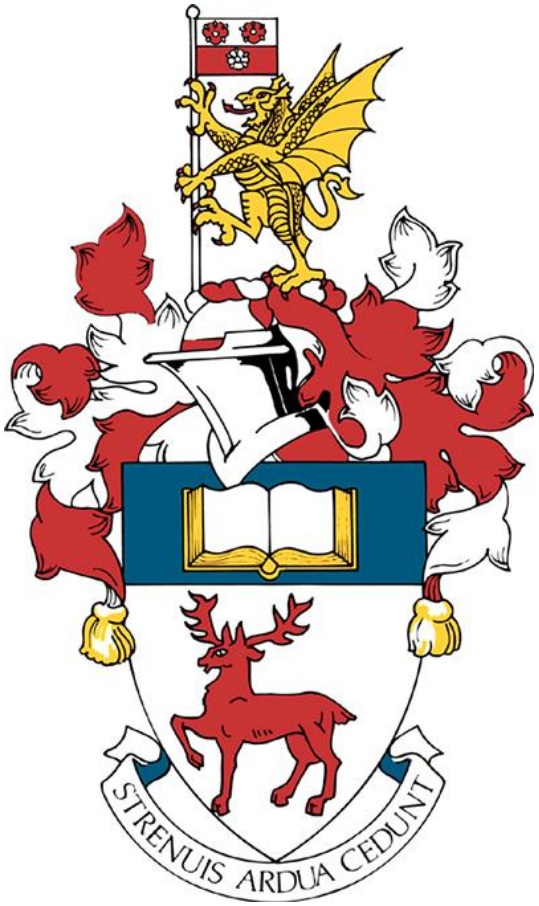
UNIVERSITY OF SOUTHAMPTON

SCHOOL OF PHYSICS AND ASTRONOMY



Evaluating compartmental models for COVID-19: A UK case study

By:
Armaan Sachdeva

Submitted in partial fulfilment of the requirements for the MSc degree in Physics of the University of Southampton

May 2025

Abstract

This report evaluates the effectiveness of multiple compartmental epidemiological models in simulating the weekly infections and deaths from the UK COVID-19 outbreak from March 21st 2020 to March 21st 2021. Model accuracy was assessed using the Root Mean Square Error (RMSE) and R^2 values. A constant parameter SIR model performed poorly ($R^2 \approx -2$), A SEIRDV model incorporating a 3 term Fourier series for transmission rate $\beta(t)$ and a multi-Gaussian function for the mortality rate $\mu(t)$ returned a much better fit for transmission rate ($R^2_{\beta(t)} \approx 0.9$) but was unable to also achieve a good fit for the mortality rate at the same time, with ($R^2_{\mu(t)} \approx 0.2$. These findings show the capability and limitations of using compartmental models in recreating the complex dynamics of a real-world epidemic.

Contents

1 Introduction

The COVID-19 pandemic highlighted how important mathematical modelling of epidemiological outbreaks is for understanding and predicting the spread of the disease. One of the most widely used of these techniques is the compartmental model. It works by splitting the population into different groups (compartments) based on their disease status. These models allow us to simulate the spread of disease over time and helps assess the projected outcome of different interventions.[13]

In this paper several different compartmental models were tested in the context of COVID-19, aiming to determine how well they simulate its spread through the initial part of the outbreak in the UK. These models include the basic SIR model (Susceptible, Infected, Recovered, model.) and more complex models such as SEIRDV which adds Deaths and Vaccinations respectively.[16]

The key challenges faced in this paper was modifying the differential equations and developing time-varying functions to model mortality and transmission rates, which are often simplified to constant parameters.

The questions that this paper hopes to address are:

- How well can a compartmental model simulate real-world epidemic data
- What are the effects of adding a time-varying transmission and mortality functions on the models

2 Methods

2.1 Covid-19 data

The goal of this project is to determine how well compartmental models can recreate data from the first year of the COVID-19 outbreak in the UK. This will be the data from March 21st 2020 to March 21st 2021.

This was acquired from (https://ourworldindata.org/covid-cases) on 22/4/2025. The following graphs will show the number of weekly cases and deaths and the daily doses of vaccine administered. From Figure 1, it can see that the daily cases follow a complex multi-peak with infections starting almost immediately before the lockdowns were enforced, then during the lockdowns over the summer period of 2020 the cases fell considerably. However this could be due to lack of reported information. a large influx of new infections towards September of 2020 when the COVID 19 lockdowns were no longer enforced can also be noted.

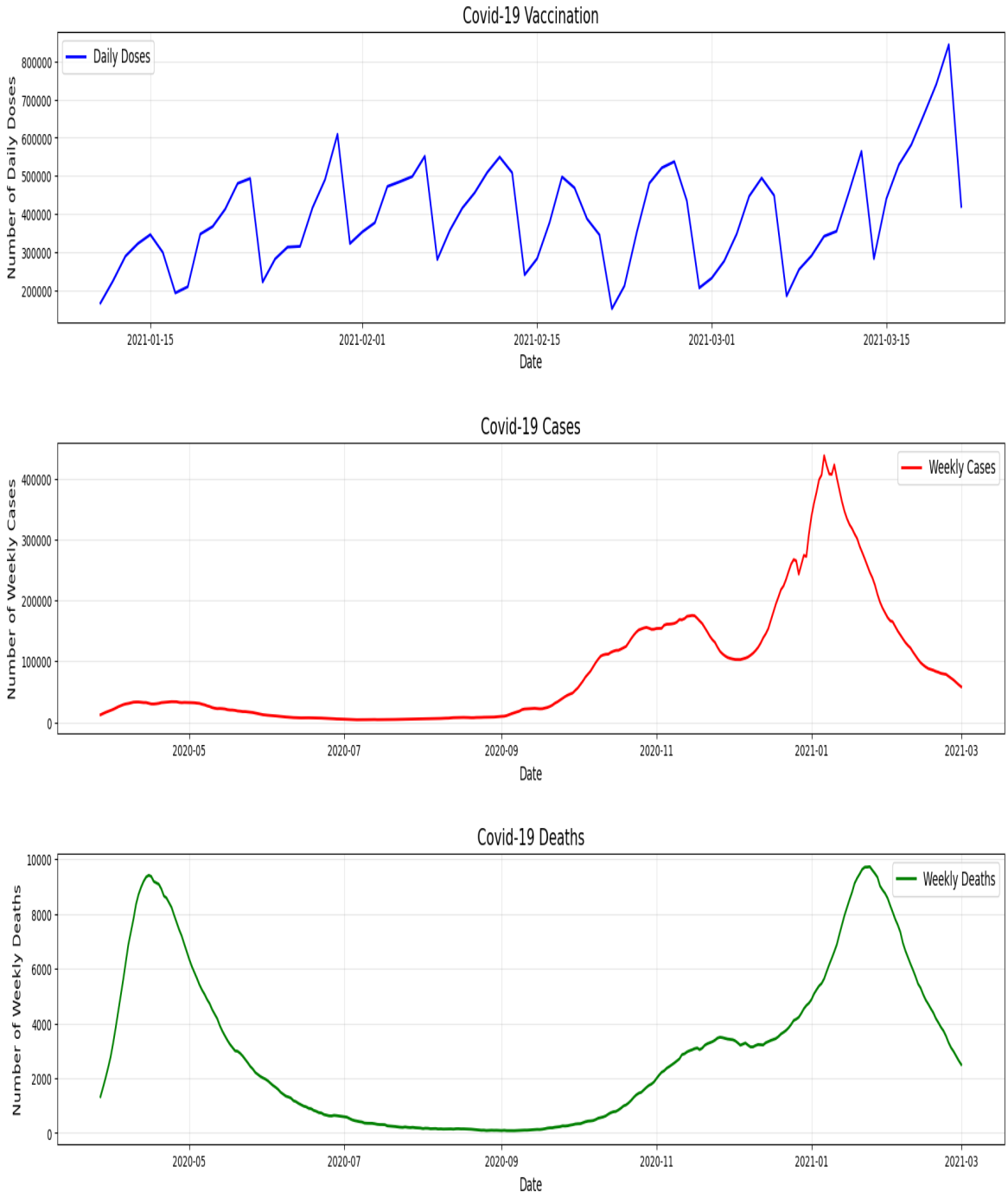


Figure 1: Overview of UK COVID-19 epidemiological and vaccination data used for model fitting (21 March 2020 - 21 March 2021). The plots illustrate (top) the rollout of daily vaccination doses starting in early 2021, (middle) the multi-wave pattern of weekly new infections, and (bottom) the corresponding trends in weekly deaths. This dataset forms the basis for evaluating compartmental model performance.

From Figure 1 it can be seen that the daily deaths had their peak at the start of the dataset in March. Before being reduced substantially by the lockdowns over the summer period before rising again when the lockdowns were lifted again, peaking again in February.

Overall these patterns indicate why real-world disease dynamics are so complex. They are affected by policy changes, public behaviour, and healthcare capacity, with limitations such as the under-reporting of cases and deaths further complicating the dynamics. Compartmental models often make assumptions such as homogeneous mixing and fixed parameters. By comparing real-world data to that of predictions from compartmental models I am to evaluate the strengths and weaknesses of these models in simulating a real outbreak.

2.2 Empirical curve fitting

For the infection and death data found. Finding the upper limits of what current mathematical models are capable of predicting gives us a starting point from which a baseline can be created to show how well the models can realistically recreate the real-world data.

The following are the mathematical models used to fit to the real world data:

2.2.1 Fourier series

This technique is an adapted version of the Fourier series of a function. It can approximate any periodic function as a sum of sines and cosines.

$$f(x) \approx a_0 + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n x}{L}\right) + b_n \sin\left(\frac{2\pi n x}{L}\right) \right)$$

- L: The length of the interval (period).
- a_0: The DC component or the average value of the function over the interval. It is calculated as $\frac{1}{L}\int_0^L f(x)dx$
- n: The harmonic number $(n = 1, 2, 3, \ldots, N) n = 1$ represents the fundamental frequency $\frac{1}{L}, n = 2$ represents twice the fundamental frequency, and so on.
- a_n: The coefficient for the n^{th} cosine term. It represents the contribution of the cosine wave at frequency $\frac{n}{L}$. It's calculated as $\frac{2}{L}\int_0^L f(x)\cos\left(\frac{2\pi n x}{L}\right)\, dx$.

- b_n: The coefficient for the n-th sine term. It represents the contribution of the sine wave at frequency $\frac{n}{L}$. It's calculated as $\frac{2}{L}\int_0^L f(x)\sin\left(\frac{2\pi nx}{L}\right)dx$.
- The combination of a_n and b_n determines the amplitude $\sqrt{a_n^2+b_n^2}$ and phase shift of the n^{th} harmonic component.

Since infinite precision (by setting N to ∞) is not achievable, N was kept at 5, yielding 5 Fourier terms. The more terms we add the lower the value additional terms will add whilst adding considerably more complexity. This diminishing return is why N was limited to just 5.

2.2.2 Multi-Gaussian approach

This approximation hopes to model the data using a sum of multiple Gaussian distributions . This approach hopes to model our infection and death curves as follows:

$$f(x) = \sum_{i=1}^{N} G(x, A_i, \mu_i, \sigma_i) = \sum_{i=1}^{N} A_i \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

Where:

- A : The amplitude of the peak of the normal distribution
- μ_i is the mean of the specified peak or where the peak is centred
- σ_i is the standard deviation of the peak which relates to the width of the peak, the greater the standard deviation the wider the Gaussian will be
- N is the number of Gaussians we'll be adding together

For modelling the infection and death curves, 3 waves were assumed , aligning with news reports that identified 3 phases of the pandemic in the first year. Those being the initial spike at the start of the period, another around August and a final one in the winter of 2020-21.

The limitation of this model is it does assume that each curve is symmetric which may not be true for a real-life example

2.2.3 Gompertz function

This approximation hopes to try and model the asymmetry seen in the infection and death curves. It works as follows

$$f(x) = A\exp(-\exp(-c(x-b)))\cdot\exp(-d(x-b)^k)\cdot[x>b]$$

where:

- A :Scales the overall amplitude (related to peak height).

- b : Represents the approximate time location of the peak.
- c : Controls the steepness of the initial rise (growth rate).
- d : Controls the rate of decay after the peak
- k : An exponent controlling the shape of the decay, in this case it was set to 1 so we could have an asymmetric peak]
- $[x > b]$: This is an indicator function to make sure that the decay occurs after the peak time b

This combination creates asymmetric growth via the Gompertz growth and then the exponential decay forces the model back down. The limitation of the model would be that it doesn't take into account multiple waves of an epidemic, instead capturing the anti symmetry present in one.

