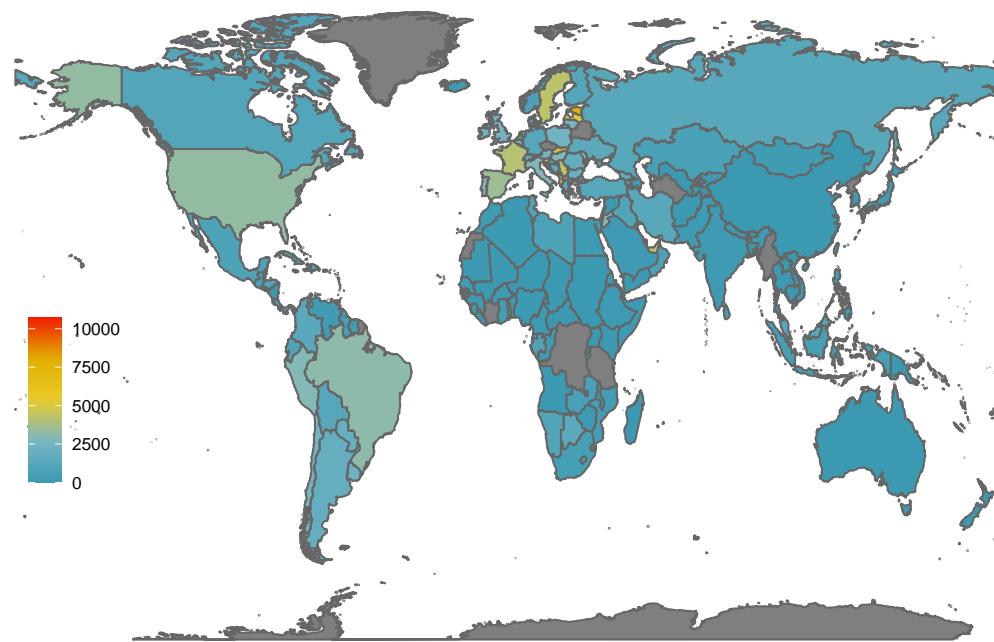
Global Total =111,285,971 as at February 23, 2021



**Figure 1:** Daily Cases Globally

We can map the cumulative number of cases per 100,000 population for each country to see the varying severity of disease spread.

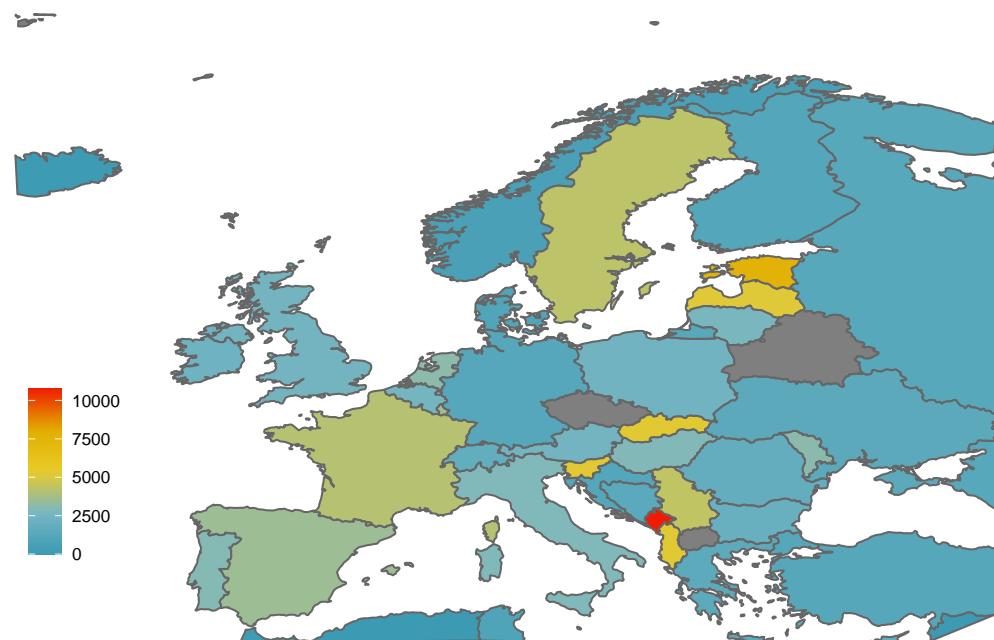
Cases per 1 million population by country  
From February 09, 2021 to February 22, 2021



**Figure 2:** Cases per 1 million population, World

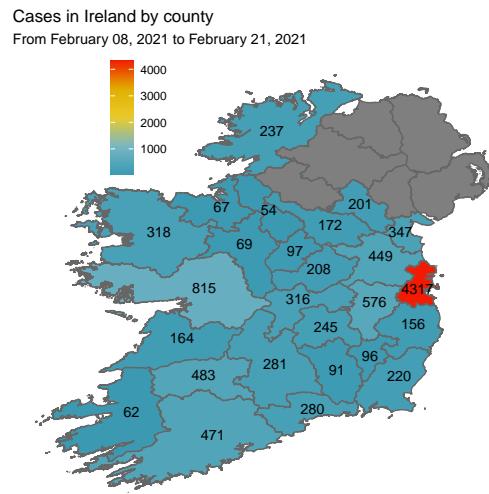
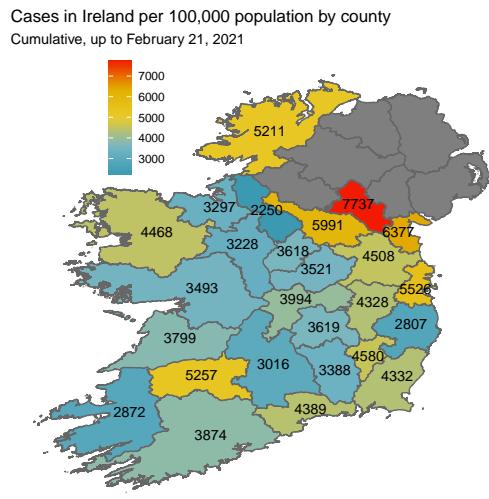
Europe is experiencing an especially high number of cases, proportionally, as well as the US.

Total cases per 1 million population by country  
From February 09, 2021 to February 22, 2021

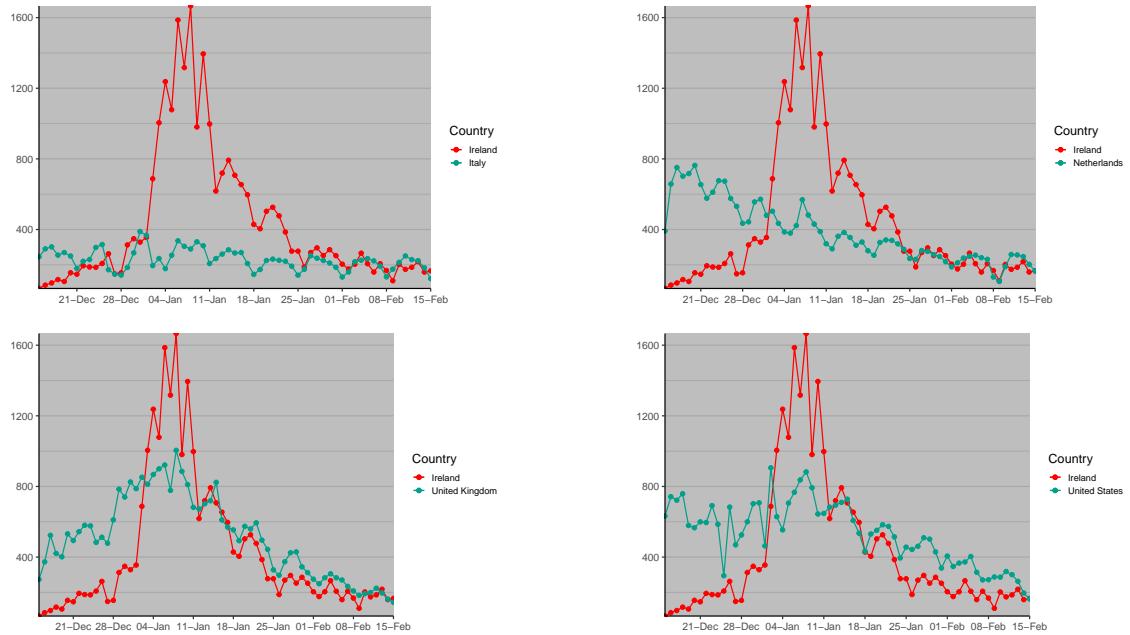


**Figure 3:** Cases per 1 million population, Europe

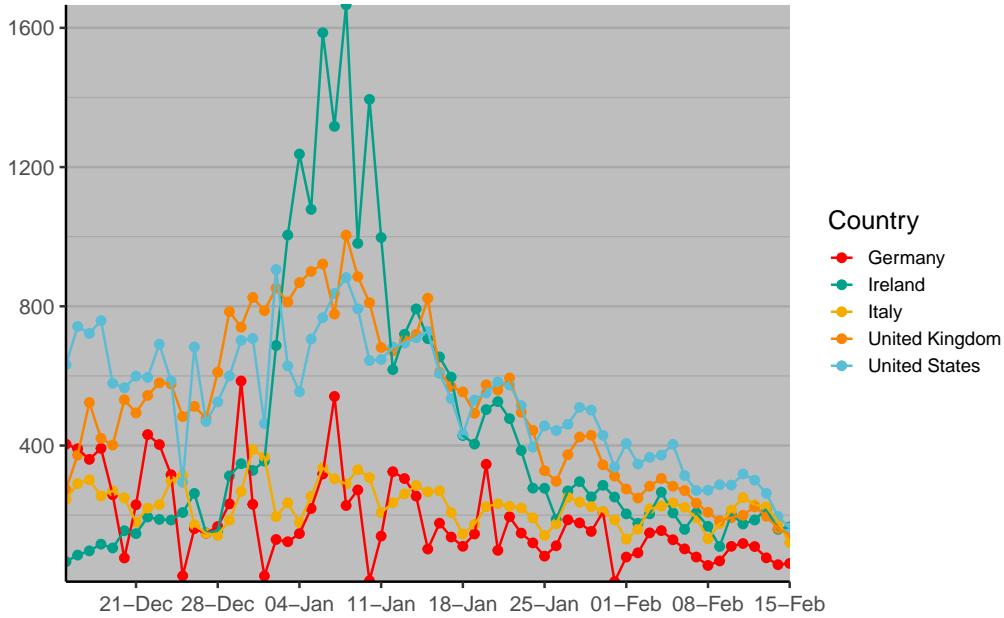
More locally, we see that Ireland also has a clear variation in concentration in cases to date, with Donegal and much of Leinster experiencing sometimes twice as many cases per 100,000 population as the rest of the country.



**Figure 4:** Situation by County, Ireland



**Figure 5:** Individual Comparison of Ireland with various countries, daily cases per million of the population



**Figure 6:** Comparison of Ireland with various countries, daily cases per million of the population

### 1.1 Previous work on Covid-19 identification and modeling

Research in the area of modeling the spread of the pandemic has been extensive and as such it would be impossible to acknowledge all the previous and ongoing work here. I would like to note a few studies (first published towards the beginning of the pandemic) that differ to my approach.

The work involving *external factors* such as government travel restrictions or full lockdowns, carried out by [3], was widely read. However, it was also criticised for their model (which tried to assign a quantitative effect of interventions on disease spread for multiple countries) lacking practical statistical distinguishability, and prompted revisions.

Artificial intelligence models have also been employed to track disease outbreaks in more local areas. The model developed by [4], which relies on phone-based surveys, certainly has the long-run potential to keep the public informed and hopefully reduce the severity of outbreaks in areas where the app is widely used. One drawback of the initial model was its estimation of the peak of case numbers (which is notoriously difficult to predict) being the highest value in the case numbers so far. This does not take into account the shape of many time series during the early stages of the virus outbreak. For example, a strictly increasing time series would have its maximum at the latest time.

### 1.2 Key aims

This project is based on the work in [1], where I attempt to reconstruct the recurrence relation to model the pandemic. This is a largely mathematical model (based on practical assumptions), but of course does not fit well in the long run. It is efficient at explaining singular phases of the pandemic (with a consistent trend), and calculating the infamous  $R_0$  number, defined below, from [5].

**Definition.** The number  $R_0$  is called the *basic reproduction number* and is unquestionably the most important quantity to consider when analyzing any epidemic model for an infectious disease. Each infective individual can be expected to infect  $R_0$  individuals.

### 1.3 Limitations

There are some limitations to the models used in this paper, some of which we can quickly see after the visualisation of the models versus actual cases reported:

- The models seem to fit well when there is an obvious trend in one direction. They do not perform well at or past a peak in case numbers.

- Countries may see their testing capabilities strained or varied, and so some reported numbers may be attributed to a later date than usual.
- Vaccinations began in late 2020 to early 2021 and will likely affect the spread of the virus
- The modeling is done on a univariate time series, whereas external factors such as population densities and other demographics. Spatial data would then be needed to model the disease spread, but this is beyond the scope of the paper.

## 2 Mathematical Models

As per the base and periodic models shown in [1].

### 2.1 Base model

#### 2.1.1 Model Assumptions

- (I) Any infected person becomes ill (symptomatic) and infectious on the  $q$ -th day after infection.
- (A) During each day, each ill person unconfined infects on average  $a$  other persons.
- (B) During each day, a proportion  $b \in (0, 1)$  of ill people loose gets isolated (hospitalized or otherwise) and withdrawn from a further spread of the epidemic.

**Remark.** The number of days before an infected person becomes infectious is called the *latent period*, and before he/she becomes symptomatically ill – the *incubation period*. Here we assume for simplicity that these two periods are equal to  $q$ .

Many models use a set of differential equations for to describe the movement of people between *groups* or *compartments*. The SIR (Susceptible–Infectious–Recovered) model, the most frequently used model in epidemiology, uses a set of 3 such differential equations.

Our main mathematical model (and even some of the statistical models) make use recurrence equations, which have some correspondence to differential equations [6].

#### 2.1.2 Notation

- $x_n$  - the number of infected people that are detected and isolated during the day  $n$ ;
- $y_n$  – the cumulative number of detected cases from the beginning of epidemic by the beginning of the day  $n$ ;
- $z_n$  – the number of ill people at large by the beginning of the day  $n$  (that is, those who were infected at least  $q$  days ago and stay unisolated);
- $u_n$  – the number of people newly infected during the day  $n$ .

We will obtain the following relation between the leading root  $r$  and the basic reproductive rate  $R_0$  that is a main characteristic of an epidemic in epidemiology:

$$r \approx R_0^{\frac{1}{2q}}. \quad (1)$$

Recurrence relation for  $z_n$ :

The number of ill people at large on day  $n + 1$  is the number at large on day  $n$ , minus the number who were detected and isolated on day  $n$ , plus the number of people who were infected  $q$  days ago, i.e.

$$z_{n+1} = z_n - x_n + u_{n-q}. \quad (2)$$

Using  $x_n = bz_n$  and  $u_{n-q} = ax_{n-q}$  we obtain the following equation for  $x_n$ :

$$x_{n+1} = (1 - b)x_n + ax_{n-q}. \quad (3)$$

It is necessary to let the model equal the actual data for the first  $q + 1$  days, as the recurrence equation needs initial data:

$$x_n = x_n^* \text{ for } n = 0, 1, \dots, q, \quad (4)$$

To fit our model we optimize against the average 1-norm:

$$\|x - x^*\| := \frac{1}{N+1} \sum_{n=0}^N |x_n - x_n^*|, \quad N \in \mathbb{N}, \quad (5)$$

Similarly we define  $\|y - y^*\|$

We choose  $N$  to be the length of the data (number of actual observations  $x_n^*$ ) in general.

In order to determine values  $a, b, q$ , we ideally want to minimize both

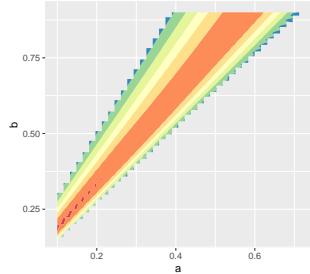
$$\|x - x^*\| \text{ and } \|y - y^*\| \quad (6)$$

While this is the ideal situation, it is far more important to minimize  $\|x - x^*\|$  as it is generally much more sensitive to variation in the parameters (such as  $a$  and  $b$ ).

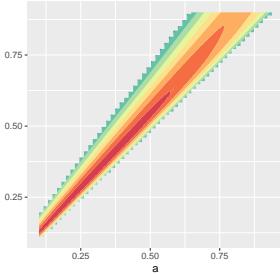
The proofs and results are in A.

### 2.1.3 How to select the best model

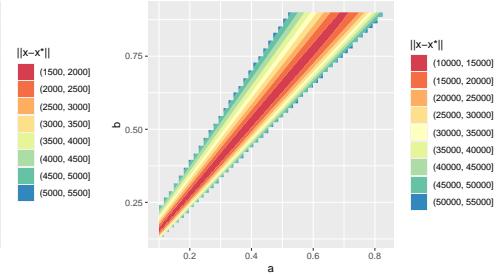
The closest-fitting model will be one that minimizes (5), which can be seen using contour plots below.



**Figure 7:** Contours, Ireland



**Figure 8:** Contours, Italy



**Figure 9:** Contours, United States

The `optim()` function in R appeared to have a danger of converging to local optima, so with some trade-off in computation time over accuracy, I have chosen to iterate the algorithm `basicmodx()` over all combinations of  $a, b, q$ .

### 2.1.4 Forecasting

The model can easily be extended  $m$  days ahead, with  $x_{N+1}, \dots, x_{N+m}$  defined recursively using 3.2.1.

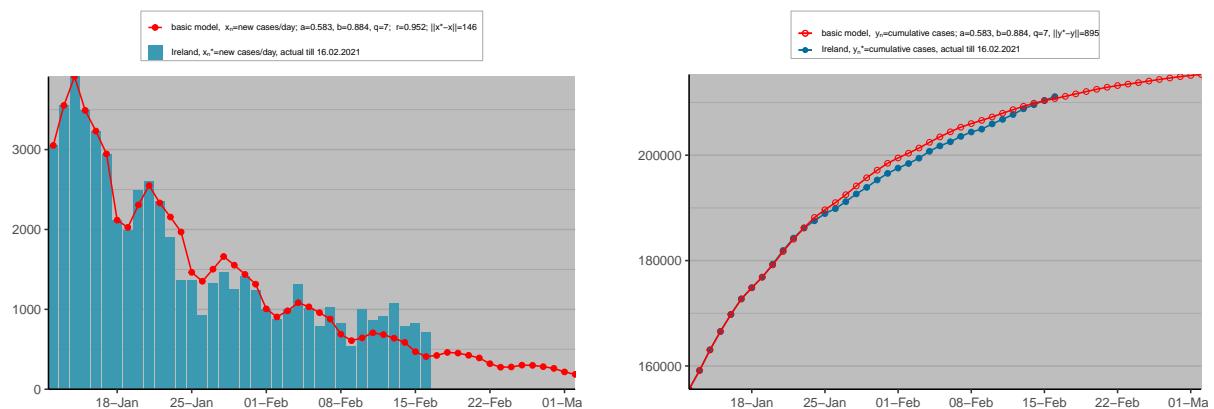
### 2.1.5 Implementation in R

```

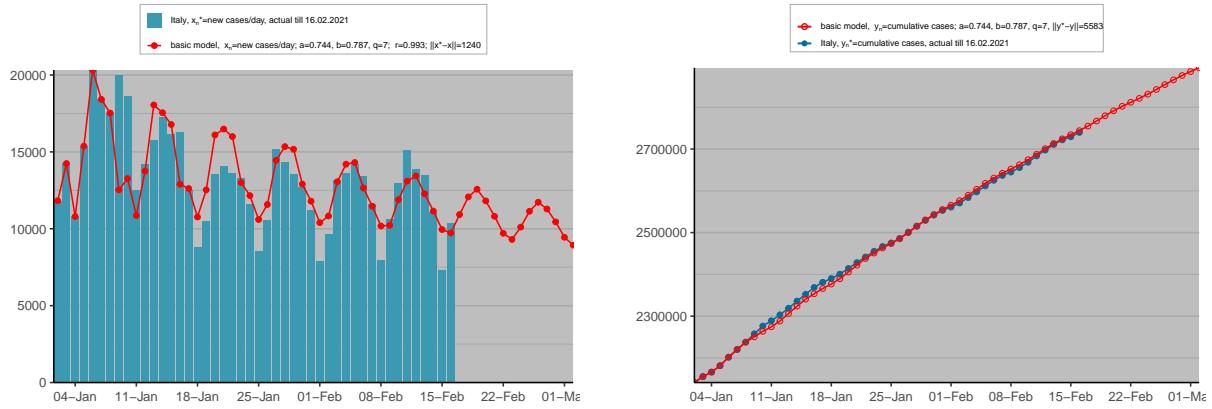
1 basicmodx <- function(x, pars, len = 0){
2   q <- floor(pars[1])
3   a <- pars[2]
4   b <- pars[3]
5   modx <- x[1:q]
6   for(i in (q+1):(length(x)+len)){
7     modx[i] <- (1-b)*modx[i-1] + a*modx[i-q]
8   }
9   return(modx)
10 }
11 basexn <- basicmodx(countrydat$xn, optimpars, len = forecastlen)
...
...
...
15 plots[["basexn"]] <- plot_basexn(countrydat, modeldat, cols, labs)
```

