Using Genetic Algorithms for Predicting the Parameters of a COVID-19 SEIR Model

Luna Pianesi, Raffaele Pojer
{luna.pianesi, raffaele.pojer}@studenti.unitn.it

***Abstract*—We investigated the use of genetic algorithms to predict the values of compartmental models parameters for COVID-19. The infection can be modelled with a mathematical abstraction that takes the form of a system of ordinary differential equations. Computational evolution strategies are able to refine the values of such parameters in a fast, efficient and accurate manner. We present some experiments made within this framework and empirically demonstrate the successful use of evolutionary computation in a setting borrowed from natural phenomena.**

I. Introduction

In the past three years, we have witnessed the spread of the COVID-19 pandemics worldwide. The mathematical modelling of this infectious disease played a fundamental role in helping governmental decisions, from the comprehension of the virus' behaviour up to the definition of exit strategies. Computational tools came in help of infectious disease experts to elaborate automated ways to predict the ongoing of the pandemic and to simulate the enforcement of response measures. Governmental decisions around the globe mostly consisted in restriction measures including social distancing, the use of facial masks and prolonged quarantine periods; for this reason, multiple computational approaches focused on the modelling and prediction of success of exit strategies from the pandemics [1][2]. Another line of research was investigating the possibility of automated modelling and prediction of the COVID-19 behaviour on a long term perspective [3][4]. This project focuses on the second framework and aims at experimenting with the use of Genetic Algorithms (GAs) and Multi-Objective Optimization (MOO) for estimating the parameters of a system of Ordinary Differential Equations (ODEs) describing the behaviour of the COVID-19 pandemics.

II. Problem definition

We define the mathematical modelling of infectious diseases as a collection of possible mathematical formulations, based on assumptions and collected data, serving as a tool to inform public health interventions and domain experts on the progress of a pandemics. One of the most straightforward approaches to a modelling that is subject to a large variety of factors is the use of compartmental models. Compartmental models allow the designers to classify a population according to compartmentalized labels, so as to simplify the task and at the same time providing a useful viewpoint on the spread of an epidemics. One particular model type is often times used to model the spread of different infections: these all share the feature of inducing a latency period from the time of infection to the time the individual becomes actually infectious. The infection caused by the virus SARS-CoV-2 falls in this category, thus it can be modelled by means of the so-called SEIR model. The name of the model is an acronym that stands for Susceptible, Exposed, Infectious, Recovered. The model is a system of ODEs where S represents the target-time histories of the susceptible cases, E the target-time histories of the exposed cases, I the target-time histories of the infectious cases and R the target-time histories of the recovered cases. The simplest possible compartmental model is the SIR model (Equation 1):

$$\begin{aligned} \frac{\partial S}{\partial t} &= -\frac{\beta I S}{N} \\ \frac{\partial I}{\partial t} &= \frac{\beta I S}{N} - \gamma I \\ \frac{\partial R}{\partial t} &= \gamma I \end{aligned} \tag{1}$$

where S indicated the population of susceptible individuals, I that of infectious individuals and R that of recovered individuals. An important feature of compartmental models is that

$$\frac{\partial S}{\partial t} + \frac{\partial I}{\partial t} + \frac{\partial R}{\partial t} = 0,$$

from which directly follows that $S(t)+I(t)+R(t) = N$, with N being the total number of individuals in the population. By tweaking the parameters of model it is possible to construct predictions over the progress of the spread of an infectious disease. Instead of manually tuning the parameters of the chosen model, we exploit the power of Evolutionary Computation (EC) to implement a Genetic Algorithm (GA) pipeline to generate, variate and select individuals in a population representing the possible parameters of the model. We then rely on `scipy.integrate.odeint()` to find the solutions to the system of ODEs. Our fitness function is a linear combination of the Root Mean Square Deviation (RMSD) of the number of deaths, infections and recoveries between the ground truth and the value predicted by our SEIR model according to the evolved parameters. We chose to experiment with a SEIRD and a modified SEIRD models, always focusing on COVID-19 data from the Autonomous Province of Trento published and daily updated by Presidenza del Consiglio dei Ministri — Department of Protezione Civile at this link[1]. We assume the population number to be 542,166 as surveyed by ISTAT on the 31st December 2020[2].

[1]https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni
[2]https://www.istat.it/it/archivio/270440

III. Related works

In the context of the COVID-19 pandemics, an almost uncountable number of publications have been worked on, using a large variety of methods to address the modelling and prediction of the ongoing epidemics.