2.2.4 Sum of Weibull distributions

This distribution is usually used to describe the probability that a random variable will take a certain value. Each peak will follow this form:

$$f(x, k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{k}\right)$$

Where:

- $k > 0$: This is the shape parameter
- $k < 1$: the probability function will decrease from infinity
- $k = 1$: it follows an exponential distribution
- $k > 1$: it starts at 0, rises to a peak then declines, the peak becomes more symmetric as k increases and at around $k \approx 3.6$ this distribution will resemble a Gaussian distribution.
- λ : This is the scale parameter that stretches the curve along the x-axis and relates the time scale

The model used for this is the sum of multiple of these curves. This follows

$$f(x) = \sum_i A_i \cdot f(x, k_i, \lambda_i)$$

Where the A_i amplitude of the i^{th} peak. similarly k_i and λ_i are the shape and scale parameters for the i^{th} peak
The reason for using this model is that it can recreate the data using multiple peaks, some of which being asymmetrical as well as symmetrical

2.2.5 Generalized logistic function

Similar to the Gompertz function from earlier this is also a another type of asymmetrical curve that grows exponentially and thus also needs an exponential decay term. The logistic function is given as

$$f(x) = K(1 + Q\exp(-r(t - t_0)))^{-1/\nu}$$

Where:

- K: Is the upper asymptote
- r : The intrinsic growth rate
- t_0 : The time of the maximum growth
- $\nu > 0$: This is the shape parameter, which affects where the inflection point occurs relative to the asymptote and thus controls the curve's asymmetry as $\nu \to 0$ it approaches a Gompertz curve when $\nu = 1$ it becomes the standard logistic curve $\nu > 0$: This is the shape parameter, which affects where the inflection point occurs relative to the asymptote and thus controls the curve's asymmetry as $\nu \to 0$ it approaches a Gompertz curve when $\nu = 1$ it becomes the standard logistic curve

2.2.6 Sum of log-normal distributions

This is also a skewed distribution, meaning it has an asymmetric shape. This is a variation of the multi Gaussian model with the difference being it doesn't assume the curves are symmetric. It follows the following form

$$f(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \text{ for } x > 0$$

Where:

- μ: The mean of the variables natural logarithm
- σ: The standard deviation of the variables natural logarithm, larger values cause more skewness

In the model implementation it uses the sum of multiple of these distributions given by

$$y(x) = \sum_i A_i \cdot f(x, \mu_i, \sigma_i)$$

Where A corresponds to the amplitude of the individual peak
This specific model is very good for modelling multiple peaks where each peak is expected to have a right skew- This is common in infection data our data especially for the infection curves as they have a very clear right skew.

2.3 The SIR model

This is the most basic of the compartmental models. It stands for Susceptible, Infected, Recovered. How these models work is by splitting the population up into different compartments and modelling how these compartments develop over time as a set of differential equations. The SIR model models each compartment as follows:

$$\frac{dS}{dt} = -\beta \cdot \frac{SI}{N}$$
$$\frac{dI}{dt} = \beta \cdot \frac{SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Where:

- S is the susceptible population
- I is the infected population
- R is the recovered population
- β is the infection rate
- γ is the recovery rate.

This coupled set of differential equations are the basis of the SIR model. In order to solve these sets of differential equations we'd have to employ a numerical solver. For this project that was done using the scipy solver, which is a numerical solver that solves Ordinary differential equations of order 1.
However this model is fairly simple and doesn't really reflect the trends shown in the real world data so more complexity is needed to be added in order for it to more accurately reflect the real world data.

2.4 The SEIRDV model

This is a more complex compartmental model given by the following coupled differential equations:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} - \nu S$$
$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E$$
$$\frac{dI}{dt} = \sigma E - \gamma I - \mu I$$
$$\frac{dR}{dt} = \gamma I$$
$$\frac{dV}{dt} = \nu S$$

Where:

- S is the susceptible population
- R is the recovered population
- β is the infection rate
- γ is the recovery rate.
- σ is the rate at which exposed individuals become infected
- μ is the disease induced mortality rate
- ν is the vaccination rate

This model is an extension of the SIR model that captures how susceptible people become exposed due to contact with infectious people. This model tracks susceptible individuals who become exposed through contact with infectious people. Exposed individuals then become infected, and infected individuals either recover or die. Susceptible individuals can also be vaccinated, moving directly to the vaccinated compartment.

It should be noted that in the time scale chosen for this project. The COVID-19 vaccines were released in December 8th 2020 for the Pfizer version and the Astra Zeneca version was released on the 4th of January 2021. [10]That means that for the vaccination compartment will only go into effect in the last 3 months of the simulation , from January to March to correspond to the real time the vaccinations were released. Furthermore the vaccination data used to model the vaccination rate only started being recorded in January with the availability of the Astra Zeneca version.

2.5 Time-varying parameters

One aspect of these models that has to be changed in order for them to be more accurate is that variables like β the infection rate, μ the death rate, γ the recovery rate, ν the vaccination rate and incubation rate σ can't all be constant values in order to achieve a good approximation of the real data. Unless it doesn't make sense for them to vary in time.

A good starting point is having these variables instead being time-varying functions. For the transmission rate (β) it's assumed that the infection rate will oscillate over the course of the first year of the pandemic. It will respond to changes in government policy, different variants, and peoples shifting attitudes.

2.5.1 Time-varying transmission rate

The following is the transmission rate initially tested.[17]

$$\beta(t) \approx \beta_0 + \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right)$$

where:

- B_0 is the baseline transmission rate.
- T is the period of the periodic events in units of days.
- a_n, b_n are the Fourier coefficients.
- t is the time in days
- n is the harmonic number

2.5.2 Time-varying mortality rate

Due to the great variation in the mortality rate across the country as well as the rate changing over time as well. This would be the other parameter that could work well with a time-varying function. From the shape of the graphs for daily deaths per million (Figure 2) it can be seen that there were 2 great waves when it comes to deaths. We could also model that as a sinusoidal function similar to the transmission function $\beta(t)$. It can also assume a different model being a sum of Gaussians

This approximation hopes to model the data using a sum of multiple Gaussian distributions . This approach hopes to model our death curve as follows

$$f(x) = \sum_{i=1}^{N} G(x, A_i, \mu_i, \sigma_i) = \sum_{i=1}^{N} A_i \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

Where:

- A : The amplitude of the peak of the normal distribution
- μ_i is the mean of the specified peak or where the peak is centred
- σ_i is the standard deviation of the peak which relates to the width of the peak, the greater the standard deviation the wider the Gaussian will be
- N is the number of Gaussians we'll be adding together

n the case of modelling the death curve we would be assuming there are 3 peaks/ waves of infection in accordance the shape of the death curve having 2 distinctive peaks, with the third Gaussian describing the data in between.

The limitation of this model is it does assume that the death rate followed a uniform structure which is very unlikely in the real world. It should however be a good approximation.

2.5.3 Vaccination rate

There is exact vaccination data available for the time period. Therefore it was used to create a time-varying function for the vaccination rate. In our vaccinated compartment we are using using the term νS which represents the rate at which the susceptible population enters the vaccinated compartment. ν represents the fraction of the current susceptible population that gets vaccinated each day. In the dataset there is the total number of people $V_{daily}(t)$. The following relationship emerges

$$\nu(t)S(t) = V_{daily}(t)$$

which can be rearranged to get

$$\nu(t) = \frac{V_{daily}(t)}{S(t)}$$

However since $S(t)$ is changes over time, the following simplification was made.Instead of using $S(t)$the total population N was used. Since the total susceptible population won't be changing substantially enough that it would cause a large impact on the vaccination result.

2.6 Constant parameters

2.6.1 Recovery rate

This was the easiest rate to use. There was no need to have a time-varying function for the recovery rate. All this variable is asking is for what percentage of the population recovers from the disease each day. this was set to 1/14 as for the majority of this period this was the NHS guideline for how long a person should self isolate to guarantee not spreading the infection further.

2.6.2 Incubation rate

The incubation rate σ was also fairly easy to find because it is known how long it takes for a infected person to become infectious after being exposed to the virus. This was known to be 5 days so the incubation rate is 1/5 or 0.2.[15]

2.7 Testing metrics

In order to stretch the limits of how well each compartmental model can represent the real world infection data. The main measures we are using to represent how well the compartmental model matches with the real data is by statistical measures of RMSE and R^2 .

2.7.1 RMSE

RMSE is the Root Mean Squared Error. It works by finding the difference between the actual data and what the model acquired $(y_i - \hat{y}_i)$. Then squaring the result $(y_i - \hat{y}_i)^2$. Then finding the mean of all the squared errors $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ before finally square rooting the result.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- The RMSE will always be a positive value.
- The lower the value of the RMSE the better the model fits to the actual data. A value of 0 indicates an exact fit
- This process is very sensitive to outliers in data due to the mean being very sensitive to outliers.

2.7.2 R^2 metric

This metric measures the proportion of the variance in the real data that is explained by the created model. It is comparing the models performance to a basic model that just predicts the mean of the real data. It is calculated as follows:

1. Calculate the sum of the squared residuals. similar to RMSE but with no averaging. $SS_{\text{Res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
2. Calculate the total sum of the squares. This step measures the total variance in the actual data. By first calculating the mean of the actual data $\bar{y}$ then finding the sum of the squared differences between each actual data point y_i and the mean. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. $SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
3. Finally calculate the R^2 using the following formula $R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{tot}}$

The R^2 is a relative metric and so the exact values of the data don't matter, only the difference. Therefore unlike RMSE, scaled data from the same source will return the same values so long as the scale change is consistent across the whole dataset in effect this means that even for data that is in units of per 1-million the R^2 will return the same value as the raw data that has not been scaled.

The metric has the following features:

- The R^2 statistic is usually a value between 0 and 1 for decent models but can be negative if the model is a bad fit for the data
- If R^2 is a good model for the data if the value is between 0 and 1. If R^2 is 0.87 for example it means the model explains 87% of the variance in the real data. The higher the value the better

- If the R^2 is less than 0 this means that the model is worse than just predicting the mean of the data. In effect if the value is 0 or less the model is a terrible fit for the data.

2.8 Computational Implementation

All simulations, data processing, and analyses were performed using the Python programming language (version 3.10.12), leveraging standard scientific libraries including NumPy (Version 2.1.2) for numerical operations, Pandas for data manipulation and time series analysis, SciPy (Version 1.15.2) for scientific and technical computing, and Matplotlib (Version 3.10.1) for generating plots.
Deterministic Model Simulation: The systems of ordinary differential equations (ODEs) describing the SIR and SEIRDV models were solved numerically using the solve ivp function from the Scipy.integrate module. The default Runge-Kutta method of order 4(5) (RK45) with adaptive step sizing was employed to ensure accurate integration over the simulation period.

Parameters for the time-varying functions $\beta(t)$ and $\mu(t)$ were estimated by fitting the deterministic SEIRDV model output to the observed weekly case and death data. This involved minimising a combined objective function, defined as the sum of the Root Mean Squared Error (RMSE) for weekly cases and the RMSE for weekly deaths, each Normalised by the mean of the respective actual data series to account for scale differences. A weighting factor was applied to the Normalised death RMSE to prioritise fitting the mortality data. The minimisation was performed using the minimise function from the scipy.optimize module, employing the L-BFGS-B algorithm, which accommodates bounds on parameters.[2]

The time-varying vaccination rate $\nu(t)$ was derived by dividing the reported daily doses by the total population N and using linear interpolation or direct lookup based on the simulation time step.

3 Results

This section presents the results of fitting various compartmental and mathematical models to the UK COVID-19 infections and death data from March 21st 2020 to March 21st 2021.

The test statistics used to measure how well the compartmental models can simulate the real world data are the RMSE and the R^2 metric.

