COVID-19 Cases Estimation in the UK using Improved SEIR Models

|  |  |  |
| --- | --- | --- |
| Shaogang Jiang *Department of Electrical Engineering and Electronics* *University of Liverpool* Liverpool, UKs.jiang19@liverpool.ac.uk | Ali Al-Ataby*Department of Electrical Engineering and Electronics* *University of Liverpool* Liverpool, UKaliataby@liverpool.ac.uk | Fawzi Al-Naima*Department of Computer Engineering* *Al-Nahrain University* Baghdad, Iraqfawzi.m.alnaima@nahrainuniv.edu.iq  |

*Abstract*—The paper suggests a machine learning algorithm with two modified SEIR models customized for the 2019-nCoV virus and vaccine uses to simulate the spread of COVID-19 in the UK (from Jan 2020 to March 2021) and make predictions of future cases. The algorithm uses COVID daily cumulative case data and second dose vaccine use data provided by the Public Health England as the training set and is capable of making relatively accurate short-term predictions of future COVID cases in the UK (before the delta and later variants of the virus starts spreading within the country). The obtained overall accuracy is above 80% for daily incremental case numbers in terms of the overall fit of the model to real-life data, and with an accuracy of more than 80% for estimation of daily incremental case numbers for 14 days period future prediction. The goal of this paper is to propose improved SEIR models capable of a more accurate simulation for COVID-19 modelling and estimation with various machine learning algorithms.

Keywords—COVID-19, SEIR, machine learning, estimation

# Introduction

The nCoV-2019 virus outbreak (aka COVID-19 pandemic) originally started in Hubei province-China in December 2019 and spread globally in early 2020. The virus is known to cause pneumonia like symptoms and is potentially fatal. The WHO declared it a global pandemic on 12th March 2020, and it is still an ongoing situation by the time this paper is being written.

The outbreak in the UK started on 29th January 2020 when two infected individuals from overseas arrived to the UK, and the virus started spreading within the country since then. The situation quickly deteriorated through February and March of the same year. On 25th March 2020, the British government announced its first national lockdown, and social distancing protocols were in place since then. The daily case number reached its initial peak in April and gradually decreased through June and July, but the situation was not completely under control in summer 2020 with a relatively small number of infected cases every day. In September 2020, the case number rose again and reached a new higher peak during Christmas period, and only started to decrease again after mid-January 2021 and in early 2021. Later, the UK saw a gradual decrease in newly reported case number from April to June, and the case number rose again partly due to new variants [1] of the virus and the reopening of the country. Fig. 1 shows UK daily new case number distribution in the period from March 2020 to July 2021.

Vaccine distribution in the UK started in late 2020 with multiple different vaccines, and they are being widely distributed from early 2021, with over 47 million people having received their first dose and around 38 million people having received the second dose by 31st July 2021 [2].

The goal of this paper is to propose an improved SEIR model capable of a more accurate simulation for COVID-19 modelling and estimation with various machine learning algorithms. This paper is organized as follows: Section II provides a literature review, Section III suggests the methodology used in this study, Section IV displays the results of the study, Section V provides discussions, and Section VI sums up the conclusions.



Fig. 1 UK daily new case number distribution [1]

# Literature Review

 Since the beginning of COVID-19 epidemic, multiple organizations in different countries have developed machine learning algorithms to estimate COVID cases. Most notably, a modified SEIR model and AI prediction of the epidemic trends of COVID-19 in China under public health interventions [3] were used to estimate the initial outbreak in China published in February 2020 using Long-Short-Term-Memory (LSTM) model method [4], which accurately estimated the turning point of the outbreak in late March 2020 and the final COVID-19 case number in China before the spread of the virus (comparison with real life Chinese COVID case number proved that the estimation was accurate). Cases in India was accurately estimated through machine learning study [4] until August (also proved accurate by real life Indian COVID cases data) [5].

 There has also been researches in the US that focuses on the estimation of COVID case number. A research carried out by CDC MInD-Healthcare Program [6] used an individual-based SEIR model to estimate the spread of COVID in the District of Columbia. This research focuses on the modelling of disease spread through a human-to-human contact network using a variation of the SEIR model that focuses on individual contact and action. The result is accurate until May 2020. Another research done in Italy and Spain [7] and the ones by Mwalili, et al. [8] and Grimm et al. [9] expanded the original 4-dimension-ODE SEIR model to higher dimensions to incorporate effects such as government intervention, social distancing, and the use of personal protective equipment (PPE). And they focus on simulating different government intervention scenarios, and added the effects of death, quarantine and the existence of pathogen in the community into the SEIR model. There has also been researches that tried to separate symptomatic and asymptomatic patients within the SEIR model [10], as these two groups behave differently during the spread cycle. Fig. 2 shows SEIR and LSTM Simulation of case number in China [3]. Fig. 3 shows the results of an individual-based SEIR simulation of COVID cases in the District of Columbia, USA [6].