**Listing 1:** Algorithm for Base Model

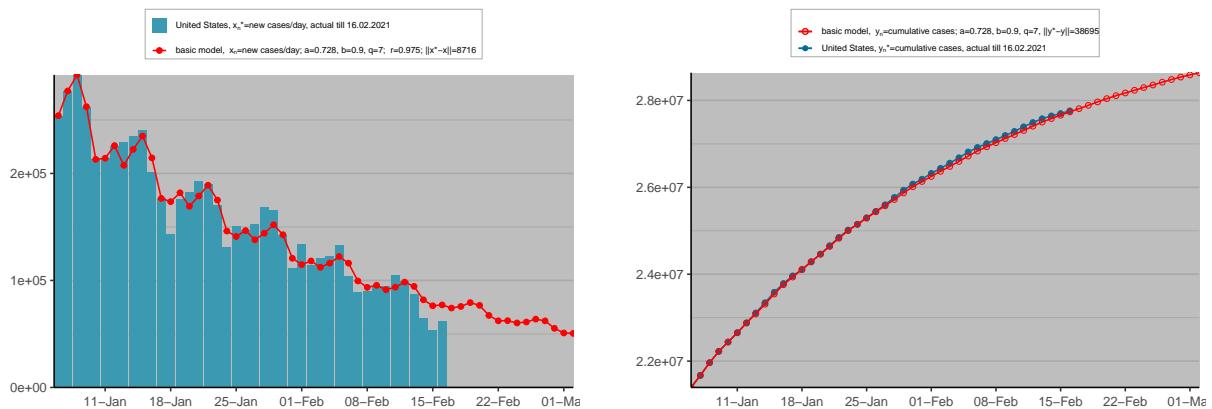
### 2.1.6 Plots



**Figure 10:** Basic model, Ireland



**Figure 11:** Basic model, Italy



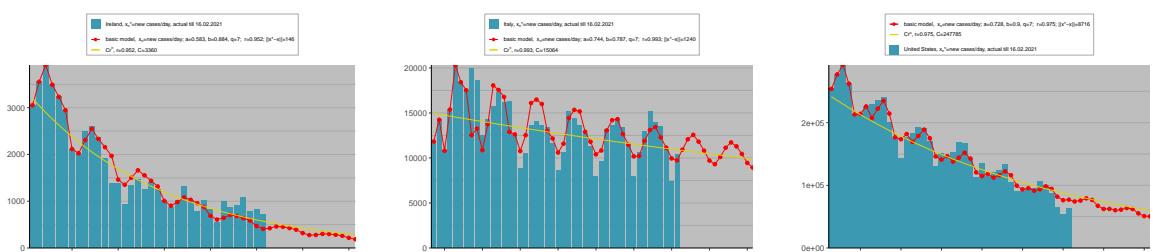
**Figure 12:** Basic model, United States

## 2.2 Limiting curve $Cr^n$

The second result in 1 is the equation (20), which defines the limiting behavior of the recurrence relation (3.2.1).

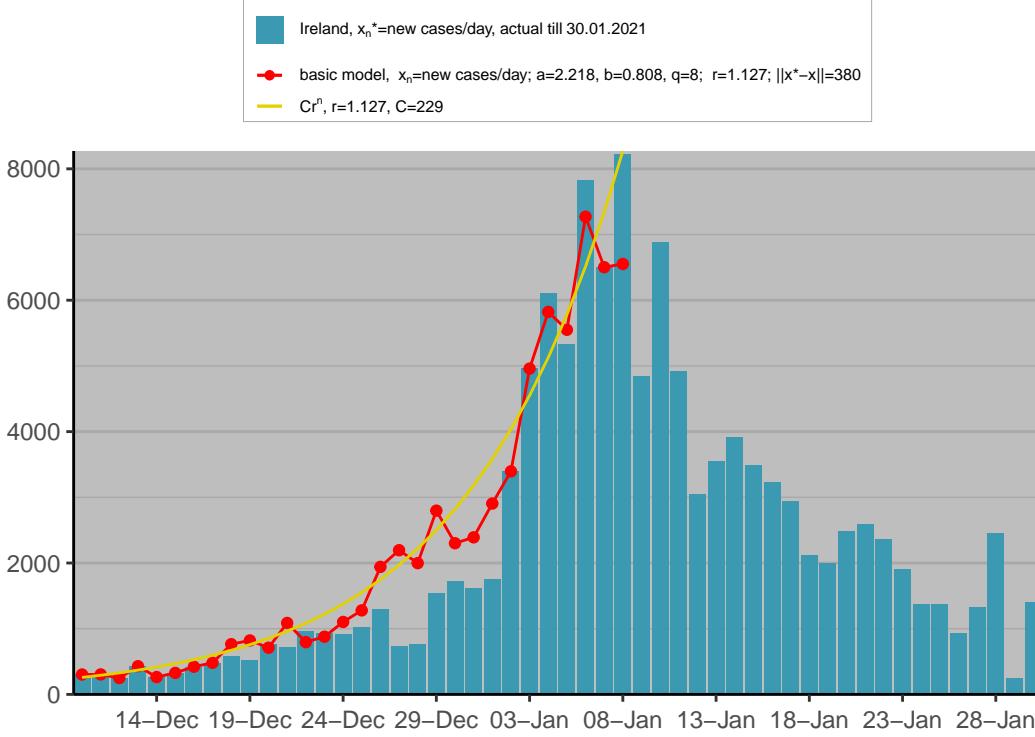
```
1 modeldat$Crn <- optimC * base_r_one^(1:nrow(modeldat))
...
...
5 plots[["Crn"]] <- plot_crn(countrydat, modeldat, cols, labs)
```

**Listing 2:** Algorithm for Limiting Curve



**Figure 13:** Limiting curve  $Cr^n$ , **Figure 14:** Limiting curve  $Cr^n$ , **Figure 15:** Limiting curve  $Cr^n$ ,  
Ireland Italy United States

The limiting curve can also show exponential growth and quickly get out of control



**Figure 16:** Limiting Curve Growing Exponentially, Ireland

### 2.3 Moving average

Define the  $2k + 1$ -day moving average of actual data  $x_n^*$  by  $x^*(k)$

$$x_n^*(1) = \frac{x_{n-1}^* + x_n^* + x_{n+1}^*}{3}, \quad 1 \leq n < N$$

$$x_0^*(1) = \frac{x_0^* + x_1^*}{2}$$

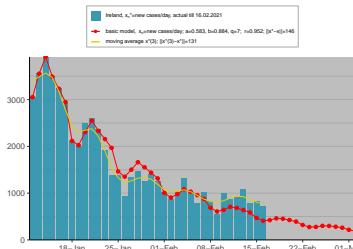
$$x_N^*(1) = \frac{x_{N-1}^* + x_N^*}{2}$$

And then

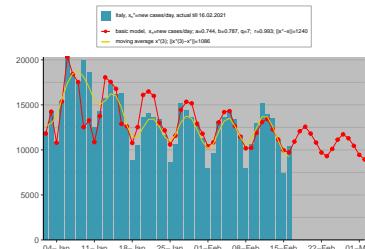
$$x_n^*(3) := x_n^*(x_n^*(1))$$

is the 7-day moving average of cases.

The 7-day moving average is a good baseline for model performance, as ideally we would want  $\|x - x^*\| \approx \|x^*(3) - x^*\|$  or better.



**Figure 17:** Moving average  $x_n^*(3)$ , **Figure 18:** Moving average  $x_n^*(3)$ , **Figure 19:** Moving average  $x_n^*(3)$ ,  
Ireland Italy United States

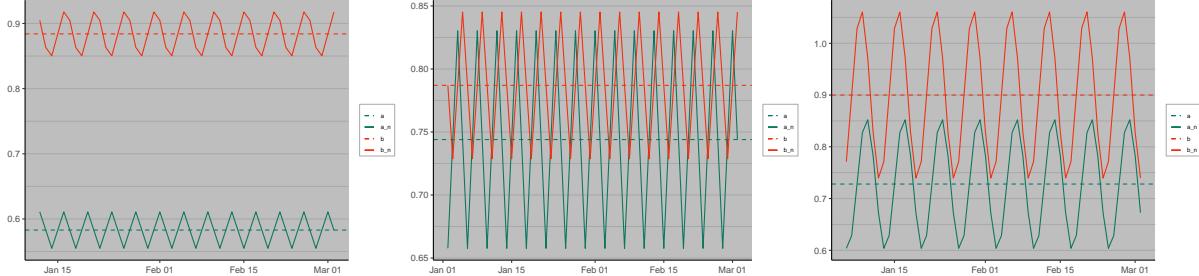


### 2.4 Periodic model

Instead of constant parameters  $a, b$ , we vary them slightly over time:

$$a_n := a \left( 1 + c_1 \left( \sin \left( \frac{2\pi}{p_1} (n - n_1) \right) \right) \right) \quad (7)$$

$$b_n := b \left( 1 + c_2 \left( \sin \left( \frac{2\pi}{p_2} (n - n_2) \right) \right) \right) \quad (8)$$



**Figure 20:** Oscillating  $a$  and  $b$  parameters, Ireland, Italy and United States

For new parameters  $c_i, p_i, n_i$ ,  $i = 1, 2$  where

$c_i \in [0.04, 0.2]$  are (small) amplitudes,

$n_i \in 1, 2, \dots, q$  are lags and

$p_i \in 1, 2, \dots, q$  are periods.

We then have to optimize (6) with respect to the 9 parameters  $a, b, q, c_1, n_1, p_1, c_2, n_2$  and  $p_2$ .

While Grigorian's paper [1] optimized with (possibly) new values of  $a$  and  $b$ , I will allow the original  $a$  and  $b$  from the earlier optimization to calculate the  $a_n$  and  $b_n$ , since the average behavior should be close to the same.

I included a sketch of the proof in the appendix 6.

This means that we can keep the original parameters  $a, b, q$  (we assume the latency period is constant), and we are instead optimizing over the 6 parameters  $c_1, n_1, p_1, c_2, n_2, p_2$

This enables the algorithm to speed up by at least an order of magnitude.

#### 2.4.1 Implementation in R

```

1 modxper <- function(par, q, x, len = 0){
2   #a,b,c1,c2,p1,p2,n1,n2
3   an  <- par[1]*(1+par[3]*sin(2*pi*(1:(length(x)+len) - par[7])/par[5]))
4   bn  <- par[2]*(1+par[4]*sin(2*pi*(1:(length(x)+len) - par[8])/par[6]))
5   modx <- x[1:q]
6   for(i in (q+1):(length(x)+len)){
7     modx[i] <- (bn[i]*(1-bn[i-1]))*modx[i-1]/bn[i-1] +
8       (an[i-q]*bn[i])*modx[i-q]/bn[i-q]
9   }
10  return(modx)
11 }
12 modeldat$periodic <- modxper(as.numeric(perooptim),countrydat$xn, q = q, forecastlen)
...
...
...
16 plots[["periodic"]] <- plot_periodic(countrydat, modeldat, cols, labs)

```

**Listing 3:** Algorithm for Periodic Model

## 2.4.2 Plots

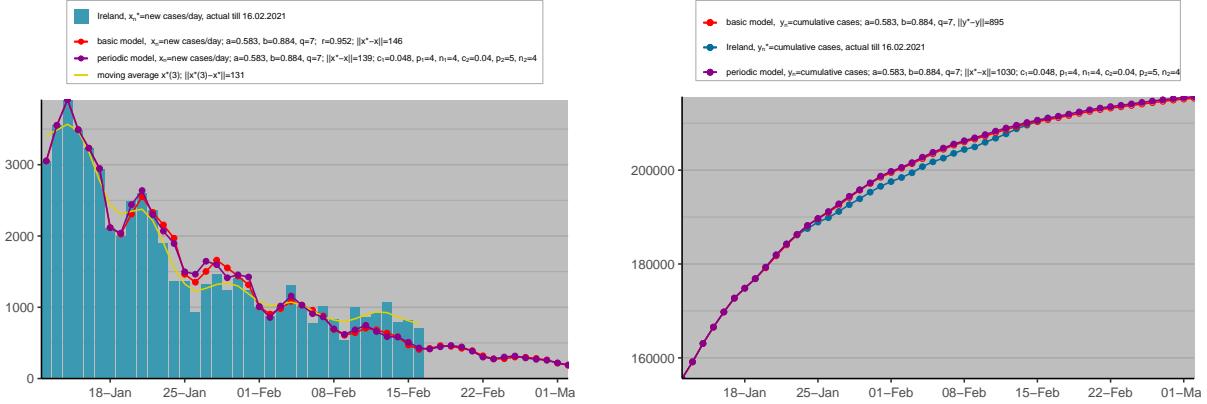


Figure 21: Periodic model, Ireland

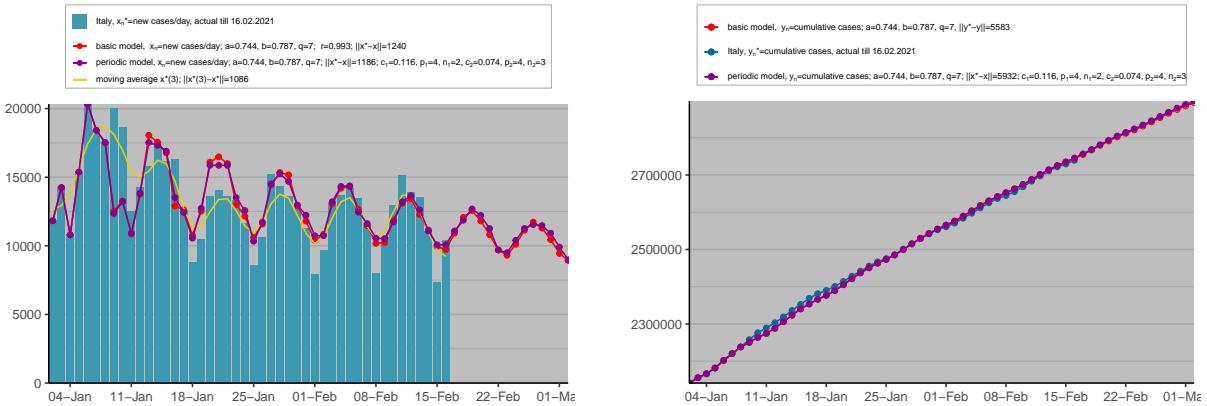


Figure 22: Periodic model, Italy

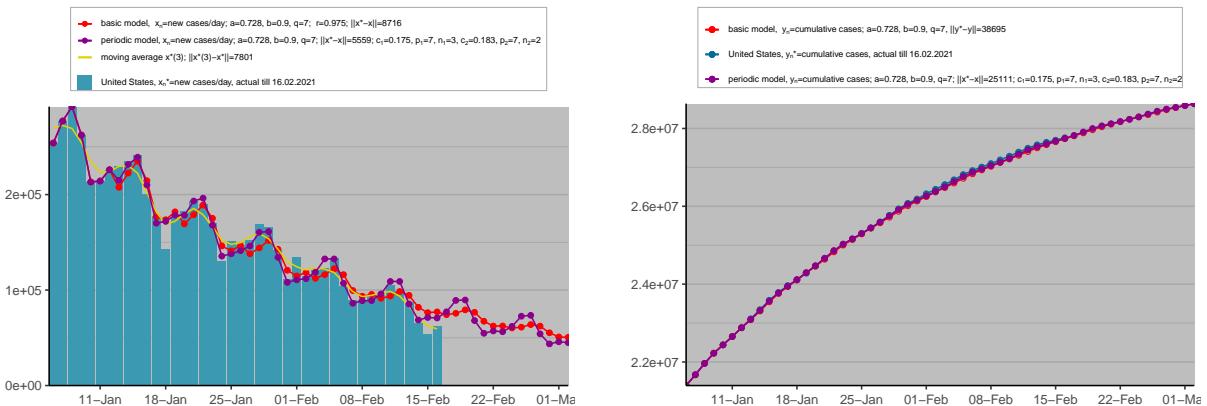


Figure 23: Periodic model, United States

## 2.5 Multi-phase model

I have generalized the two-phase model formulated in [1] to any number of phases, both in the equations below and the code.

Suppose there are  $M$  phases of the pandemic, and we wish to compute the parameters  $a, b, q$  separately for each phase.

Therefore we have phases

- (1) parameters  $a_1, b_1, q_1$  and model  $x_n$  for  $0 \leq n \leq N_1$
- (2) parameters  $a_2, b_2, q_2$  and model  $x_n$  for  $N_1 + 1 \leq n \leq N_2$
- $\vdots$
- (M) parameters  $a_M, b_M, q_M$  and model  $x_n$  for  $N_{M-1} + 1 \leq n \leq N_M = N$

Grigorian's paper uses a form of smoothing of parameter  $b$ , which is using  $\frac{b_{\text{new}}}{b_{\text{old}}}$  for a few days around the change point to deal with the transition between phase.

This perturbation is already dealt with sufficiently in the periodic version so I have excluded this in the event that  $\frac{b_{\text{new}}}{b_{\text{old}}}$  is much larger than 1.

The model parameters are computed separately for each phase, with the values  $x_n$  feeding into the next phase to continue the recursive definition (e.g. phase 2 needs some values from phase 1 of the model).

The model parameters are chosen to minimize (5) for each phase, and the overall model performance displayed is computed for the model as a whole against the data as a whole.

### 2.5.1 Plots

The multi-phase model without periodicity can be sharp and unrealistic, and is purely for demonstration

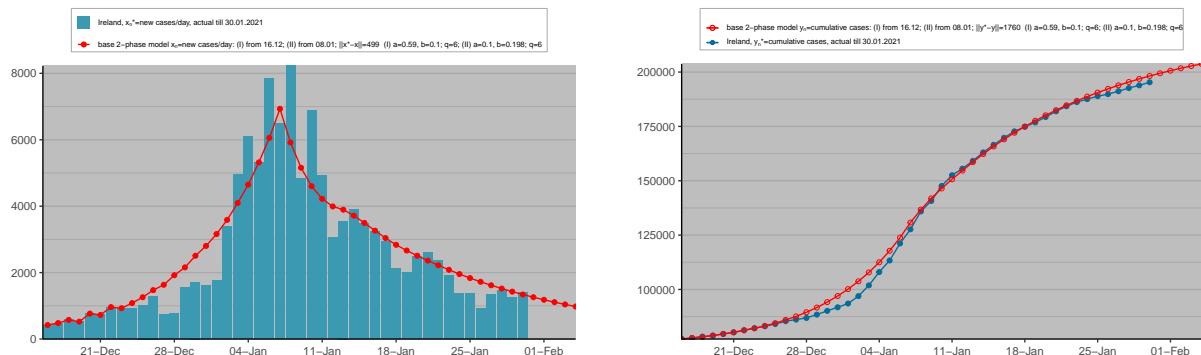


Figure 24: Multi-phase model, Ireland

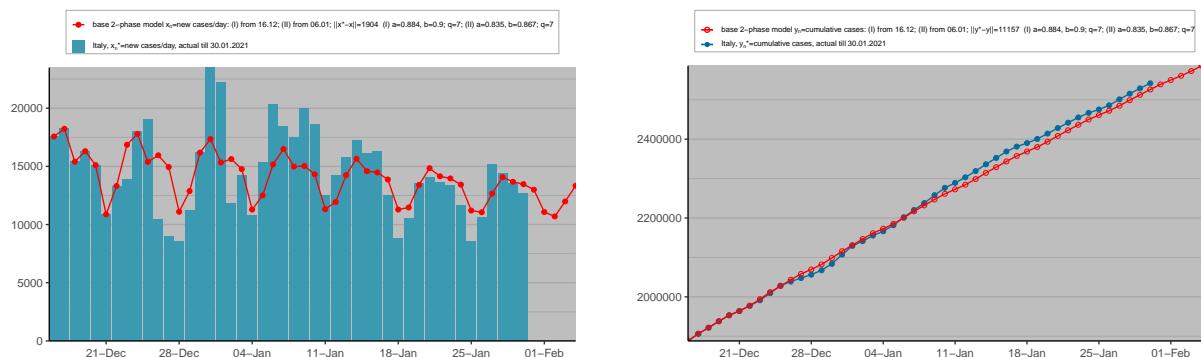
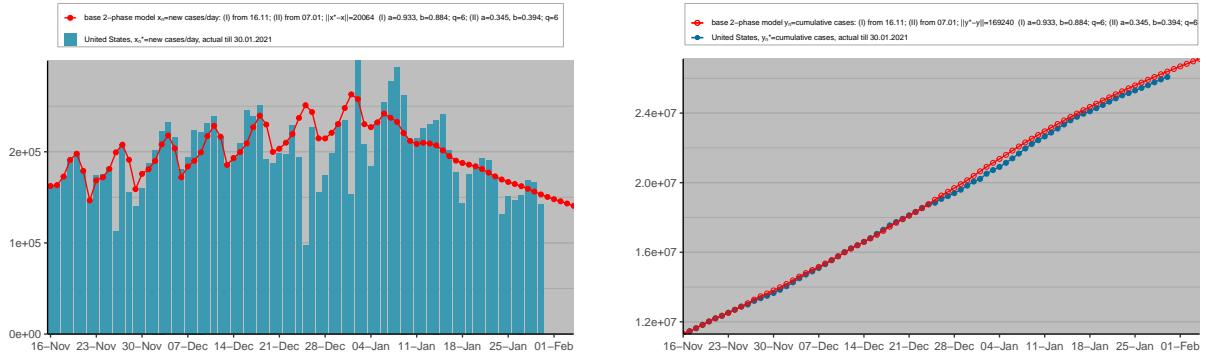
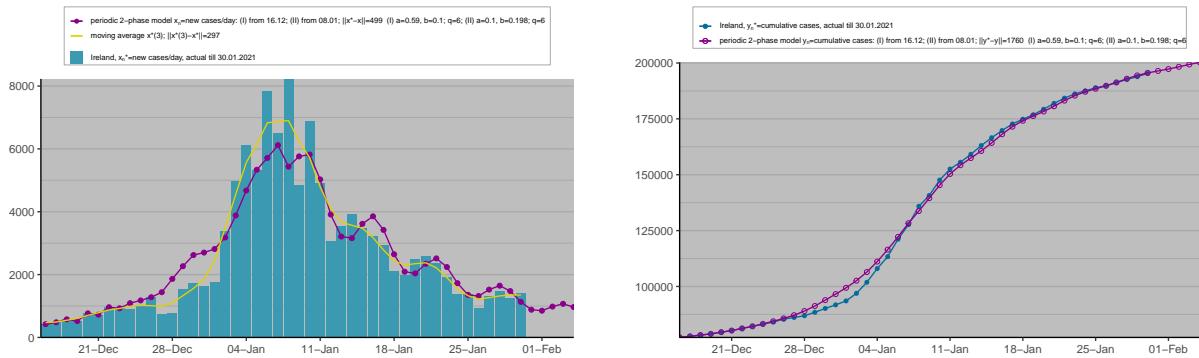


