Final Year Project

Anthony Gibbons

February 19, 2021

Abstract

We construct various models of the Covid-19 pandemic over various periods of 2020. We first construct simple model of a the epidemic by using a recurrence equation. We also add a periodic complexity to these simpler models. We then use more statistical methods, modelling using time series forecasting methods such as HoltWinters, ARIMA and Neural Network methods. All of this is with the aim of predicting the course of the epidemic.

Keywords— ARIMA, Autoregressive model, COVID-19; Coronavirus, Forecasting, Mathematical model, Neural Network, Pandemic, Parameter estimation, SARS-CoV-2, Statistical Model.

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.

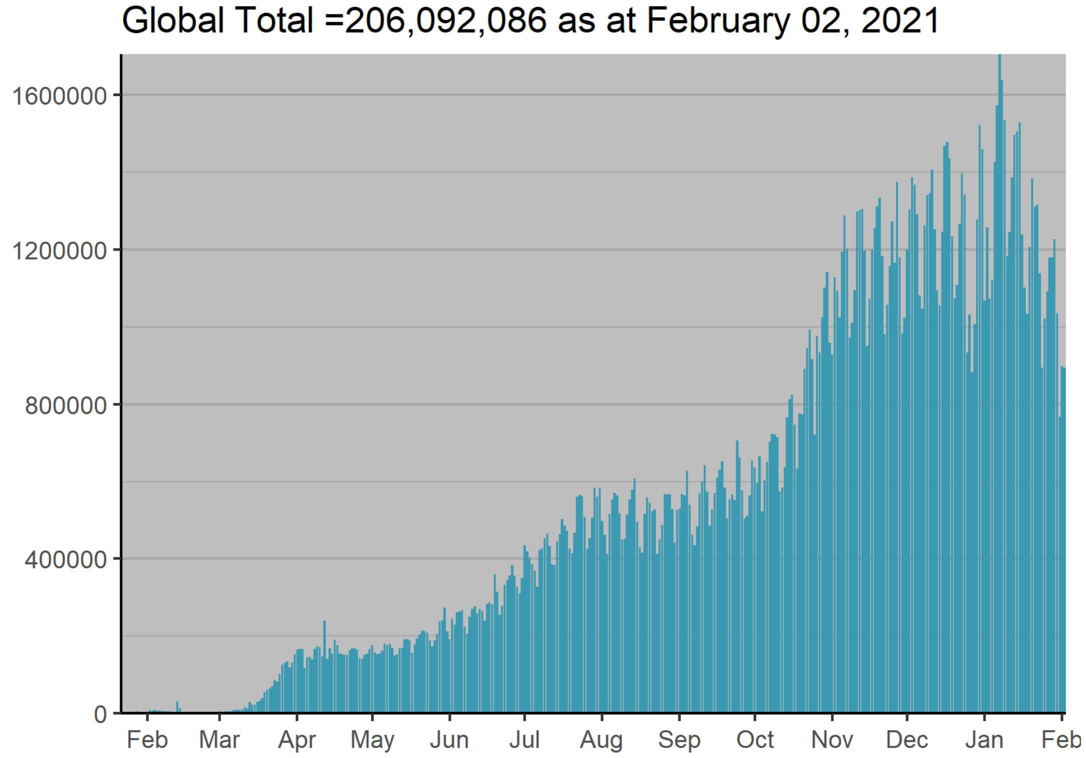
Signed: Anthony Gibbons
Date: 03/02/2021

Contents

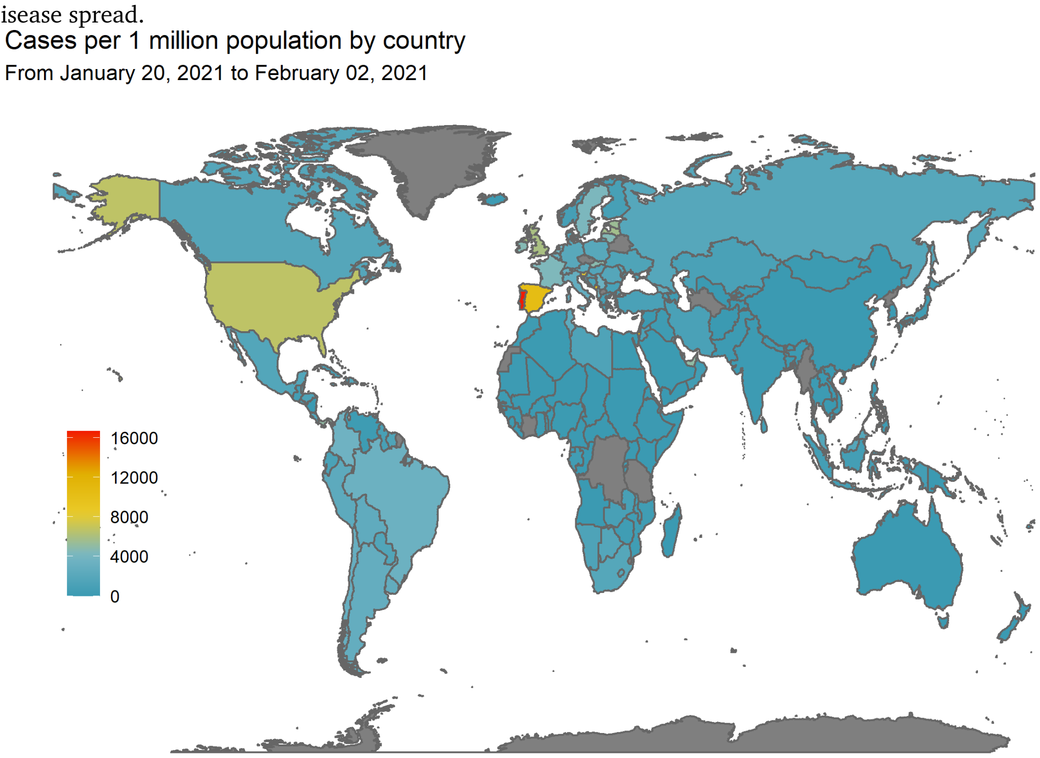
List of Figures

1 Introduction

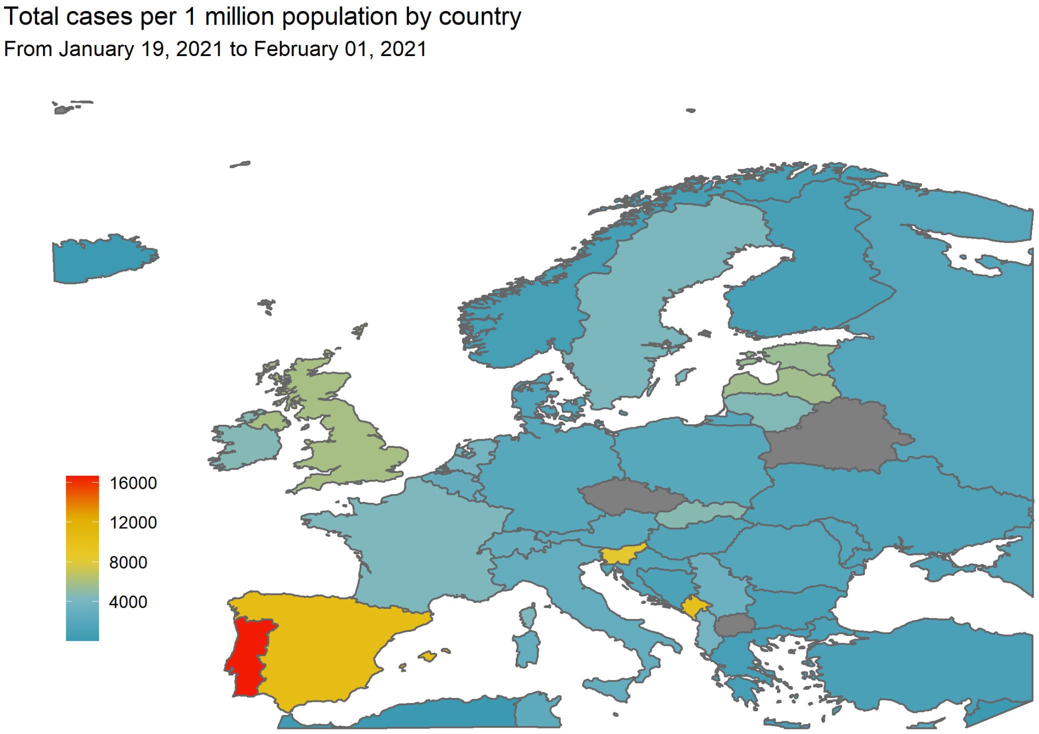
The Coronavirus disease (COVID-19) was first characterized by the World Health Organisation as pandemic on 11th March 2020 [14]. The outbreak has affected almost every aspect of human life throughout 2020, and is expected to continue for much of 2021.



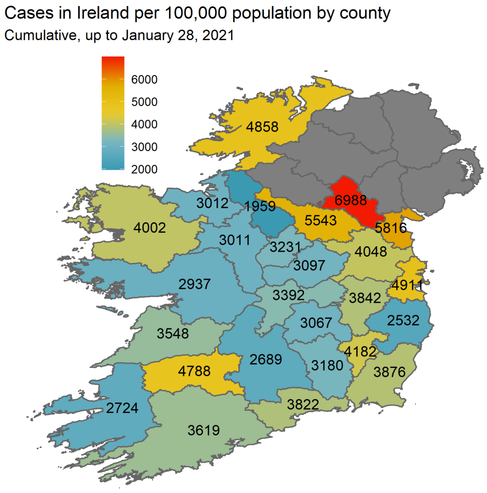
We can map the cumulative number of cases per 100,000 population for each country to see the varying severity of disease spread.

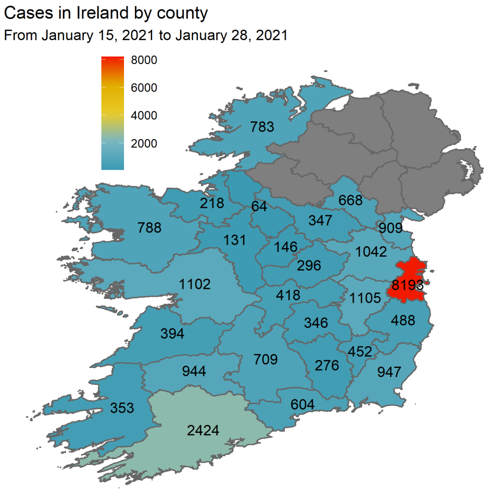

Cases per 1 million population by country
From January 20, 2021 to February 02, 2021
16000
12000
8000
4000
0

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


Total cases per 1 million population by country
From January 19, 2021 to February 01, 2021

More locally, we see that Ireland also has a clear variation in concentration of 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.


Cases in Ireland per 100,000 population by county
Cumulative, up to January 28, 2021


Cases in Ireland by county
From January 15, 2021 to January 28, 2021

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 [13], 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 [12].

Artificial intelligence models have also been employed to track disease outbreaks in more local areas. The model developed by [21], which relies on phone–based surveys, certainly has the long-run potential to keep the public informed and hopefully reduce the 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 [15], 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 [2].

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.

2 Mathematical Model

As per the base and periodic models shown in [15].