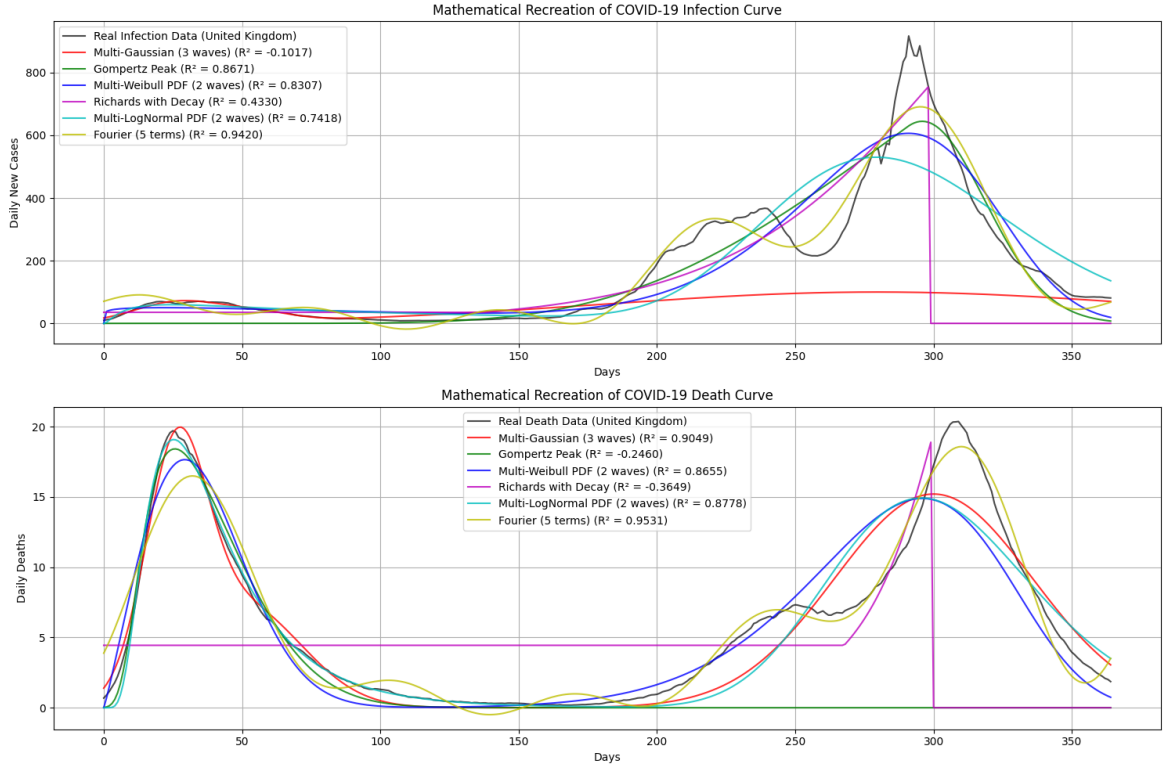


Figure 2: Performance of selected mathematical functions in empirically fitting observed UK COVID-19 daily infection data (top) and daily death data (bottom) from March 2020 to March 2021. These fits, with R^2 values shown, serve as a baseline to assess the maximum achievable goodness-of-fit by purely mathematical functions, prior to applying mechanistic compartmental models.

3.1 Empirical fitting

Table 1: The optimised curve fitting metrics.

Function	RMSE Infection	R^2 Infection	RMSE Death	R^2 Death
Multi-Gaussian	215.0	−0.100	1.81	0.900
Gompertz peak	74.8	0.860	6.58	−0.250
Multi Weibull	84.5	0.830	2.16	0.860
Richards with decay	154.0	0.430	6.88	−0.370
Multi logarithmic Gaussian	104.0	0.740	2.06	0.880
Fourier series (5 terms)	49.4	0.940	1.28	0.953

From Figure 2 and Table 1, it can be seen that the model that fit the best for the infection curve was the 5 term Fourier series. With an RMSE of only 49.4 and an R^2 of 0.94 . Indicating a very good fit. The model that works best with the deaths curve is also the 5 term Fourier series. with an RMSE of 1.28 and a R^2 of 0.9531, being the best fitting in the table. It should be noted that this graph was taken from Normalised data and so the RMSE is lower than for the other graphs, However the

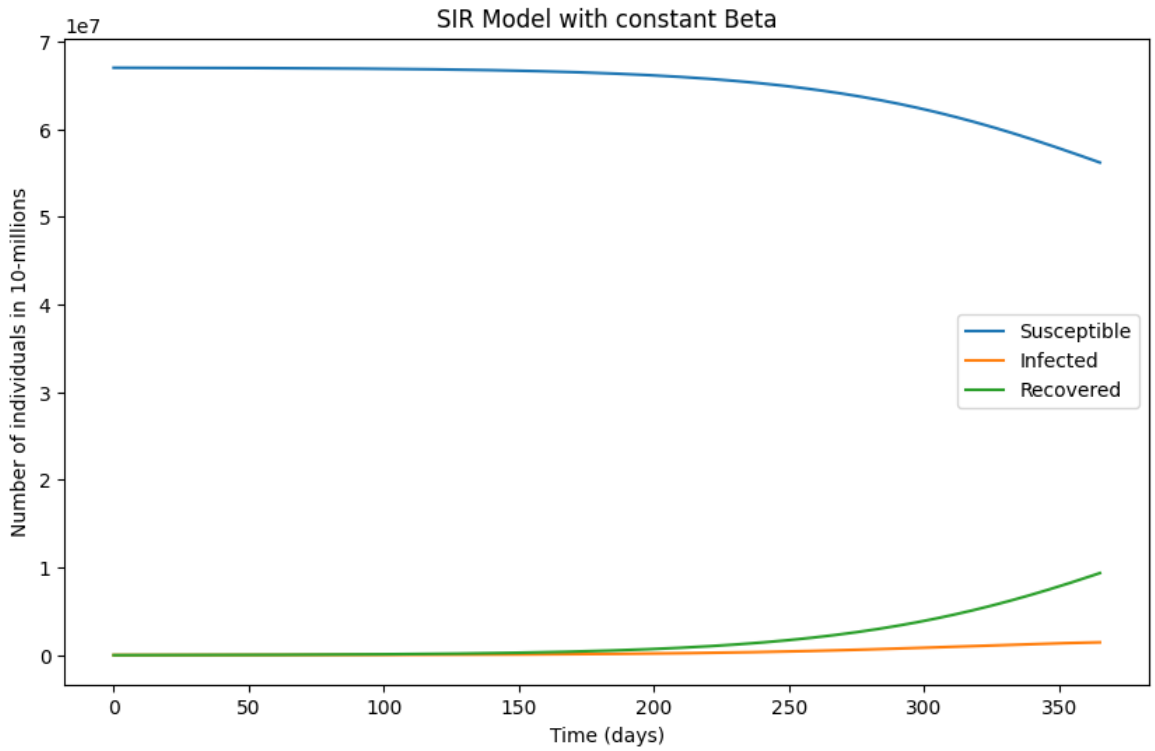


Figure 3: Dynamics of Susceptible, Infected, and Recovered populations predicted by a basic SIR model assuming a constant transmission rate (β). This simulation illustrates the typical epidemic curve produced by a simple SIR model and serves as a baseline for comparison with models incorporating time-varying parameters

R^2 values are the same due to that metric being relative and not absolute like RMSE. The multi-Gaussian approach was the worst model by far for the transmission rate $\beta(t)$. Evident by its the only result for transmission with a negative R^2 value of -0.1 and the highest RMSE at 215, more than 1.5 times the next highest model, Richards with decay at an RMSE of 154. The worst fit for the mortality rate $\mu(t)$ was the Richards with decay model. It had the most negative R^2 with a value of -0.37 although it wasn't far off from the Gompertz peak's -0.250, as it also had a negative value for the same metric. The RMSE for both of these models was also almost 3-times higher than the next worst performing model, being multi-Weibull at 2.16 compared to these models, who were at 6.88 and 6.58 respectively.

3.2 The SIR model

From Figure 3 It can be seen that the in the context of the entire country only around 10 million people would be infected in total in the first year of the outbreak, with the total value of the infected compartment staying fairly stable throughout the whole year, as shown by the infected curve being almost a straight line. Figure 4 then compares the infected compartment with the real recorded covid cases in the UK over the 1 year time period. It does manage to increase similarly to the infected curve for COVID data but it doesn't manage to capture the peaks shown in the real

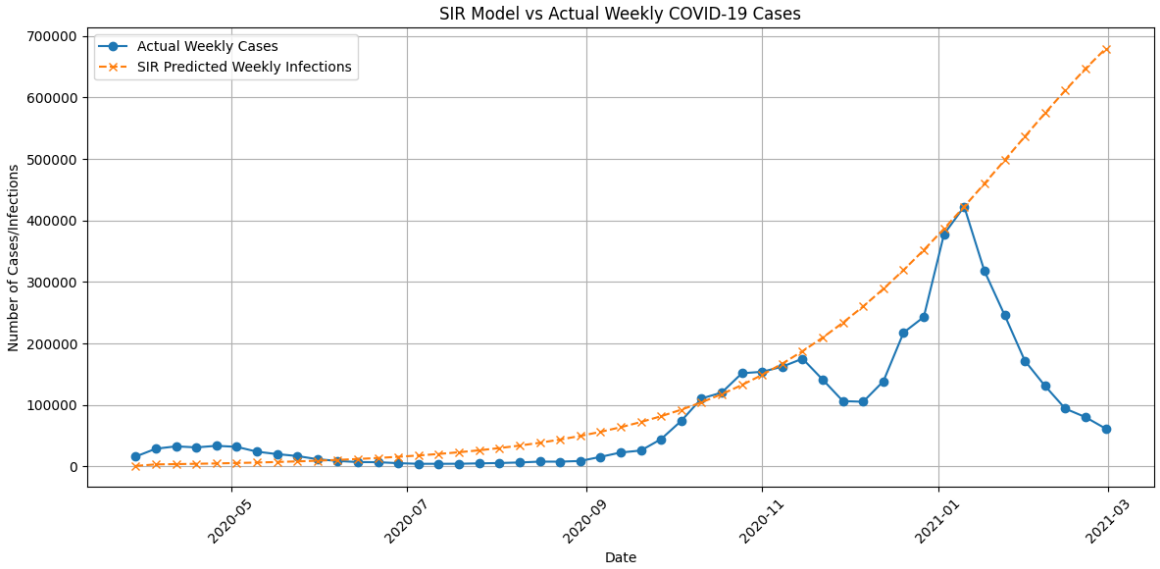


Figure 4: Comparison of the infected compartment from an SIR model with a constant transmission rate (orange dashed line) against actual weekly COVID-19 cases in the UK (blue line) from March 2020 to March 2021. The model exhibits a poor fit, with an RMSE of 174,597.89 and an R^2 of -2.012, failing to capture the multi-wave nature of the observed data.

data, the number of cases only continue to increase. With an RMSE of 174,597.89 (greater than the axis of the graph itself) and a negative R^2 value of -2.012.

3.3 SIR model with time-varying transmission

Figure 5 shows the effect of adding a time-varying transmission function $\beta(t)$ to the the SIR model. The transmission function was modelled after a 2 term Fourier series.

$$\beta(t) \approx \beta_0 + \left(a_1 \cos\left(\frac{2\pi t}{T}\right) + b_1 \sin\left(\frac{2\pi t}{T}\right)\right) + \left(a_2 \cos\left(\frac{4\pi t}{T}\right) + b_2 \sin\left(\frac{4\pi t}{T}\right)\right)$$

Figure 5 shows a much better fitting than in the previous Figure 4. The infection curve follows the 2nd peak of the real case data. This is expressed quantitatively by the much better R^2 value of 0.86.

Table 2: Parameters for the 2-term Fourier series transmission rate $\beta(t)$ optimised to fit the real cases curve.