Figure 25: Multi-phase model, Italy

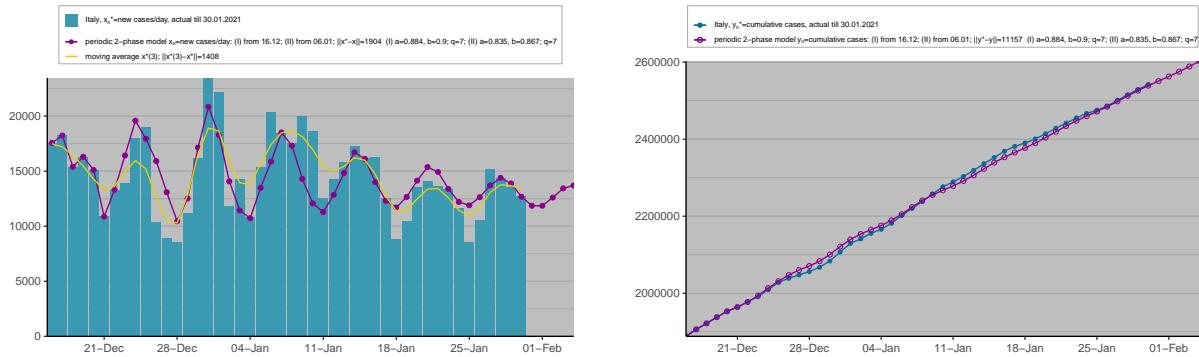


**Figure 26:** Multi-phase model, United States

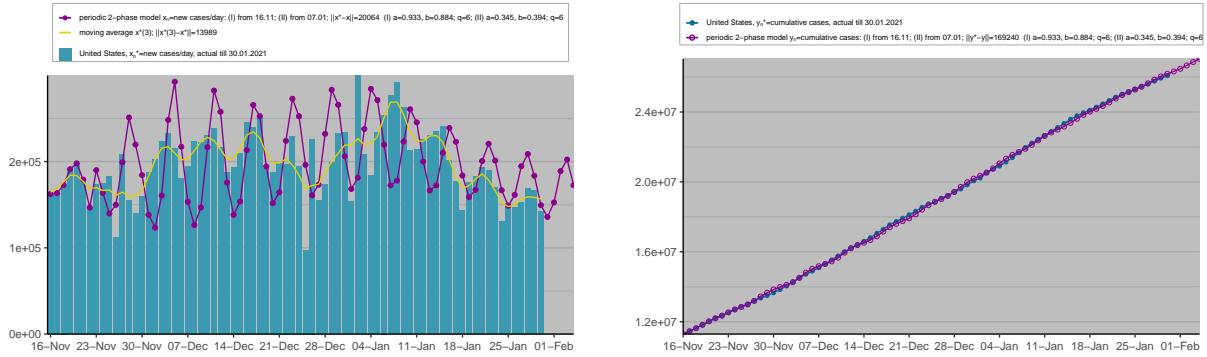
The periodic model is a similar extension as in 2.4 and often performs better.



**Figure 27:** Multi-phase periodic model, Ireland



**Figure 28:** Multi-phase periodic model, Italy



**Figure 29:** Multi-phase periodic model, United States

### 3 Statistical Models

Primary source for this was Hyndman and Athanasopoulos [7].

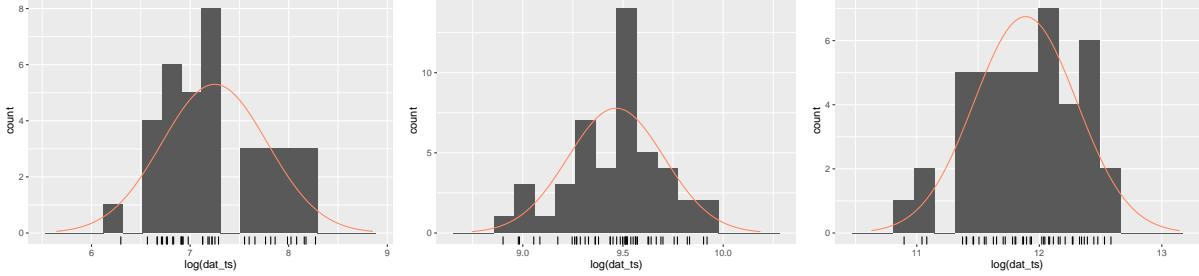
Some of our statistical models require *homoscedasticity*, i.e., that the model errors are identically distributed with have the same variance  $\sigma^2$ .

We can check this by plotting histograms and checking that they are centred around zero and approximately fit the overlaying normal curve.

Using

```
1 forecast::gghistogram(log(dat_ts), add.normal = TRUE, bins = 10)
```

with the full code in 8.



**Figure 30:** Normality checks, Ireland, Italy and United States

#### 3.1 Holt-Winters' seasonal method

The Holt-Winters seasonal exponential smoothing algorithm models three aspects of the time series: the average value at different levels, the trend over time, and any seasonality.

##### 3.1.1 Definitions and Theory

Suppose there are  $N$  observations.

Initial step:

$$\begin{aligned} L_s &= \frac{1}{s} \sum_{i=1}^s x_i \\ b_s &= \frac{1}{s} \left[ \frac{x_{s+1}-x_1}{s} + \frac{x_{s+2}-x_2}{s} + \dots + \frac{x_{2s}-x_s}{s} \right] \\ S_n &= x_n - L_s, \quad n = 1, \dots, s \end{aligned}$$

and choose parameters  $0 \leq \alpha \leq 1$ ,  $0 \leq \beta \leq 1$  and  $0 \leq \gamma \leq 1$

Then compute for  $s < n \leq N$ :

$$\begin{aligned} \text{Level} \quad L_n &= \alpha(x_n - S_{n-s}) + (1-\alpha)(L_{n-1} + b_{n-1}) \\ \text{Trend} \quad b_n &= \beta(L_n - L_{n-1}) + (1-\beta)b_{n-1} \\ \text{Seasonal} \quad S_n &= \gamma(x_n - L_n) + (1-\gamma)S_{n-s} \\ \text{Forecast} \quad F_{n+1} &= L_n + b_n + S_{n+1-s} \end{aligned}$$

For subsequent observations,

$$F_{N+k} = L_N + k \cdot b_N + S_{N+k-s}$$

**Figure 31:** Seasonal Holt-Winters' Additive Model Algorithm (denoted SHW<sub>+</sub>)

The parameters  $\alpha, \beta, \gamma$  are selected using Maximum Likelihood Estimation.

##### 3.1.2 Implementation in R

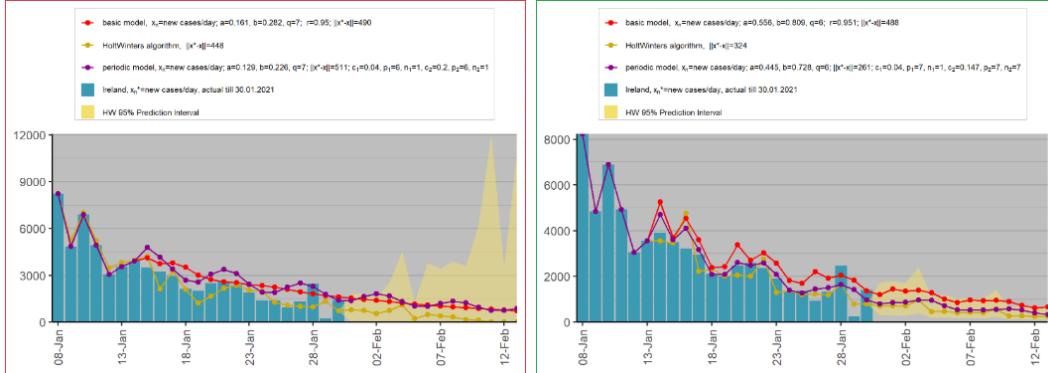
```
1 #lambda=0 ensures values stay positive
2 hwfct <- forecast::hw(dat_ts, h = forecastlen, seasonal = hwmethod, lambda = 0)
...
...
...
```

```
6 plots[["hw"]]  
  <- plot_hw(countrydat, modeldat, cols, labs)
```

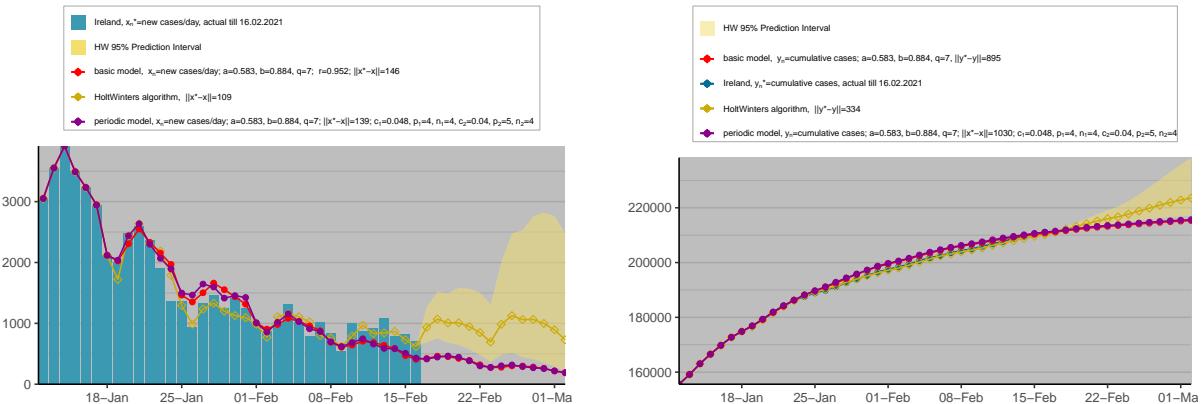
**Listing 4:** Algorithm for HoltWinters Model

### 3.1.3 Plots

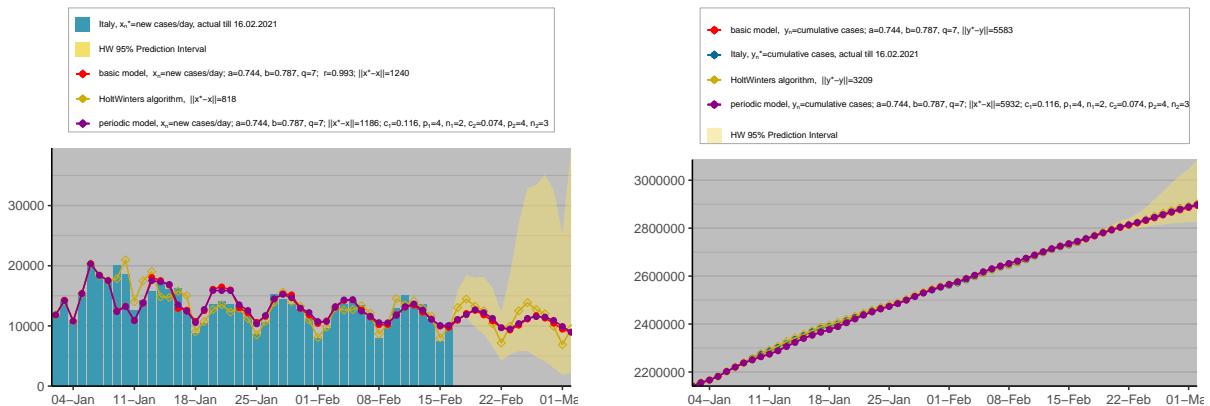
We see that the additive seasonal method is a better choice for both model fit and confidence interval size.



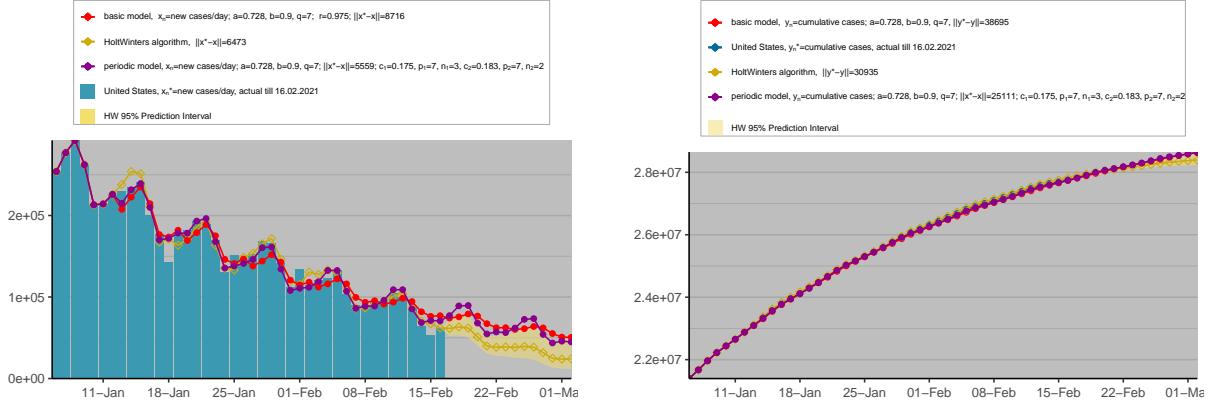
**Figure 32:** Comparison of HoltWinters multiplicative (left) and additive (right) algorithms



**Figure 33:** HoltWinters model, Ireland



**Figure 34:** HoltWinters model, Italy



**Figure 35:** HoltWinters model, United States

## 3.2 ARIMA models

### 3.2.1 Definitions and Theory

**Definition 1.** The *backshift operator*  $B$  is a function on a time series  $(x_n)_{n \geq 1}$  such that  $Bx_n = x_{n-1}$  and more generally:

$$B^k x_n = x_{n-k}, \quad n > k$$

And similarly for the independent errors  $\varepsilon_n$ :

$$B^k \varepsilon_n = \varepsilon_{n-k}, \quad n > k$$

We must first define the each component of a non-seasonal ARIMA model (suitable for time series with a trend).

- An  $AR(p)$  model, or an autoregressive model of order  $p$  of a time series  $x_1, \dots, x_N$  states that each  $x_n$  is a linear function of  $x_{n-p}, x_{n-p+1}, \dots, x_{n-1}$  and an error term, i.e.

$$x_n = \phi_0 + \phi_1 x_{n-1} + \phi_2 x_{n-2} + \dots + \phi_p x_{n-p} + \varepsilon_n, \quad n > p, \quad \varepsilon_n \sim N(0, \sigma^2)$$

Where there are constraints on the  $\phi_0, \phi_1, \dots, \phi_p \in \mathbb{R}$

We can simplify using the backshift operator  $B$ :

$$\begin{aligned} x_n &= \phi_0 + \phi_1 Bx_n + \phi_2 B^2 x_n + \dots + \phi_p B^p x_n + \varepsilon_n \\ &= \phi_0 + (\phi_1 B + \phi_2 B^2 + \dots + \phi_p B^p) x_n + \varepsilon_n \end{aligned} \tag{9}$$

- An  $MA(k)$  model, or a moving average model of order  $q$  of a time series  $x_1, \dots, x_N$  states that each  $x_n$  is a linear function of the  $q$  previous errors  $\varepsilon_{n-k}, \varepsilon_{n-k+1}, \dots, \varepsilon_{n-1}$ , plus the current error  $\varepsilon_n$ , i.e.

$$x_n = \psi_0 - \psi_1 \varepsilon_{n-1} - \psi_2 \varepsilon_{n-2} - \dots - \psi_k \varepsilon_{n-k} + \varepsilon_n, \quad n > p$$

By convention we use minus signs in the coefficients  $\psi_1, \dots, \psi_k$ . We can simplify using the backshift operator  $B$ :

$$\begin{aligned} x_n &= \psi_0 - \psi_1 B \varepsilon_n - \psi_2 B^2 \varepsilon_n - \dots - \psi_k B^k \varepsilon_n + \varepsilon_n \\ &= \psi_0 + (1 - \psi_1 B - \psi_2 B^2 - \dots - \psi_k B^k) \varepsilon_n \end{aligned} \tag{10}$$

- The first order differencing of the time series,  $I(1)$ , is evaluated as

$$\begin{aligned} x'_n &= x_n - x_{n-1} \\ &= x_n - Bx_n \\ &= (1 - B) x_n \end{aligned} \tag{11}$$

More generally, the differencing of order  $d$ , denoted  $I(d)$  is

$$(1 - B)^d x_n$$

This only affects the  $x_n$  (although constants are differenced to zero) and the errors  $\varepsilon_n$  are unchanged. Therefore, an ARIMA( $p, d, k$ ) model can be evaluated by combining the  $AR(p)$ ,  $I(d)$  and  $MA(k)$

$$\begin{aligned} (1 - B)^d x_n &= \phi_0 + (1 - B)^d (\phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p) x_n + \psi_0 + (\psi_1 B + \psi_2 B^2 + \cdots + \psi_k B^k) \varepsilon_n \\ (1 - B)^d x_n + (1 - B)^d (-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) x_n &= \phi_0 + \psi_0 + (1 - \psi_1 B - \psi_2 B^2 - \cdots - \psi_k B^k) \varepsilon_n \\ (1 - B)^d (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) x_n &= c + (1 - \psi_1 B - \psi_2 B^2 - \cdots - \psi_k B^k) \varepsilon_n \end{aligned} \tag{12}$$

where  $c = \phi_0 + \psi_0$  (it is zero if  $d \geq 1$ ).