 The research in this paper is inspired by these previous researches and aims to extend the SEIR models in a similar fashion to incorporate multiple aspects uncovered by the original SEIR model, especially the effect of nationwide vaccination in the UK [1].



Fig. 2 SEIR and LSTM simulation of the case number in China [3]



Fig. 3 Results of an individual based SEIR simulation of COVID cases in the District of Columbia, USA [6]

# Methodology

This study aims at developing an improved version of the SEIR model specifically for COVID simulation. For implementation of the models with machine learning algorithm, this study develops the algorithm in MATLAB environment with its ode45 function that uses Levenberg-Marquardt learning method [11] for the context of this work. For the simulation of vaccine use number, a polynomial function was used with relatively high accuracy (above 92%) compared to the vaccine use data from March to April 2021.

## SEIR Model

The standard SEIR model stands for Susceptible, Exposed, Infected and Removed model [12]. It is one of the models derived from the SI model [13], with two additional factors: exposed and removed for simulation of viruses that behaves like nCoV-2019. Mathematically, the model can be expressed with a system of four first-order Ordinary Differential Equations (ODE):

 $\frac{dS}{dt}=-βIS$

$\frac{dI}{dt}=βαE-λI$

$\frac{dE}{dt}=βIS-αE$

$\frac{dR}{dt}=λI$

Where *S*, *E*, *I*, *R* denote the number of susceptible, exposed, infected, and removed patients, and *β*, *α*, *λ* are coefficients related to the rate of infection, virus latency and rate of removing. This work has developed improved versions of the SEIR model, called SEIRM (without vaccine use) and SEIRV (with vaccine use).

## SEIRM Model

In SEIRM, a new factor was introduced: the reinfection rate *τ*. This is because some COVID cases can be released from isolation with false negative test result, or that patients who have recovered from COVID may fail to develop antibodies or may come in contact with a different strain of the nCoV-2019 virus that the patient is not immune to, causing the recovered individual leaving the removed population. Mathematically, the changes were made to equations (1) and (2), the susceptible and infected population.

 $\frac{dS}{dt}=-βIS+τR$ (5)

 $\frac{dI}{dt}=βαE-λI+τR$ (6)

Where *τ* is the reinfection rate. An individual from the *R* population has a chance to fail in developing antibodies or come in contact with a different strain, this would make the individual susceptible to the virus again; this transition is represented by the term *τR* added to the *S* population rate of change. If the individual is released from isolation with false negative teste result, then the individual is still carrying the virus and is still capable of infecting others, moving the individual back to the *I* population, hence the term *τR* added to the *I* population rate of change. In the SEIRM model, the transmission cycle does not end at *R* population, and any member of this population has a small chance to go back to previous population groups, which can result in a slower end to the outbreak. In a completely uncontrolled environment without vaccination and containment effort, the spread of the virus continues indefinitely. This is because the removed cases can always become susceptible and infected again due to virus mutation, so the spread will never stop and the population will never be immune to the pathogen and will be repeatedly infected by new strains. However, not all individuals go directly back to the S population after recovery from the disease, in SEIRM, only a small portion of the recovered individual will become susceptible again at a time, which is the case for COVID-19 at this stage of mutation.

## SEIRV Model

Starting from late 2020, a few COVID vaccines started to see distribution in the UK, and the British government spent a lot of effort in the mass distribution of vaccines within the country. This could have a significant effect on the spread of the vaccine, thus a separate model with a vaccine factor was developed to simulate the spread of COVID-19 cases starting from January 2021. The only difference between SEIRV model and SEIRM model is in the *S* population. The model can be mathematically expressed as:

$\frac{dS}{dt}=-βIS+τR-R\_{o}V$

Where *Ro* is the vaccine rollout and *V* is the size of vaccinated population. If an individual that is susceptible to the pathogen received vaccine, there is a chance that the individual will no longer be infected by the pathogen even though the individual has never gone through the S-E-I-R transmission cycle. This is represented by the term *RoV* subtracted from the susceptible population change *dS/dt*. By decreasing the *S* population size, the outbreak can be contained quickly with less people being infected. In this study, for the sake of simplicity and stability, individuals who have completed COVID-19 vaccination process regardless of the type of vaccine are all put into one single *V* group, with vaccine rollout *Ro* calculated from the rollout of all vaccines currently being distributed in the UK. Comparing the result between SEIRV and SEIRM, it can be observed that the introduction of vaccine will cause a lower number of peak exposed, infected and removed cases, and a slower increase in these cases, and will cause a quicker decrease of susceptible cases, which helps with the containment of a pathogen. Fig. 4 shows a comparison of the behavior curves for SEIRM, SEIRV, and SEIR models, in terms of susceptible, removed, exposed, and infected populations.