2.1 Base model

2.1.1 Definitions and Theory

2.1.2 How to select the best model

2.1.3 Forecasting

2.1.4 Implementation in R

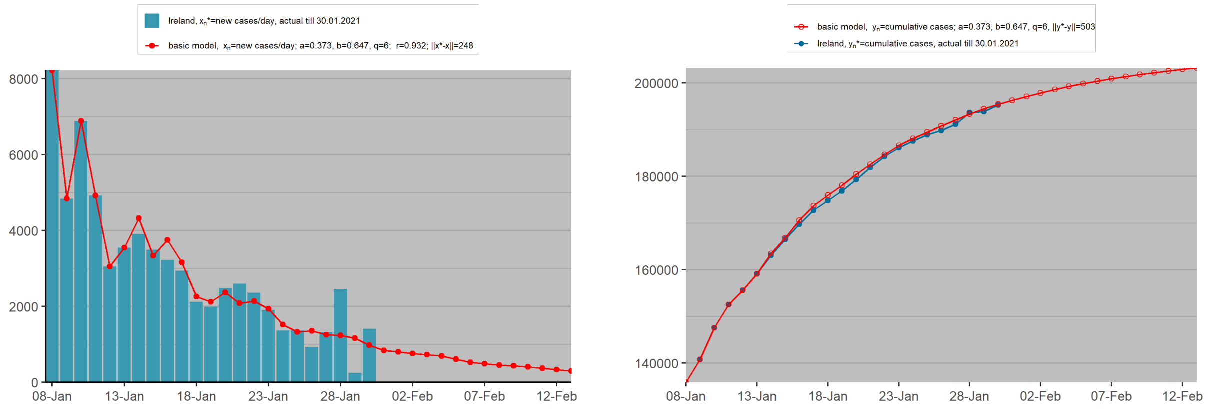


Figure 1: Basic model, Ireland

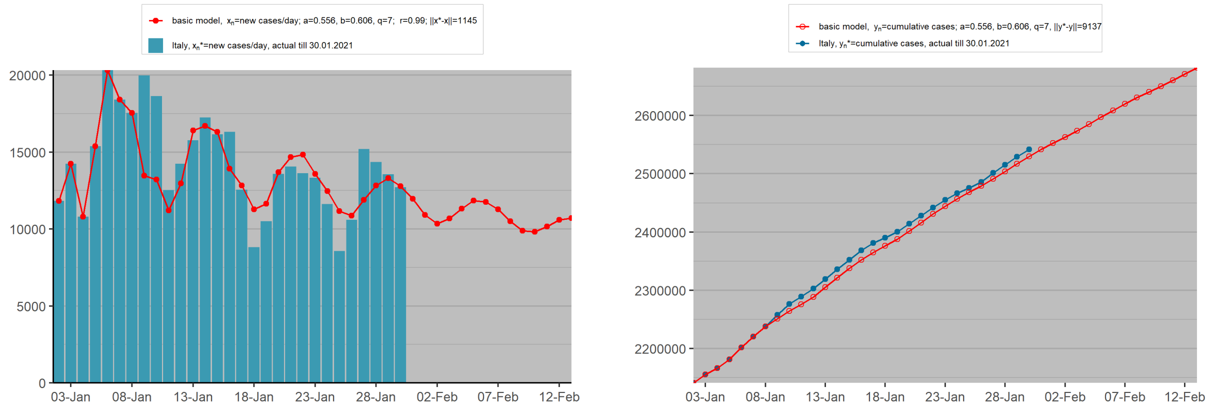


Figure 2: Basic model, Italy

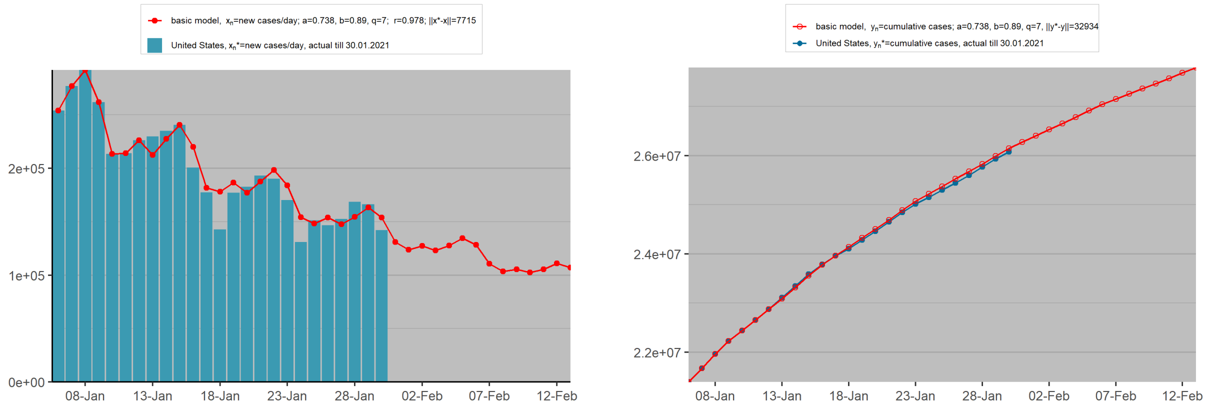


Figure 3: Basic model, United States

2.2 Limiting curve

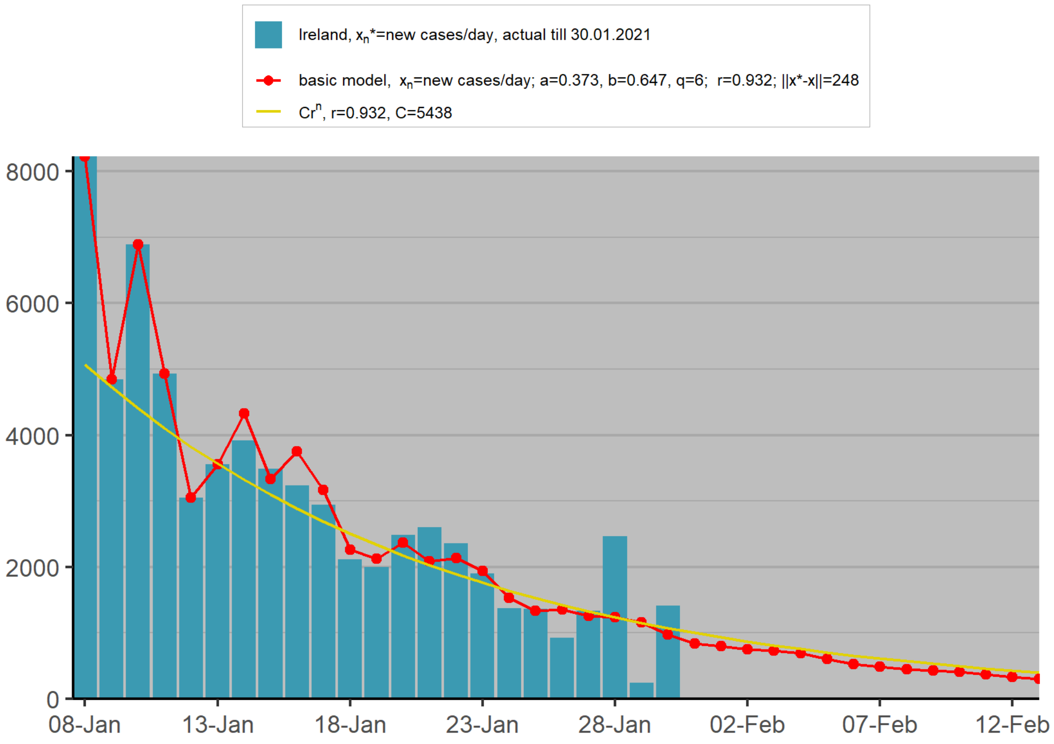


Figure 4: Comparison of x_n^*, x_n and the limiting curve Cr^n, Ireland

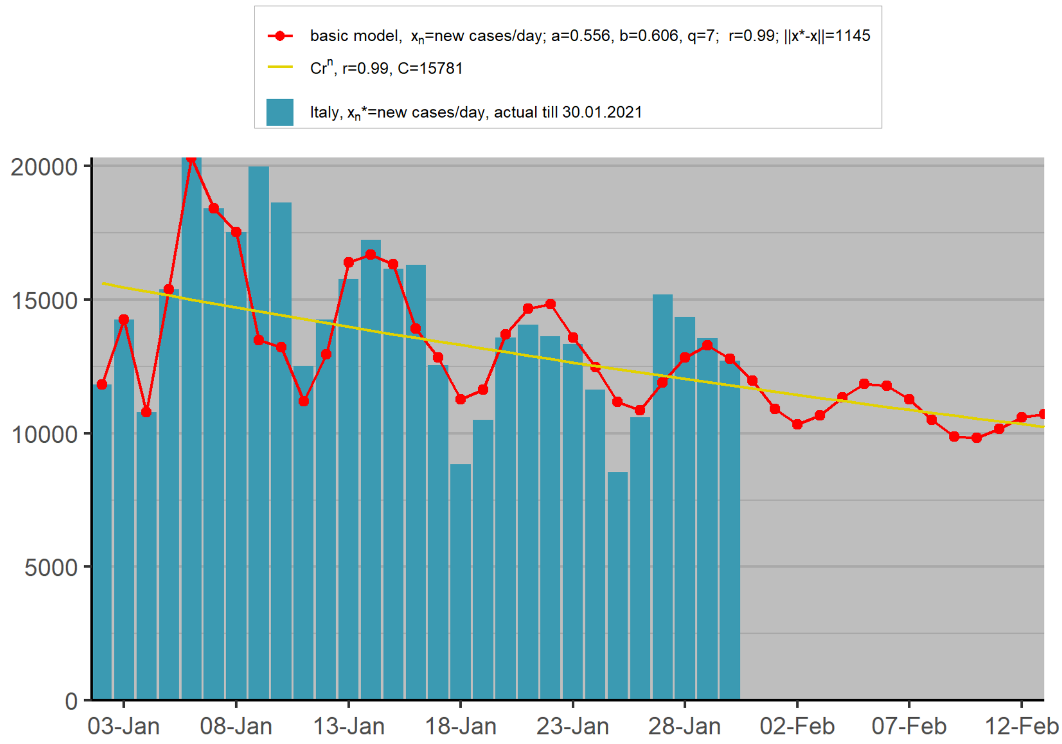


Figure 5: Comparison of $x_n^*,\ x_n$ and the limiting curve Cr^n, Italy

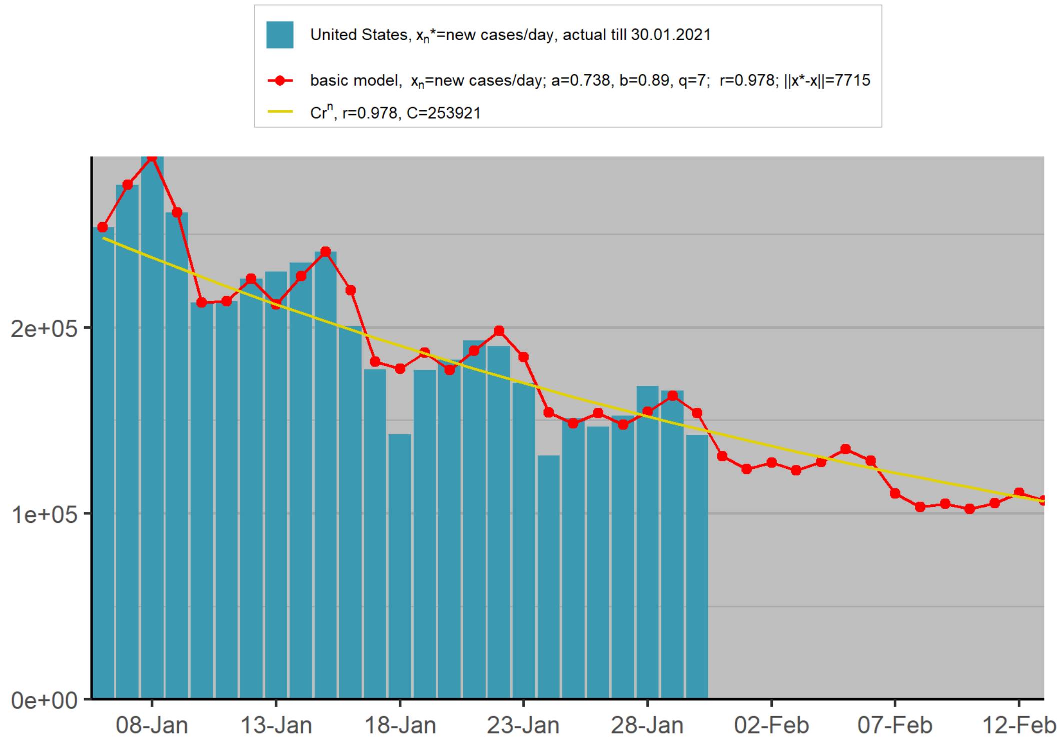


Figure 6: Comparison of $x_n^*,\ x_n$ and the limiting curve Cr^n, United States

2.3 Moving average

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

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

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

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

And then

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

is the 7-day moving average of cases.