Parameter	Value
a_0	0.1561
a_1	-0.0486
b_1	0.0162
a_2	-0.0289
b_2	-0.0150

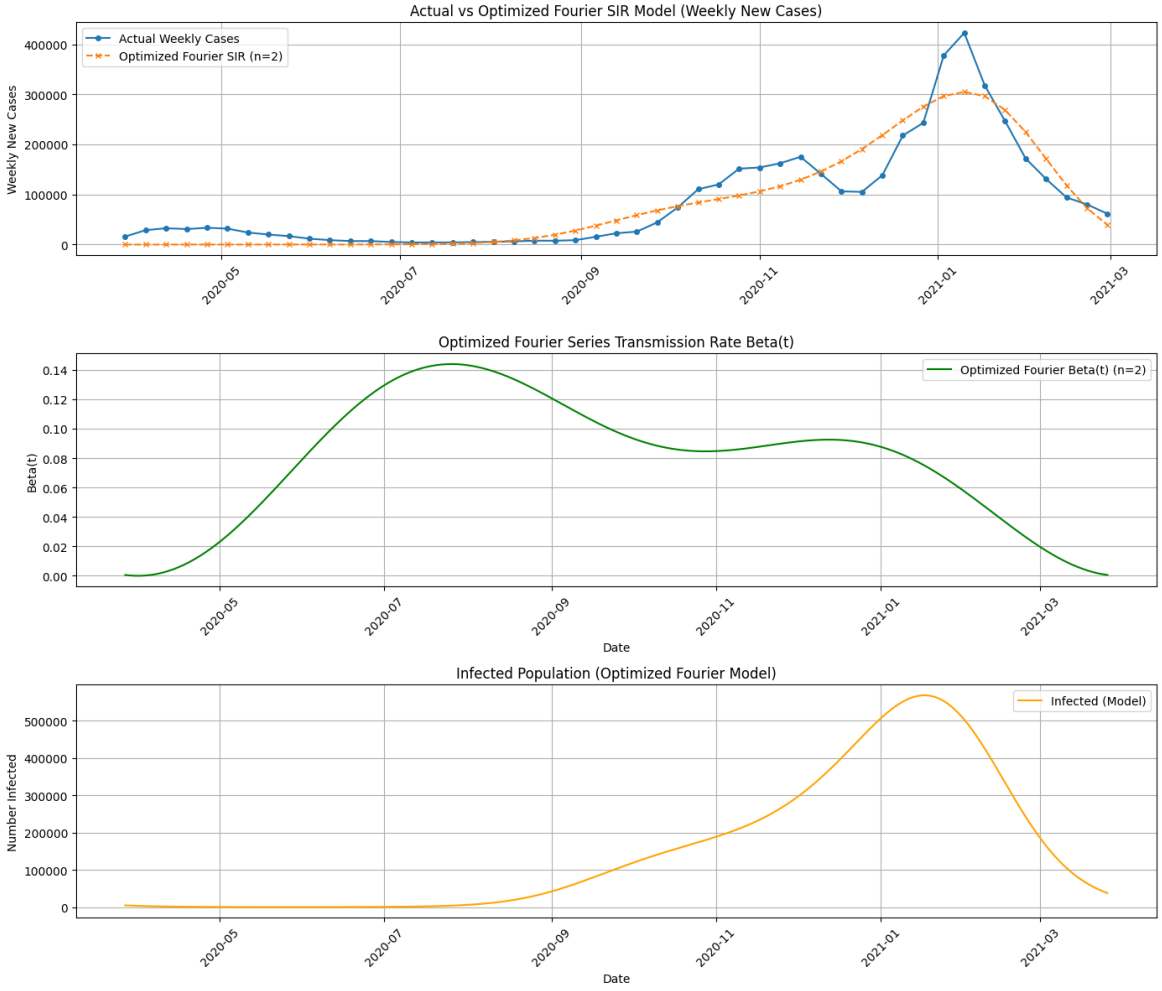


Figure 5: Impact of a time-varying transmission rate, $\beta(t)$, modelled as an optimised 2-term Fourier series, on SIR model predictions. (Top) Comparison of model-predicted weekly new cases (orange dashed line) against actual weekly COVID-19 cases (blue line). (Middle) The optimised 2-term Fourier series $\beta(t)$ function. (Bottom) The corresponding simulated infected population over time. Model fit: $R^2 \approx 0.86, RMSE \approx 37,087$.

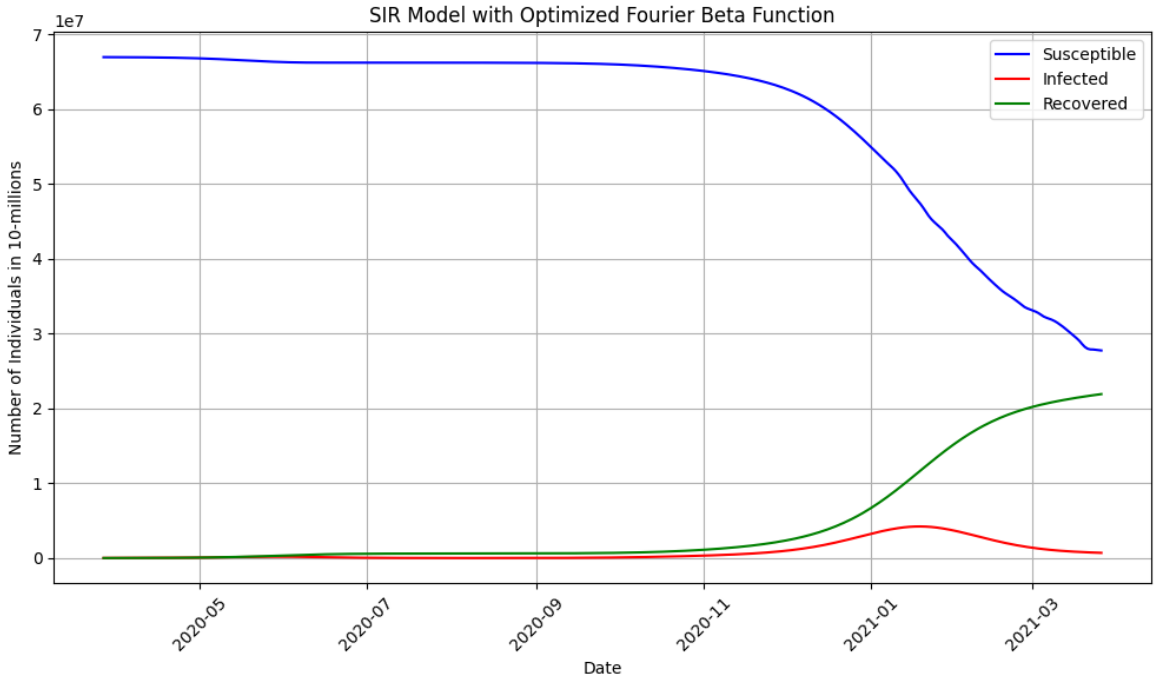


Figure 6: Simulated changes in the Susceptible (blue), Infected (red), and Recovered (green) compartments of the SIR model over the study period (May 2020 - March 2021), utilizing the optimised 2-term Fourier series for the time-varying transmission rate $\beta(t)$ shown previously in Figure [number of Figure 5, where $\beta(t)$ is plotted].

Figure 6 shows that there is a small fluctuation in the number of infected people at around January-February of 2021. There are considerably more people leaving the susceptible population in this version of the SIR model, as well as around 20 million people in the recovered compartment at the end of the first year of the model.

3.4 SEIRDV model with Fourier transmission and Gaussian mortality

Figure 7 has 3 more compartments than the previous SIR model. The added compartments allowed for the amount of people to leave the susceptible compartment fall the most out out of all the other models.Dropping to only 25 million by the end of the year. The recovered compartment started from 0 and ended at 30 million people. In the 3 months that the vaccinations were being undertaken during the year just over 10 million people were in the vaccinated compartment.The deaths remained very low, barely reaching 2 million by the end of the year, similarly the infections followed a similar curve to the previous SIR implementation, falling to a small number by the end of the year.From this it can be seen where the 67 million people in the UK's population would be during the simulation. This version of the SEIRDV has a more complex 3 term Fourier terms than the previous SIR model. However it has now also been optimied to find parameters for a sum of Gaussian mortality rate. The vaccination rate was calculated from the real world vaccination

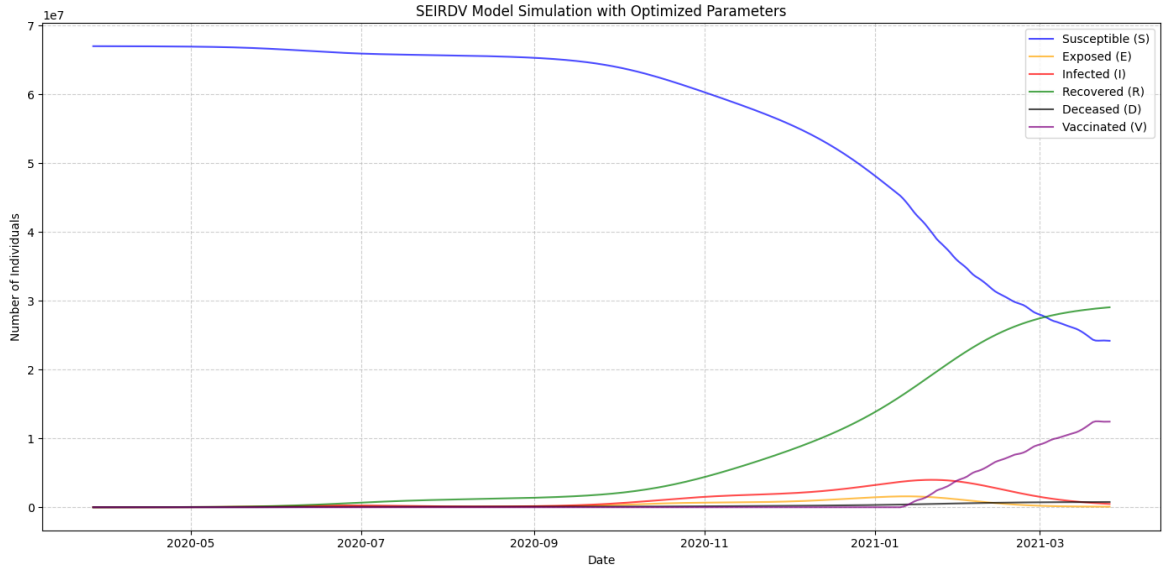


Figure 7: Simulated evolution of all compartments in the SEIRDV model (Susceptible, Exposed, Infected, Recovered, Deceased, Vaccinated) using optimised time-varying functions for the transmission rate $\beta(t)$ as a 3-term Fourier series) and mortality rate ($\mu(t)$ as a sum of Gaussians), along with a data-derived vaccination rate ($\nu(t)$. The simulation spans the period May 2020 to March 2021.

data available. The Fourier transmission rate was able to closely recreate the peak shape seen in the infection curve. However it couldn't quite manage to reach the real peaks of the COVID pandemic in the UK. The death curve on the other hand was not as good of a fit. It was unable to recreate the sharp initial spike of deaths at the start of the period. Only managing to recreate that of the 2nd peak in the winter of 2020-21. As a result it had a poor R^2 fitting at 0.202.

3.5 SEIRDV with Fourier death and transmission rates

This version of the SEIRDV model has both the transmission $\beta(t)$ and the mortality functions be Fourier series with 3 terms. In Figure 12 The infection curve still has a respectable R^2 of 0.8 and an RMSE of 44892. However the mortality rate still had a poor fit. While the fit is better than that of the previous Gaussian model it did considerably lower the goodness of fit of the transmission function in the process. The mortality rate had a R^2 of 0.2 and an RMSE of 2437.

In Figure 12 the time-varying functions for the 3 term Fourier series functions is plotted for both the transmission rate and mortality rate as well as for the vaccination rate which is based on daily vaccination data. It should be noted that in this approximation the optimiser saw fit to set the transmission rate to 0 for a period of time , shown by the function being at 0-or close to it- for several months

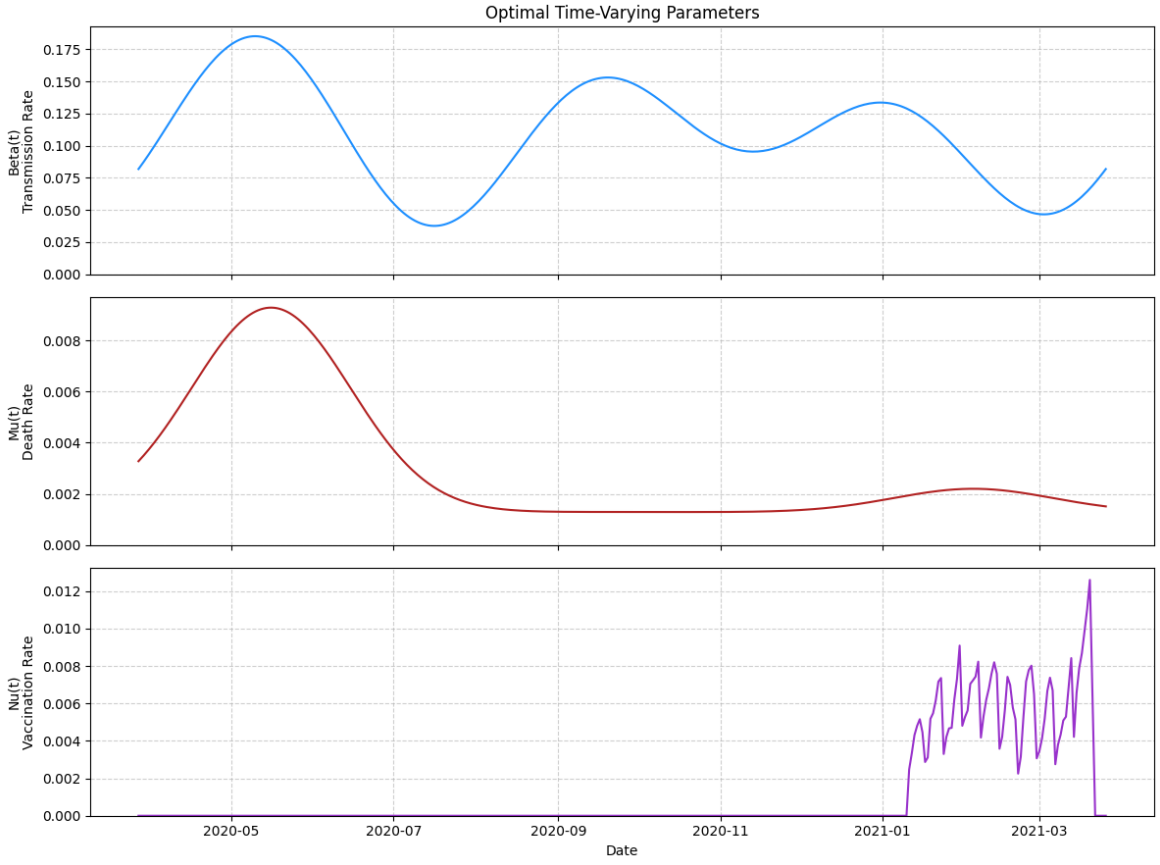


Figure 8: Optimised time-varying parameter functions used as inputs for the SEIRDV model (corresponding to the results in Figures 7 and 9). (Top) Transmission rate, $\beta(t)$, modelled as a 3-term Fourier series. (Middle) Mortality rate, $\mu(t)$, modelled as a sum of three Gaussian functions. (Bottom) Daily vaccination rate, $\nu(t)$, derived from UK government data.