We also need the seasonal components for an ARIMA( $p, d, k$ ) $(P, D, K)_s$

Suppose a time series  $x_n$  has period  $s$  (seasonal pattern every  $s$  values)

- An  $AR(P)_s$  model, or a seasonal autoregressive model of order  $P$  of a time series  $x_1, \dots, x_N$  states that each  $x_n$  is a *linear function* of  $x_{n-Ps}, x_{n-(P-1)s}, \dots, x_{n-s}$  and an error term, i.e.

$$x_n = \beta_0 + \beta_1 x_{n-s} + \beta_2 x_{n-2s} + \cdots + \beta_P x_{n-Ps} + \varepsilon_n$$

We can simplify using the backshift operator  $B$ :

$$x_n = \beta_0 + (\beta_1 B^s + \beta_2 B^{2s} + \cdots + \beta_P B^{Ps}) x_n \tag{13}$$

- An  $MA(K)_s$  model, or a seasonal moving average model of order  $K$  of a time series  $x_1, \dots, x_N$  states that each  $x_n$  is a *linear function* of the  $K$  errors  $\varepsilon_{n-Ks}, \varepsilon_{n-(K-1)s}, \dots, \varepsilon_{n-s}$ , plus the current error  $\varepsilon_n$ , i.e.

$$x_n = \gamma_0 - \gamma_1 \varepsilon_{n-s} - \gamma_2 \varepsilon_{n-2s} - \cdots - \gamma_K \varepsilon_{n-Ks} + \varepsilon_n$$

Again, by convention we use minus signs in the coefficients  $\gamma_1, \dots, \gamma_K$

We can simplify using the backshift operator  $B$ :

$$\begin{aligned} x_n &= \gamma_0 - \gamma_1 \varepsilon_{n-s} - \gamma_2 \varepsilon_{n-2s} - \cdots - \gamma_K \varepsilon_{n-Ks} + \varepsilon_n \\ &= \gamma_0 + (1 - \gamma_1 B^s - \gamma_2 B^{2s} - \cdots - \gamma_K B^{Ks}) \varepsilon_n \end{aligned} \tag{14}$$

- The first order seasonal differencing of the time series,  $I_s(1)$ , is evaluated as

$$x_n - x_{n-s} = (1 - B^s) x_n$$

More generally, the seasonal differencing of order  $D$ , denoted  $I_s(D)$  is

$$(1 - B^s)^D x_n$$

The purpose of this is to make the time series stationary in mean

Then we can similarly compose our seasonal components (with seasonal period  $s$ ) with the previous ARIMA( $p, d, k$ ) to get the definition of an ARIMA( $p, d, k$ ) $(P, D, K)_s$  model

$$\begin{aligned} &\underbrace{(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)}_{AR(p)} \underbrace{(1 - \beta_1 B^s - \beta_2 B^{2s} - \cdots - \beta_P B^{Ps})}_{AR_s(P)} \underbrace{(1 - B)^d}_{I(d)} \underbrace{(1 - B^s)^D}_{I_s(D)} x_n = \\ &c + \underbrace{\left(1 - \psi_1 B - \psi_2 B^2 - \cdots - \psi_k B^k\right)}_{MA(k)} \underbrace{\left(1 - \gamma_1 B^s - \gamma_2 B^{2s} - \cdots - \gamma_K B^{Ks}\right)}_{MA_s(K)} \varepsilon_n \end{aligned} \tag{15}$$

where the constant  $c$  is some function of the constants  $\phi_0, \psi_0, \beta_0$  and  $\gamma_0$

The orders  $p, d, k$  and  $P, D, K$  are computed by analysing the correlation functions (ACF and PACF).

The  $P + K + p + k + 1$  coefficients  $c, \phi_1, \dots, \phi_p, \psi_1, \dots, \psi_k, \beta_1, \dots, \beta_P, \gamma_1, \dots, \gamma_K$  are computed using Maximum Likelihood Estimation.

Again, (15) can similarly be used to compute forecasted values  $x_{N+1}, \dots, x_{N+m}$ .

**Remark.** Most of the literature, for example [7], use  $q$  for non-seasonal and  $Q$  for seasonal Moving Average models. I am using  $k$  and  $K$  respectively to avoid confusion with  $q$  in the mathematical models .

### 3.2.2 Implementation in R

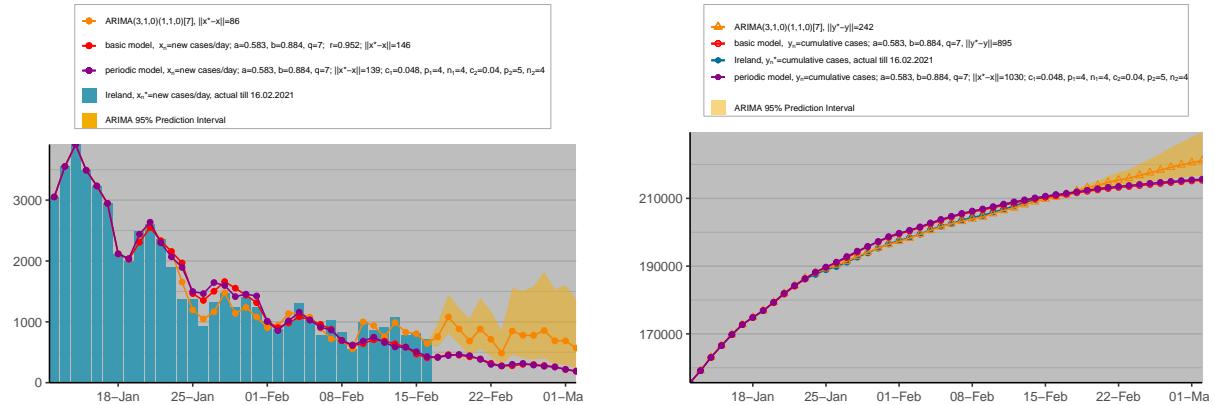
```

1 #lambda=0 ensures values stay positive
2 auto.fit <- auto.arima(dat_ts, lambda = 0)
...
...
6 plots[["arima"]] <- plot_arima(countrydat, modeldat, cols, labs)

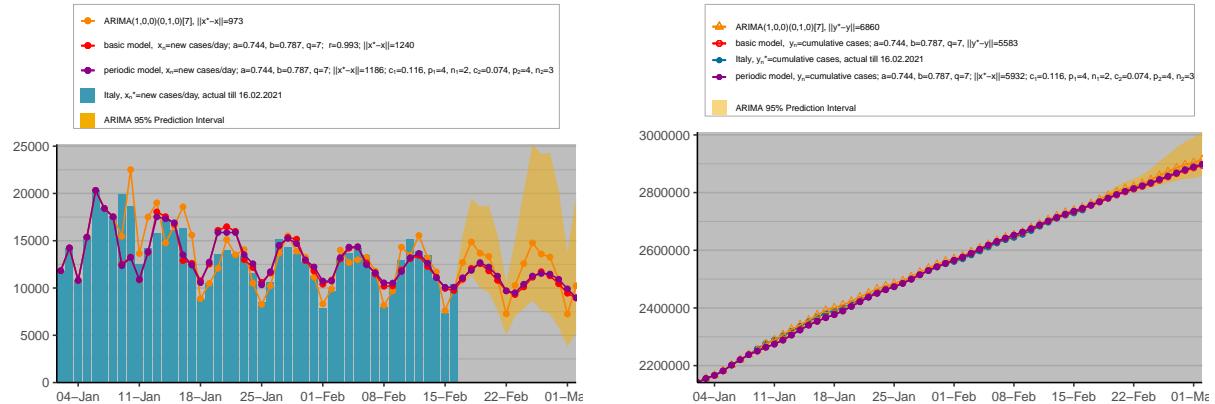
```

**Listing 5:** Algorithm for ARIMA Model

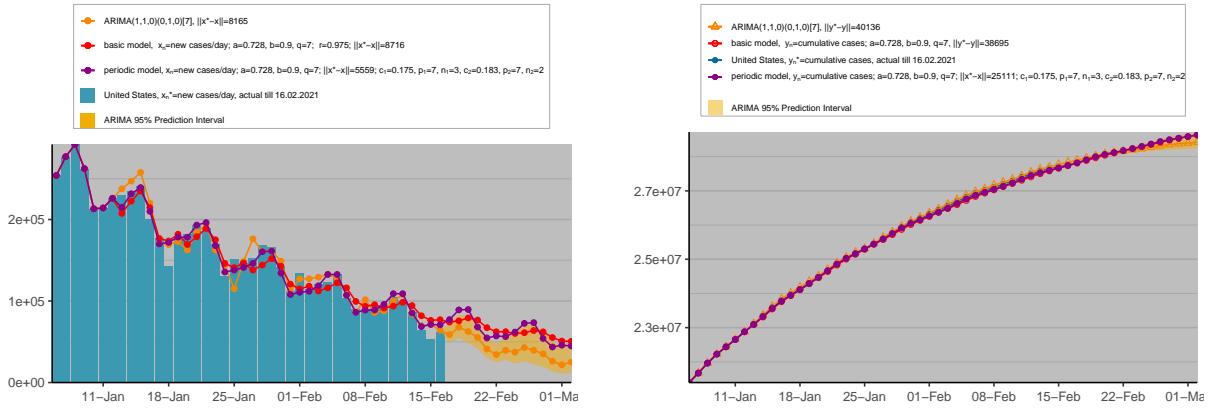
### 3.2.3 Plots



**Figure 36:** ARIMA model, Ireland



**Figure 37:** ARIMA model, Italy

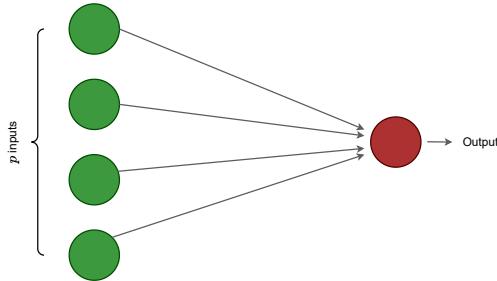


**Figure 38:** ARIMA model, United States

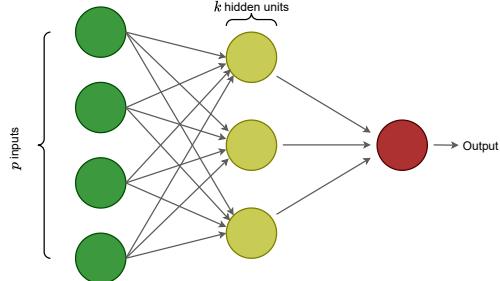
### 3.3 Neural network models

An ARIMA( $p, 0, 0$ ) has inputs  $x_{n-1}, \dots, x_{n-p}$  and parameters  $\phi_1, \dots, \phi_p$  in order to compute  $x_n$  via a linear combination.

Similarly, an ARIMA( $p, 0, 0$ )( $P, 0, 0$ )<sub>s</sub> has inputs  $x_{n-1}, \dots, x_{n-p}, x_{n-s}, \dots, x_{n-Ps}$  and parameters  $\phi_1, \dots, \phi_p, \beta_1, \dots, \beta_P$  in order to compute  $x_n$  via another linear combination.



**Figure 39:** A linear regression model, or ARIMA( $p, 0, 0$ ) model.



**Figure 40:** A neural network with  $p$  inputs and one hidden layer with  $k$  hidden neurons.

With a Neural Network Auto-regressive model, a *hidden layer* of  $k$  inputs is introduced.

The model then becomes non-linear and allows for interactions between inputs (previous values in the time series).

The inputs to the hidden layer are then transformed using a sigmoid function

$$g(z) = \frac{1}{1 + \exp(z)} \quad (16)$$

The diagram 40 is a *multilayer feed-forward network*, as each consecutive layer (left to right) of nodes takes inputs from the previous layer, adding complexity.

The `forecast::nnetar()` function only allows for one layer to reduce the likelihood of overfitting of a univariate time series.

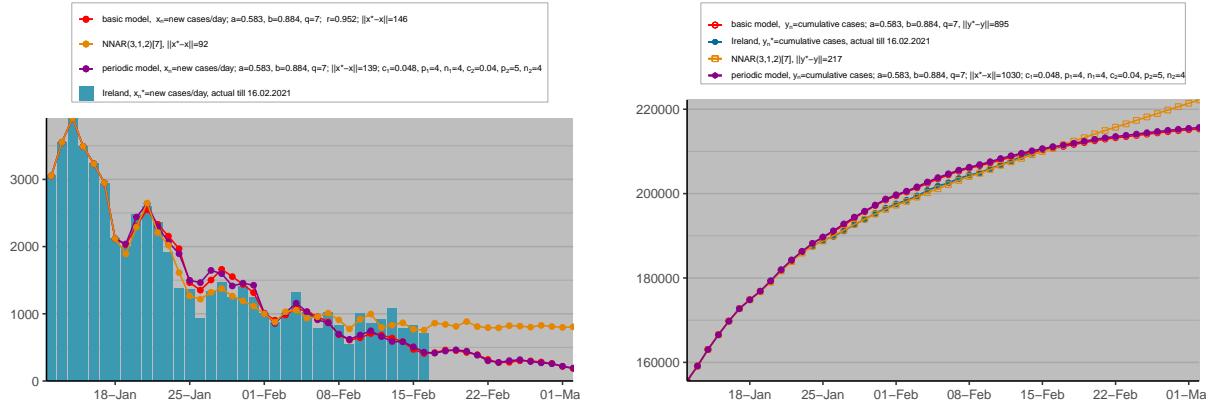
#### 3.3.1 Implementation in R

```
1 #lambda=0 ensures values stay positive
2 #size is the number of nodes in the hidden layer
3 nnfit <- nnetar(dat_ts, p = auto.fit$arma[1], P = auto.fit$arma[3], size = nHidden, lambda = 0,
   repeats = 20, maxit = 50)
...
...
...
```

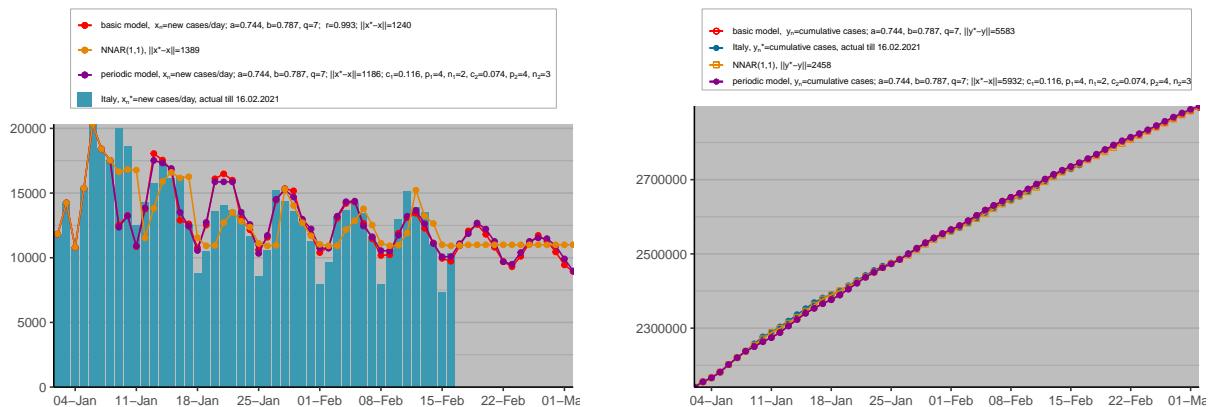
```
7 plots[["nn"]]
```

**Listing 6:** Algorithm for NNAR Model

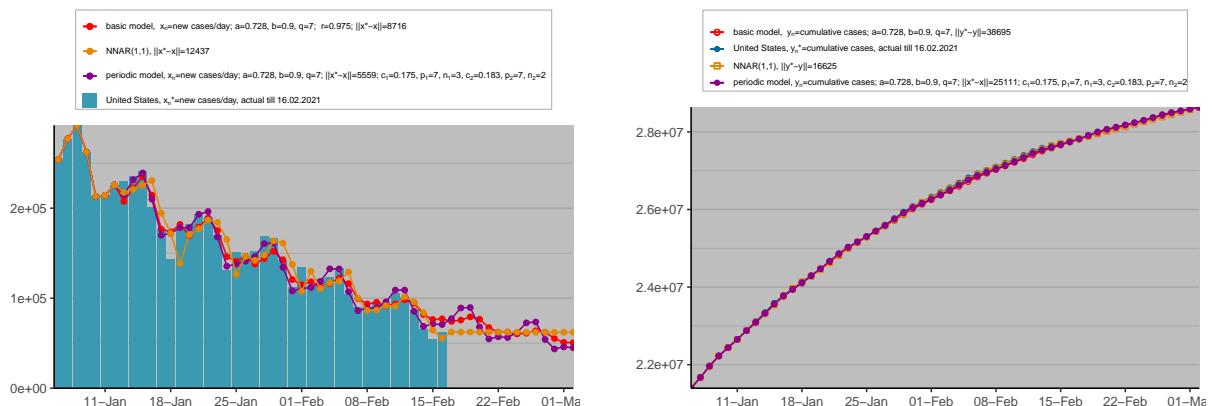
### 3.3.2 Plots



**Figure 41:** Neural Network Autoregression model, Ireland



**Figure 42:** Neural Network Autoregression model, Italy



**Figure 43:** Neural Network Autoregression model, United States

## 4 Model performance with Train/Test sets

To be more rigorous about judging model performance, and to avoid overfitting, we can fit our models using a *training set* and evaluate it on new data in the *test set*.

### 4.1 Summary of model parameters and results

Below is a summary of the parameters for the mathematical models, also located in the plots in the body of the report

	$a$	$b$	$q$	$r$	$c_1$	$p_1$	$n_1$	$c_2$	$p_2$	$n_2$
1	0.57	0.86	7.00	0.95	0.04	4.00	4.00	0.04	5.00	4.00

**Table 1:** Ireland: mathematical models parameters

	$a$	$b$	$q$	$r$	$c_1$	$p_1$	$n_1$	$c_2$	$p_2$	$n_2$
1	0.71	0.76	7.00	0.99	0.15	4.00	2.00	0.10	4.00	3.00

**Table 2:** Italy: mathematical models parameters

	$a$	$b$	$q$	$r$	$c_1$	$p_1$	$n_1$	$c_2$	$p_2$	$n_2$
1	0.71	0.88	7.00	0.97	0.13	7.00	3.00	0.15	7.00	2.00

**Table 3:** United States: mathematical models parameters

Below is a summary of results of the applied model for the primary date range for the training set (at least 4 weeks). The test set is then the following fourteen days of cases.

	model	$\ x - x^*\ _{train}$	$\ y - y^*\ _{train}$	$\ x - x^*\ _{test}$	$\ y - y^*\ _{test}$
1	Basic Recursion	148	916	350	2902
2	Moving Average	130	126	75	109
3	Periodic	140	1049	350	2576
4	HoltWinters	109	334	274	1124
5	ARIMA	86	255	135	306
6	Neural Network	91	221	163	556

**Table 4:** Ireland, 2021-01-12 to 2021-02-16

	model	$\ x - x^*\ _{train}$	$\ y - y^*\ _{train}$	$\ x - x^*\ _{test}$	$\ y - y^*\ _{test}$
1	Basic Recursion	1247	5093	5126	34386
2	Moving Average	1072	1030	1034	1175
3	Periodic	1183	5119	4767	26857
4	HoltWinters	818	3209	4357	18774
5	ARIMA	973	6172	3834	14472
6	Neural Network	1388	2460	4623	24717

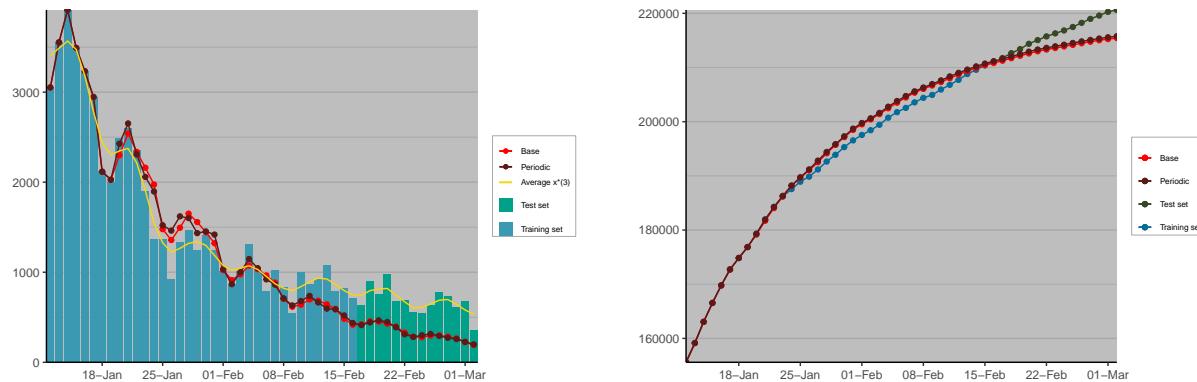
**Table 5:** Italy, 2021-01-02 to 2021-02-16

	model	$\ x - x^*\ _{train}$	$\ y - y^*\ _{train}$	$\ x - x^*\ _{test}$	$\ y - y^*\ _{test}$
1	Basic Recursion	9049	45009	7882	42965
2	Moving Average	7869	8525	3926	5445
3	Periodic	5613	25841	8785	22280
4	HoltWinters	5763	22833	21703	93870
5	ARIMA	8223	37447	24889	107978
6	Neural Network	12586	16902	8945	47923

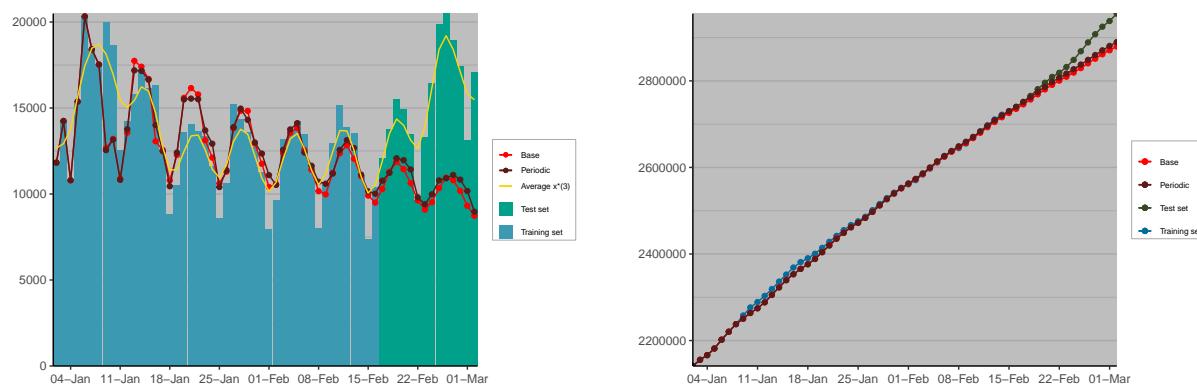
**Table 6:** United States, 2021-01-06 to 2021-02-16

The moving average is only here for comparison and usually outperforms all other models.