Good baseline for model performance

(want $||x - x^*|| \approx ||x^*(3) - x^*||$ or better)

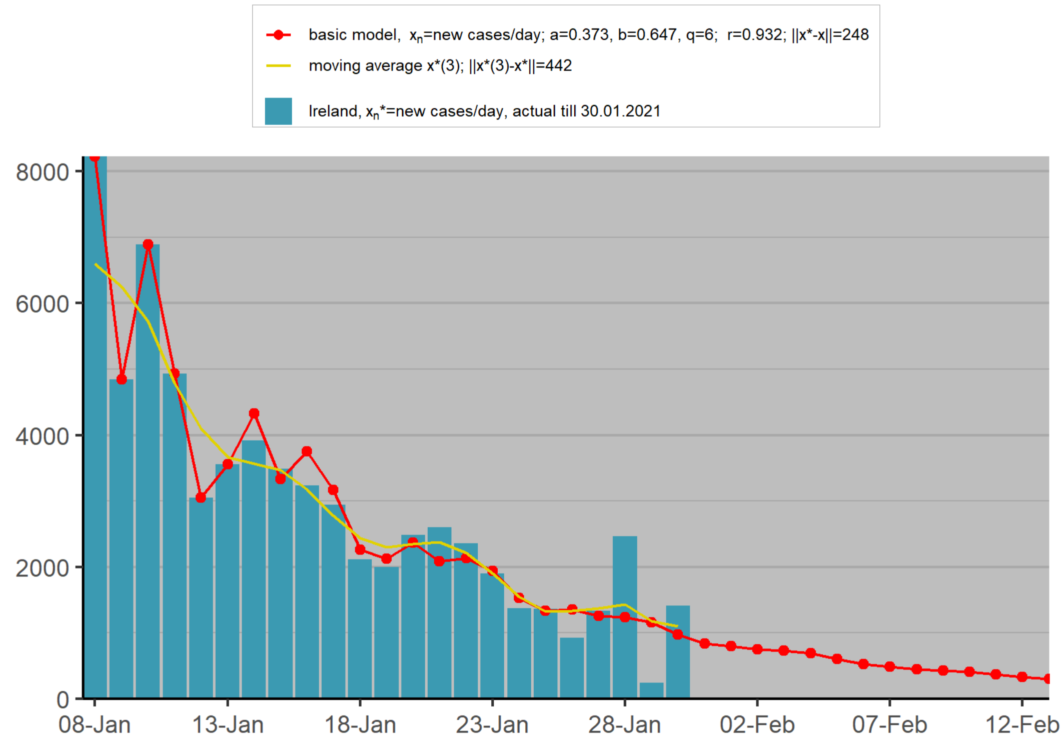


Figure 7: Moving average $x_n^*(3)$, Ireland

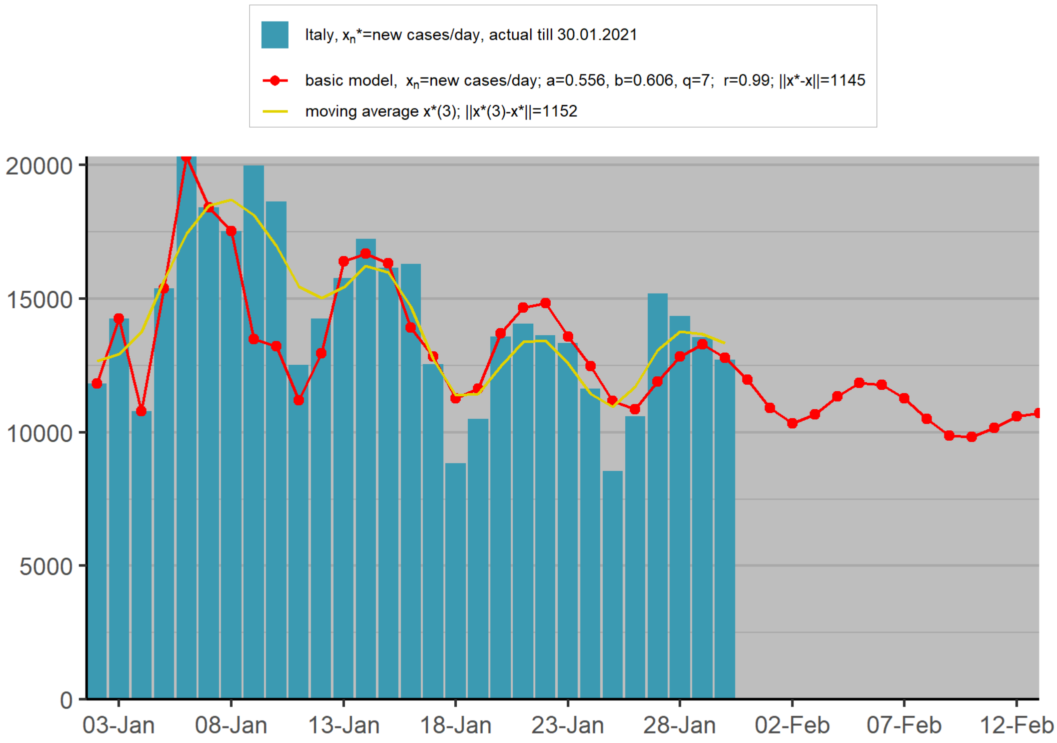


Figure 8: Moving average $x_n^*(3)$, Italy

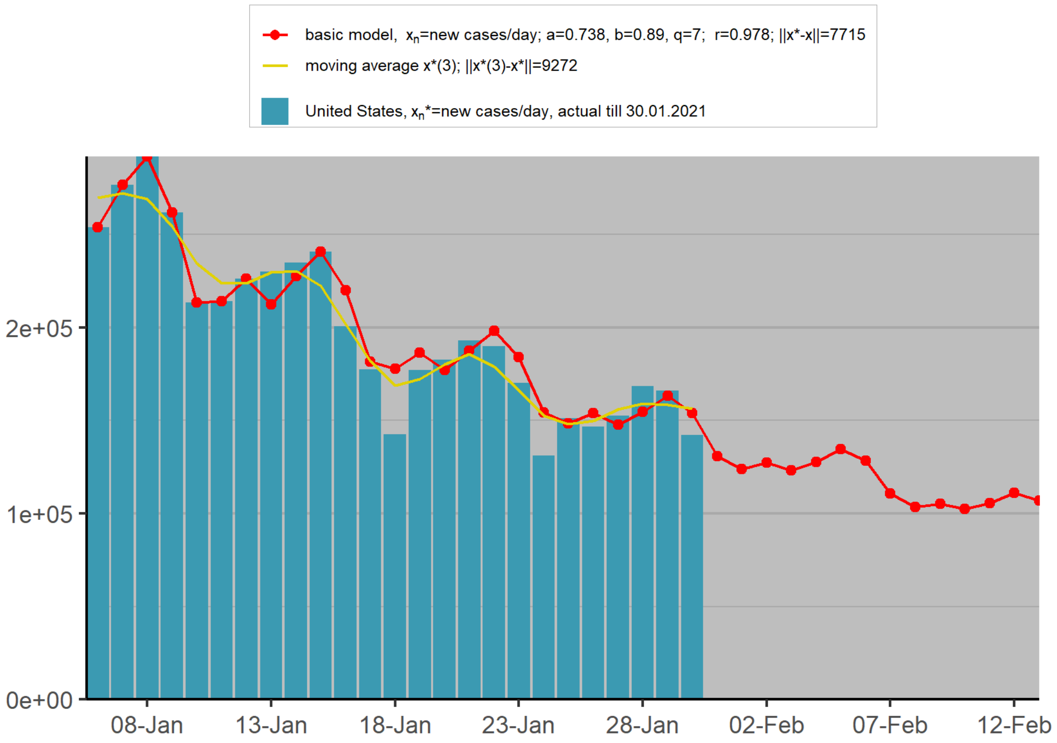


Figure 9: Moving average $x_n^*(3)$, United States

2.4 Periodic model

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

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

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

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

$c_i \in [0.04, 0.2]$ small

$n_i \in 1, 2, \ldots, q$

$p_i \in 1, 2, \ldots, q$

2.4.1 Definitions and Theory

2.4.2 How to select the best model

2.4.3 Forecasting

2.4.4 Implementation in R

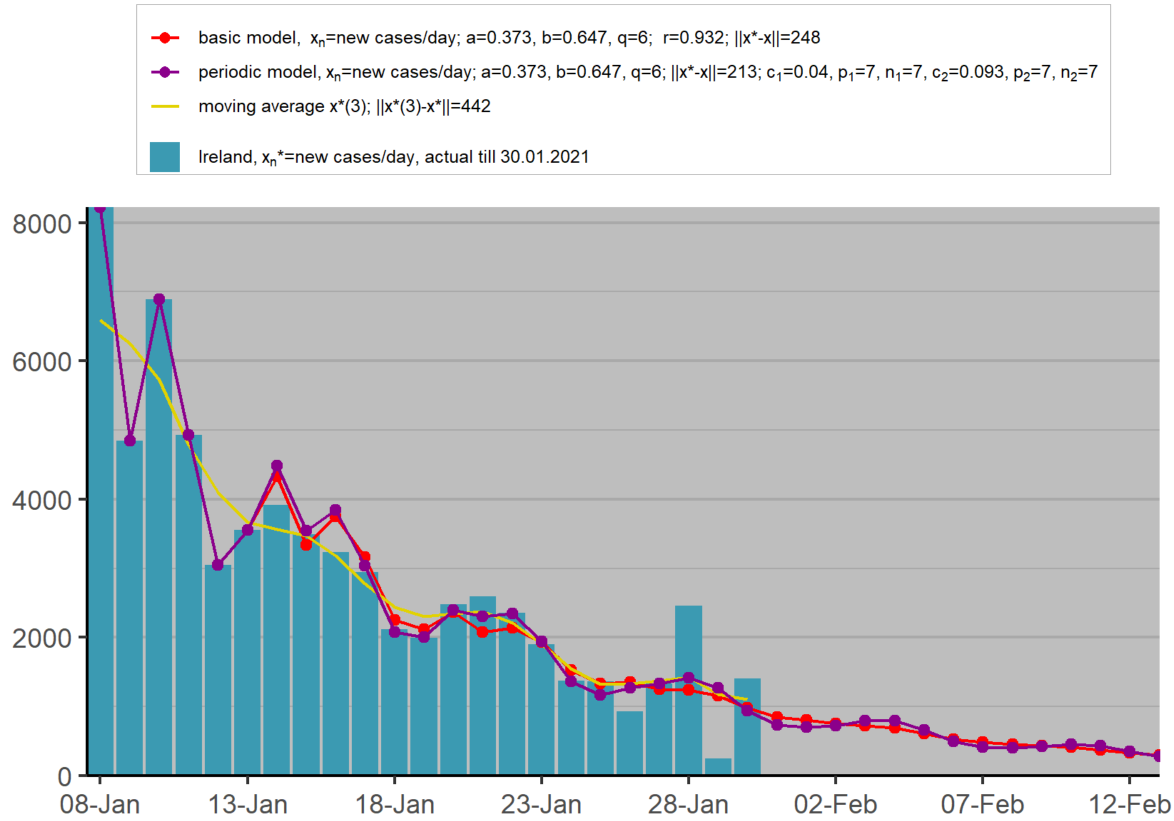


Figure 10: Periodic Model, Ireland

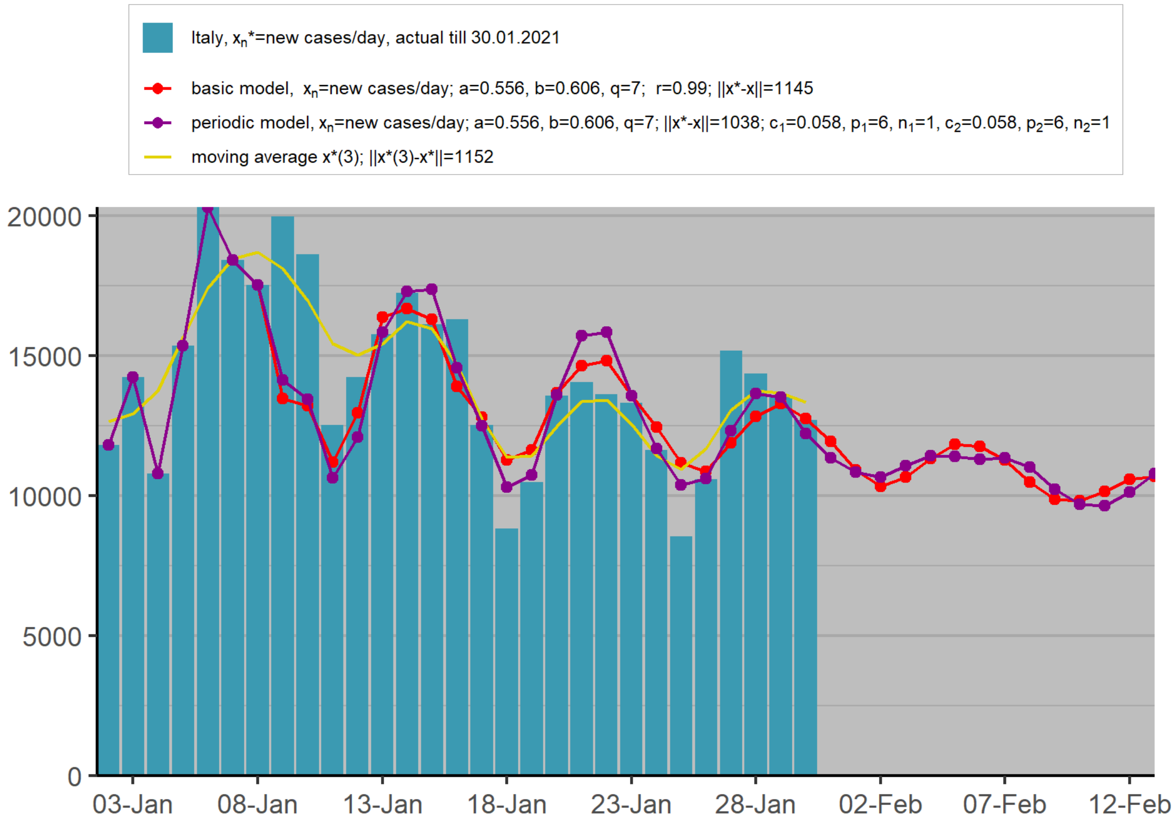


Figure 11: Periodic Model, Italy

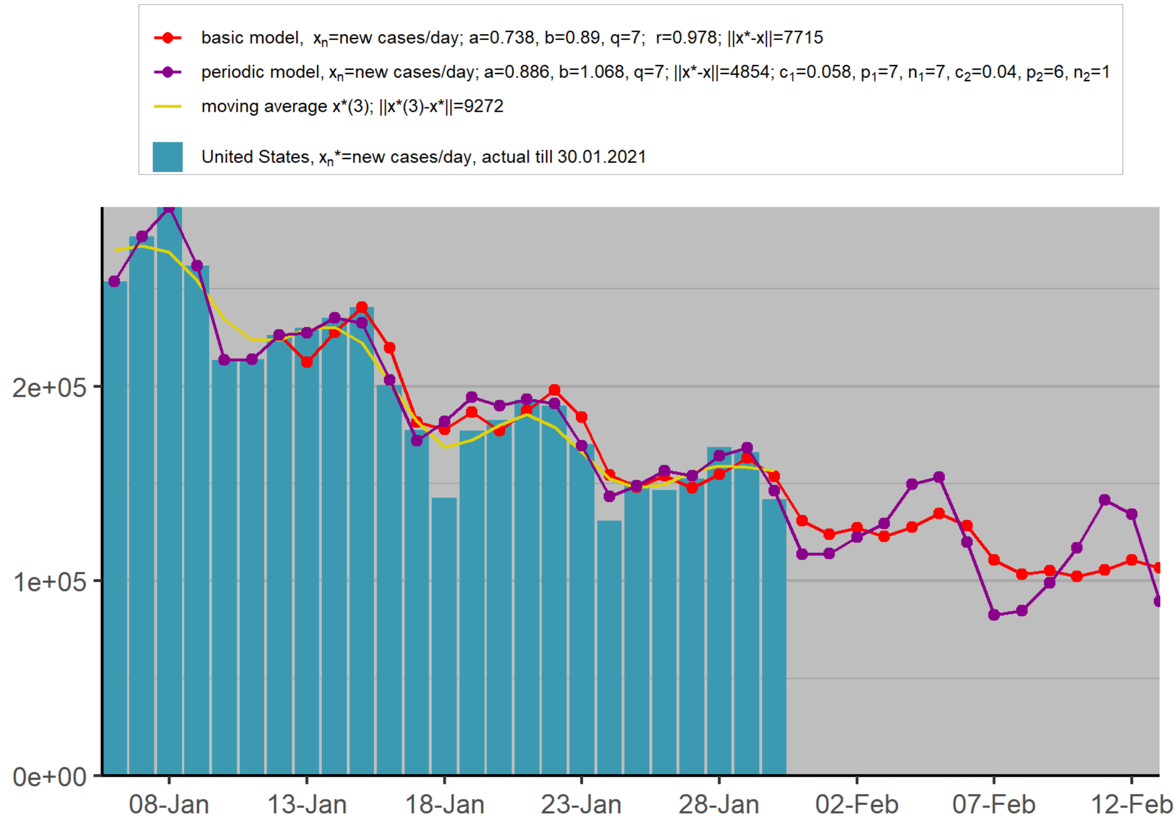


Figure 12: Periodic Model, Ireland