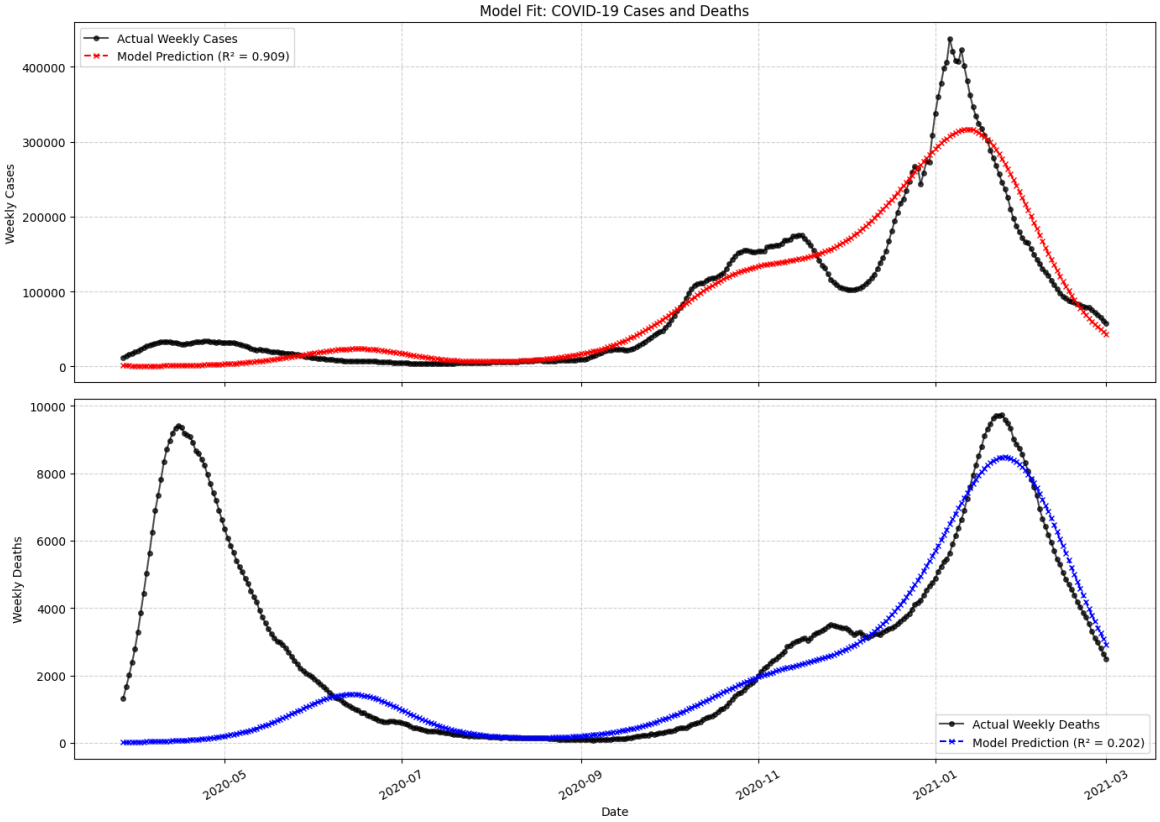


Figure 9: Comparison of the SEIRDV model predictions (using an optimised 3-term Fourier series for transmission rate $\beta(t)$ and a multi-Gaussian function for mortality rate $\mu(t)$ against actual UK COVID-19 data from May 2020 to March 2021. (Top) Model fit to weekly cases (red dashed line, $R^2 = 0.909$. (Bottom) Model fit to weekly deaths (blue dashed line, $R^2 = 0.202$. Actual data shown in black.

Table 3: Optimised parameters for the time-varying mortality rate $\mu(t)$ (Table A, left) and transmission rate $\beta(t)$ (Table B, right) for the SEIRDV model with Fourier transmission and Gaussian mortality.

Table A: $\mu(t)$ parameters

Parameter	Value for $\mu(t)$
A_1	0.00800
$\bar{\mu}_1$	50.0
σ_1	30.0
A_2	0.000100
$\bar{\mu}_2$	183
σ_2	30.0
A_3	0.000917
$\bar{\mu}_3$	315
σ_3	30.0
Offset	0.00108
RMSE	2570
R^2	0.203

Table B: $\beta(t)$ parameters

Parameter	Value for $\beta(t)$
β_0	0.108
a_1	0.00260
b_1	0.299
a_2	0.300
b_2	-0.0216
a_3	0.0783
b_3	-0.300
RMSE	30290
R^2	0.909

4 Discussions

The goal of this study was to evaluate the ability of compartmental models, from basic SIR to SEIRDV with optimised time-varying parameters, to replicate UK COVID-19 infections and deaths from 21st March 2020 to 21st of March 2021. Empirical curve fitting using functions like Multi Gaussian and Fourier series accurately managed to describe the shape of the data, standard compartmental models needed time-varying parameters to even get close to some of the results. The optimised SEIRDV model with a 3 term Fourier series for transmission $\beta(t)$ and a multi -Gaussian function for mortality $\mu(t)$ achieved the best fit to the observed weekly cases ($R^2 \approx 0.91$) but struggled to replicated the magnitude of the observed death peaks ($R^2 \approx 0.2$)

Table 4: Optimised Fourier series coefficients for the time-varying mortality rate $\mu(t)$ and transmission rate $\beta(t)$ used in the SEIRDV model configuration where both rates were modelled as Fourier series.

Parameter	Value for $\mu(t)$	Value for $\beta(t)$
Baseline (μ_0/β_0)	0.0100	0.120
a_1	0.500	-0.209
b_1	0.500	0.280
a_2	0.165	0.494
b_2	-0.133	-0.145
a_3	0.500	-0.500
b_3	0.395	-0.0276

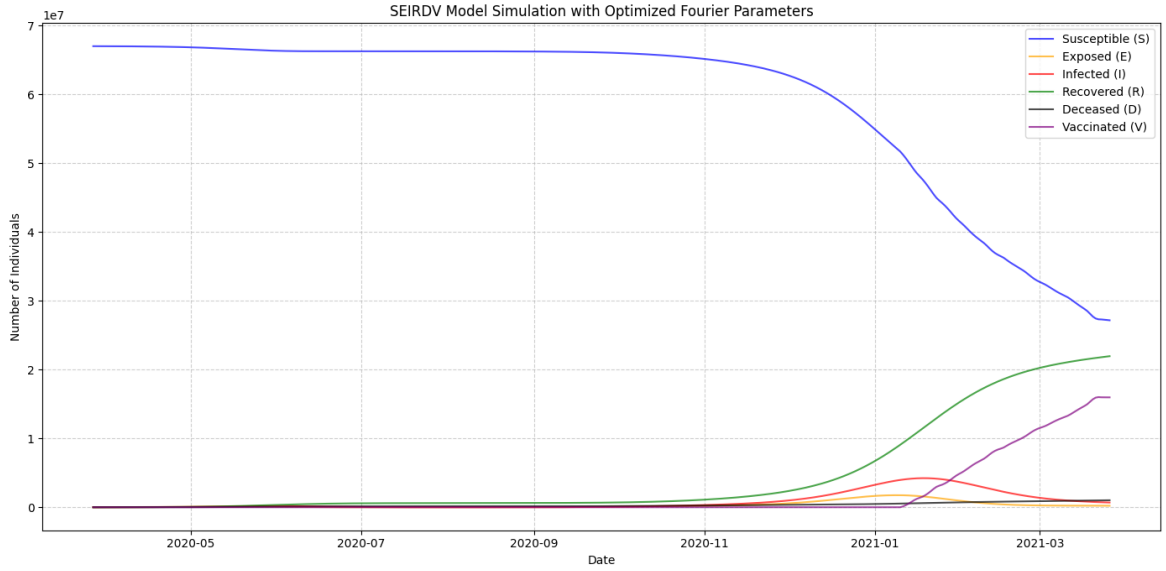


Figure 10: Simulated changes in Susceptible (S), Exposed (E), Infected (I), Recovered (R), Deceased (D), and Vaccinated (V) populations using the SEIRDV model where both the transmission rate $\beta(t)$ and the mortality rate $\mu(t)$ were modelled as optimised 3-term Fourier series. The simulation covers the period from May 2020 to March 2021.

Despite the fact many efforts for parameter tuning were tried such as changing the weighting of the optimisation function.

4.1 Insights gained from empirical curve fitting

The data in Figure 2 shows the goodness of fit for all the different mathematical models. For the mortality rate there is a pattern amongst the models. The models which were distinctly asymmetric in nature were the models which had the worst fittings. From table 1 these are Gompertz peak with an RMSE of 6.58 and an R^2 of -0.25. The other being the Richards with decay model with an RMSE of 6.88 and an R^2 of -0.37. These results show clearly that the real-world death data is not asymmetric in nature. However this may be a generalisation as there is only 1 of each of these functions as opposed to the others such as multi-Weibull and multi-Gaussian. Perhaps these models would have provided better fittings if there were multiple being added together.

On the other hand for the transmission rate the worst performing was the multi-Gaussian technique. This is likely due to the fact that the real infection curves can't be modelled by symmetric distributions. This is further shown by the 2 asymmetric functions having better fittings than the multi-Gaussian technique. These being the Gompertz and Richards with decay.Furthermore the models which incorporate both symmetric and asymmetric functions such as the multi-Weibull and the log-Normal distribution model had much better fitting metrics.