The proposed algorithm uses the SEIRM/SEIRV models in a 7-stage simulation of UK COVID-19 spread. Thus, the situation in the UK is divided into 7 different stages from 12th Jan 2020 to 3rd Apr 2021. For estimation purposes, this paper will focus on the last stage, from 11th Jan 2021 till 5th April 2021.

## Randomizer and Adjustment

Adjustments were made based on the situation during March and April of 2021 so that the simulated result can fit more closely to the actual situation after 5th Mar. This is done by combining Logistic Growth Model [14] with a randomizer to simulate additional cases caused by the rise in case number in Europe.

Mathematically, the randomizer can be expressed as:

 $R\_{m}=Random\left(1,k\right)×\left(δ-\frac{δ}{1+e^{-μt}}\right)$

where *δ* and *µ* are the logistic growth coefficients, *k* is maximum random imported case number and Random (1, *k*) is a random integer ranging from 1 to *k*.

The coefficient *δ* and *µ* and the maximum imported case number are manually adjusted to optimize the accuracy of a 14-day simulation and also a 26-day simulation. Currently, the parameters are: *δ* = 5, *µ* = 1/7, and *k* = 500. The logistic growth model used in the randomizer is inverted, so that as time progresses, the random additional case number decreases until it reaches to 0. For smaller amount of random infected cases, the randomizer can be used directly on the *I* population via addition of *Rm* to the *I* population in the ODEs. For stability issues, the randomizer is added to the *R* population for estimation of cases in March and April 2021.









Fig. 4 Behavior curves of SEIRM, SEIRV, and SEIR models for susceptible, exposed, infected, and removed populations

# Results

The algorithm is run with dataset updated on 5th March 2021 from source [1]. Simulated results may vary with different dataset updated at different time or from different countries, and accuracy can be affected by changes to the situation that the algorithm is unable to adapt to, so the result in this section is not final and is subjected to change in the future with different situation. Due to the changes in the COVID-19 situation in the UK in late March 2021 with rise in case numbers in Europe, and the third big rise in case number in July 2021, the current algorithm may not be able to accurately simulate recent COVID-19 spread in April 2021 without some extended modification. The dataset used for the machine learning algorithm was updated on 5th Mar 2021, and simulation was run for 14 and 26 days.

## Reference Data

This is the real-life case number from source [1] about UK COVID-19 case number from 6th Mar to 31st Mar 2021. The daily cases dataset used in this study is case number by case specimen date [1]. Table 1 provides this data set.

Table 1. Incremental Real-life Case Number During the 26-day Simulation [1]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Date** | 06-03-21 | 07-03-21 | 08-03-21 | 09-03-21 | 10-03-21 | 11-03-21 | 12-03-21 |
| **Cases** | 4,529 | 4,336 | 6,837 | 6,366 | 6,166 | 5,914 | 5,702 |
| **Date** | 13-03-21 | 14-03-21 | 15-03-21 | 16-03-21 | 17-03-21 | 18-03-21 | 19-03-21 |
| **Cases** | 4,271 | 4,374 | 6,681 | 5,809 | 5,912 | 5,611 | 5,096 |
| **Date** | 20-03-21 | 21-03-21 | 22-03-21 | 23-03-21 | 24-03-21 | 25-03-21 | 26-03-21 |
| **Cases** | 4,139 | 5,480 | 6,427 | 5,283 | 5,978 | 5,404 | 4,455 |
| **Date** | 27-03-21 | 28-03-21 | 29-03-21 | 30-03-21 | 31-03-21 |  |  |
| **Cases** | 3,533 | 4,954 | 3,956 | 3,487 | 3,444 |  |  |

## Simulation Results

The results of a 14-day simulation until 19th of March are shown in Table 2.