2.5 Multi-phase model

2.5.1 Definitions and Theory

2.5.2 How to select the best model

2.5.3 Forecasting

2.5.4 Implementation in R

The multi-phase model without periodicity can be sharp and unrealstic

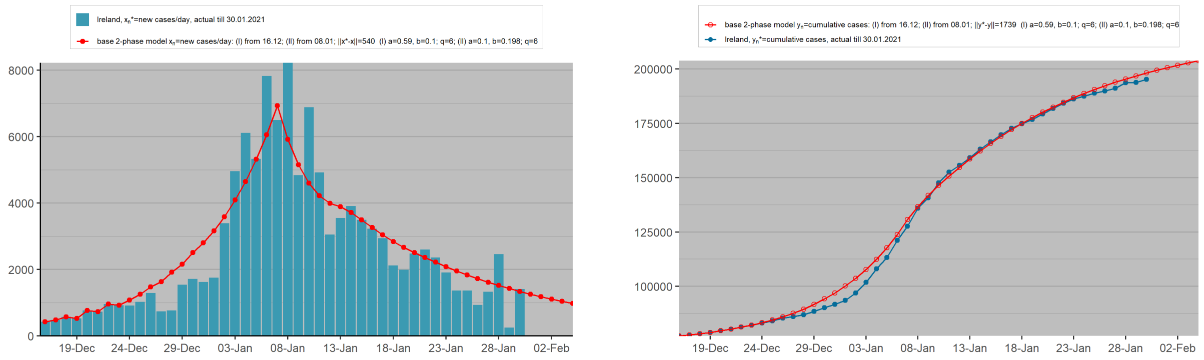


Figure 13: Multi-phase model, Ireland

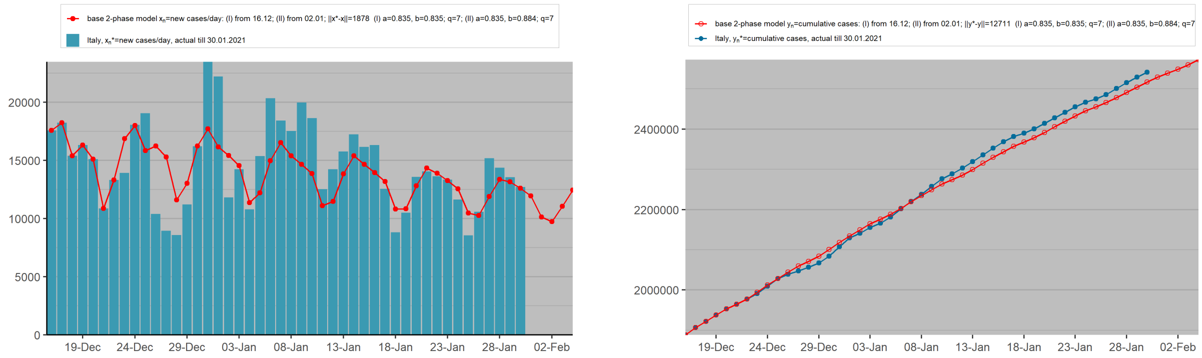


Figure 14: Multi-phase model, Italy

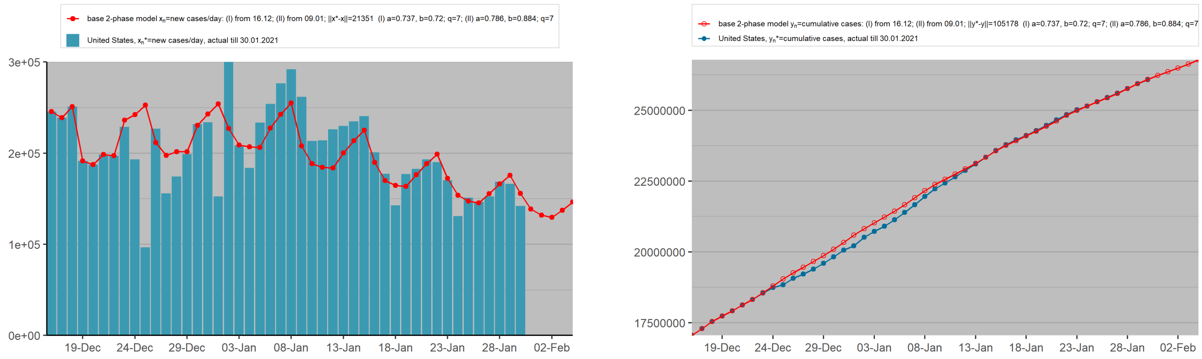


Figure 15: Multi-phase model, United States

The periodic model often performs much better

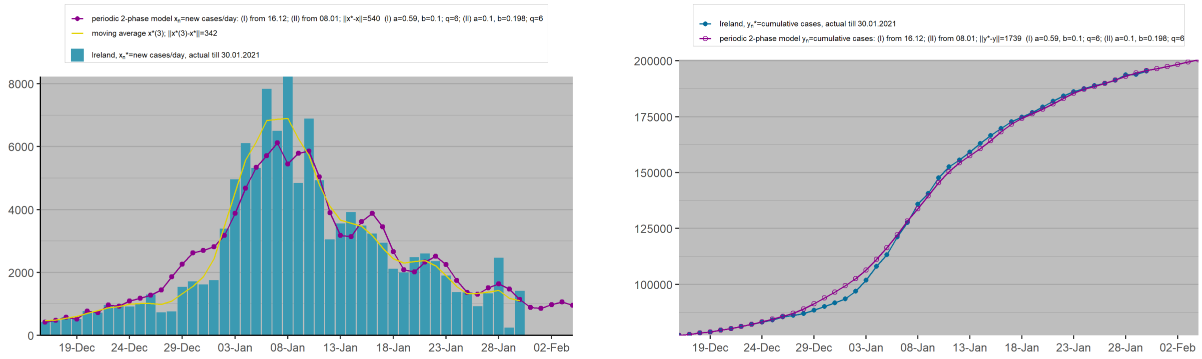


Figure 16: Multi-phase periodic model, Ireland

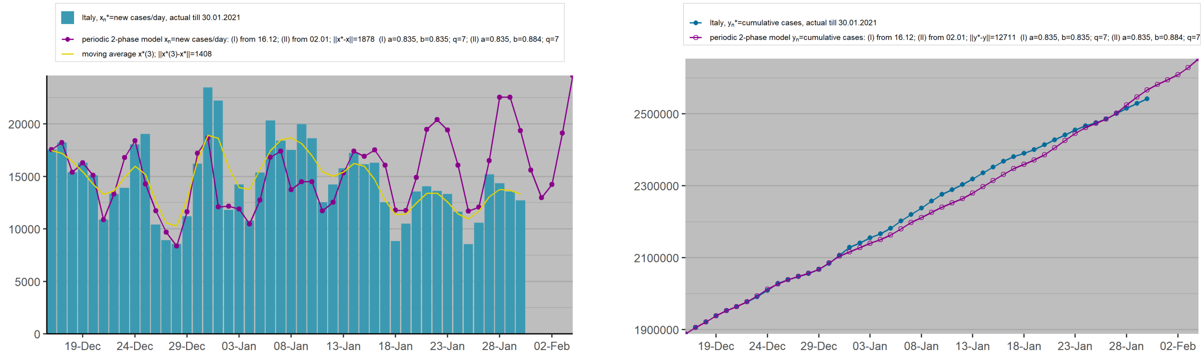


Figure 17: Multi-phase periodic model, Italy

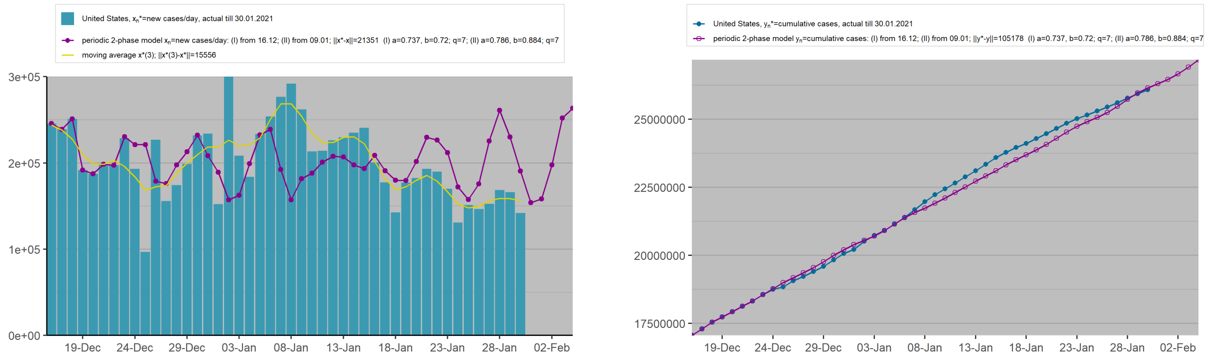


Figure 18: Multi-phase periodic model, United States

3 Theorems

3.1 Model Assumptions

(I) Any infected person becomes ill (symptomatic) and infectious on the q-th day after infection.[1]

(A) During each day, each ill person unconfined infects on average a other persons.