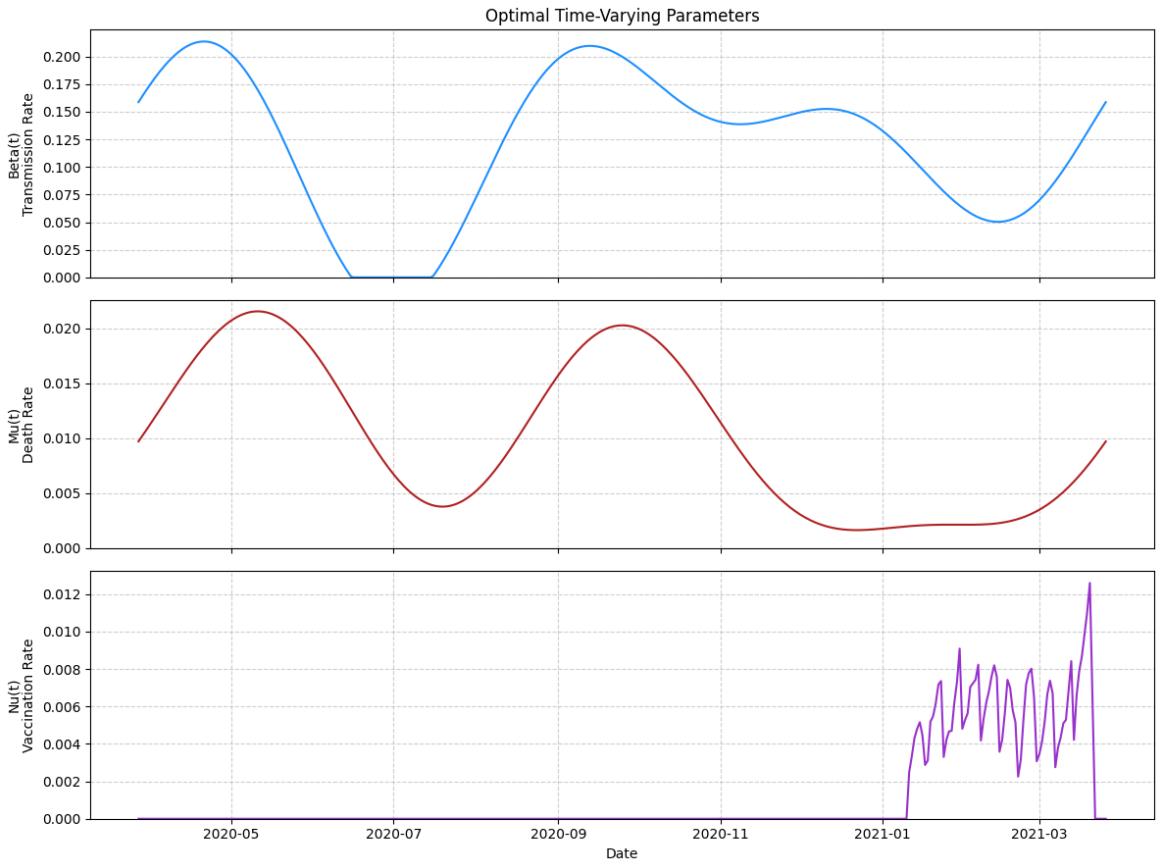


Figure 11: Optimised time-varying parameter functions for the SEIRDV model where both transmission rate $\beta(t)$ and mortality rate $\mu(t)$ were modeled as 3-term Fourier series (parameters in Table [relevant table number]). (Top) Transmission rate, $\beta(t)$. (Middle) Mortality rate, $\mu(t)$. (Bottom) Data-derived daily vaccination rate, $\nu(t)$. These parameters correspond to the model fit shown in Figure 12.

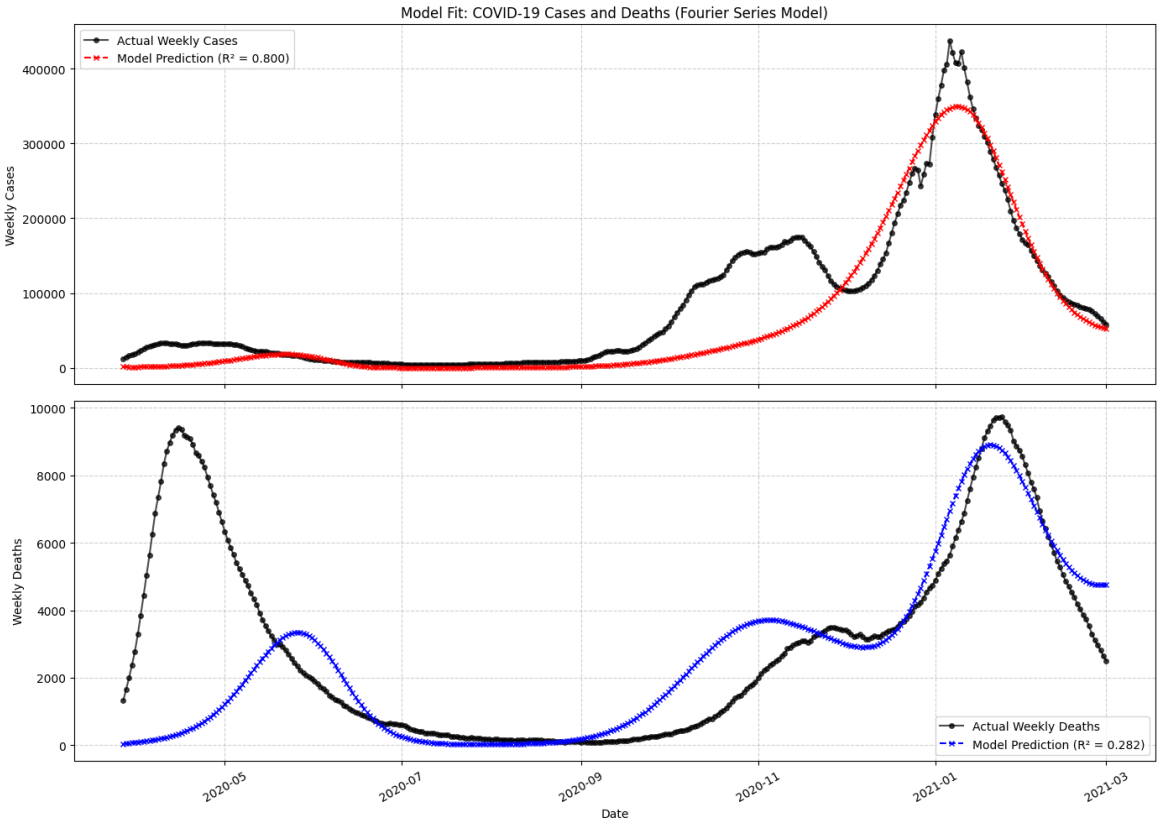


Figure 12: Model fit of the SEIRDV simulation, where both the transmission rate $\beta(t)$ and mortality rate $\mu(t)$ were modeled as optimised 3-term Fourier series (parameter functions shown in Figure 11; coefficients in Table [relevant table number]), against actual UK COVID-19 data. (Top) Fit to weekly cases (red dashed line, $R^2 = 0.800$. (Bottom) Fit to weekly deaths (blue dashed line, $R^2 = 0.282$. Actual data shown in black.

The best fitting for both the number of cases and number of deaths was the 5-term Fourier series. However this could most likely be attributed to the fact the Fourier series had more precision due to the higher number of terms. For example the multi-Gaussian model only had 9 parameters being optimised in contrast to the 5-term Fourier series which had 11 parameters being optimised. This slight edge in the number of optimised parameters may have been the reason as to why the Fourier series had the best fitting.

In the future keeping the number of parameters fixed in order to model the data would help in more accurately determining which model had the best fittings.

However the empirical fitting results do still indicate that the infections curve was characterised by a more definite asymmetry, while the deaths curve from the real-data proved the opposite, being better described by the more symmetric models. Furthermore the best fit being the Fourier series for both the datasets implies there is some underlying periodic or complex oscillatory component in both the infections and death trends.

4.2 Inadequacy of the constant parameter SIR model

The simple models like SIR with constant parameters failed to recreate the real world data. constant parameters like transmission rate β is simply not enough to capture the complex actions that occur in real-life that end up shaping how an epidemic will pan out. The only situation where the constant parameters can work is for very short time frames right at the start of an epidemic. Or for a disease that has such a high transmission rate that it can feasibly run out of people to infect within a population. The longer the epidemic is being simulated the worse the constant parameters would be. Constant parameters will be unable to model any kind of multi peak profile, they will only increase up until some peak infection rate then decline when new people can no longer be infected, usually be recovering from the disease. The constant parameter model will look like a Gaussian distribution.

4.3 Improvements made by adding time-varying parameters

The effect of simply making the transmission a time-varying function was very apparent and can be seen simply in the difference between the visual plots of Figure 4 and Figure 5. By adding a time-varying function for the transmission the model was able to follow the downward trend in the real-world data. This can also be seen quantitatively with the test statistics. The SIR model with constant parameters had a negative R^2 value of ≈ -2 which constitutes a bad fit. The mean is a better way to explain the variance in the real data by 200%. Not to mention the $RMSE \approx 175000$ which is greater than the y axis of the graph. When you compare this to Figure 5

which had a much more reasonable value of R^2 as $R^2 \approx 0.86$. Thereby saying that the model was able to recreate roughly 86% of the variance in the infections. Furthermore a much more reasonable RMSE of roughly 37000. which is almost 5 times less than the constant parameter SIR model.

4.4 Improvements and limitations of time-varying parameters and more compartments

The SEIRDV models tried to expand on this further. By adding time-varying functions to the mortality rate $\mu(t)$ and the transmission rate $\beta(t)$. By further increasing the number of terms for the fit as well from 2 Fourier terms to 3 the hope was that it would be able to get an even better result. This was true for the transmission function. By doing this it was shown that the model produced a better fit by having an even lower RMSE of roughly 30,000 compared the basic SIR model with a similar time-varying function with an RMSE of 37,000. This also translates to the R^2 value which was improved to $R^2 \approx 0.91$ While not as big of a difference as between the basic SIR model with constant parameters and the SIR model with a 2 term Fourier series transmission function. It was still an improvement on that.

However the main problem with this model was that the fitting for the mortality curve $\mu(t)$ wasn't very good. Visually it can be seen in Figure 9. The mortality curve simply couldn't capture the initial deaths peak at all. While it did manage to mostly capture the 2nd one. It still meant that the model had very low fitting metrics. With a $R^2 \approx 0.2$ This means the model could only capture 20% of the real variance in the deaths. Initially it was assumed that this was simply a limitation of the model used to fit the deaths curve to. Therefore another model was used. This being the same implementation as the transmission curve, being a 3 term Fourier series.

This however proved to be unsuccessful as the 3 term Fourier series was worse overall for the fitting. This can be seen in Figure 11. While the new mortality function $\mu(t)$ improved the R^2 by 8% going from $R^2 = 0.2$ to $R^2 = 0.28$ it caused the transmission function $\beta(t)$ do decrease by almost 11%. $R^2 \approx 0.91$ to $R^2 \approx 0.8$ which was an overall decrease in the effective fitting compared to the original implementation.

Even changing the way the optimization function worked by skewing the weighting for function to favor reduction for the RMSE for the mortality rate $\mu(t)$ caused no significant improvements in the fittings for the models. This highlighted a shortcoming of the model. It struggled to find a way to give good fits for both the transmission and mortality rates. In effect maybe improving the fit for the cases decreased the effectiveness for the deaths. On the other hand did the results gained make sense?

4.5 Questioning the validity of the time-varying parameters

4.5.1 SIR model with time-varying transmission

Starting with the SIR model with time-varying transmission. According to Figure 5, the time with which the infection rate was highest was around August of 2020 It was also at 14% which is a bit higher than the published data would suggest, on the other hand the real case data was always flawed due to under reporting so this may be a reasonable estimate for the peak infection rate. This is an expected time for which transmission might have peaked. The lockdowns in the UK were eased in June and almost entirely relaxed by the end of July. Considering the fact that people were meeting in public again and more social places like pubs were open again. It is expected that people suddenly having more contact again would cause infection rates to spike again. From that point it started to decrease again. Decreasing gradually over time as the UK started increasing restrictions again until January 2021. Before falling again considerably. In January of 2021 which coincidentally is also when the Vaccine rollout started in the UK for the general public and when the 2nd UK lockdown started going into effect. Overall the SIR model with 2 term Fourier transmission has very good timings for changes in the infection rate. It managed to capture the effects of the changing UK policy fairly accurately.

4.5.2 SEIRDV model with Fourier transmission and multi Gaussian mortality

The same cannot be said for this model. The SEIRDV model said that the transmission rate spiked in May. This is very unlikely as May was when the lockdowns in the UK were enforced. Furthermore the transmission rate was found to be considerably higher than what was ever recorded in the UK. Spiking at 17.5%. It then preceded to decrease substantially. Reaching a minimum in August. Which is when lockdown restrictions were almost lifted. Which is the exact opposite of what you'd expect to see. Therefore, while the result looks very good, The optimization itself doesn't really make sense when compared to the real timeline of the first year of the pandemic.