Table 2. Daily Case Number From 14-day Simulation Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 06-03-21 | 07-03-21 | 08-03-21 | 09-03-21 | 10-03-21 | 11-03-21 | 12-03-21 |
| Cases | 6,874 | 6,609 | 6,872 | 6,131 | 6,303 | 6,017 | 6,124 |
| Date | 13-03-21 | 14-03-21 | 15-03-21 | 16-03-21 | 17-03-21 | 18-03-21 | 19-03-21 |
| Cases | 5,667 | 5,354 | 5,415 | 5,139 | 5,095 | 4,913 | 4,635 |

Accordingly, the average error of a 14-day simulation is 17.2%, with average accuracy of 82.8%.

The results of a 26-day simulation give simulation of the case number after 19th March are shown in Table 3.

Table 3. Daily Case Number From 20th to 31st for 26-day simulation results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 20-03-21 | 21-03-21 | 22-03-21 | 23-03-21 | 24-03-21 | 25-03-21 | 26-03-21 |
| Cases | 4,613 | 4,403 | 4,280 | 4,107 | 3,960 | 3,902 | 3,734 |
| Date | 27-03-21 | 28-03-21 | 29-03-21 | 30-03-21 | 31-03-21 |  |  |
| Cases | 3,594 | 3,479 | 3,350 | 3,223 | 3,120 |  |  |

Accordingly, the average error combined with simulation from the previous 14 days (6th Mar to 19th Mar) were 18.1%, with an average accuracy of 81.9%. Also, because the involvement of a randomizer, 10 test runs were conducted and the average error from the 10 runs was 18.4%, with an average accuracy of 81.6%. Fig. 5 shows the simulated number of daily cases versus the actual daily cases for comparison purposes.



Fig. 5 Daily cases simulation versus actual case number for the 26-day simulation period

# Discussion

Since the dataset used in this study was updated on 5th Mar 2021, by the time this work is being written (Apr 2021), there are not enough data to analyze the accuracy of long-term estimation. The model used in the simulation of the last stage is the SEIRV model. Mathematically for a SEIRV model, once new case number starts to fall, there will be an exponential decrease until there are no newer cases. The rate of decrease and how soon will the outbreak stop largely depends on the infection rate, removal rate and vaccine use number.

For this simulation, the vaccine use was estimated using a polynomial function, which will increase rapidly until all 66 million UK residents are vaccinated. With the current polynomial, the entire UK population will be all vaccinated before 1st July 2021. This is not achieved as only around 50% of UK population has been fully vaccinated (second dose) by 21st July [2]. The difference mainly comes from the use of a simple polynomial function for the estimation of vaccine use, which cannot accurately simulate the complicated situation with vaccine use in the UK. Because the result of long-term estimation largely depends on the vaccine use, so depending on different scenarios related to vaccine distribution, there are different results. For this, the simulation was run a few times with different settings:

1. Setting A: Current situation, the same as the algorithm explained previously.
2. Setting B: Faster vaccine rollout at 140% of the original rollout.
3. Setting C: Slower vaccine distribution with lower effectiveness, with the vaccine rollout at 60% of the original rollout.
4. Setting D: Same vaccine setting as setting A, but with another randomizer operating on the *I* population size to simulate imported cases from abroad at later stage.

For stability issues, the simulation of a faster or slower vaccine distribution is done by adjusting the rollout instead of the vaccine simulation model. Mathematically with the SEIRV model, they have the same effect.

The results of simulations obtained from each setting will be published in a future publication.

It is worth mentioning that due to the COVID outbreak being an ongoing situation, by the time this paper is prepared, some aspect (especially the dataset and estimated cases) could be outdated, and the situation cannot be guaranteed to be going as the estimation shows. For example, due to the recent “Delta” variant of the virus, there has been a big rise in case numbers in European nations [15], which is not foreseen when the learning algorithm is being developed in early 2021, thus this cannot be simulated. But this is mostly due to the structure of the machine learning algorithm, and the model is applicable to the new situation with proper algorithm.

# Conclusions

In this study, a machine learning algorithm was developed that is able to predict future COVID-19 cases in the UK, and also compatible with dataset from other countries. The accuracy for a 14-day and 26-day simulation are all above 80%, which exceeds the target set in the specification (80%). Also, along with the machine learning algorithm, modifications were done to the SEIR model; the SEIRM and SEIRV models were developed specifically for COVID-19 simulation that can simulate situation that is more complex, such as infected case coming from abroad, virus mutation that makes recovered cases susceptible again, etc., and most importantly, the SEIRV model that brings the use of vaccine to the SEIR model. Many previous research studies about COVID-19 prediction were done when COVID vaccine has not been developed, so introducing the SEIRV model can help in future studies in the same field after vaccine starts to be distributed in certain countries.