(B) During each day, a fraction b of ill people loose gets isolated (hospitalized or otherwise) and withdrawn from a further spread of the epidemic.

Many models use a set of differential equations for to describe the movement of people between *groups* or *compartments*[19, 4, 11]. The SIR (Susceptible–Infectious–Recovered) model, the most frequently used model in epidemiology, uses a set of 3 such differential equations [3, 5].

Our main mathematical model (and even some of the statistical models) make use recurrence equations, which have some correspondence to differential equations [1].

3.2 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:

$$r \approx R_0^{\frac{1}{2q}}. \tag{1}$$

Recurrence relation for z_n:

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

Using $x_n = bz_n$ we obtain the following equation for x_n:

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

We let the model equal the actual data for the first $q+1$ days

$$x_n = x_n^* \text{ for } n = 0, 1, \ldots, q, \tag{4}$$

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

$$||x - x^*|| := \frac{1}{N+1} \sum_{n=0}^{N} |x_n - x_n^*|, \tag{5}$$

Similarly we define $||y - y^*||$
In order to determine values a, b, q, we want to minimize both

$$||x - x^*|| \text{ and } ||y - y^*|| \tag{6}$$

3.2.1 Why minimize both diatances?

Do 3 pairs of plots: - xn/yn for just x-norm - xn/yn for just y-norm - xn/yn for both x-norm and y-norm

3.2.2 Recurrence equation

This is our general linear recurrence equation with constant coefficients:

$$x_{n+1} = a_0x_n + a_1x_{n-1} + a_2x_{n-2} + \cdots + a_qx_{n-q} \tag{7}$$

The characteristic polynomial of 7

$$f(\lambda) = \lambda^{q+1} - a_0\lambda^q - a_1\lambda^{q-1} - a_2\lambda^{q-2} - \cdots - a_{q-1}\lambda - a_q. \tag{8}$$

Definition 1. A root λ of f with the maximal absolute value $|\lambda|$ will be referred to as a leading root of the general linear recurrence relation 7.

[1] 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.

3.3 Theorems

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

(a) (Cauchy, 1829) The polynomial $f(\lambda)$ from 8 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 7.

(b) For any positive solution x_n of 7, there exists $C > 0$ such that

$$x_n \sim Cr^n \text{ as } n \to \infty. \tag{9}$$

It follows from 9 that if $r < 1$ then the epidemic fades away, whereas if $r > 1$ then it spreads unlimited.

Proof:

(a) Although this statement is not new, we give here the proof as it is quite simple and a part of the argument will be used below. 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 λ^{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

$$1 = \underbrace{\frac{a_0}{\lambda} + \frac{a_1}{\lambda^2} + \frac{a_2}{\lambda^3} + \cdots + \frac{a_{q-1}}{\lambda^q} + \frac{a_q}{\lambda^{q+1}}}_{g(\lambda)} \tag{10}$$

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

- $\lim\limits_{\lambda \to 0^+} g(\lambda) = +\infty$
- $\lim\limits_{\lambda \to +\infty} g(\lambda) = 0^+$

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 $\lambda \in \mathbb{C}\backslash\{0\}$ be another root of f. We obtain from 10 (using the triangle inequality) that

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

And so $g(r) \leq g(|\lambda|)$ which implies $|\lambda| \leq r$ by the definition of decreasing functions.

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

$$f'(\lambda) = (q+1)\lambda^q - qa_0\lambda^{q-1} - (q-1)a_1\lambda^{q-2} - (q-2)a_2\lambda^{q-3} - \cdots - a_{q-1} - 0$$

$$\frac{1}{q+1}f'(\lambda) = \lambda^q - \frac{q}{q+1}a_0\lambda^{q-1} - \frac{q-1}{q+1}a_1\lambda^{q-2} - \frac{q-2}{q+1}a_2\lambda^{q-3} - \cdots - \frac{1}{q+1}a_{q-1}$$

$$\frac{1}{q+1}f'(r') = (r')^q - \frac{q}{q+1}a_0(r')^{q-1} - \frac{q-1}{q+1}a_1(r')^{q-2} - \frac{q-2}{q+1}a_2(r')^{q-3} - \cdots - \frac{1}{q+1}a_{q-1}$$

$$0 = (r')^q - \frac{q}{q+1}a_0(r')^{q-1} - \frac{q-1}{q+1}a_1(r')^{q-2} - \frac{q-2}{q+1}a_2(r')^{q-3} - \cdots - \frac{1}{q+1}a_{q-1}$$

$$(r')^q = \frac{q}{q+1}a_0(r')^{q-1} + \frac{q-1}{q+1}a_1(r')^{q-2} + \frac{q-2}{q+1}a_2(r')^{q-3} + \cdots + \frac{1}{q+1}a_{q-1}$$

dividing both sides by $(r')^q > 0$

$$\begin{aligned}
1 &= \frac{qa_0}{(q+1)r'} + \frac{(q-1)a_1}{(q+1)(r')^2} + \cdots + \frac{a_{q-1}}{(q+1)(r')^q} \\
&= \left(\frac{q+1-1}{q+1}\right)\frac{a_0}{r'} + \left(\frac{q+1-2}{q+1}\right)\frac{a_1}{(r')^2} + \cdots + \left(\frac{q+1-q}{q+1}\right)\frac{a_{q-1}}{(r')^q} \\
&= \left(1 - \frac{1}{q+1}\right)\frac{a_0}{r'} + \left(1 - \frac{2}{q+1}\right)\frac{a_1}{(r')^2} + \cdots + \left(1 - \frac{q}{q+1}\right)\frac{a_{q-1}}{(r')^q} \\
&< \frac{a_0}{r'} + \frac{a_1}{(r')^2} + \cdots + \frac{a_{q-1}}{(r')^q}
\end{aligned}$$

So $g(r') > 1$, but $g(r) = 1$
$\implies g(r') > g(r) \implies r' < r$ by the definition of decreasing functions.

(b) Let $\lambda_1, \lambda_2, \dots$ be all other distinct roots of f apart from r (so that λ_k are negative or imaginary). Any solution x_n of 7 has the form

$$x_n + Cr^n + \tilde{x}_n \tag{11}$$

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

$$|\tilde{x}_n| = o(r^n) \text{ as } n \to \infty \tag{12}$$

Since $x_n > 0$, it follows from 11 and 12 that $C \geq 0$. Let us verify that $C > 0$, which will finish the proof. It is tempting to say that if $C = 0$ then $x_n = \tilde{x}_n$ is a linear combination of terms of the form $n^j \rho^n \sin(\phi n)$ and $n^j \rho^n \cos(\phi n)$ and, therefore, cannot stay positive. However, it is not easy to make this argument rigorous because different roots of f may have the same absolute value ρ and an uncontrollable cancellation of the terms can occur. We employ here a different, simpler approach that takes advantage of nonnegative coefficients a_k. To that end, consider a new sequence

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

This satisfies the equation

$$X_{n+1} = A_0 X_0 + A_1 X_{n-1} + \cdots + A_q X_{n-q} \tag{13}$$

with $A_k = \frac{a_k}{r^{k+1}}$. 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 10, and $g(r) = 1$ that

$$A_0 + A_1 + \cdots + A_q = 1 \tag{14}$$

Set $c := \min(X_1, \dots, X_{q+1}) > 0$ since x_n have positive initial values. Then we obtain from 13 and 14 by induction that $X_n \geq c$ for all $n \in \mathbb{N}$, which implies
$x_n \geq cr^n$
as required.

□

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

$$\min\left(1, \left(\frac{a}{b}\right)^{1/m}\right) \leq r \leq \max\left(1, \left(\frac{a}{b}\right)^{1/m}\right) \tag{15}$$

Remark 1. Although there are in the literature plenty of estimates of the leading roots of polynomial (see, for example, [2]), none of them seems to imply 15. The latter is very useful for a basic model as we will see below in an example.

Proof:

(a) We have
$f(1) = 1 - a_0 - a1 - \cdots - a_q$
$= \underbrace{(1 - a_0)}_{b} - \underbrace{(a_1 + \cdots + a_q)}_{a}$
$= b - a$
We know f is increasing.
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 $a_1, \ldots, a_{m-1}$ are all zero
$\implies r^{q+1} - a_0 r^q - a_m r^{q-m} - a_{m+1} r^{q-m-1} - \cdots - a_{q-1} r - a_q = 0$
$\implies r^{q+1} - (1-b) r^q - a_m r^{q-m} - \cdots - a_q = 0$
$\implies r^{q+1} - r^q + b r^q - a_1 r^{q-m} - \cdots - a_q = 0$
$\implies r^{q+1} - r^q = -b r^q + a_m r^{q-m} + \cdots + a_q$
If $r > 1$ then $r^{q+1} > r^q$ and so $r^{q+1} - r^q > 0$
and so
$0 < -b r^q + a_m r^{q-m} + \cdots + a_q$

$$\begin{aligned} \implies b r^q &< a_m r^{q-m} + \cdots + a_q \\ &\leq a_m r^{q-m} + \cdots + a_q r^{q-m} \\ &= (a_m + \cdots + a_q) r^{q-m} \\ &= a r^{q-m} \end{aligned}$$

So $b r^q < a r^{q-m} \iff r^m = \frac{a}{b} \iff r < \left(\frac{a}{b}\right)^{1/m}$
And if $r < 1$ we get $r < \left(\frac{a}{b}\right)^{1/m}$.
We can combine both cases with $a \leq \max(()1, a)$ and $a \geq \min(1, a)$ to get 15, as required.

□

Lemma 3. For the model described by equation 7 we have

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

Proof: Let u be the number of people infected on some day, say 0. On the day $k = 1, \ldots, q$ 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

$$a c_k u + a(1-b) c_k u + a(1-b)^2 c_k u + \cdots = a c_k u \sum_{n=0}^{\infty} (1-b)^n = \frac{a c_k u}{1-(1-b)} = \frac{a}{b} c_k u$$

since $0 < 1 - b < 1$.

other people.

Hence, the initial group of u people infects in total

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

So we know R_0 is the unit reprodiction number per infected person ($u = 1$).
And so we get the result $R_0 = \frac{a}{b}$ as required. □

4 Statistical Models

Primary source for this was Hyndman-et-al-2018 [18].