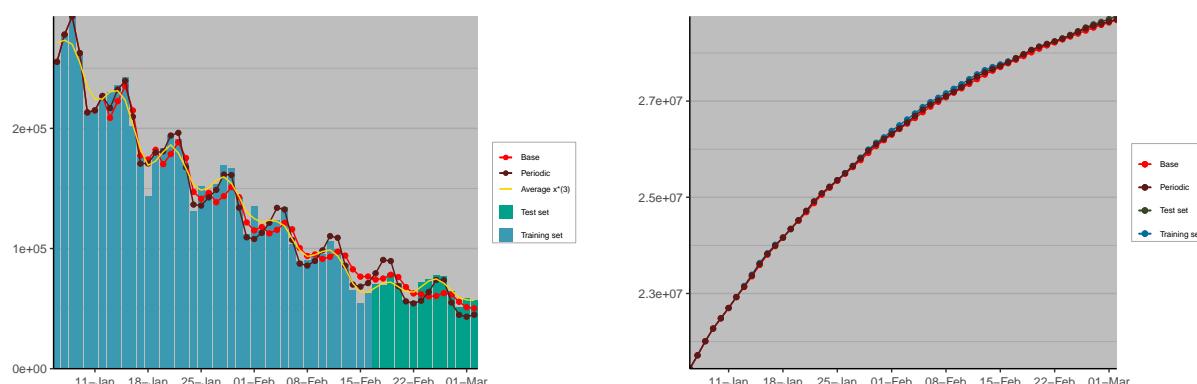
## 4.2 Base and Periodic models



**Figure 44:** Periodic model with train/test split, Ireland

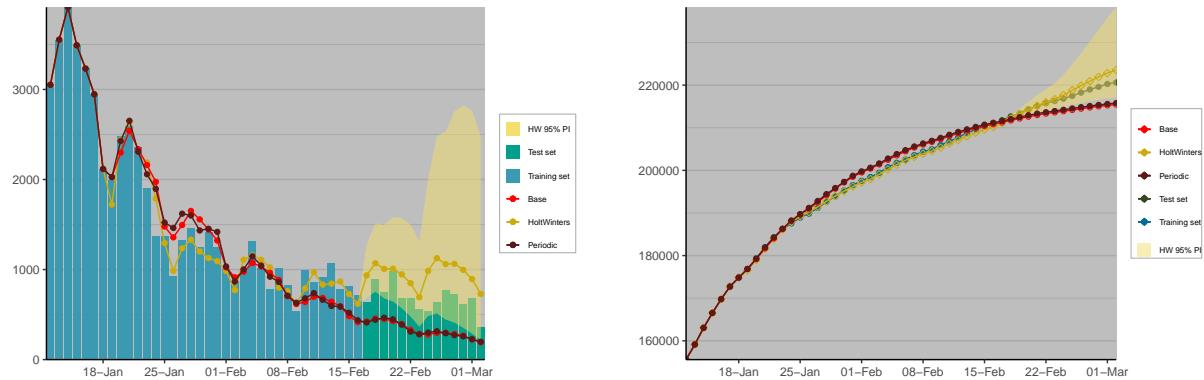


**Figure 45:** Periodic model with train/test split, Italy

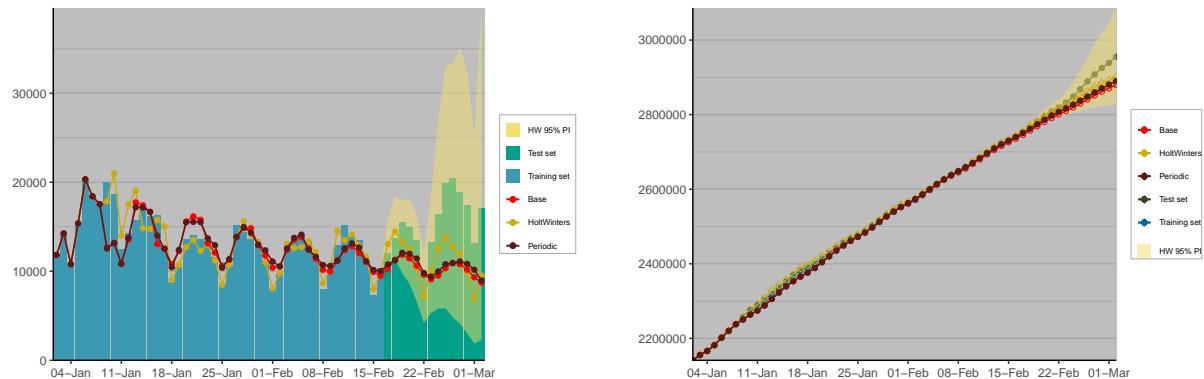


**Figure 46:** Periodic model with train/test split, United States

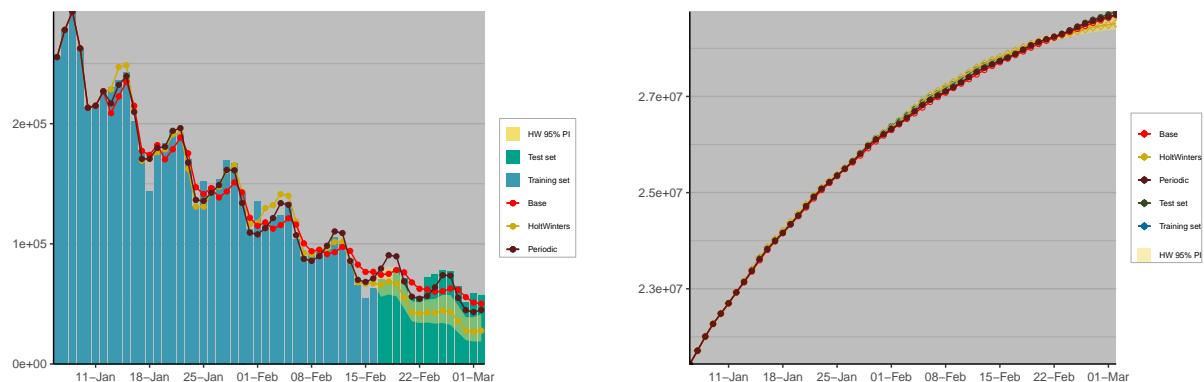
### 4.3 HoltWinters model



**Figure 47:** HoltWinters model with train/test split, Ireland

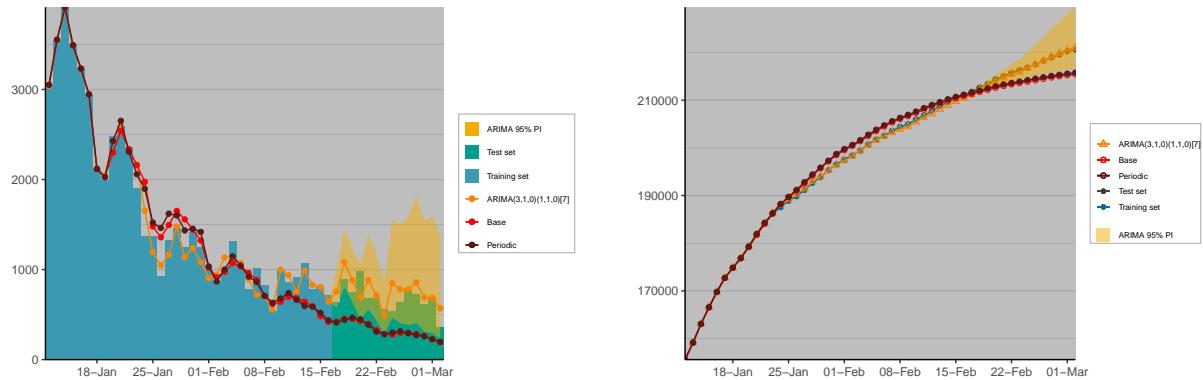


**Figure 48:** HoltWinters model with train/test split, Italy

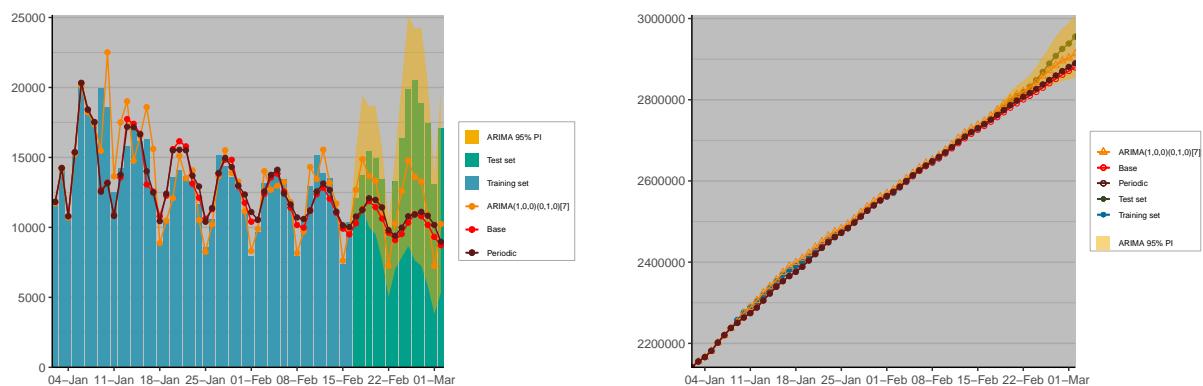


**Figure 49:** HoltWinters model with train/test split, United States

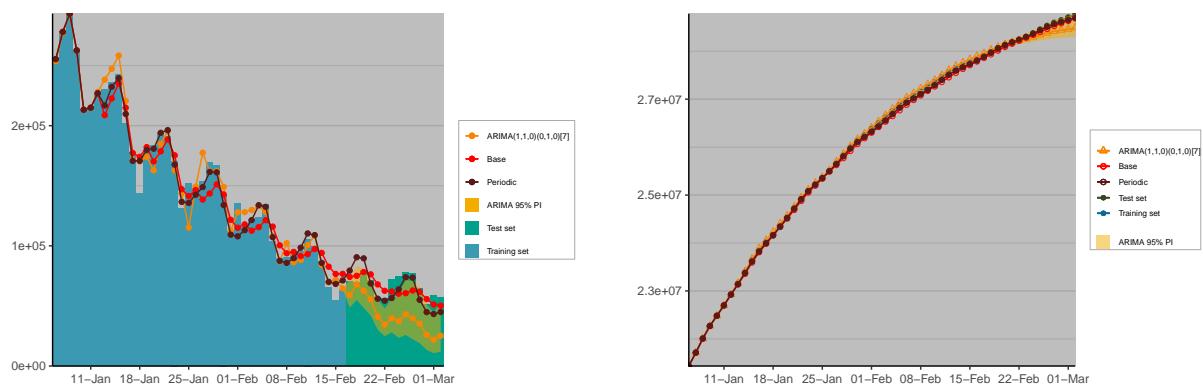
## 4.4 ARIMA model



**Figure 50:** ARIMA model with train/test split, Ireland

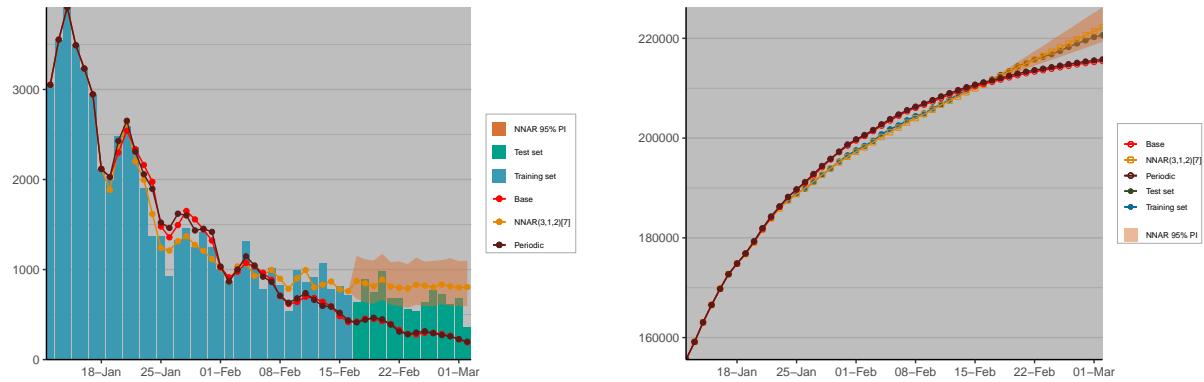


**Figure 51:** ARIMA model with train/test split, Italy

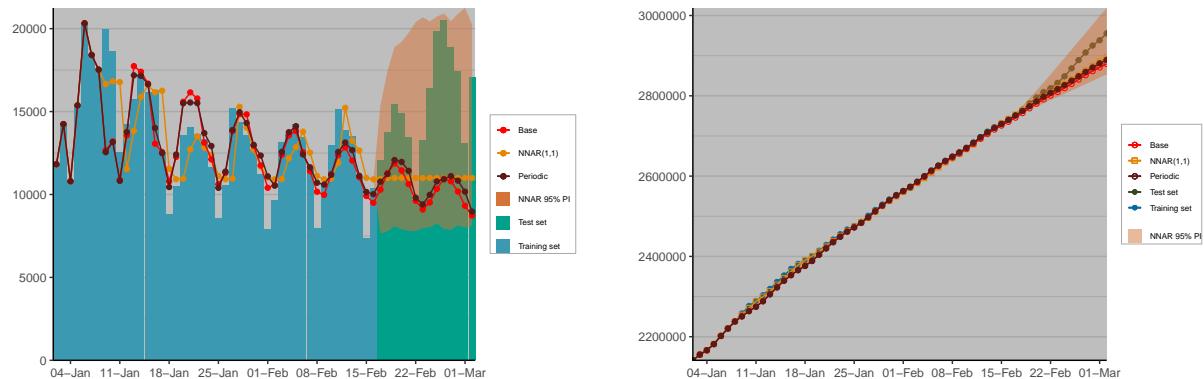


**Figure 52:** ARIMA model with train/test split, United States

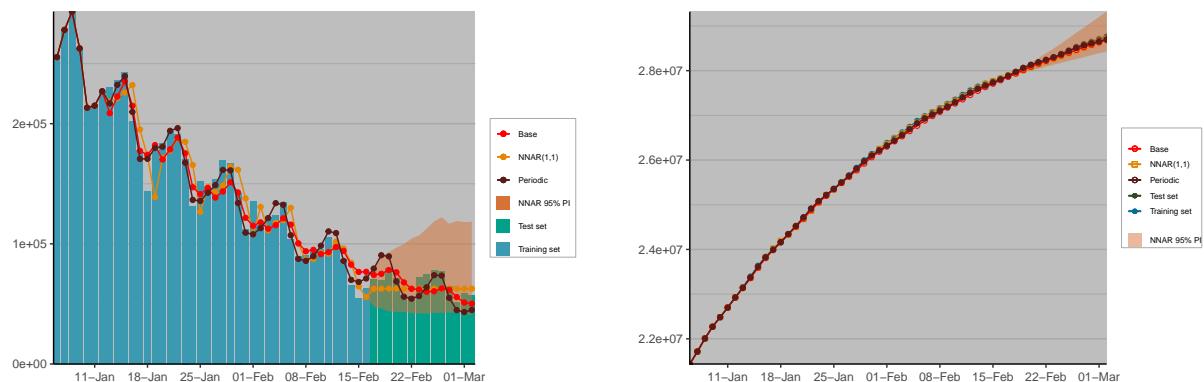
## 4.5 NNAR model



**Figure 53:** Neural Network Autoregression model with train/test split, Ireland



**Figure 54:** Neural Network Autoregression model with train/test split, Italy



**Figure 55:** Neural Network Autoregression model with train/test split, United States

## A Definitions and Theorems for Mathematical Models

### A.1 Recurrence equation

This is our general linear recurrence equation with constant coefficients:

$$x_{n+1} = a_0 x_n + a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_q x_{n-q} \quad (17)$$

The characteristic polynomial of (17)

$$f(\lambda) = \lambda^{q+1} - a_0 \lambda^q - a_1 \lambda^{q-1} - a_2 \lambda^{q-2} - \cdots - a_{q-1} \lambda - a_q. \quad (18)$$

**Definition 2.** A root  $\lambda$  of  $f$  with the maximal absolute value  $|\lambda|$  will be referred to as a leading root of the general linear recurrence relation (17).

**Definition 3.** The geometric mean of  $n$  values  $x_1, \dots, x_n$  is

$$\left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n} \quad (19)$$

### A.2 Results

**Theorem 1.** Let  $a_k \geq 0$  for all  $k \in \{0, \dots, q\}$  and  $a_{k_0} > 0$  for some  $k_0 \in \{0, \dots, q\}$ .

- (a) (Cauchy, 1829) The polynomial  $f(\lambda)$  from (18) has exactly one positive real root  $r$ . Besides, the root  $r$  is simple and, for any other root  $\lambda \in \mathbb{C}$ , we have  $|\lambda| < r$ . Consequently,  $r$  is the leading root of (17).
- (b) For any positive solution  $x_n$  of (17), there exists  $C > 0$  such that

$$x_n \sim C r^n \text{ as } n \rightarrow \infty. \quad (20)$$

It follows from (20) that if  $r < 1$  then the epidemic fades away, whereas if  $r > 1$  then it will spread indefinitely.

**Proof:**

- (a) The equation  $f(\lambda) = 0$  is equivalent to

$$0 = \lambda^{q+1} - a_0 \lambda^q - a_1 \lambda^{q-1} - a_2 \lambda^{q-2} - \cdots - a_{q-1} \lambda - a_q$$

dividing across by  $\lambda^{q+1}$

$$= 1 - \frac{a_0}{\lambda} - \frac{a_1}{\lambda^2} - \frac{a_2}{\lambda^3} - \cdots - \frac{a_{q-1}}{\lambda^q} - \frac{a_q}{\lambda^{q+1}}$$

And so

$$1 = \underbrace{\frac{a_0}{\lambda} + \frac{a_1}{\lambda^2} + \frac{a_2}{\lambda^3} + \cdots + \frac{a_{q-1}}{\lambda^q} + \frac{a_q}{\lambda^{q+1}}}_{g(\lambda)} \quad (21)$$

Since  $a_{k_0} > 0$  for some  $k_0$ , and the remaining  $a_k$  are non-negative,  $g(\lambda)$  is strictly monotone decreasing in  $\lambda > 0$ , and we have the limits

- $\lim_{\lambda \rightarrow 0^+} g(\lambda) = +\infty$
- $\lim_{\lambda \rightarrow +\infty} g(\lambda) = 0^+$

Hence, there is exactly one positive value  $\lambda = r$  that satisfies this  $g(r) = 1$ , that is,

$$1 = \frac{a_0}{r} + \frac{a_1}{r^2} + \frac{a_2}{r^3} + \cdots + \frac{a_{q-1}}{r^q} + \frac{a_q}{r^{q+1}}.$$

Now, let  $\lambda \in \mathbb{C} \setminus \{0\}$  be another root of  $f$ . We obtain from (21) (using the triangle inequality) that

$$1 \leq \frac{a_0}{|\lambda|} + \frac{a_1}{|\lambda|^2} + \frac{a_2}{|\lambda|^3} + \cdots + \frac{a_{q-1}}{|\lambda|^q} + \frac{a_q}{|\lambda|^{q+1}}$$

Since  $g$  is strictly decreasing,

$g(r) \leq g(|\lambda|)$  which implies  $|\lambda| \leq r$ .

We next need to show that the root  $r$  is simple. Denote by  $r'$  the largest non-negative root of the derivative  $f'(\lambda)$  that exists for the following reason. If  $a_k > 0$  for some  $k < q$  then the polynomial  $\frac{1}{q+1} f'(\lambda)$  satisfies the

hypotheses of the present theorem and, by the above argument,  $f'(\lambda)$  has exactly one positive root, that is  $r'$ . If  $a_k = 0$  for all  $k < q$  then  $f'(\lambda) = (q+1)\lambda^q$  has the only root 0, and, hence,  $r' = 0$ .