A. SEIR-modelling epidemics with Genetic Algorithms

Many efforts have been directed towards the design of automated methods to predict the ongoing of the pandemics. Many publications were focused on the use of GAs to to adjust the population-dependent parameters of some version of the SEIR model to ground truth data. One particular contribution by Yarsky (2021) [3] was centered on the prediction of cases and deaths caused by the virus. A standard GA pipeline is implemented, including: definition of genome and fitness function, selection, reproduction and mutation steps. The author observes how the model tends to under-predict the number of cases and casualties early in the outbreak, probably due to a bias present in the data. Nonetheless, results show that the algorithm is able to predict a reasonable curve that very much follows the one given by real data; this demonstrates the usefulness of embedding some kind of evolutionary computation when dealing with domains that are borrowed or belong to natural evolution of any kind.

B. SEIR-modelling epidemics with Swarm Intelligence

One of the main contributions over the classical SEIR model is provided by Godio et al. (2020) [5], whose innovation resides in the use of the stochastic approach known as Particle Swarm Optimization (PSO) to solve the model and to assess the propagation of the uncertainties of the model solution. The main objective of this work is indeed to improve the classical SEIR model by means of a stochastic solver which identifies a set of possible solutions (or most probable scenarios) predicting the epidemic evolution with the associated uncertainty assessment. The authors use a least squares criterion to fit the parameters to the observed data, first using a deterministic approach, then using a PSO solver. The results of the study suggest that a deterministic approach is not well-fitted to this underdetermined problem, while a stochastic approach like PSO provides better performances being able to handle this characteristic.

IV. Method

A. Data

Regional data from the official repository of Department of Protezione Civile is retrieved automatically for an arbitrarily specified timeframe. The data is cached without being fully downloaded in order not to clog the memory availability of the machine. The data in the subfolder `dati-regioni` is presented in the form of a .csv file, where each line corresponds to a code ranging from 1 to 22 uniquely assigned to an Italian region. P.A.Trento is assigned the code 22. Each file in the folder then is composed of 24 columns presenting a large amount of information. For the sake of this project, we apply a simple pre-processing pipeline consisting in the removal of any file column except for `data`, `totale_positivi`, `deceduti`, `dimessi_guariti`, `totale_casi`, which respectively correspond to the date and time the data in the file correspond to, the cumulative number of infectious people, the cumulative number of casualties, the cumulative number of recovered and the total overall number of cases. We then proceed with feeding the data to our models. The code used to generate results can be found of GitHub[3].

B. SEIRD model

For this project we took inspiration from one of the models presented in [5], defined as (Equation 2):

$$\begin{aligned} \frac{\partial S}{\partial t} &= -\beta I \frac{S}{N} \\ \frac{\partial E}{\partial t} &= \beta I \frac{S}{N} - \sigma E \\ \frac{\partial I}{\partial t} &= \sigma E - (\lambda + \kappa) I \\ \frac{\partial R}{\partial t} &= (\lambda + \kappa) I \end{aligned} \tag{2}$$

where β is the infection rate, γ is the inverse of the average latent time of the infection, λ and κ are respectively the recovery rate and death rate.
We extend over the simplest available SEIR model (the *SIR* model) and over the first version of the model originally presented by Kermack and McKendrick in [6]. The model for our first experiment is a SEIRD model (Equation 3) [7]:

$$\begin{aligned} \frac{\partial S}{\partial t} &= -\beta I \frac{S}{N} \\ \frac{\partial E}{\partial t} &= \beta I \frac{S}{N} - \sigma E \\ \frac{\partial I}{\partial t} &= \sigma E - \frac{1}{t_{inf}} I \\ \frac{\partial R}{\partial t} &= \frac{1-f}{t_{inf}} I \\ \frac{\partial D}{\partial t} &= \frac{f}{t_{inf}} I \end{aligned} \tag{3}$$

in which we introduce slight variations to the parameters definitions with respect to Equation 2: an equation is added to the system specifically modelling the compartment reserved to casualties; a new parameter f is added representing the death rate; a constant t_{inf} is added representing the time in days in which an infected individual is actually infectious for other individuals.
Our first experiment consisted in estimating the SEIRD model's β, σ and f using an EC strategy with a single-objective optimization algorithm as evaluator, that we fed with the linear combination of two objective functions — RMSD (Root Mean Square Deviation) of the number of deaths and RMSD of the number of infections — respectively weighted 0.9 and 0.1. We chose to assign a much larger weight to the first objective function because we empirically observed, and further on found to be validated by literature [7], that one of