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

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

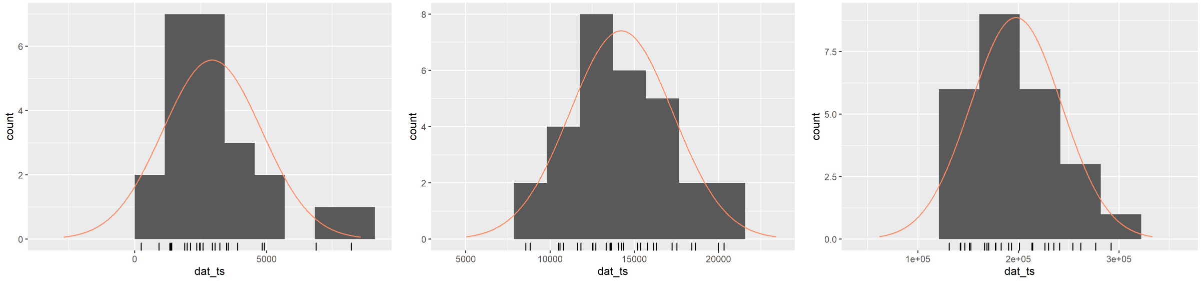


Figure 19: Normality checks, Ireland, Italy and United States

4.1 Holt-Winters' seasonal method

4.1.1 Definitions and Theory

Suppose there are N observations.
Initial step:

$$\left|\begin{array}{l} L_s = \frac{1}{s}\sum_{i=1}^{s} x_i \\ b_s = \frac{1}{s}\left[\frac{x_{s+1}-x_1}{s} + \frac{x_{s+2}-x_2}{s} + \cdots + \frac{x_{2s}-x_s}{s}\right] \\ S_n = x_n - L_s,\ n = 1, \ldots, s \end{array}\right.$$

and choose parameters $0 \leq \alpha \leq 1,\ 0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$
Then compute for $s < n \leq N$:

$$\left|\begin{array}{lll} \text{Level} & L_n & = \alpha(x_n - S_{n-s}) + (1-\alpha)(L_{n-1} + b_{n-1}) \\ \text{Trend} & b_n & = \beta(L_n - L_{n-1}) + (1-\beta)b_{n-1} \\ \text{Seasonal} & S_n & = \gamma(x_n - L_n) + (1-\gamma)S_{n-s} \\ \text{Forecast} & F_{n+1} & = L_n + b_n + S_{n+1-s} \end{array}\right.$$

For subsequent observations,
$F_{N+k} = L_N + k \cdot b_N + S_{N+k-s}$

Figure 20: Seasonal Holt Winter's Additive Model Algorithm (denoted SHW$_+$)

4.1.2 How to select the best model

4.1.3 Forecasting

4.1.4 implementation in R

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

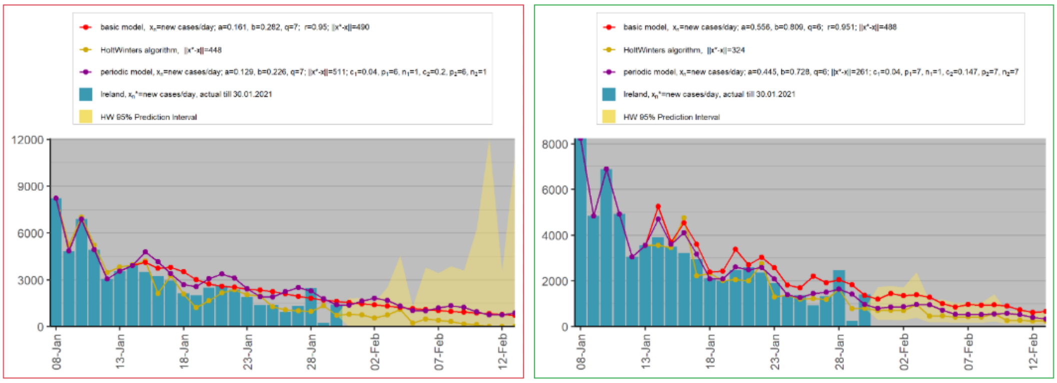


Figure 21: Comparison between HoltWinters multiplicative (left, red outline) and additive (right, green outline)algorithms, at some point during the research

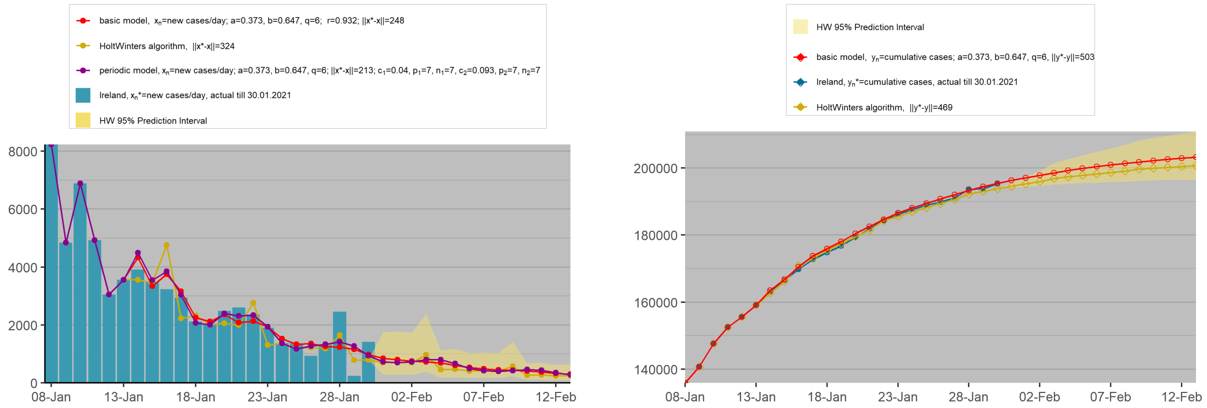


Figure 22: HoldWinters model, Ireland

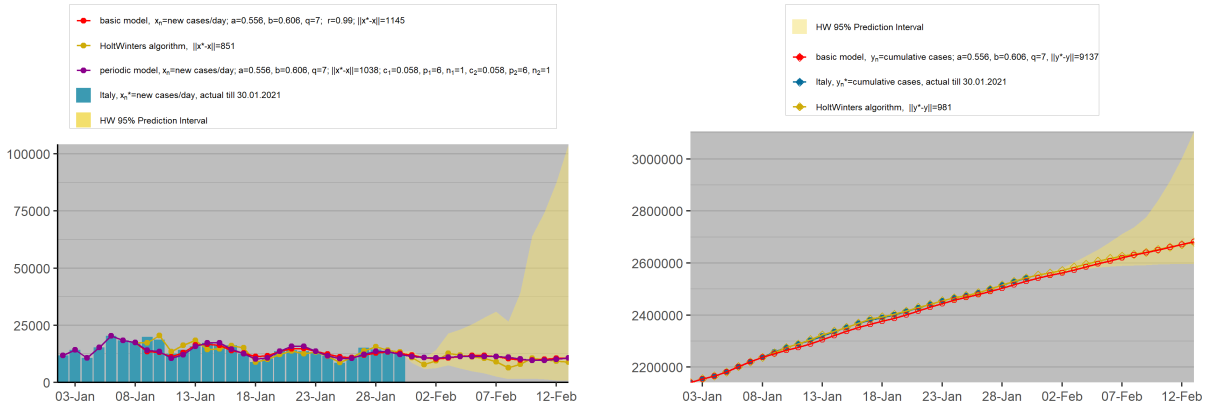


Figure 23: HoldWinters model, Italy

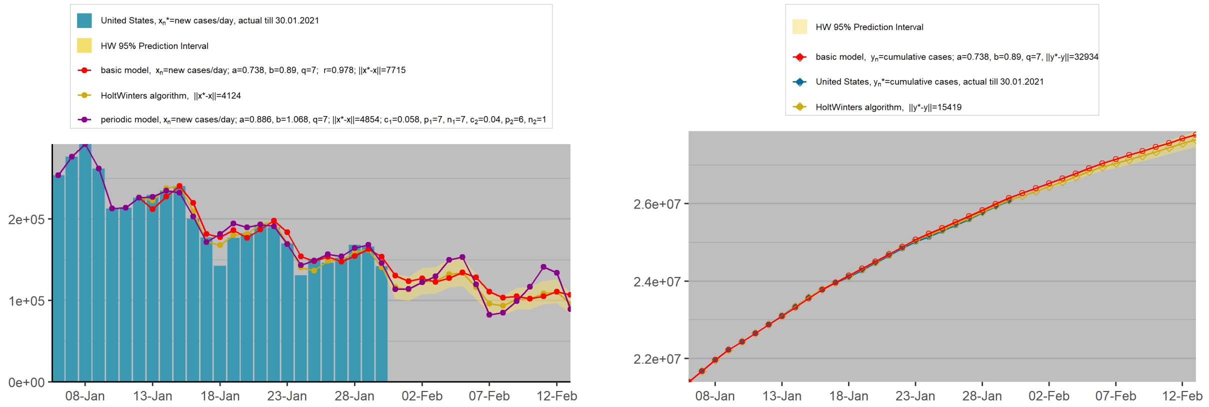


Figure 24: HoldWinters model, United States

4.2 ARIMA models

4.2.1 Definitions and Theory

Definition 2. 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 genrerally:

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

And similarly for the independent errors ε_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 $x_1, \ldots, x_N$ states that each x_n is a *linear function* of $x_{n-p}, x_{n-p+1}, \ldots, x_{n-1}$ and an error term, i.e.

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

  We can simplify using the backshift operator B:

  $$\begin{aligned} x_n &= \phi_0 + \phi_1 B x_n + \phi_2 B^2 x_n + \cdots + \phi_p B^p x_n + \varepsilon_n \\ &= \phi_0 + \left(\phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p\right) x_n + \varepsilon_n \end{aligned} \tag{16}$$

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

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

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

  $$\begin{aligned} x_n &= \psi_0 - \psi_1 B \varepsilon_n - \psi_2 B^2 \varepsilon_n - \cdots - \psi_q B^q \varepsilon_n + \varepsilon_n \\ &= \psi_0 + \left(1 - \psi_1 B - \psi_2 B^2 + \cdots - \psi_q B^q\right) \varepsilon_n \end{aligned} \tag{17}$$

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

  $$\begin{aligned} x'_n &= x_n - x_{n-1} \\ &= x_n - Bx_n \\ &= (1 - B)\, x_n \end{aligned} \tag{18}$$

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 ε_n are unchanged.

Therefore, an ARIMA(p,d,q) model can be evaluated by combining the $AR(p)$, $I(d)$ and $MA(q)$

$$(1-B)^d x_n = \phi_0 + (1-B)^d \left(\phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p\right) x_n + \psi_0 + \left(\psi_1 B + \psi_2 B^2 + \cdots + \psi_q B^q\right) \varepsilon_n$$

$$\begin{aligned}(1-B)^d x_n + (1-B)^d \left(-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) x_n &= \phi_0 + \psi_0 + \left(1 - \psi_1 B - \psi_2 B^2 + \cdots - \psi_q B^q\right) \varepsilon_n \\ (1-B)^d \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) x_n &= c + \left(1 - \psi_1 B - \psi_2 B^2 + \cdots - \psi_q B^q\right) \varepsilon_n \end{aligned} \tag{19}$$

where $c = \phi_0 + \psi_0$ (it is zero if $d \geq 1$).
We also need the seasonal components for an ARIMA$(p,d,q)(P,D,Q)_s$
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 $x_1, \ldots, x_N$ states that each x_n is a *linear function* of $x_{n-Ps}, x_{n-(P-1)s}, \ldots, x_{n-s}$ 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:

  $$x_n = \beta_0 + \left(\beta_1 B^s + \beta_2 B^{2s} + \cdots + \beta_P B^{Ps}\right) x_n \tag{20}$$

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

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

  Again, by convention we use minus signs in the coefficients $\gamma_1, \ldots, \gamma_Q$
  We can simplify using the backshift operator B:

  $$\begin{aligned} x_n &= \gamma_0 - \gamma_1 \varepsilon_{n-s} - \gamma_2 \varepsilon_{n-2s} - \cdots - \gamma_Q \varepsilon_{n-Qs} + \varepsilon_n \\ &= \gamma_0 + \left(1 - \gamma_1 B^s - \gamma_2 B^{2s} + \cdots - \gamma_Q B^{Qs}\right) \varepsilon_n \end{aligned} \tag{21}$$