Let us verify that  $r' < r$ , which will also imply that  $r$  is simple. If  $r' = 0$  then it is clear. If  $r' > 0$  then it follows from  $f'(r') = 0$  that

$$\begin{aligned} f'(\lambda) &= (q+1)\lambda^q - qa_0\lambda^{q-1} - (q-1)a_1\lambda^{q-2} - (q-2)a_2\lambda^{q-3} - \cdots - a_{q-1} - 0 \\ \frac{1}{q+1}f'(\lambda) &= \lambda^q - \frac{q}{q+1}a_0\lambda^{q-1} - \frac{q-1}{q+1}a_1\lambda^{q-2} - \frac{q-2}{q+1}a_2\lambda^{q-3} - \cdots - \frac{1}{q+1}a_{q-1} \\ \frac{1}{q+1}f'(r') &= (r')^q - \frac{q}{q+1}a_0(r')^{q-1} - \frac{q-1}{q+1}a_1(r')^{q-2} - \frac{q-2}{q+1}a_2(r')^{q-3} - \cdots - \frac{1}{q+1}a_{q-1} \\ 0 &= (r')^q - \frac{q}{q+1}a_0(r')^{q-1} - \frac{q-1}{q+1}a_1(r')^{q-2} - \frac{q-2}{q+1}a_2(r')^{q-3} - \cdots - \frac{1}{q+1}a_{q-1} \\ (r')^q &= \frac{q}{q+1}a_0(r')^{q-1} + \frac{q-1}{q+1}a_1(r')^{q-2} + \frac{q-2}{q+1}a_2(r')^{q-3} + \cdots + \frac{1}{q+1}a_{q-1} \\ \text{dividing both sides by } (r')^q &> 0 \\ 1 &= \frac{qa_0}{(q+1)r'} + \frac{(q-1)a_1}{(q+1)(r')^2} + \cdots + \frac{a_{q-1}}{(q+1)(r')^q} \\ &= \left(\frac{q+1-1}{q+1}\right) \frac{a_0}{r'} + \left(\frac{q+1-2}{q+1}\right) \frac{a_1}{(r')^2} + \cdots + \left(\frac{q+1-q}{q+1}\right) \frac{a_{q-1}}{(r')^q} \\ &= \left(1 - \frac{1}{q+1}\right) \frac{a_0}{r'} + \left(1 - \frac{2}{q+1}\right) \frac{a_1}{(r')^2} + \cdots + \left(1 - \frac{q}{q+1}\right) \frac{a_{q-1}}{(r')^q} \\ &< \frac{a_0}{r'} + \frac{a_1}{(r')^2} + \cdots + \frac{a_{q-1}}{(r')^q} \end{aligned}$$

So  $g(r') > 1$ , but  $g(r) = 1$

Again, since the function  $g$  is strictly decreasing,  $g(r') > g(r) \implies r' < r$

Therefore  $r$  is simple, and is the leading root of (17).

- (b) Let  $\lambda_1, \lambda_2, \dots$  be all other distinct roots of  $f$  apart from  $r$  (so that  $\lambda_k$  are negative or imaginary). Any solution  $x_n$  of (17) has the form

$$x_n = Cr^n + \tilde{x}_n \quad (22)$$

where  $\tilde{x}_n$  is a linear combination of the functions  $n^j \lambda_k^n$ . Since by (a) we have  $|\lambda_k| < r$ , it follows that

$$|\tilde{x}_n| = o(r^n) \text{ as } n \rightarrow \infty \quad (23)$$

Since  $x_n > 0$ , it follows from (22) and (23) that  $C \geq 0$ .

We need to show that  $C > 0$ .

Consider a new sequence

$$X_n = \frac{x_n}{r^n}.$$

This satisfies the equation

$$X_{n+1} = A_0 X_n + A_1 X_{n-1} + \cdots + A_q X_{n-q} \quad (24)$$

with  $A_k = \frac{a_k}{r^{k+1}}$ .

Since

$$\begin{aligned} A_0 X_0 + A_1 X_{n-1} + \cdots + A_q X_{n-q} &= \frac{a_0}{r^1} \frac{x_n}{r^n} + \frac{a_1}{r^2} \frac{x_{n-1}}{r^{n-1}} + \cdots + \frac{a_q}{r^{q+1}} \frac{x_{n-q}}{r^{n-q}} \\ &= \frac{a_0 x_n}{r^{n+1}} + \frac{a_1 x_{n-1}}{r^{n+1}} + \cdots + \frac{a_q x_{n-q}}{r^{n+1}} \\ &= \frac{1}{r^{n+1}} (a_0 x_n + a_1 x_{n-1} + \cdots + a_q x_{n-q}) \\ &= \frac{1}{r^{n+1}} (x_{n+1}) \quad \text{from (17)} \\ &= \frac{x_{n+1}}{r^{n+1}} (x_{n+1}) \\ &= X_{n+1} \end{aligned}$$

Since  $r$  is a root of  $f$ , we have

$$\begin{aligned} A_0 + A_1 + \cdots + A_q &= \frac{a_0}{r^1} + \frac{a_1}{r^2} + \cdots + \frac{a_q}{r^{q+1}} \\ &= g(r) \end{aligned}$$

This implies, by (21), and  $g(r) = 1$  that

$$A_0 + A_1 + \cdots + A_q = 1 \quad (25)$$

Set  $c := \min(X_1, \dots, X_{q+1}) > 0$  since  $x_n$  have positive initial values. Then we obtain from (24) and (25) by induction that  $X_n \geq c$  for all  $n \in \mathbb{N}$ , which implies

$$x_n \geq cr^n$$

But equating with (22) we get

$$\begin{aligned} Cr^n + \tilde{x}_n &\geq cr^n \\ C + \frac{\tilde{x}_n}{r^n} &\geq c \end{aligned}$$

Taking the limit as  $n \rightarrow \infty$

$$C \geq c > 0 \quad (26)$$

Therefore  $C > 0$ , as required.  $\square$

**Theorem 2.** Let  $a_k \geq 0$  for all  $k = 0, \dots, q$ . Denote  $a = a_1 + \cdots + a_q, b = 1 - a_0$  and assume that  $a > 0, b > 0$ .

- (a) We have the equivalences:  $r < 1 \iff a < b$  and  $r > 1 \iff a > b$ .
- (b) Let  $m \geq 1$  be such that  $a_1 = \cdots = a_{m-1} = 0$  and  $a_m > 0$ . Then

$$\min\left(1, \left(\frac{a}{b}\right)^{1/m}\right) \leq r \leq \max\left(1, \left(\frac{a}{b}\right)^{1/m}\right) \quad (27)$$

**Proof:**

- (a) We have

$$\begin{aligned} f(1) &= 1 - a_0 - a_1 - \cdots - a_q \\ &= \underbrace{(1 - a_0)}_b - \underbrace{(a_1 + \cdots + a_q)}_a \\ &= b - a \end{aligned}$$

We know  $f$  is increasing.

If  $r = 1$  then the inequality obviously holds.

So if  $r < 1$ , we have  $f(1) > 0$  and then  $b - a > 0 \implies a < b$ .

And if  $r > 1$ , we have  $f(1) < 0$  and then  $b - a < 0 \implies a > b$

- (b)  $f(r) = 0$  is equivalent to

$$r^{q+1} - a_0 r^q - a_1 r^{q-1} - a_2 r^{q-2} - \cdots - a_{q-1} r - a_q = 0$$

But any  $a_1, \dots, a_{m-1}$  are all zero

$$\implies r^{q+1} - a_0 r^q - a_m r^{q-m} - a_{m+1} r^{q-m-1} - \cdots - a_{q-1} r - a_q = 0$$

$$\implies r^{q+1} - (1 - b) r^q - a_m r^{q-m} - \cdots - a_q = 0$$

$$\implies r^{q+1} - r^q + br^q - a_m r^{q-m} - \cdots - a_q = 0$$

$$\implies r^{q+1} - r^q = -br^q + a_m r^{q-m} + \cdots + a_q$$

If  $r > 1$  then  $r^{q+1} > r^q$  and so  $r^{q+1} - r^q > 0$

and so

$$0 < -br^q + a_m r^{q-m} + \cdots + a_q$$

$$\implies br^q < a_m r^{q-m} + \cdots + a_q$$

$$\leq a_m r^{q-m} + \cdots + a_q r^{q-m}$$

$$= (a_m + \cdots + a_q) r^{q-m}$$

$$= ar^{q-m}$$

So  $br^q < ar^{q-m} \iff r^m = \frac{a}{b} \iff r < \left(\frac{a}{b}\right)^{1/m}$

And if  $r < 1$  we get  $r < \left(\frac{a}{b}\right)^{1/m}$ .

We can combine both cases with  $x \leq \max(1, r)$  and  $x \geq \min(1, r)$  (for any  $x \in \mathbb{R}$ ) to get (27), as required.

□

**Remark.** The inequality (27) will be useful to help approximate the leading root  $r$  in 4.

**Lemma 3.** For the model described by recurrence equation (17) we have

$$R_0 = \frac{a}{b}$$

**Proof:** Let  $u$  be the number of people infected on some day, set this to be day 0. On the day  $k = 1, \dots, q$  the number  $c_k u$  of them become ill and can infect other people. On the day  $k + 1$  they infect  $a c_k u$  people while  $b c_k u$  of them get isolated. On the day  $k + 1$ , the remaining  $(1 - b)c_k u$  people infect further  $a(1 - b)c_k u$  people. Continuing this way, we obtain that this group of  $c_k u$  people infects in total

$$\begin{aligned} ac_k u + a(1 - b)c_k u + a(1 - b)^2 c_k u + \dots &= ac_k u \sum_{n=0}^{\infty} (1 - b)^n \\ &= \frac{ac_k u}{1 - (1 - b)} \quad \text{since } 0 < 1 - b < 1 \\ &= \frac{a}{b} c_k u \end{aligned}$$

other people.

Hence, the initial group of  $u$  people infects in total

$$\sum_{k=0}^q \frac{a}{b} c_k u = \frac{a}{b} u \sum_{k=0}^q c_k = \frac{a}{b} u$$

other people.

And we defined  $R_0$  as being the unit reproduction number per infected person ( $u = 1$ ).

And so we get the result  $R_0 = \frac{a}{b}$  as required. □

We can then apply Newton's method [8] to find a better approximation for  $r$ .

**Definition 4.** Let  $\{r_n\}_{n \geq 0}$  be a sequence defined by

$$r_{n+1} = r_n - \frac{f(r_n)}{f'(r_n)}, \quad n \geq 0 \quad (28)$$

Then the limit converges to the leading root  $r$ , i.e.

$$\{r_n\} \rightarrow r \quad \text{as } n \rightarrow \infty \quad (29)$$

**Result 4.** A good approximation for the leading root  $r$  is

$$r \approx \frac{q \left(\frac{a}{b}\right)^{\frac{1}{2q}} - (q-1)(1-b) + (ab)^{-\frac{1}{2}}}{(q+1) - q(1-b) \left(\frac{a}{b}\right)^{-\frac{1}{2q}}} \quad (30)$$

**Proof:** Our characteristic polynomial (via (18)) for our recurrence equation (3.2.1) is

$$f(\lambda) = \lambda^{q+1} - (1-b)\lambda^q - a \quad (31)$$

Let  $r$  be its leading root, i.e.  $f(r) = 0$

Then by 2, we get

$$\min \left( 1, \left(\frac{a}{b}\right)^{1/q} \right) \leq r \leq \max \left( 1, \left(\frac{a}{b}\right)^{1/q} \right) \quad (32)$$

since  $m = q$  in our polynomial.

Taking the *geometric mean* of the bounds, defined as (19), we get an approximation for  $r$

$$r_0 = \left( 1 \cdot \left(\frac{a}{b}\right)^{1/q} \right)^{\frac{1}{2}} = \left(\frac{a}{b}\right)^{\frac{1}{2q}} \quad (33)$$

The derivative of our characteristic polynomial (31) is

$$f'(\lambda) = (q+1)\lambda^q - q(1-b)\lambda^{q-1} \quad (34)$$

Then we apply Newton's method (28) once to get

$$\begin{aligned}
r_1 &= r_0 - \frac{f(r_0)}{f'(r_0)} \\
&= r_0 - \frac{r_0^{q+1} - (1-b)r_0^q - a}{(q+1)r_0^q - q(1-b)r_0^{q-1}} \\
&= \left(\frac{a}{b}\right)^{\frac{1}{2q}} - \frac{\left(\left(\frac{a}{b}\right)^{\frac{1}{2q}}\right)^{q+1} - (1-b)\left(\left(\frac{a}{b}\right)^{\frac{1}{2q}}\right)^q - a}{(q+1)\left(\left(\frac{a}{b}\right)^{\frac{1}{2q}}\right)^q - q(1-b)\left(\left(\frac{a}{b}\right)^{\frac{1}{2q}}\right)^{q-1}} \\
&= \left(\frac{a}{b}\right)^{\frac{1}{2q}} - \frac{\left(\frac{a}{b}\right)^{\frac{q+1}{2q}} - (1-b)\left(\frac{a}{b}\right)^{\frac{1}{2}} - a}{(q+1)\left(\frac{a}{b}\right)^{\frac{1}{2}} - q(1-b)\left(\frac{a}{b}\right)^{\frac{q-1}{2q}}} \\
&= \left(\frac{a}{b}\right)^{\frac{1}{2q}} - \frac{\left(\frac{a}{b}\right)^{-\frac{1}{2}} \cdot \left(\frac{a}{b}\right)^{\frac{q+1}{2q}} - (1-b)\left(\frac{a}{b}\right)^{\frac{1}{2}} - a}{\left(\frac{a}{b}\right)^{-\frac{1}{2}} \cdot (q+1)\left(\frac{a}{b}\right)^{\frac{1}{2}} - q(1-b)\left(\frac{a}{b}\right)^{\frac{q-1}{2q}}} \\
&= \left(\frac{a}{b}\right)^{\frac{1}{2q}} - \frac{\left(\frac{a}{b}\right)^{\frac{1}{2q}} - (1-b) - a\left(\frac{a}{b}\right)^{-\frac{1}{2}}}{(q+1) - q(1-b)\left(\frac{a}{b}\right)^{-\frac{1}{2q}}} \\
&= \left(\frac{a}{b}\right)^{\frac{1}{2q}} - \frac{\left(\frac{a}{b}\right)^{\frac{1}{2q}} - (1-b) - a\left(\frac{a}{b}\right)^{-\frac{1}{2}}}{(q+1) - q(1-b)\left(\frac{a}{b}\right)^{-\frac{1}{2q}}} \\
&= \frac{(q+1)\left(\frac{a}{b}\right)^{\frac{1}{2q}} - q(1-b) - \left(\frac{a}{b}\right)^{\frac{1}{2q}} + (1-b) + a\left(\frac{a}{b}\right)^{-\frac{1}{2}}}{(q+1) - q(1-b)\left(\frac{a}{b}\right)^{-\frac{1}{2q}}} \\
&= \frac{q\left(\frac{a}{b}\right)^{\frac{1}{2q}} - (q-1)(1-b) + a\left(\frac{b}{a}\right)^{-\frac{1}{2}}}{(q+1) - q(1-b)\left(\frac{a}{b}\right)^{-\frac{1}{2q}}} \\
&= \frac{q\left(\frac{a}{b}\right)^{\frac{1}{2q}} - (q-1)(1-b) + (ab)^{-\frac{1}{2}}}{(q+1) - q(1-b)\left(\frac{a}{b}\right)^{-\frac{1}{2q}}}
\end{aligned}$$

□

While one iteration of Newton's method is often enough with our initial choice  $r_0 = (R_0)^{\frac{1}{2q}}$ , we will show that it converges to  $r$  in the limit

**Lemma 5.** Let  $r'$  be the largest non-negative root of  $f'(\lambda)$ , the derivative of (18).

If we choose an initial  $r_0 > r'$  then  $\lim_{n \rightarrow \infty} r_n = r$ .

**Proof:** Since  $f(\lambda)$  defined above has leading root  $r$  and has leading coefficient 1, we know that  $f(\lambda) > 0$  for  $\lambda > r$ .

Also, since  $r$  is the only positive root, by part (a) of 1, then  $f(\lambda) < 0$  for  $0 < \lambda < r$ .

We have

$$r_1 = r_0 - \frac{f(r_0)}{f'(r_0)} \tag{35}$$

We require  $r_0$  and  $r_1$  to be in the open interval  $(r', \infty)$  in order for  $r_n$  to converge to  $r$ .

There are two cases to consider:

- If our chosen  $r_0$  is less than  $r$ , then  $f(r_0) < 0$  and  $f'(r_0) > 0$ .

$$r_1 = r_0 - \underbrace{\frac{f(r_0)}{f'(r_0)}}_{<0} > r_0 > r'$$

Then  $r_1 > r'$  and the sequence will converge.

- If our chosen  $r_0$  is greater than  $r$ , then  $f(r_0) > 0$  and  $f'(r_0) > 0$ .

By a similar argument to above,

$$r_1 < r_0$$

so we need to do more work to show  $r_1 > r'$  in this case

Since  $f$  is continuous in  $[r_1, r_0]$  and differentiable in  $(r_1, r_0)$ , we can apply the *Mean Value Theorem* to imply that there exists some point  $c \in (r_1, r_0)$  such that

$$f'(c) = \frac{f(r_0) - f(r_1)}{r_0 - r_1} \quad (36)$$

We can rearrange this equation

$$f(r_1) = f(r_0) + f'(c)(r_1 - r_0)$$

But from (35) we have

$$r_1 - r_0 = -\frac{f(r_0)}{f'(r_0)}$$

And so

$$f(r_1) = f(r_0) - f'(c) \frac{f(r_0)}{f'(r_0)} \quad (37)$$

$$= f(r_0) \left( 1 - \frac{f'(c)}{f'(r_0)} \right) \quad (38)$$

But since the function  $f$  is convex in the open interval  $(r', \infty)$ , we know that  $0 < f'(c) < f'(r_0)$ . Therefore

$$0 < f(r_1) < f(r_0)$$

which gives

$$r < r_1 < r_0$$

and finally

$$r' < r < r_1 \quad (39)$$

And so  $r_1 > r$  in both cases, then Newton's method ensures  $\lim_{n \rightarrow \infty} r_n = r$ , as required.  $\square$

**Definition 5.** The function  $f$  takes on the average value between points  $x_1$  and  $x_2$ ,  $f_{\text{avg}}$  given by the formula

$$f_{\text{avg}} = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} f(x) dx \quad (40)$$

**Proposition 6.** The average value of the sequence  $a_0, \dots, a_N$  defined by (7) can be reasonably estimated by our previous parameter  $a$ .

Similarly, the average value of the sequence  $b_0, \dots, b_N$  defined by (8) can be reasonably estimated by our previous parameter  $b$ .

**Proof:** Let  $a(x)$  be the continuous extension of  $a_0, a_1, \dots, a_N$ , i.e.

$$a(x) = a \left( 1 + c_1 \left( \sin \left( \frac{2\pi}{p_1} (x - n_1) \right) \right) \right), \quad 0 \leq x \leq N \quad (41)$$

Then, for  $x_1 = 0$  and  $x_2 = N$

$$\begin{aligned} a_{\text{avg}} &= \frac{1}{N - 0} \int_0^N a(x) dx \\ &= \frac{1}{N} \int_0^N a \left( 1 + c_1 \left( \sin \left( \frac{2\pi}{p_1} (x - n_1) \right) \right) \right) dx \\ &= \frac{1}{N} \int_0^N adx + \frac{ac_1}{N} \int_0^N \sin \left( \frac{2\pi}{p_1} (x - n_1) \right) dx \end{aligned}$$

Then, using  $\sin(A - B) = \sin(A)\cos(B) - \cos(A)\sin(B)$

$$\begin{aligned}
a_{\text{avg}} &= \frac{1}{N} ax \Big|_0^N + \frac{ac_1}{N} \int_0^N \left( \sin\left(\frac{2\pi x}{p_1}\right) \cos\left(\frac{2\pi n_1}{p_1}\right) - \cos\left(\frac{2\pi x}{p_1}\right) \sin\left(\frac{2\pi n_1}{p_1}\right) \right) dx \\
&= \frac{aN}{N} + \frac{ac_1}{N} \cos\left(\frac{2\pi n_1}{p_1}\right) \int_0^N \sin\left(\frac{2\pi x}{p_1}\right) dx - \frac{ac_1}{N} \sin\left(\frac{2\pi n_1}{p_1}\right) \int_0^N \cos\left(\frac{2\pi x}{p_1}\right) dx \\
&= a + \frac{ac_1}{N} \cos\left(\frac{2\pi n_1}{p_1}\right) \cdot \left( -\frac{p_1}{2\pi} \cos\left(\frac{2\pi x}{p_1}\right) \right) \Big|_0^N - \frac{ac_1}{N} \sin\left(\frac{2\pi n_1}{p_1}\right) \cdot \left( \frac{p_1}{2\pi} \sin\left(\frac{2\pi x}{p_1}\right) \right) \Big|_0^N \\
&= a - \frac{ac_1}{N} \cos\left(\frac{2\pi n_1}{p_1}\right) \frac{p_1}{2\pi} \cos\left(\frac{2\pi N}{p_1}\right) + \frac{ac_1}{N} \cos\left(\frac{2\pi n_1}{p_1}\right) \frac{p_1}{2\pi} \cdot 1 - \frac{ac_1}{N} \sin\left(\frac{2\pi n_1}{p_1}\right) \frac{p_1}{2\pi} \sin\left(\frac{2\pi N}{p_1}\right) + 0 \\
&= a - \frac{ac_1 p_1}{2\pi N} \left( \cos\left(\frac{2\pi n_1}{p_1}\right) \cos\left(\frac{2\pi N}{p_1}\right) + \cos\left(\frac{2\pi n_1}{p_1}\right) - \sin\left(\frac{2\pi n_1}{p_1}\right) \sin\left(\frac{2\pi N}{p_1}\right) \right)
\end{aligned}$$

We use a similar identity  $\cos(A + B) = \cos(A)\cos(B) - \sin(A)\sin(B)$  to get

$$a_{\text{avg}} = a - \frac{ac_1 p_1}{2\pi N} \left( \cos\left(\frac{2\pi(N + n_1)}{p_1}\right) + \cos\left(\frac{2\pi n_1}{p_1}\right) \right)$$

Since  $\cos(\cdot)$  is bounded between  $\pm 1$  and

$c_1 p_1 < 0.2 \cdot 8 = 1.6$  and  $N > 50$  usually, we see that

$$\left| \frac{c_1 p_1}{2\pi N} \left( \cos\left(\frac{2\pi(N + n_1)}{p_1}\right) + \cos\left(\frac{2\pi n_1}{p_1}\right) \right) \right| < \frac{1.6}{300} \cdot (1 + 1) = \frac{4}{375}$$

So the average value of  $a(x)$  is approximately within  $4/375 \approx 1.07\%$  of the value of  $a$ .