Also, the change in parameters shows that although government effort has lowered the infection rate β greatly, there is still work to be done to increase the removal rate. The result of the study shows that it is important to test enough people in order to contain the outbreak. These findings can help in determining future tactics for disease intervention. In addition, the modified SEIR models can hopefully provide insights for future modification to model more accurately the nCoV-2019 or other SARS like pathogens.

Last but not least, due to the COVID outbreak being an ongoing and constantly changing situation, by the time this paper is written, some aspect (e.g. dataset and estimated cases) could be outdated, and the situation cannot be guaranteed to be going as anticipated. Moreover, due to the recent new variants of the virus, there has been a big rise in case numbers in many countries, which is not foreseen when the learning algorithm is being developed. However, the model could be adapted to the new situation by updating the structure of the machine learning algorithm.

##### References

|  |  |
| --- | --- |
| [1]  | GOV.UK, “Coronavirus (COVID-19) in the UK,” GOV.UK, 21st July 2021. [Online]. Available: https://coronavirus.data.gov.uk/details/cases. [Accessed 21st July 2021]. |
| [2]  | “Coronavirus (COVID-19) Vaccinations,” Statistics and Research, 31st Jul 2021. [Online]. Available: https://ourworldindata.org/covid-vaccinations?country=GBR. [Accessed 31st Jul 2021]. |
| [3]  | Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, and M. Zanin, “Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions,” *Journal of Thoracic Disease,* vol. 12, no. 3, pp. 165-174, 2020.  |
| [4]  | A. Croeze, L. Pittman, and W. Reynolds, “Nonlinear Least-Squares Problems with the Gauss-Newton and Levenberg-Marquardt Methods,” 2018. [Online]. Available: http://fourier.eng.hmc.edu/e176/lectures/NM/node36.html. [Accessed 10 Apr 2021]. |
| [5]  | H. Verma, A. Gupta, and U. Niranjan, “Analysis of COVID-19 cases in India through Machine Learning: A Study of Intervention,” *Europe PMC,* p. PPR: PPR270998, 2020.  |
| [6] | D. K. Sewell, and A. Miller, “Simulation-free Estimation of an Individual-based SEIR Model for Evaluating Nonpharmaceutical Interventions with an Application to COVID-19 in the District of Columbia,” CDC MInD-Healthcare Program. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241949. |
| [7] | L. López, and X. Rodó, “A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics,” Results in Physics, Volume 21, 2021, Elsevier, [Online]. Available: https://doi.org/10.1016/j.rinp.2020.103746 |
| [8] | S. Mwalili, M. Kimathi, V. Ojiambo, D. Gathungu, and R. Mbogo*.* “SEIR model for COVID-19 dynamics incorporating the environment and social distancing,” *BMC Res Notes* 13**,**352 (2020). [Online]. Available: https://doi.org/10.1186/s13104-020-05192-1. |
| [9] | V. Grimm, F. Mengel, and M. Schmidt, “Extensions of the SEIR model for the analysis of tailored social distancing and tracing approaches to cope with COVID-19,” *Sci Rep* 11**,**4214 (2021). [Online], Available: <https://doi.org/10.1038/s41598-021-83540-2>. |
| [10] | J. P. Arcede, Randy L. Caga-anan, Ch. Q. Mentuda, and Y. Mammeri, “Accounting for Symptomatic and Asymptomatic in a SEIR-type model of COVID-19,”Math. Model. Nat. Phenom. 15 34 (2020), [Online], Available: https://doi.org/10.1051/mmnp/2020021. |
| [11]  | H. P. Gavin, “The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems,” 18 September 2020. [Online]. Available: http://people.duke.edu/~hpgavin/ce281/lm.pdf. [Accessed 08 Mar 2021]. |
| [12]  | P. Chausse, “The Susceptible, Exposed, Infectious, Removed (SEIR) Model,” University of Waterloo, 2020. [Online]. Available: http://www.arts.uwaterloo.ca/~pchausse/seir. [Accessed 20 May 2021]. |
| [13]  | M. Mohtashemi, P. Szolovits, J. Dunyak, and K. Mandl, “A susceptible-infected model of early detection of respiratory infection outbreaks on a background of influenza,” *J Theor Biol,* vol. 4, no. 241, pp. 954-963, 2006.  |
| [14]  | “Logistic Growth Model,” Duke University, 2000. [Online]. Available: https://services.math.duke.edu/education/ccp/materials/diffeq/logistic/logi1.html. [Accessed 19 Mar 2021]. |
| [15]  | WHO, “WHO Coronavirus (COVID-19) Dashboard,” WHO, 2021. [Online]. Available: https://covid19.who.int/. [Accessed 2nd Aug 2021]. |