- The first order seasonal differencing of the time series, $I_s(1)$, is evalueated 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 the previous ARIMA(pd,q) to get the definition of an ARIMA$(p,d,q)(P,D,Q)_s$ model

$$\begin{aligned} &\underbrace{\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)}_{AR(p)} \underbrace{\left(1 - \beta_1 B^s - \beta_2 B^{2s} - \cdots - \beta_P B^{Ps}\right)}_{AR_s(P)} \underbrace{(1-B)^d}_{I(d)} \underbrace{(1-B^s)^D}_{I_s(D)} x_n = \\ &\qquad c + \underbrace{\left(1 - \psi_1 B - \psi_2 B^2 - \cdots - \psi_q B^q\right)}_{MA(q)} \underbrace{\left(1 - \gamma_1 B^s - \gamma_2 B^{2s} - \cdots - \gamma_Q B^{Qs}\right)}_{MA_s(Q)} \varepsilon_n \end{aligned} \tag{22}$$

where the constant c is some function of the constants ϕ_0, ψ_0, β_0 and γ_0

4.2.2 How to select the best model

4.2.3 Forecasting

4.2.4 implementation in R

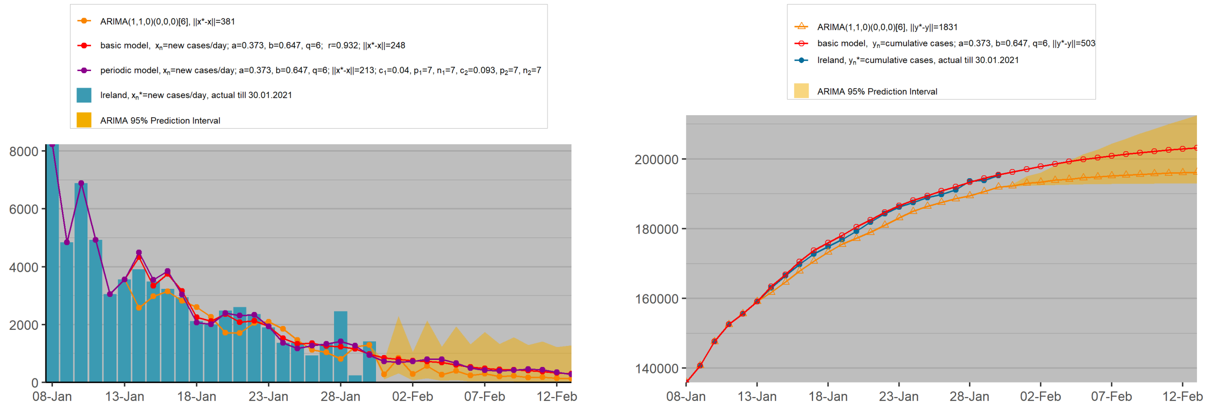


Figure 25: ARIMA model, Ireland

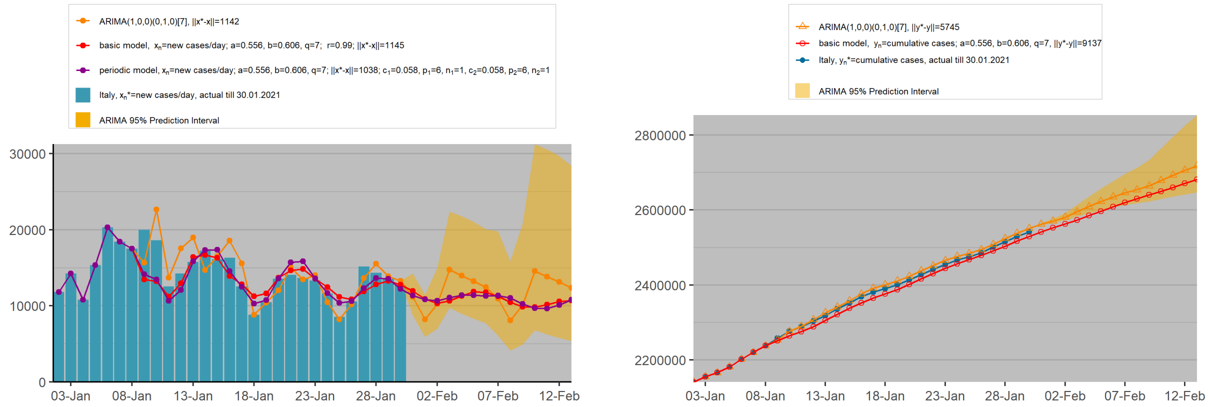


Figure 26: ARIMA model, Italy

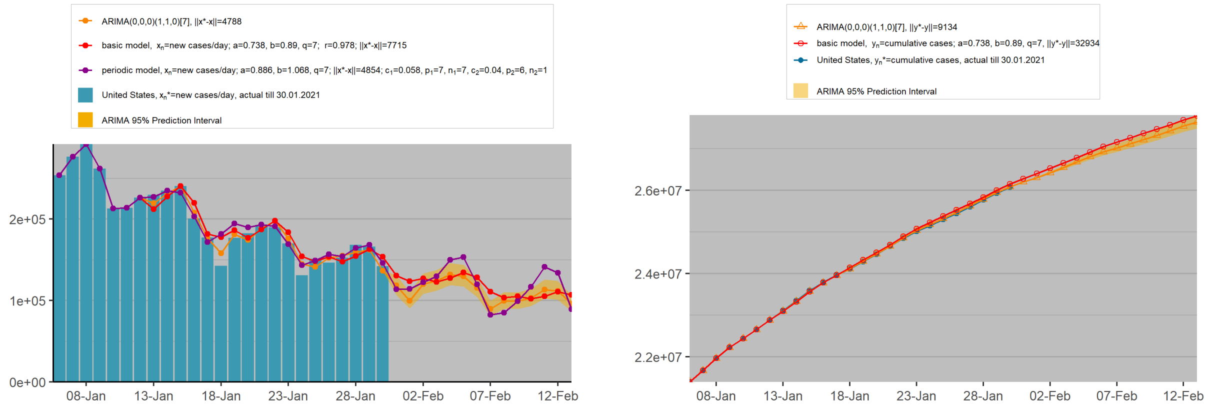


Figure 27: ARIMA model, United States

4.3 Neural network models

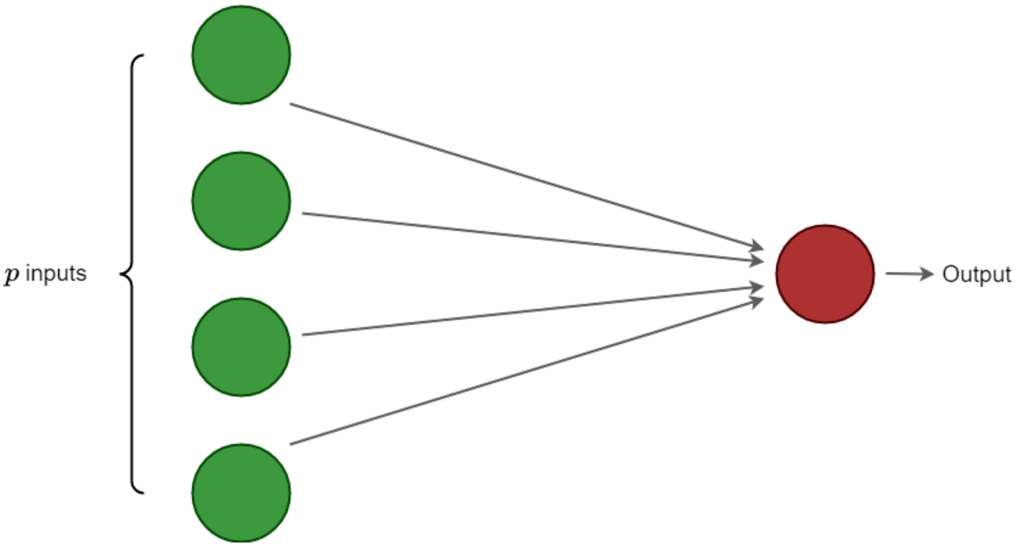


Figure 28: A linear regression model, or ARIMA$(p, 0, 0)$ model.

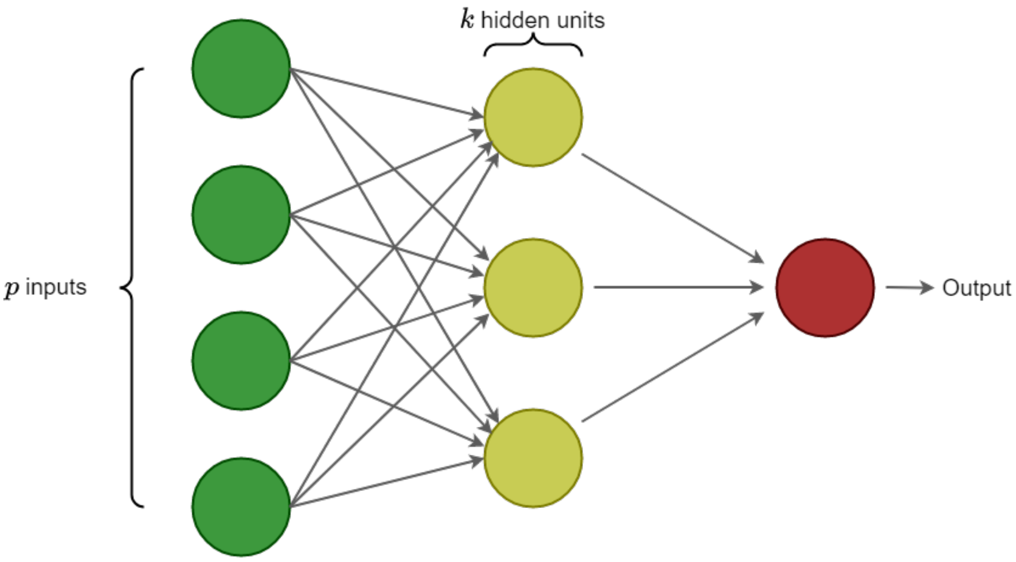


Figure 29: A neural network with p inputs and one hidden layer with k hidden neurons.

4.3.1 Definitions and Theory

4.3.2 How to select the best model

4.3.3 Forecasting

4.3.4 implementation in R

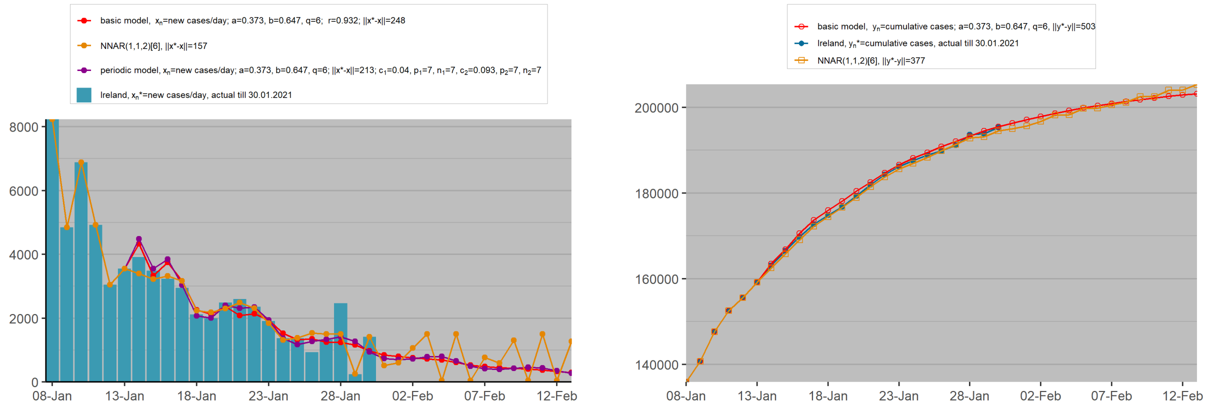


Figure 30: Neural Network model, Ireland

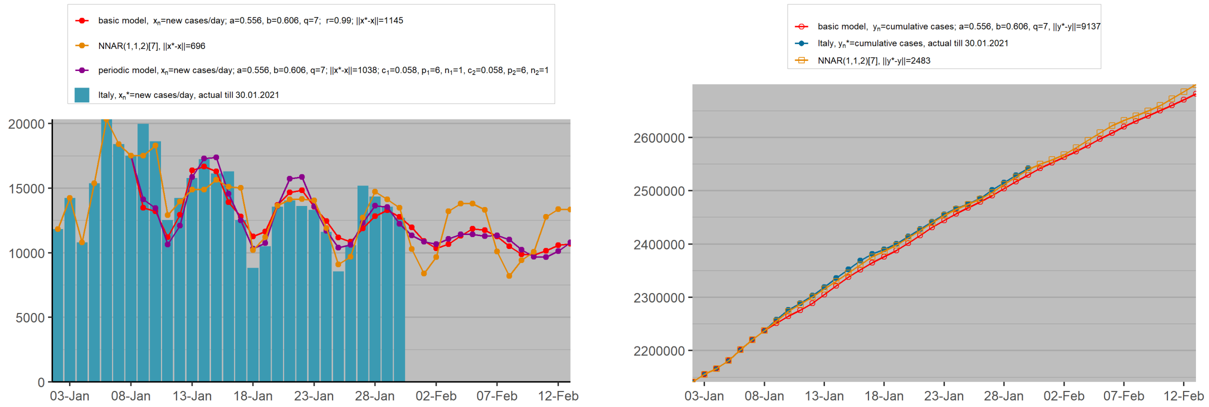


Figure 31: Neural Network model, Italy

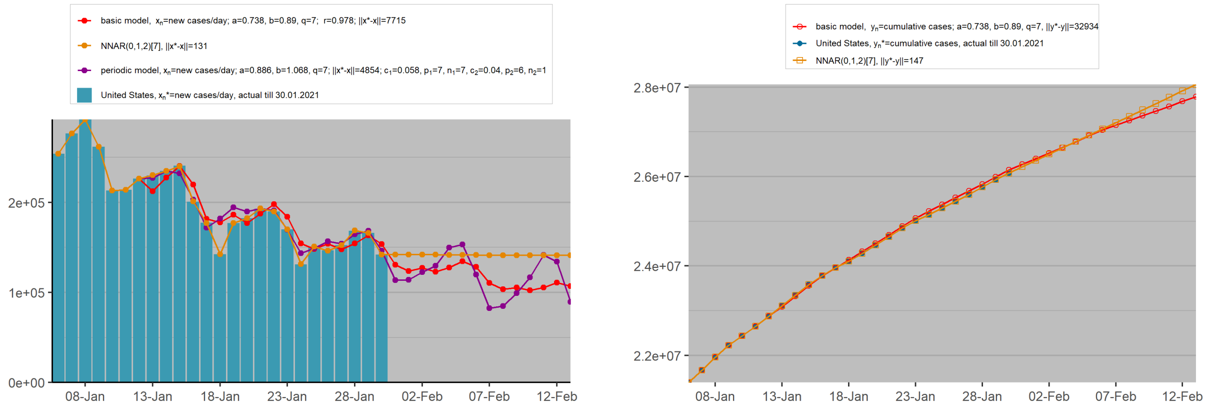


Figure 32: Neural Network model, United States

5 R 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 [15].

5.1 R packages

`ggplot2` [22] is widely used for easily plotting and visualising the models. `rgdal` [6] allows geospatial `.shp` files to be read into R. `raster` [16] allows this data to be manipulated and plotted. `dplyr` [23] provides useful data manipulation functions, both for models and geospatial mapping. Statistical models (HoltWinters, ARIMA and Neural Network Regression) were readily implemented from `forecast` [**forecasting**].