Therefore  $a$  is a reasonable estimate for the average  $a_{\text{avg}}$  of the periodic parameter  $a(x)$ .

Similarly, exchanging  $c_1, n_1, p_1$  for  $c_2, n_2, p_2$  we have a definition for  $b(x)$  and hence  $b$  is a reasonable estimate for the average  $b_{\text{avg}}$ .

□

## B Source Code and Data Sources

Much of the code was written from scratch for this project, or is a close to direct translation of the formulas described in papers such as Grigorian's [1].

### B.1 R packages

`ggplot2` [9] is widely used for easily plotting and visualising the models. `rgdal` [10] allows geospatial .shp files to be read into R.

`raster` [11] allows this data to be manipulated and plotted.

`dplyr` [12] provides useful data manipulation functions, both for models and geospatial mapping.

Statistical models (HoltWinters, ARIMA and Neural Network Regression) were readily implemented from `forecast` [13].

The majority of the colours for plotting are selected using colour palettes from the `wesanderson` package [14].

### B.2 Shapefiles and Datasets

This data includes the geospatial vector data which can be used to draw country (and county) coastlines and borders.

World country shape data was obtained from [15], while the more detailed county-level shapefile for Ireland was downloaded from [16].

Country-level data:

I have chosen to use the data from [17], which has detailed daily data for nearly 200 countries and territories.

Ireland county-level data:

Downloaded from [18] and used to visualise the 14-day incidence rate per 1 million of the population, for each county.

### B.3 Source Code

All the code, plots, and even this report, are available at my [GitHub](#).

I will include the main code files in text below.

```
1 modnorm <- function(x,modx) return(floor(sum(abs(x-modx))/length(x)))
2
3 xntoyn <- function(xn) return(cumsum(xn))
4
5 basicmodx <- function(x, pars, len = 0){
6   q <- floor(pars[1])
7   a <- pars[2]
8   b <- pars[3]
9   modx <- x[1:q]
10  for(i in (q+1):(length(x)+len)){
11    modx[i] <- (1-b)*modx[i-1] + a*modx[i-q]
12  }
13  return(modx)
14 }
15
16 norm <- function(par, x) return(modnorm(x,basicmodx(x, par)))
17
18 normy <- function(par, x, y) return(modnorm(y,xntoyn(basicmodx(x, par))))
19
20 normalize <- function(x) return((x-min(x))/(max(x)-min(x)))
21
22 basef <- function(lambda,par) {
23   return(lambda^(par[1]+1)-(1-par[3])*lambda^(par[1])-par[2])
24 }
25
26 basefprime <- function(lambda,par) {
27   return((par[1]+1)*lambda^(par[1])-(1-par[3])*par[1]*lambda^(par[1]-1))
28 }
29
30 normC <- function(par,x,r){
31   return(modnorm(x, par*r^(0:(length(x)!is.na(x))-1)))
32 }
33
34 movingavg <- function(x){
35   mavgx <- (x[1]+x[2])/2
36   for(i in 2:(length(x)-1)){
```

```

37     mavgx[i] <- sum(x[(i-1):(i+1)]) / 3
38   }
39   mavgx[length(x)] <- (x[length(x)-1] + x[length(x)]) / 2
40   return(mavgx)
41 }
42
43 normper <- function(par, q, x) return(modnorm(x, modxper(par, q, x)))
44
45 modxper <- function(par, q, x, len = 0){
46   #a,b,c1,c2,p1,p2,n1,n2
47   an <- par[1] * (1 + par[3] * sin(2 * pi * (1:(length(x)+len) - par[7]) / par[5]))
48   bn <- par[2] * (1 + par[4] * sin(2 * pi * (1:(length(x)+len) - par[8]) / par[6]))
49   bn[bn > 1] <- 1
50   modx <- x[1:q]
51   for(i in (q+1):(length(x)+len)){
52     modx[i] <- (bn[i] * (1 - bn[i-1])) * modx[i-1] / bn[i-1] +
53       (an[i-q] * bn[i]) * modx[i-q] / bn[i-q]
54   }
55   return(modx)
56 }
```

**Listing 7:** Model helper functions

```

1 require(ggplot2)
2 require(forecast)
3 require(dplyr)
4 require(wesanderson)
5 require(gridExtra)
6 require(xtable)
7
8 owiddat <- read.csv("Data/owid-covid-data.csv")
9 owiddat$date <- as.Date(owiddat$date, tryFormats = c("%Y-%m-%d"))
10
11 plotslist <- list()
12
13 source("Code/covid-plotutils.R")
14 source("Code/covid-modelutils.R")
15
16 covidPlots <- function(country, dateBounds, data){
17   plots <- list()
18   dateBounds <- as.Date(dateBounds)
19   countryrows <- grep(country, data$location)
20   countrydat <- data.frame(date = data$date[countryrows],
21                             xn = data$new_cases[countryrows],
22                             yn = data$total_cases[countryrows])
23
24   forecastlen <- 14
25
26   prevcases <- countrydat$yn[countrydat$date == as.Date(dateBounds[1])-1]
#Specific dates
28   countrydat$test <- countrydat[countrydat$date >= dateBounds[1] & countrydat$date <= (dateBounds
29     [2] + forecastlen),]
30   countrydat <- countrydat[countrydat$date >= dateBounds[1] & countrydat$date <= dateBounds[2],]
31   latest_date <- dateBounds[2]
32   testBounds <- dateBounds[2] + c(1:forecastlen)
33
34   countrydesc <- paste0(country, " ", dateBounds[1], " to ", dateBounds[2])
35
36   cols <- list(
37     xn      = wes_palettes$Zissou1[1],
38     xntest  = wes_palettes$Darjeeling1[2],
39     yn      = wes_palettes$Darjeeling2[2],
40     yntest  = wes_palettes$BottleRocket2[4],
41     basexn  = wes_palettes$Darjeeling1[1],
42     baseyn  = wes_palettes$Darjeeling1[1],
43     x3      = wes_palettes$BottleRocket2[1],
44     Crn     = wes_palettes$FantasticFox1[2],
45     arima    = wes_palettes$Darjeeling1[4],
46     arimapi  = wes_palettes$Darjeeling1[3],
47     hw      = wes_palettes$Moonrise1[2],
48     hwpi    = wes_palettes$Moonrise1[1],
49     periodic = wes_palettes$GrandBudapest1[3], "#magenta",
50     nn      = wes_palettes$FantasticFox1[4],
51     nnpi    = wes_palettes$GrandBudapest1[4]
```

```

51  )
52
53 labs <- list(
54   train = "Training set",
55   test = "Test set",
56   base = "Base",
57   x3 = "Average x*(3)",
58   periodic = "Periodic",
59   hw = "HoltWinters",
60   hwpi = "HW 95% PI",
61   arimapi = "ARIMA 95% PI",
62   nnpi = "NNAR 95% PI"
63 )
64
65 #Basic model: need a,b,q,r using ||x-x*|| and ||y-y*||
66 ##q - Any infected person becomes ill and infectious on the q-th day after infection.
67 ##a - During each day, each ill person at large infects on average a other persons.
68 ##b - During each day, a fraction b of ill people at large gets isolated
69
70 aseq <- seq(from = 0.1, to = 1.5, length.out = 40)
71 bseq <- seq(from = 0.1, to = 0.9, length.out = 40)
72 qseq <- 6:8
73 normdat <- expand.grid(q = qseq, a = aseq, b = bseq)
74 abnorm <- apply(normdat, 1, function(x) norm(x, countrydat$xn))
75
76 normdat$abnorm <- abnorm
77 normdat$abnormn <- normalize(abnorm)
78
79 newnormdat <- normdat %>% top_n(abnorm, n = -0.07*nrow(.))
80
81 abnormy <- apply(newnormdat, 1, function(x) normy(x, countrydat$xn, countrydat$yn))
82 newnormdat$abnormyn <- normalize(abnormy)
83
84 newnormdat$combnorm <- 1*newnormdat$abnormn + 0.00*newnormdat$abnormyn
85
86 col_grad <- wes_palette("Zissou1", 20, type = "continuous")
87
88 #newnormdat$abnormy <- apply(newnormdat, 1, function(x) normy(x, countrydat$xn, countrydat$yn))
89
90 tileoptim <- newnormdat[which.min(newnormdat$abnorm), 1:3]
91 optimpars <- c(tileoptim$q, tileoptim$a, tileoptim$b)
92 plots[["combnorm"]] <- ggplot(newnormdat, aes(x = a, y = b, z = abnorm)) +
93   geom_contour_filled() + labs(fill = "||x-x*||") +
94   scale_fill_brewer(palette = "Spectral")
95
96 q <- optimpars[1]
97 a <- optimpars[2]
98 b <- optimpars[3]
99
100 basexn <- basicmodx(countrydat$xn, optimpars, len = forecastlen)
101
102 modeldat <- data.frame(date = c(countrydat$date, latest_date + 1:forecastlen),
103                           basexn = basexn, baseyn = xntoy(baseyn) + prevcases,
104                           xn = countrydat$xn, yn = countrydat$yn)
105
106 modeldat$tgroup <- ifelse(modeldat$date <= dateBounds[2], "train", "test")
107
108 #Newtons method, r_1 = r_0 - f(r_0)/f'(r_0)
109 base_r_zero <- (a/b)^(1/(2*q))
110 base_r_one <- base_r_zero -basef(base_r_zero, optimpars)/basefprime(base_r_zero, optimpars)
111 base_r_one <- round(base_r_one, 3)
112
113 trainrange <- 1:nrow(countrydat)
114 testrange <- nrow(countrydat) + 1:forecastlen
115
116 roptimpars <- round(optimpars, 3)
117 xnorm <- norm(optimpars, countrydat$xn)
118 xnormtest <- modnorm(basexn[testrange], countrydat$xn[testrange])
119
120 ynorm <- modnorm(countrydat$yn, modeldat$baseyn[1:length(countrydat$yn)])
121 ynormtest <- modnorm(modeldat$baseyn[testrange], countrydat$yn[testrange])
122
123 plots[["basexn"]] <- plot_basexn(modeldat, cols, labs)
124 plots[["baseyn"]] <- plot_baseyn(modeldat, cols, labs)

```

```

125| summarydf <- data.frame(model = "Basic Recursion", xnorm = xnorm, ynorm = ynorm,
126|                               xnormt = xnormtest, ynormt = ynormtest)
127|
128| optimC <- optim(par = countrydat$xn[1], normC, method = "Brent",
129|                   lower = 1, upper = 2*max(countrydat$xn[!is.na(countrydat$xn)]),
130|                   x = basexn[trainrange], r = base_r_one$par
131|
132| modeldat$Crn <- optimC*base_r_one^(1:nrow(modeldat))
133|
134| labs$Crn <- list(bquote(Cr~n^*, r=~*(base_r_one)^*, C=~*(floor(optimC))))
135|
136| plots[["Crn"]] <- plot_crn(modeldat, cols, labs)
137|
138| mavgx1 <- movingavg(modeldat$xn[!is.na(modeldat$xn)])
139|
140| modeldat$mavgx3 <- movingavg(mavgx1)
141| modeldat$mavgx3y <- xtoyn(modeldat$mavgx3) + prevcases
142|
143| x3norm <- modnorm(modeldat$xn[trainrange], modeldat$mavgx3[trainrange])
144| x3normt <- modnorm(modeldat$xn[testrange], modeldat$mavgx3[testrange])
145| x3normy <- modnorm(modeldat$yn[trainrange], modeldat$mavgx3y[trainrange])
146| x3normyt <- modnorm(modeldat$yn[testrange], modeldat$mavgx3y[testrange])
147|
148| plots[["mavgx3"]] <- plot_mavgx3(modeldat, cols, labs)
149|
150|
151| summarydf <- rbind(summarydf, data.frame(model = "Moving Average", xnorm = x3norm, ynorm = x3
152|                               normy,
153|                               xnormt = x3normt, ynormt = x3normyt))
154|
155| #parameters of the form (ci, pi, ni)
156| ##an = a(1 + c1 sin(2pi/p1 (n - n1)))
157| c_1seq <- c_2seq <- seq(0.04, 0.2, length.out = 17)
158| n_1seq <- n_2seq <- 1:q
159| p_1seq <- p_2seq <- 1:q
160| normdatp <- expand.grid(a = a, b = b,
161|                           c1 = c_1seq, c2 = c_2seq,
162|                           p1 = p_1seq, p2 = p_2seq,
163|                           n1 = n_1seq, n2 = n_2seq)
164|
165| pernorm <- apply(normdatp, 1, function(x) normper(x, q = q, countrydat$xn))
166| normdatp$pernorm <- pernorm
167|
168| peroptim <- normdatp[which.min(pernorm), 1:8]
169| pernorm <- normdatp[which.min(pernorm), 9]
170|
171| modeldat$periodic <- modxper(as.numeric(peropty), countrydat$xn, q = q, forecastlen)
172| modeldat$periodicy <- xtoyn(modeldat$periodic) + prevcases
173|
174| pernormy <- modnorm(countrydat$yn, modeldat$periodicy[trainrange])
175| pernormt <- modnorm(modeldat$xn[testrange], modeldat$periodic[testrange])
176| pernormyt <- modnorm(modeldat$yn[testrange], modeldat$periodicy[testrange])
177|
178| peropty <- round(as.numeric(peropty), 3)
179|
180| perparamdat <- data.frame(
181|   x = modeldat$date,
182|   an = peropty[1]*(1+peropty[3]*sin(2*pi*(1:(nrow(modeldat)) - peropty[7])/peropty[5])),
183|   bn = peropty[2]*(1+peropty[4]*sin(2*pi*(1:(nrow(modeldat)) - peropty[8])/peropty[6]))
184| )
185|
186| plots[["perparam"]] <- ggplot(perparamdat) +
187|   geom_line(aes(x = x, y = an, col = "a_n")) +
188|   geom_line(aes(x = x, y = bn, col = "b_n")) +
189|   geom_hline(aes(yintercept = peropty[1], col = "a"), linetype = "dashed") +
190|   geom_hline(aes(yintercept = peropty[2], col = "b"), linetype = "dashed") +
191|   xlab("date") + ylab("") +
192|   scale_color_manual(values = wes_palettes$Rushmore1[c(3,3,5,5)]) +
193|   guides(colour = guide_legend	override.aes = list(linetype =
194|     c("a"="dashed", "a_n"="solid", "b"="dashed", "b_n"="solid")))) +
195|   xntheme() + theme(legend.position = "right")
196|
197| #a,b,c1,c2,p1,p2,n1,n2

```

```

198 mathperamdat <- data.frame(a = peroptim[1], b = peroptim[2], q = q, r = base_r_one,
199                                     c1 = peroptim[3], p1 = peroptim[5], n1 = peroptim[7],
200                                     c2 = peroptim[4], p2 = peroptim[6], n2 = peroptim[8])
201
202 colnames(mathperamdat) <- c("$a$", "$b$", "$q$",
203                             "$r$",
204                             "$c_1$",
205                             "$p_1$",
206                             "$n_1$",
207                             "$c_2$",
208                             "$p_2$",
209                             "$n_2$")
210
211 #Statistical methods using timeseries forecasting
212
213 dat_ts <- ts(data = countrydat$xn, frequency = q)
214
215 g1 <- autoplot(dat_ts) + theme(axis.title = element_blank())
216 g2 <- ggAcf(dat_ts) + ggtitle(" ")
217 g3 <- ggPacf(dat_ts) + ggtitle(" ")
218
219 plots[["tsdisplay"]] <- grid.arrange(grobs = list(g1,g2,g3), layout_matrix = rbind(c(1, 1),
220                                         , 3)))
221
222 plots[["residuals"]] <- gghistogram(log(dat_ts), add.normal = TRUE, bins=10)
223
224 plots[["tsdecompose"]] <- autoplot(decompose(dat_ts))
225
226 if(any(countrydat$xn <= 0)){
227   plots[["hw"]]<- plots[["hwy"]]<- error_plot()
228 } else{
229   hwmethod <- "additive"
230   #lambda=0 ensures values stay positive
231   hwfct <- forecast::hw(dat_ts, h = forecastlen, seasonal = hwmethod, lambda = 0)
232
233   hwfct$fitted[1:q] <- countrydat$xn[1:q]
234   modeldat$hwxn <- c(hwfct$fitted, hwfcst$mean)
235   modeldat$hwlo <- c(hwfct$fitted, hwfcst$lower[,2])
236   modeldat$hwhi <- c(hwfct$fitted, hwfcst$upper[,2])
237
238   modeldat$hwyn <- xtoyn(modeldat$hwxn) + prevcases
239   modeldat$hwylo <- xtoyn(modeldat$hwlo) + prevcases
240   modeldat$hwyhi <- xtoyn(modeldat$hwhi) + prevcases
241
242   hwnorm <- modnorm(countrydat$xn, hwfcst$fitted)
243   hwnormt <- modnorm(modeldat$xn[testrange], hwfcst$mean)
244   hwnormy <- modnorm(countrydat$yn, modeldat$hwyn[trainrange])
245   hwnormyt <- modnorm(modeldat$yn[testrange], modeldat$hwyn[testrange])
246
247   plots[["hw"]]<- plot_hw(modeldat, cols, labs)
248   plots[["hwy"]]<- plot_hwy(modeldat, cols, labs)
249 }
250
251 auto.fit <- auto.arima(dat_ts, lambda = 0) #keep values positive
252
253 getArmaModel <- function(arma, pdq = c(1,6,2), PDQ = c(3,7,4), s=5){
254   return(paste0("ARIMA(", paste0(arma[pdq], collapse = ","), ")",
255                 paste0(arma[PDQ], collapse = ","), "[", arma[s], "])"))
256 }
257
258 summarydf <- rbind(summarydf, data.frame(model = "HoltWinters", xnorm = hwnorm,
259                           ,
260                           xnormt = hwnormt, ynormt = hwnormyt))
261
262 arima.fcst <- forecast(auto.fit, level = c(80, 95), h = forecastlen)
263 arima.fcst$fitted[1:q] <- countrydat$xn[1:q]
264
265 arimanorm <- modnorm(countrydat$xn, arima.fcst$fitted)
266 arimanormt <- modnorm(modeldat$xn[testrange], arima.fcst$mean)
267
268 arimalabs <- getArmaModel(auto.fit$arma)
269 labs$arima <- arimalabs
270
271 modeldat$arimaxn <- c(auto.fit$fitted, arima.fcst$mean)