For the multi Gaussian mortality rate on the other hand. The time variation looks like it would make a little bit more sense. In Figure 8 , the mortality rate $\mu(t)$ shows a sharp initial spike at the start of the simulation, towards a fairly high death rate of 8%. before decreasing to a more stable amount for the rest of the simulation time at around 2%. This makes more sense as in order to explain the initial sharp peak in deaths as seen in Figure 9. When case numbers are low, the mortality rate would need to be increased so that the deaths can be increased. Then when case numbers start to increase more the mortality rate can decrease because a lower proportion of a large number is usually considerably higher than a larger portion of a much smaller number. This is why this model was able to accurately model the 2nd peak in deaths towards the end of the simulation period in January to March of 2021. However despite the fairly accurate timings of the death curve it was unable to model the initial spike in deaths due to the low case numbers of the time. This could potentially be

due to under reporting at the start of the pandemic leading to considerably lower case numbers than what was actually available. The mismatch in the data at the time making it difficult for the optimiser to accurately manage both.[12]

However it is more likely to be an issue with the way the optimization scheme was setup. The objective function to be minimised was the RMSE, combined for both transmission and mortality. However the absolute number of deaths pales in comparison to the number of cases in the dataset, therefore the optimiser likely put a lot more weighting for fitting the transmission well and not the mortality rate.

4.5.3 SEIRDV with both Fourier transmission and mortality rate

A similar problem can be seen for the Fourier series transmission and mortality rate. In general however. This model performed worse in terms applicability to the real world. This can be seen in the Figure 11. The time-varying transmission rate initially peaks in May. Which was when the first UK lockdown was in full effect. Therefore the transmission rate should not be peaking in this period. Then in August of 2020 the transmission rate goes to 0. Which is not feasible at all. This result makes no sense and only exists because the optimiser used had a check to make sure the transmission rate couldn't be below 0. The fact that the transmission rate is 0 at any period of time makes no sense at all for any type of epidemiological model because such a result could only be achieved if every individual in a population isolates themselves from everyone else. After this minimum it rises again reaching another peak in October. This result makes more sense, but in October the UK was starting to see higher levels of restrictions after the summer period. If anything the transmission rate should be decreasing in this period. After October the transmission curve does start to decrease. Reaching another local minimum in February of 2021. Before starting to increase again at the end of the simulation. This increase in transmission could be attributed to the rise of different strains of COVID which were circulating at this period such as the alpha variant which was discovered in October of 2020 in the UK. Overall it should be noted that this time-varying transmission model did not produce a fitting that that aligned with the history of the pandemic in the UK.

The mortality curve $\mu(t)$ on the other hand is a slightly better result due to the fact that the peaks align more with the timeline of COVID in the UK. This can be seen in Figure 11. The mortality rate initially starts at 1% in March of 2020 and as a result. It peaked in May at just over 2%. This is slightly lagging behind the death data. That had its first initial peak in April, which can be seen in Figure 11. It then decreased, reaching a local minimum in August. Which does also coincide with the data in Figure 11. Where the weekly deaths in the UK from COVID was at it's lowest in this period from around July to late October. In October the mortality rate then increases again to around 2% and this is a good time to do this as this is the time which deaths start increasing again, as seen in Figure 11. The mortality rate then falls again, going as low as 0.25% at the start of 2021 before returning to 1% at the

end of the simulation, a result which follows the 2nd deaths peak falling again in an around this period as well. Taking this into account, despite the overall worse fitting for the model, the mortality rate does agree with the timeline of the pandemic in the UK.

4.6 Shortcomings of the model

4.6.1 Limitations of the model itself

This model makes many assumptions that limit the efficacy of its results. Firstly the model assumes homogenous mixing [21], in effect this means that everyone interacts equally, In real-life this is not true as some people are considerably more social than others. Next we have the fact that the compartments themselves have no age stratification in them. This implies that the model assumes the entire population is around the same age, such that age doesn't play a factor in the transmission or mortality rates. This is not true as both factors did depend heavily on age. Mortality especially was considerably higher amongst the elder population, while for children and young adults COVID-19 was a much milder illness. Furthermore there was no stratification for different ethnicities as well [26]. With people in the BAME (Black, Asian Minority Ethnicities) being more at risk to dying from COVID than white ethnic groups.

There was also the fact that for the SEIRDV models specifically used a fairly simple vaccination compartment. It assumed that people in the vaccinated compartment were unable to carry the disease or spread it to anyone else. This was also not true in real-life. The vaccination reduced the chance of serious complications from the virus. Thereby reducing the mortality rate. However vaccinated people were still liable to spread the disease to others. It did not give full immunity. The vaccinations themselves also waned in effectiveness over time as well. This is why multiple doses needed to be administered. Furthermore due to COVID being a virus. The mutations it undergoes leading to different strains needed new vaccinations to be developed. None of this was accounted for, instead adopting for a simple vaccination rate calculated from the real world daily vaccination data provided by UK government. Leading to the next limitation

4.6.2 Limitations of the dataset

The dataset itself had its own problems. Firstly we have the fact that the data is very unreliable at the start of the pandemic in March of 2020. Lateral flow tests were not readily available to the public at this time and would not be so until April of 2021, so case numbers at the time were only for confirmed cases in hospitals. Which is most likely considerably lower than what was really happening across the UK, especially since many people who carried the disease were asymptomatic and as a result unless they took a test they would have no idea they are infected. The changes in the levels of testing over the course of the time period used (March 2020 to March

2021) would have most likely lead to large discrepancies in the numbers of cases. This is most apparent in the raw data for cases and deaths. If we look at it directly at the cases and deaths in May of 2020 In Figure 1. It can be seen that the number of weekly cases is around 50 thousand cases. Meanwhile the deaths peaked at almost 10 thousand cases in the previous weeks. Indicating that a considerable amount of the people who were getting infected were dying from COVID. This was most likely not the case , instead the under reporting of cases most likely made the situation look considerably worse than what it was.

For mortality on the other hand this was documented much better due to the fact that people dying in the hospitals has no grounds for being misinterpreted like for cases. For COVID cases many people would be infected and still not report that they were. Most people would however go to the hospital if they were at risk of dying from COVID, thereby meaning the death data is considerably more reliable than the case data.

4.7 Problems with model calculation

4.7.1 Parameter estimation

The optimization was built around the idea of trying to minimise the RMSE. This was a fairly good approach for the basic SIR model. This is due to the fact that fitting to a single curve and only having a few parameters to optimise allows for the optimization algorithm to accurately find the best values for the parameters.

However for more complex cases like the SERIDV models, fitting 2 curves with around 16 parameters to optimise from the transmission and mortality rates due to the fittings chosen. Is a considerably more complex problem. There are many more local minima which the optimiser could get stuck in. As a result of this it was also very sensitive to the initial conditions and the bounds set. Requiring multiple runs with different initial guesses for the parameters and varying bounds to see if the final result was actually the best result it could have found or if the optimiser was being constrained by the bounds set.

Furthermore RMSE as the metric to optimise is probably not the best choice for a combined fitting for both the transmission and mortality rates due to RMSE being an absolute metric. The scales of cases and deaths were not the same. There were considerably more cases than deaths meaning when trying to optimise for both functions. It would more heavily skew towards optimising for the transmission and mortality. This was attempted to be rectified by adding weightings to the objective function which was made up of the RMSE of both the cases and deaths of the models. By trying to skew more heavily towards fitting the deaths curve by using some scaler multiple.

4.7.2 Fixed parameters

The values of certain parameters such as the incubation rate σ and the recovery rate γ were assumed to be fixed. However they could have varied over the course of the year due to different variants such as the alpha variant which was discovered in October of 2020 in the UK. Furthermore because of the lack of age and ethnic stratification within the compartments there was no further refinement into different rates for different groups as they could have also been subject to different rates.

4.8 Future work

In order to improve upon the work done in this study one could do any of the following

- Enhance the model: This could be done by adding stratification to each of the compartments to incorporate age, ethnicity or both. Adding new compartments like hospitalizations or using a more realistic effect of vaccination by incorporating waning effectiveness, varying effectiveness of the vaccine through different individuals in the population-could a person take the vaccine but fail to receive any of the benefits?
- Different parameter functions: Using different models for the mortality and transmission rate. such as any of the different models used for empirical fitting, even the ones which didn't fit well to the data, maybe they would work better for either of the rates.
- Optimization: using different optimization algorithms such as Bayesian hopping. Or perhaps changing the objective function to have different weightings or even changing the metric from RMSE to R^2 instead as that is a relative measure.
- Use more data: Incorporate different data streams into the modelling effort, data by age, infections by variant, larger time frames for analysis, data from other countries or locations within a country.
- Simulated noise functions: adding stochastic terms to the differential equation to try and model some of the real-life effects these models often over look.

5 Conclusion

In conclusion, this report has shown that compartmental models are indeed able to simulate real world data. Furthermore time-varying functions for the parameters within the differential equations that make up the compartmental models are needed to make sure the model can return a good result. Finally that adding additional compartments to the model does increase how well the compartmental model can recreate real world data. However there is still a lot more work that can be done to make these models even better. They are very much a key component in the

modelling of future epidemics despite the fact that they will never be perfect due to how complex the world truly is.

6 Annex

A Numerical Techniques and Uses

This annex provides a brief overview of several fundamental numerical techniques, outlining their basic principles and illustrating their specific application across various academic and non-academic domains.

A.1 Numerical Integration

Numerical integration, or quadrature, encompasses algorithms designed to approximate the value of a definite integral, $\int_a^b f(x)\,dx$. These methods are crucial when analytical solutions are unavailable, perhaps because the function $f(x)$ is too complex, or when $f(x)$ is known only at discrete points (e.g., from experimental measurements). Common techniques include the Trapezium rule, which approximates the area under the curve using straight-line segments between points, forming trapezoids, and Simpson's rule, which uses parabolic segments passing through sets of three points for a generally more accurate approximation, especially for smoother functions. These methods work by dividing the integration interval $[a, b]$ into smaller subintervals and summing the areas of the simple geometric shapes (trapezoids or parabolic sections) used to approximate the function over each subinterval. As the number of subintervals increase so too does the accuracy of the integral.

A.1.1 Academic Applications

1. **Physics and Engineering (Centre of Mass/Work):** Calculating physical quantities often involves integration. For instance, finding the centre of mass of an object with a non-uniform density distribution requires integrating the density multiplied by the position vector over the object's volume ($\boldsymbol{r}_{cm} = \frac{1}{M}\int \boldsymbol{r} dm$). Similarly, calculating the work done by a variable force $F(x)$ involves integrating the force over the displacement ($W = \int F(x)dx$). When the density function or the force function is complex or defined by data points, analytical integration is impossible, necessitating numerical methods like Simpson's rule or the Trapezoidal rule to approximate these integrals and obtain crucial physical parameters, applying the fundamental concepts outlined in physics texts [7].

2. **Probability and Statistics (Cumulative Probabilities):** Determining the probability that a continuous random variable X falls within a certain range ($P(c \leq X \leq d)$) involves integrating its probability density function (PDF), $f(x)$, over that range $\int_c^d f(x)\,dx$. This gives the area under the PDF curve, representing the cumulative probability. When the PDF is complex (e.g., derived empirically from data or a non-standard theoretical distribution) and lacks an analytical integral, numerical integration techniques are required for calculating these

probabilities or related quantities like expected values $E[X] = \int_{-\infty}^{\infty} f(x)\,dx$, forming the basis for applying probability theory [23].

3. **Computational Chemistry (Electron Density):** Calculating molecular properties often requires integrating functions over spatial coordinates. For example, determining the total number of electrons in a molecule involves integrating the electron density function, $\rho(\boldsymbol{r})$, over all space ($\int \rho(\boldsymbol{r})d\tau$). Evaluating electrostatic potentials or interaction energies also involves complex integrals. Due to the complexity of electron wave functions (especially for many-electron systems), these integrals are almost always computed numerically using sophisticated quadrature schemes (like multi-dimensional Gaussian quadrature) tailored to the specific forms of functions used in quantum chemistry [9].

A.1.2 Non-Academic Applications

1. **Finance (Option Pricing):** Pricing financial derivatives often involves calculating the expected value of their future payoff under a risk-neutral probability measure. For many options, especially exotic types (like Asian commodity options whose payoff depends on an average price), this expectation takes the form of an integral involving the probability distribution of the underlying asset. Since these distributions can be complex (e.g., incorporating stochastic volatility or jumps) and lack simple analytical forms, numerical integration methods (like Monte Carlo simulation or deterministic quadrature) are essential tools used by financial institutions to accurately value these instruments and manage risk [8].

2. **Computer Graphics (Global Illumination):** Creating realistic images requires simulating how light propagates and reflects within a scene. Calculating the illumination at a surface point involves integrating the incoming light radiance from all directions over the hemisphere above the point (the rendering equation). Due to the complexity of scenes (arbitrary geometry, materials), this integral is almost always intractable analytically. Monte Carlo integration techniques, such as path tracing, are widely used to stochastically sample light paths and approximate the integral, enabling realistic rendering effects like soft shadows and indirect lighting for a variety of applications such as games and animated movies. An example of this would be the advent of ray tracing in most modern games. [22].