5.2 Plotting and colour

wesanderson [20]

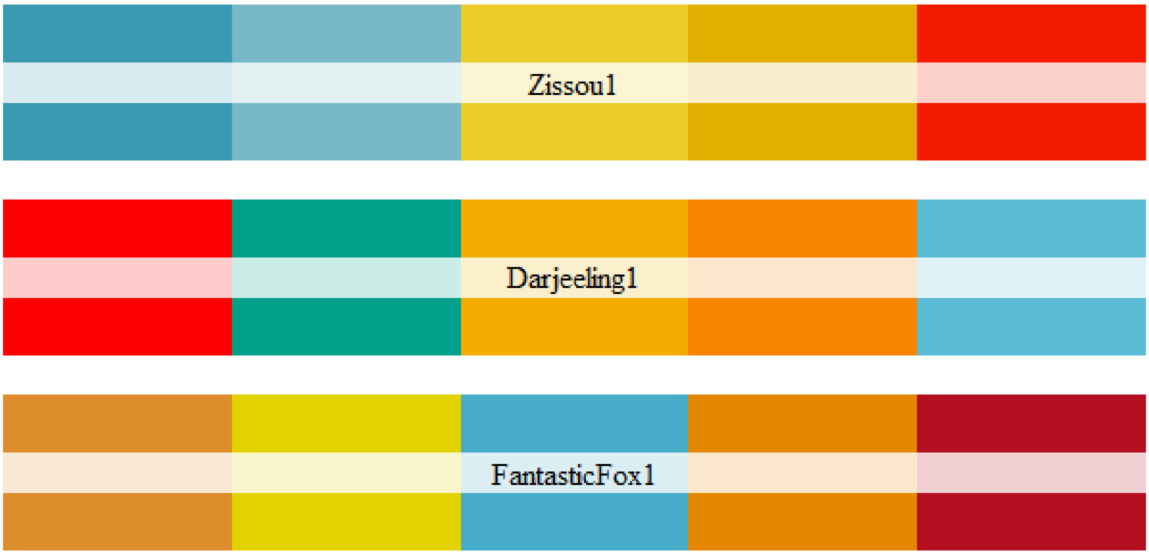


Figure 33: Wes Anderson Palettes

5.3 Shapefiles

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

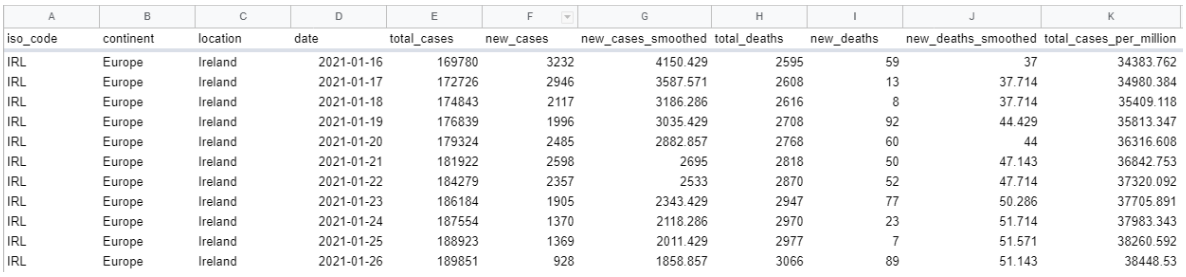
World country shape data was obtained from [17], while the more detailed county-level shapefile was downloaded from [8].

5.4 Datasets

Country-based data:

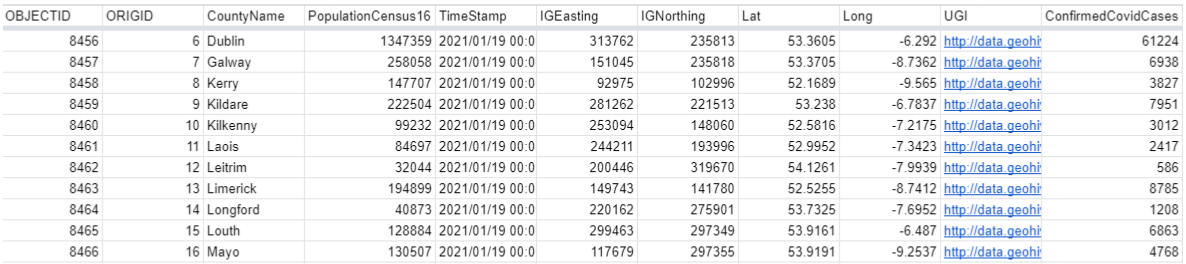
Originally used data from [10], but the ECDC switched from a daily to a weekly update from 14 December 2020. Therefore, I have chosen to use the data from [9], which has remained daily

Ireland cases by county Downloaded from [7].



A	B	C	D	E	F	G	H	I	J	K
iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million
IRL	Europe	Ireland	2021-01-16	169780	3232	4150.429	2595	59	37	34383.762
IRL	Europe	Ireland	2021-01-17	172726	2946	3587.571	2608	13	37.714	34980.384
IRL	Europe	Ireland	2021-01-18	174843	2117	3186.286	2616	8	37.714	35409.118
IRL	Europe	Ireland	2021-01-19	176839	1996	3035.429	2708	92	44.429	35813.347
IRL	Europe	Ireland	2021-01-20	179324	2485	2882.857	2768	60	44	36316.608
IRL	Europe	Ireland	2021-01-21	181922	2598	2695	2818	50	47.143	36842.753
IRL	Europe	Ireland	2021-01-22	184279	2357	2533	2870	52	47.714	37320.092
IRL	Europe	Ireland	2021-01-23	186184	1905	2343.429	2947	77	50.286	37705.891
IRL	Europe	Ireland	2021-01-24	187554	1370	2118.286	2970	23	51.714	37983.343
IRL	Europe	Ireland	2021-01-25	188923	1369	2011.429	2977	7	51.571	38260.592
IRL	Europe	Ireland	2021-01-26	189851	928	1858.857	3066	89	51.143	38448.53

Figure 34: OWID World data extract



OBJECTID	ORIGID	CountyName	PopulationCensus16	TimeStamp	IGEasting	IGNorthing	Lat	Long	UGI	ConfirmedCovidCases
8456	6	Dublin	1347359	2021/01/19 00:0	313762	235813	53.3605	-6.292	http://data.geohi	61224
8457	7	Galway	258058	2021/01/19 00:0	151045	235818	53.3705	-8.7362	http://data.geohi	6938
8458	8	Kerry	147707	2021/01/19 00:0	92975	102996	52.1689	-9.565	http://data.geohi	3827
8459	9	Kildare	222504	2021/01/19 00:0	281262	221513	53.238	-6.7837	http://data.geohi	7951
8460	10	Kilkenny	99232	2021/01/19 00:0	253094	148060	52.5816	-7.2175	http://data.geohi	3012
8461	11	Laois	84697	2021/01/19 00:0	244211	193996	52.9952	-7.3423	http://data.geohi	2417
8462	12	Leitrim	32044	2021/01/19 00:0	200446	319670	54.1261	-7.9939	http://data.geohi	586
8463	13	Limerick	194899	2021/01/19 00:0	149743	141780	52.5255	-8.7412	http://data.geohi	8785
8464	14	Longford	40873	2021/01/19 00:0	220162	275901	53.7325	-7.6952	http://data.geohi	1208
8465	15	Louth	128884	2021/01/19 00:0	299463	297349	53.9161	-6.487	http://data.geohi	6863
8466	16	Mayo	130507	2021/01/19 00:0	117679	297355	53.9191	-9.2537	http://data.geohi	4768

Figure 35: ArcGIS Ireland data extract

References

[1] 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. URL: http://www.sciencedirect.com/science/article/pii/S0377042701004320.

[2] 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.

[3] Roy M. Anderson. "Discussion: The Kermack-McKendrick epidemic threshold theorem". In: *Bulletin of Mathematical Biology* 53 (Mar. 1, 1991). ISSN: 522-9602. DOI: 10.1007/BF02464422. URL: https://doi.org/10.1007/BF02464422.

[4] Derdei Bichara, Abderrahman Iggidr, and Gauthier Sallet. "Global analysis of multi-strains SIS, SIR and MSIR epidemic models". In: *Journal of Applied Mathematics and Computing* 44 (Feb. 2014), pp. 273–292. DOI: 10.1007/s12190-013-0693-x.

[5] Alexander Bird. *A simple introduction to epidemiological modelling—the SIR model*. https://philosophyandmedicine.org/wp-content/uploads/2020/04/Introduction-to-epidemiological-modelling.pdf. 2020.

[6] 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.

[7] 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.

[8] © OpenStreetMap contributors. *Townlands*. 2020. URL: https://www.townlands.ie/page/download/.

[9] Our World in Data. *The complete Our World in Data COVID-19 dataset*. 2020. URL: https://covid.ourworldindata.org/data/owid-covid-data.csv.

[10] European Centre for Disease Prevention and Control. *Historical data on the daily number of new reported COVID-19 cases and deaths worldwide*. 2020. URL: https://opendata.ecdc.europa.eu/covid19/casedistribution/.

[11] Jesús Fernández-Villaverde and Charles I Jones. *Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities*. Working Paper 27128. National Bureau of Economic Research, May 2020. DOI: 10.3386/w27128. URL: http://www.nber.org/papers/w27128.

[12] Seth Flaxman. "Reply to: The effect of interventions on COVID-19". In: *Nature* 588.1 (Dec. 1, 2020), pp. 29 –32. ISSN: 1476-4687. DOI: 10.1038/s41586-020-3026-x. URL: https://doi.org/10.1038/s41586-020-3026-x.

[13] 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. URL: https://doi.org/10.1038/s41586-020-2405-7.

[14] 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.

[15] Alexander Grigorian. *Mathematical riddles of COVID-19*. June 2020. URL: https://www.math.uni-bielefeld.de/~grigor/corv.pdf.

[16] Robert J. Hijmans. *raster: Geographic Data Analysis and Modeling*. R package version 3.4-5. 2020. URL: https://CRAN.R-project.org/package=raster.

[17] ArcGIS Hub. *UNIGIS Geospatial Education Resources*. 2020. URL: https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08_1.

[18] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice, 2nd edition*. OTexts.com/fpp2. OTexts, 2018.

[19] Roni Parshani, Shai Carmi, and Shlomo Havlin. "Epidemic Threshold for the Susceptible-Infectious-Susceptible Model on Random Networks". In: *Physical Review Letters* 104.25 (June 2010). ISSN: 1079-7114. DOI: 10.1103/physrevlett.104.258701. URL: http://dx.doi.org/10.1103/PhysRevLett.104.258701.

[20] Karthik Ram. *Wes Anderson Palettes*. 2013. URL: https://github.com/karthik/wesanderson.

[21] 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. URL: https://pubmed.ncbi.nlm.nih.gov/32122430.

[22] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[23] 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.