```

```

270 modeldat$arimalo <- c(auto.fit$fitted, arima.fcst$lower[,2])
271 modeldat$arimahi <- c(auto.fit$fitted, arima.fcst$upper[,2])
272
273 modeldat$arimayn <- xtoyn(modeldat$arimaxn) + prevcases
274 modeldat$arimaylo <- xtoyn(modeldat$arimalo) + prevcases
275 modeldat$arimayhi <- xtoyn(modeldat$arimahi) + prevcases
276
277 arimanormy <- modnorm(countrydat$yn, modeldat$arimayn[trainrange])
278 arimanormyt <- modnorm(modeldat$yn[testrange], modeldat$arimayn[testrange])
279
280 plots[["arima"]] <- plot.arima(modeldat, cols, labs)
281 plots[["arimay"]] <- plot.arimay(modeldat, cols, labs)
282 plots[["hwarima"]] <- plot.hwarima(modeldat, cols, labs)
283
284 summarydf <- rbind(summarydf, data.frame(model = "ARIMA",
285   , xnrm = arimanorm, ynm = arimanormy,
286   , xnrmmt = arimanormmt, ynmmt = arimanormyt))
287
288 nHidden <- max(1, floor(0.5*(1+auto.fit$arma[1]+auto.fit$arma[3])))
#Box-Cox transformation with lambda=0 to ensure the forecasts stay positive.
289 nnfit <- nnetar(dat_ts, p = auto.fit$arma[1], P = auto.fit$arma[3], size = nHidden, lambda =
290   0, repeats = 20, maxit = 50)
nn.fcst <- forecast(nnfit, PI=TRUE, h = forecastlen)
labs$nn <- nnfit$method
292
293 nn.fcst$fitted[1:q] <- countrydat$xn[1:q]
294
295 modeldat$nnxn <- c(nn.fcst$fitted, nn.fcst$mean)
296 modeldat$nnlo <- c(nn.fcst$fitted, nn.fcst$lower[,2])
297 modeldat$nnhi <- c(nn.fcst$fitted, nn.fcst$upper[,2])
298
299 modeldat$nnyn <- xtoyn(modeldat$nnxn) + prevcases
300 modeldat$nnnylo <- xtoyn(modeldat$nnlo) + prevcases
301 modeldat$nnnyhi <- xtoyn(modeldat$nnhi) + prevcases
302
303 nnnorm <- modnorm(countrydat$xn, nn.fcst$fitted)
304 nnnormtest <- modnorm(modeldat$xn[testrange], nn.fcst$mean)
305 nnnormy <- modnorm(countrydat$yn, modeldat$nnyn[trainrange])
306 nnnormytest <- modnorm(modeldat$yn[testrange], modeldat$nnny[testrange])
307
308 plots[["nn"]] <- plot.nn(modeldat, cols, labs)
309 plots[["nnny"]] <- plot.nny(modeldat, cols, labs)
310
311 summarydf <- rbind(summarydf, data.frame(model = "Neural Network",
312   , xnrm = nnorm, ynm = nnormy,
313   , xnrmmt = nnormtest, ynmmt = nnormytest))
314
315 summarydf$xnrm <- as.integer(summarydf$xnrm)
316 summarydf$ynrm <- as.integer(summarydf$ynrm)
317 summarydf$xnrmmt <- as.integer(summarydf$xnrmmt)
318 summarydf$ynrmmt <- as.integer(summarydf$ynrmmt)
319
320 colnames(summarydf) <- c("model", "$||x-x^*||_{-train}$", "$||y-y^*||_{-train}$",
321   "$||x-x^*||_{-test}$", "$||y-y^*||_{-test}$")
322 return(list("plots" = plots, "summary" = summarydf, "desc" = countrydesc,
323   "dat" = modeldat, "mathmodelpars" = mathperamdat))
324 }
325
326 grigorDates <- c("2020-04-26", "2020-06-09")
327 datebounds <- list(
328   "Italy" = c("2021-01-02", "2021-02-16"),
329   "United States" = c("2021-01-06", "2021-02-16"),
330   "Ireland" = c("2021-01-12", "2021-02-16")
331 )
332
333 owiddat <- owiddat[!is.na(owiddat$new_cases),]
334 totaldat <- owiddat[owiddat$location == "World",]
335 latest_date <- totaldat$date[nrow(totaldat)]
336 wt_title <- sprintf('Global Total =%s as at %s',
337   format(sum(totaldat$new_cases), big.mark=",", scientific=FALSE),
338   format(latest_date, "%B %d, %Y"))
339
340 plotslist[["WorldTotal"]][["xn"]] <- plot.worldtotal(totaldat)
341 for(country in names(datebounds)){

```

```

341 | plotslist [[country]] <- covidPlots(country, datebounds[[country]], owiddat)
342 }

```

**Listing 8:** Main algorithm

```

1 source("Code/covid-plotutils.R")
2 source("Code/covid-modelutils.R")
3
4 owiddat      <- read.csv("Data/owid-covid-data.csv")
5 owiddat$date <- as.Date(owiddat$date, tryFormats = c("%Y-%m-%d"))
6 multilist <- list()
7
8 multidates <- list(
9   "Italy"       = list(c("2020-12-16", "2021-01-05"),
10    c("2021-01-06", "2021-01-30")),
11   "United States" = list(c("2020-11-16", "2021-01-06"),
12    c("2021-01-07", "2021-01-30")),
13   "Ireland"     = list(c("2020-12-16", "2021-01-07"),
14    c("2021-01-08", "2021-01-30"))
15 )
16
17 multiphasePlots <- function(country, dates, data){
18   plots <- list()
19   crows <- grep(country, data$location)
20   countrydat <- data.frame(date = data$date[crows],
21    xn = data$new_cases[crows], yn = data$total_cases[crows])
22   if(nrow(countrydat[countrydat$date < dates[[1]][1],]) == 0)
23     beforeecumcases <- 0
24   else
25     beforeecumcases <- sum(countrydat$xn[countrydat$date < dates[[1]][1]])
26
27   countrydat <- countrydat[countrydat$date >= dates[[1]][1],]
28   countrydat <- countrydat[countrydat$date <= dates[[length(dates)]][2],]
29   latest_date <- countrydat$date[nrow(countrydat)]
30
31   forecastlen <- 5
32
33   multimodx <- function(x, multix, pars, oldp = rep(1,10), start=FALSE, len = 0){
34     q <- floor(pars[1])
35     a <- pars[2]
36     b <- pars[3]
37     fitstd <- length(multix)+1
38
39     if(start){
40       multix[fitstd:(fitstd+q-1)] <- x[1:q]
41       for(i in (fitstd+q):(fitstd+length(x)+len-1)){
42         multix[i] <- (1-b)*multix[i-1] + a*multix[i-q]
43       }
44     } else {
45       for(i in (fitstd):(fitstd+length(x)+len-1)){
46         multix[i] <- (1-b)*multix[i-1] + a*multix[i-q]
47       }
48     }
49     return(multix)
50   }
51
52   multimodxper <- function(par, q=7, x, multix, oldp = rep(1,10), start=FALSE, len = 0){
53   #a,b,c1,c2,p1,p2,n1,n2
54   #first day of this phase
55   fitstd <- length(multix)+1
56   an <- par[1]*(1+par[3]*sin(2*pi*(1:(fitstd+length(x)+len-1) - par[7])/par[5]))
57   bn <- par[2]*(1+par[4]*sin(2*pi*(1:(fitstd+length(x)+len-1) - par[8])/par[6]))
58
59   if(start){
60     multix[fitstd:(fitstd+q-1)] <- x[1:q]
61     for(i in (fitstd+q):(fitstd+length(x)+len-1)){
62       multix[i] <- (bn[i]*(1-bn[i-1]))*multix[i-1]/bn[i-1] +
63       (an[i-q]*bn[i])*multix[i-q]/bn[i-q]
64     }
65   } else {
66     for(i in fitstd:(fitstd+length(x)+len-1)){
67       multix[i] <- (bn[i]*(1-bn[i-1]))*multix[i-1]/bn[i-1] +
68       (an[i-q]*bn[i])*multix[i-q]/bn[i-q]
69     }

```

```

70    }
71    return(multix)
72  }
73
74 #Specific dates
75 multimodel <- c()
76 multimodelp <- c()
77 phasepars <- list()
78 for(i in 1:length(dates)){
79   phase <- dates[[i]]
80   phasedat <- countrydat[countrydat$date >= phase[1] & countrydat$date <= phase[2],]
81
82   #get basic model for each phase first in order to get
83   # starting a and b to guess for periodic an and bn
84
85   aseq <- seq(from = 0.1, to = 2.5, length.out = 50)
86   bseq <- seq(from = 0.1, to = 0.9, length.out = 50)
87   qseq <- 6:8
88   normdat <- expand.grid(q = qseq, a = aseq, b = bseq)
89
90   if(i == 1)
91     abnorm <- apply(normdat, 1, function(x) modnorm(multimodx(phasedat$xn, multimodel, x, start=TRUE), phasedat$xn))
92   else
93     abnorm <- apply(normdat, 1, function(x) modnorm(multimodx(phasedat$xn, multimodel, x, oldp = phasepars[[i-1]])[length(multimodel) + 1:nrow(phasedat)], phasedat$xn))
94
95   normalize <- function(x){
96     return((x-min(x))/(max(x)-min(x)))
97   }
98   normdat$abnorm <- abnorm
99
100  tileoptim <- normdat[which.min(normdat$abnorm),1:3]
101  optimpars <- c(tileoptim$q, tileoptim$a, tileoptim$b)
102
103  q <- optimpars[1]
104  a <- optimpars[2]
105  b <- optimpars[3]
106
107  phasepars[[i]] <- round(optimpars,3)
108
109  if(i == 1 & i != length(dates))
110    multimodel <- multimodx(phasedat$xn, multimodel, optimpars, start=TRUE)
111  if(i == 1 & i == length(dates))
112    multimodel <- multimodx(phasedat$xn, multimodel, optimpars, start=TRUE, len=forecastlen)
113  if(i > 1 & i == length(dates))
114    multimodel <- multimodx(phasedat$xn, multimodel, optimpars, oldp = phasepars[[i-1]], len=forecastlen)
115  if(length(dates) > 2 & i %in% 2:(length(dates)-1))
116    multimodel <- multimodx(phasedat$xn, multimodel, optimpars, oldp = phasepars[[i-1]])
117
118  aseqper <- a*seq(from = 0.7, to = 1.3, length.out = 10)
119  bseqper <- b*seq(from = 0.7, to = 1.3, length.out = 10)
120  c_1seq <- c_2seq <- seq(0.04, 0.2, length.out = 10)
121  n_1seq <- n_2seq <- c(1,7)
122  p_1seq <- p_2seq <- 6:7
123  normmdatp <- expand.grid(a = aseqper, b = bseqper,
124                           c1 = c_1seq, c2 = c_2seq,
125                           p1 = p_1seq, p2 = p_2seq,
126                           n1 = n_1seq, n2 = n_2seq)
127
128  if(i == 1)
129    pernorm <- apply(normmdatp, 1, function(par) modnorm(multimodxper(par, q = optimpars[1], phasedat$xn, multimodelp, start=TRUE), phasedat$xn))
130  else
131    pernorm <- apply(normmdatp, 1, function(par) modnorm(multimodxper(par, q = optimpars[1], phasedat$xn, multimodelp)[length(multimodelp)+1:nrow(phasedat)], phasedat$xn))
132
133  normmdatp$pernorm <- pernorm
134
135  newnormmdatp <- normmdatp %>% top_n(pernorm, n = -25)
136
137  if(i == 1)
138    pernormy <- apply(newnormmdatp, 1, function(par) modnorm(beforecumcases+xntoyn(multimodxper(

```

```

139      par, q = optimpars[1], phasedat$xn, multimodelp, start=TRUE)), phasedat$yn))
140  else
141    pernormy <- apply(newnormmdatp, 1, function(par) modnorm(beforecumcases+xntoyn(multimodxper(
142      par, q = optimpars[1], phasedat$xn, multimodelp))[length(multimodelp)+1:nrow(phasedat)
143      ], phasedat$yn)))
144
145  newnormmdatp$pernormy <- pernormy
146  newnormmdatp$combnorm <- normalize(newnormmdatp$pernorm) + normalize(newnormmdatp$pernormy)
147
148  peroptim <- as.numeric(newnormmdatp[which.min(newnormmdatp$combnorm), 1:8])
149
150  phasepars[[i]] <- c(phasepars[[i]], round(eroptim, 3))
151
152  if(i == 1 & i != length(dates))
153    multimodelp <- multimodxper(eroptim, q = q, phasedat$xn, multimodelp, start=TRUE)
154  if(i == 1 & i == length(dates))
155    multimodelp <- multimodxper(eroptim, q = q, phasedat$xn, multimodelp, start=TRUE, len=
156      forecastlen)
157  if(i > 1 & i == length(dates))
158    multimodelp <- multimodxper(eroptim, q = q, phasedat$xn, multimodelp, len=forecastlen)
159  if(length(dates) > 2 & i %in% 2:(length(dates)-1))
160    multimodelp <- multimodxper(eroptim, q = q, phasedat$xn, multimodelp)
161
162  cols <- list(
163    xn = wes_palettes$Zissou1[1],
164    yn = wes_palettes$Darjeeling2[2],
165    multixn = wes_palettes$Darjeeling1[1],
166    multiyn = wes_palettes$Darjeeling1[1],
167    multip = "magenta4",
168    x3 = wes_palettes$FantasticFox1[2]
169  )
170
171  labs <- list(
172    xn = list(bquote(.(country)*,"`x[n]*`=new cases/day, actual till`~.(format.Date(latest_date
173      , "%d.%m.%Y"))),
174    yn = list(bquote(.(country)*,"`y[n]*`=cumulative cases, actual till`~.(format.Date(latest_
175      date, "%d.%m.%Y"))))
176  )
177
178  multimodnormval <- modnorm(multimodel[1:length(countrydat$xn)], countrydat$xn)
179  multimodnormyval <- modnorm(beforecumcases+xntoyn(multimodel[1:length(countrydat$xn)], 
180    countrydat$yn))
181
182  b.roman <- function(x){ return(paste0("(", as.roman(x), ")"))}
183
184  multilabd <- c()
185  for(i in 1:length(dates)){
186    multilabd <- c(multilabd, paste(b.roman(i), "from", format.Date(as.Date(as.character(dates[[i
187      ]][1])), "%d.%m")))
188  }
189  multilabd <- paste0(multilabd, collapse = "; ")
190
191  multilabp <- c()
192  for(i in 1:length(dates)){
193    multilabp <- c(multilabp, paste0(b.roman(i), " a=", phasepars[[i]][[2]], ", b=", phasepars[[i
194      ]][[3]]))
195  }
196  multilabp <- paste0(multilabp, paste0("; q=", phasepars[[i]][[1]]), collapse = "; ")
197
198  labs$multixn <- list(bquote("base`~.(length(dates))*`-phase model`~`x[n]*`=new cases/day: `*
199      .(multilabd)*
200      `; ||x-x||="`*(.multimodnormval)*` `*
201      .(multilabp)))
202  labs$multiyn <- list(bquote("base`~.(length(dates))*`-phase model`~`y[n]*`=cumulative cases: `*
203      .(multilabd)*
204      `; ||y-y||="`*(.multimodnormyval)*` `*
205      .(multilabp)))
206
207  mavgx1 <- movingavg(countrydat$xn[!is.na(countrydat$xn)])
208
209  countrydat$mavgx3 <- movingavg(mavgx1)
210
211  x3norm <- modnorm(countrydat$xn, countrydat$mavgx3)

```

```

204 labs$x3 <- list(bquote("moving average x*(3); ||x*(3)-x*||="*.(x3norm)))
205
206 modeldat <- data.frame(date = c(countrydat$date, as.Date(latest_date) + 1:forecastlen),
207                         multixn = multimodel,
208                         multiyn = beforecumcases + xntoyn(multimodel),
209                         multipxn = multimodelp,
210                         multipyn = beforecumcases + xntoyn(multimodelp))
211
212 plots[["xn"]] <- plot_multixn(countrydat, modeldat, cols, labs) + theme(legend.position = "top")
213
214 plots[["yn"]] <- plot_multiyn(countrydat, modeldat, cols, labs) + theme(legend.position = "top")
215
216 multilabp <- c()
217 for(i in 1:length(dates)){
218   multilabp <- c(multilabp, paste0(b.roman(i), " a=", phasepars[[i]][[2]], ", b=", phasepars
219   [[i]][[3]]))
220 }
221 multilabp <- paste0(multilabp, paste0("; q=", phasepars[[i]][[1]]), collapse = "; ")
222
223 labs$multipxn <- list(bquote("periodic"~.(length(dates))*"phase model"~x[n]*"=new cases/day: "
224   *
225   .(multilabd)*
226   ";" ||x*-x||="*.(multimodnormval)*" "*"
227   .(multilabp)))
228
229 labs$multipyn <- list(bquote("periodic"~.(length(dates))*"phase model"~y[n]*"=cumulative cases
230   : "*"
231   .(multilabd)*
232   ";" ||y*-y||="*.(multimodnormyval)*" "*"
233   .(multilabp)))
234
235 plots[["perxn"]] <- plot_multiperxn(countrydat, modeldat, cols, labs) + theme(legend.position =
236   "top")
237
238 plots[["peryn"]] <- plot_multiperyn(countrydat, modeldat, cols, labs) + theme(legend.position =
239   "top")
240
241 return(plots)
242 }
243
244 for(country in names(multidates)){
245   multilist[[country]] <- multiphasePlots(country, multidates[[country]], owiddat)
246 }
```

**Listing 9:** Multi-phase models

## References

- [1] Alexander Grigorian. *Mathematical riddles of COVID-19*. June 2020. URL: <https://www.math.uni-bielefeld.de/~grigor/corv.pdf>.
- [2] Dr Tedros Adhanom Ghebreyesus. *WHO Director-General's opening remarks at the media briefing on COVID-19*. World Health Organization. Mar. 11, 2020. URL: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>.
- [3] Seth Flaxman and Imperial College COVID-19 Response Team. "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe". In: *Nature* 584.7820 (2020), pp. 257–261. ISSN: 1476-4687. DOI: [10.1038/s41586-020-2405-7](https://doi.org/10.1038/s41586-020-2405-7). URL: <https://doi.org/10.1038/s41586-020-2405-7>.
- [4] Arni S. R. Srinivasa Rao and Jose A. Vazquez. "Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine". eng. In: *Infection control and hospital epidemiology* 41.7 (2020). 32122430[pmid], pp. 826–830. ISSN: 1559-6834. DOI: [10.1017/ice.2020.61](https://doi.org/10.1017/ice.2020.61). URL: <https://pubmed.ncbi.nlm.nih.gov/32122430>.
- [5] L.J.S. Allen et al. *Mathematical Epidemiology*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2008. ISBN: 9783540789109. URL: <https://books.google.ie/books?id=gcP5l1a22rQC>.
- [6] Ravi Agarwal et al. "Dynamic equations on time scales: a survey". In: *Journal of Computational and Applied Mathematics* 141.1 (2002). Dynamic Equations on Time Scales, pp. 1 –26. ISSN: 0377-0427. DOI: [https://doi.org/10.1016/S0377-0427\(01\)00432-0](https://doi.org/10.1016/S0377-0427(01)00432-0). URL: <http://www.sciencedirect.com/science/article/pii/S0377042701004320>.
- [7] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice, 2nd edition*. OTexts.com/fpp2. OTexts, 2018.
- [8] Springer Verlag GmbH, European Mathematical Society. *Encyclopedia of Mathematics, Newton method*. Website. Accessed on 2021-03-19. URL: [https://encyclopediaofmath.org/index.php?title=Newton\\_method](https://encyclopediaofmath.org/index.php?title=Newton_method).
- [9] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: <https://ggplot2.tidyverse.org>.
- [10] Roger Bivand, Tim Keitt, and Barry Rowlingson. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.5-18. 2020. URL: <https://CRAN.R-project.org/package=rgdal>.
- [11] Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling*. R package version 3.4-5. 2020. URL: <https://CRAN.R-project.org/package=raster>.
- [12] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.3. 2021. URL: <https://CRAN.R-project.org/package=dplyr>.
- [13] Rob J Hyndman and Yeasmin Khandakar. *Automatic time series forecasting: the forecast package for R*. 3. 2008, pp. 1–22. URL: <https://www.jstatsoft.org/article/view/v027i03>.
- [14] Karthik Ram. *Wes Anderson Palettes*. 2013. URL: <https://github.com/karthik/wesanderson>.
- [15] ArcGIS Hub. *UNIGIS Geospatial Education Resources*. 2020. URL: [https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08\\_1](https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08_1).
- [16] © OpenStreetMap contributors. *Townlands*. 2020. URL: <https://www.townlands.ie/page/download/>.
- [17] Our World in Data. *The complete Our World in Data COVID-19 dataset*. 2020. URL: <https://covid.ourworldindata.org/data/owid-covid-data.csv>.
- [18] GeoHive Open Data Catalogue. *Covid-19 Daily Statistics for Ireland by County polygon as reported by the Health Surveillance Protection Centre*. 2020. URL: [https://opendata-geohive.hub.arcgis.com/datasets/d9be85b30d7748b5b7c09450b8aede63\\_0](https://opendata-geohive.hub.arcgis.com/datasets/d9be85b30d7748b5b7c09450b8aede63_0).