3. **Medical Physics (Radiotherapy Dose Calculation):** In planning radiation therapy for cancer treatment, such as radiotherapy, it is critical to calculate the radiation dose delivered throughout the patient's body. Sophisticated algorithms compute the dose distribution by effectively integrating the energy deposited by radiation beams as they pass through tissues of varying densities (obtained from CAT scans). This involves complex integrals accounting for scattering and attenuation. Numerical methods are essential for accurately performing these integrations and ensuring the prescribed dose targets the tumour while minimising damage to healthy organs [14].

A.2 RK Methods for ODEs

Ordinary Differential Equations (ODEs) describe how quantities change over time or with respect to another independent variable, such as $\frac{dy}{dt} = f(t, y)$. Runge-Kutta (RK) methods are a cornerstone family of numerical techniques for approximating the solution $y(t)$ given an initial condition $y(t_0) = y_0$. They work by taking discrete steps Δt and using a weighted average of several evaluations of the derivative function $f(t, y)$ within each step to estimate the change in y. This use of intermediate derivative evaluations allows RK methods (especially higher-order ones like the popular RK4 or even the RK45 method used in this paper) to achieve greater accuracy and often better stability compared to simpler methods like Euler's method, which only uses the derivative at the beginning of the step. Finite difference methods are related, often used for discretising derivatives in space for solving Partial Differential Equations (PDEs), which can then be solved using ODE techniques in time.

A.2.1 Academic Applications

1. **Celestial Mechanics (Orbit Prediction):** Simulating the motion of celestial bodies (planets, moons, satellites, asteroids,etc) under mutual gravitational forces requires solving Newton's laws of motion, which form a system of second-order ODEs, $\frac{d^2\boldsymbol{r}}{dt^2} = \frac{\boldsymbol{F}(\boldsymbol{r})}{m}$. To predict trajectories accurately over long periods, especially in complex scenarios like multi-body systems or missions requiring high precision, robust numerical integration methods like high-order Runge-Kutta schemes (often with adaptive step-size control) are standard tools used by astronomers and space agencies [18].

2. **Chemical Kinetics (Reaction Simulation):** Modelling the time evolution of concentrations of different chemical species during a reaction involves solving a system of coupled first-order ODEs derived from the reaction rate laws, $\frac{d[C_i]}{dt} = \sum_j v_{ij} r_j$. For complex mechanisms, especially those involving chemicals with vastly different reaction rates (stiff systems), standard explicit RK methods may be unstable or inefficient. Specialized implicit RK methods or other stiff solvers are often required to accurately simulate the reaction progress and understand the underlying mechanism [1].

3. **Population Dynamics (Ecosystem Modelling):** Studying how populations of interacting species change over time often involves using systems of ODEs, such as the Lotka-Volterra predator-prey model: $\frac{dx}{dt} = \alpha x - \beta xy, \quad \frac{dy}{dt} = \delta xy - \gamma y$. Numerical methods like RK4 allow biologists to simulate these non-linear systems, explore scenarios beyond analytical tractability, observe population cycles, investigate the effects of parameter changes (e.g., carrying capacity, interaction strengths), and assess the stability of ecological models [19]. These equations were actually the starting point for the ODE's that defined the compartmental models discussed in this paper.

A.2.2 Non-Academic Applications

1. **Weather Forecasting and Climate Modelling:** Predicting weather and climate involves solving the complex PDEs (like the Navier-Stokes equations adapted for atmospheric/oceanic flow) that govern fluid motion, heat transfer, and other physical processes. These PDEs are typically discretized on a spatial grid (using finite differences, finite volumes, or spectral methods), resulting in a very large system of coupled ODEs describing the state variables (temperature, pressure, wind velocity, etc.) at each grid point over time. Numerical weather prediction centres run sophisticated models that integrate these ODE systems forward in time using stable and efficient time-stepping schemes, often related to RK or multi-step methods [11].

2. **Electrical Engineering (Circuit Simulation):** Analysing the transient behaviour of electrical circuits containing energy storage elements (capacitors and inductors) requires solving systems of ODEs derived from Kirchhoff's laws (e.g., $i = C\frac{dv}{dt}$, $v = L\frac{di}{dt}$). Circuit simulation programs like SPICE (Simulation Program with Integrated Circuit Emphasis) formulate these circuit equations and use robust numerical integration methods (often implicit methods like Trapezoidal rule or Backward Differentiation Formulas, related to implicit RK) to solve for voltages and currents over time, enabling engineers to verify designs before fabrication [24].

3. **Game Development (Physics Engines):** Real-time physics simulation is crucial for immersive video games. Physics engines simulate the motion, collision response, and constraints of objects (characters, vehicles, debris) by solving Newton's second law ($F = ma$, a second-order ODE). To achieve interactive frame rates, these engines typically use fast, explicit integration methods like Verlet integration or the semi-implicit Euler method. While potentially less accurate than high-order RK methods per step, they offer a good balance of performance and stability suitable for real-time simulation within the broader game engine architecture [4]. This method of calculation is one of the key reasons games were able to reach performance values above 120 frames per second.

A.3 Fourier Transforms&Time-Series

The Fourier Transform (FT) is a powerful mathematical operation that decomposes a signal (like a sound wave or an economic time series) into a spectrum of its constituent frequencies. It essentially reveals "how much" of each frequency is present in the original signal, transforming the signal from the time domain (amplitude vs. time) to the frequency domain (amplitude/phase vs. frequency). This is invaluable for identifying periodic patterns or analysing the frequency content of signals. The Fast Fourier Transform (FFT) is an extremely efficient algorithm for computing the Discrete Fourier Transform (DFT), which is the version applied to digitally sampled signals, making frequency analysis computationally feasible for practical applications.

A.3.1 Academic Applications

1. **Signal Processing Research (Filtering and Analysis):** The FT is a cornerstone of digital signal processing. Researchers use it extensively to analyse signals by examining their frequency spectra, identify dominant frequencies or noise components, and design digital filters. For example, a low-pass filter can be designed by specifying the desired cut-off frequency in the frequency domain, converting this specification back to the time domain (an impulse response) often using the inverse FT, and then implementing the filter via convolution [20].

2. **Astrophysics (Periodicity Search):** Analysing time-series data from astronomical observations, such as the brightness measurements of a star over time (its light curve), frequently employs Fourier techniques to search for periodicities. Detecting a stable period in a star's light curve can reveal its rotation rate, pulsations, or the presence of an orbiting companion (like an eclipsing binary star or an exoplanet). Specialized algorithms like the Lomb-Scargle periodogram are often used for unevenly sampled astronomical data [25].

3. **Quantum Mechanics (Momentum Representation):** In quantum mechanics, a particle's state can be described by a wave function in position space, $\psi(x)$, or equivalently by a wave function in momentum space, $\phi(p)$. These two representations are fundamentally linked by the Fourier transform: $\phi(p)$ is the Fourier transform of $\psi(x)$, and vice versa. This relationship is crucial for understanding the Heisenberg Uncertainty Principle and for solving problems where calculations are simpler in one representation than the other [5].

A.3.2 Non-Academic Applications

1. **Audio Engineering and Music Technology (Equalization/Compression):** The FFT (Fast Fourier Transform) is ubiquitous in audio processing. Audio equalizers work by transforming a segment of audio into the frequency domain using FFT, multiplying the frequency components by desired gains (boosting bass, cutting treble), and transforming back using inverse FFT. Lossy audio compression formats like MP3 and AAC also rely heavily on frequency transforms (often the Modified Discrete Cosine Transform, related to FFT) to identify and discard frequency components that are less perceptible to human hearing, achieving significant data reduction [27]. Leading to smaller file sizes for audio files using those extensions.

2. **Image Processing (Compression/Filtering):** The 2D Fourier Transform is applied in image processing for tasks like compression and filtering. The JPEG compression standard uses the Discrete Cosine Transform (DCT, related to FT) on 8×8 image blocks to concentrate most image energy into a few low-frequency coefficients, which can be stored more efficiently. Filtering in the frequency domain (e.g., multiplying the 2D FFT of an image by a filter mask and

inverse transforming) can be efficient for tasks like removing periodic noise patterns or performing certain types of blurring or sharpening [3].

3. **Medical Imaging (MRI Reconstruction):** Magnetic Resonance Imaging (MRI) works by exciting atomic nuclei in a magnetic field and measuring the radio frequency signals they emit as they relax. The clever use of magnetic field gradients ensures that the frequency and phase of these signals encode spatial information about their origin. The raw data collected exists in ”k-space,” which is essentially the 2D or 3D spatial frequency domain of the image. The final, visible anatomical image is reconstructed by applying an Inverse Fourier Transform (computed via IFFT) to this k-space data.[6].

References

[1] Peter Atkins and Julio de Paula. *Atkins' Physical Chemistry*. Oxford University Press, 2014. pages 39

[2] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5), 1995. doi: 10.1137/0916069. pages 16

[3] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 2017. pages 42

[4] Jason Gregory. *Game Engine Architecture*. CRC Press, 2014. pages 40

[5] David J. Griffiths and Darrell F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, 2018. pages 41

[6] E. Mark Haacke, Robert W. Brown, Michael R. Thompson, and Ramesh Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley-Liss, 1999. pages 42

[7] David Halliday, Robert Resnick, and Jearl Walker. *Fundamentals of Physics*. Wiley, 2013. pages 37

[8] John C. Hull. *Options, Futures, and Other Derivatives*. Pearson, 2018. pages 38

[9] Frank Jensen. *Introduction to Computational Chemistry*. Wiley, 2017. pages 38

[10] Helen Johnson. The UK's Covid vaccination rollout timeline at a glance, 01 2021. URL `https://www.yorkshirepost.co.uk/health/coronavirus/the-uks-covid-vaccination-rollout-timeline-at-a-glance-3091431`. pages 12

[11] Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003. pages 40

[12] Matthew James Keeling, Louise Dyson, G. F. Guyver-Fletcher, A. Holmes, Malcolm G. Semple, Michael J. Tildesley, and Edward M. Hill. Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number. *Statistical Methods in Medical Research*, 29(11), Nov 2020. doi: 10.1177/0962280220942655. pages 32

[13] William Ogilvy Kermack and Anderson G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772), 1927. doi: 10.1098/rspa.1927.0118. pages 5

[14] Faiz M. Khan and John P. Gibbons. *Khan's The Physics of Radiation Therapy*. Lippincott Williams & Wilkins, 2014. pages 38

[15] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172, 03 2020. doi: 10.7326/M20-0504. pages 14

[16] Rui Li, Chunlin Lei, and Jing Han. Compartmental models of the COVID-19 pandemic: A systematic literature review. *International Journal of Environmental Research and Public Health*, 20(5), Mar 2023. doi: 10.3390/ijerph20054372. pages 5

[17] Hridoy MB. An exploration of modeling approaches for capturing seasonal transmission in stochastic epidemic models. *Mathematical Biosciences and Engineering*, 21(5), 2024. doi: 10.3934/mbe.2024339. Also available as arXiv:2410.16664 [q-bio.PE]. pages 13

[18] Carl D. Murray and Stanley F. Dermott. *Solar System Dynamics*. Cambridge University Press, 1999. pages 39

[19] James D. Murray. *Mathematical Biology I: An Introduction*. Springer, 2002. pages 39

[20] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 2009. pages 41

[21] Lorenzo Pellis, Frank Ball, Shweta Bansal, Ken Eames, Thomas House, Valerie Isham, and Pieter Trapman. Eight challenges for network epidemic models. *Epidemics*, 10, Mar 2015. doi: 10.1016/j.epidem.2014.07.003. pages 33

[22] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory To Implementation*. Morgan Kaufmann, 2016. pages 38

[23] Sheldon M. Ross. *A First Course in Probability*. Pearson, 2014. pages 38

[24] Matthew N. O. Sadiku and Charles K. Alexander. *Fundamentals of Electric Circuits*. McGraw-Hill Education, 2012. pages 40

[25] Christiaan Sterken and Carlos Jaschek, editors. *Light Curves of Variable Stars: A Pictorial Atlas*. Cambridge University Press, 2005. pages 41

[26] UCL. Lessons must be learnt from Covid-19's unequal impact on minority groups, 07 2023. URL `https://www.ucl.ac.uk/news/2023/jul/lessons-must-be-learnt-covid-19s-unequal-impact-minority-groups`. pages 33

[27] Udo Zölzer, editor. *DAFX: Digital Audio Effects*. Wiley, 2011. pages 41