[3]https://github.com/LunaBaozi/SEIRD-covid-models-with-GAs

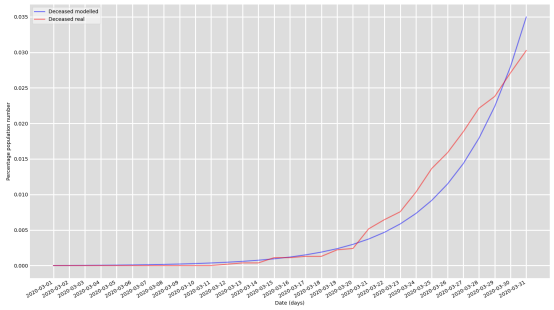

Fig. 1. Red line: deceased people from 01-03-2020 to 01-04-2020. Blue line: deceased people as modelled by the Genetic Algorithm in the same time span.

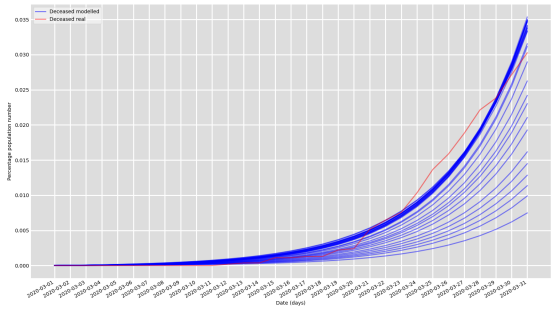

Fig. 2. Red line: deceased people from 01-03-2020 to 01-04-2020. Blue lines: time series of deceased people as modelled by NSGA2 in the same time span.

the peculiarities of the Italian situation during the COVID-19 pandemics was the numerical instability of the amount of individuals belonging to the S, E, I and R compartments; instead, the numbers of the D compartment tended to remain predictable across short periods of time and thus are better suited for fitting and modelling purposes. Table I reports the values of the best individual of the population as found by the genetic algorithm and the best fitness value obtained by the individual, while Figure 1 shows the modelled curve plotted against the real curve of casualties in the period between the 1st March of 2020 and the 1st of April 2020.

TABLE I
ESTIMATED PARAMETERS AND BEST FITNESS VALUE OF FIGURE 1.

β	σ	f	t_{inf}	Best fitness
0.47	1.52	0.37	5.1	12.50

For our second experiment we chose to use a different genetic algorithm to generate individuals, NSGA-II. In this setting we also chose to use a multi-objective optimization approach to find the best fitness values of the RMSD of deaths and the RMSD of infections functions. We also enforced some simple constraints to the evaluation of the objective functions targeted on the predicted number of individuals on each compartment (S+E+I+R+D=N), the value of the reproduction number R0 (R0$\geq$ 2), the percentage of deaths ($<$ 15%, otherwise the estimation continues to follow the exponential trend) and the percentage of infections ($>$ 10%, empirically obtained from data). Results of the obtained time series are plotted in Figure 2, while Figure 3 shows the Pareto front of the found solution.

C. Modified SEIRD model

During the initial phases of the COVID-19 pandemics in Italy, domain experts had a hard time calculating and predicting the real number of infected, exposed and recovered people. In an attempt towards a neat modelling of the compartments' populations, we experiment with a modified version of the SEIRD model presented in Equation 3, where α is the IFR (Infectious Fatality Rate), β the infectious rate (with $\beta = R0 \cdot \gamma$), σ the incubation period and γ the duration of the illness, defined as:

$$\begin{aligned}
\frac{\partial S}{\partial t} &= -\beta I \frac{S}{N} \\
\frac{\partial E}{\partial t} &= \beta I \frac{S}{N} - \sigma E \\
\frac{\partial I}{\partial t} &= \sigma E - (\gamma + \alpha) I \\
\frac{\partial R}{\partial t} &= \gamma I \\
\frac{\partial D}{\partial t} &= \alpha I
\end{aligned} \tag{4}$$

An important feature of models in both Equation 4 and 3 is that they do not take into consideration any quarantine or mitigation measures, thus the predictions they provide are somewhat approximate. For this experiment, together with α, β, γ and σ, we decided to predict also the number of exposed individuals, starting from the ground truth initial number of infectious and recovered people. We focused on one of the later waves of the pandemics, starting from the 1st of March 2021; we evolved and learnt the parameters of the model up to the 1st of May 2021. We noticed how important the initialization of the parameters bounds was, since we obtained highly different results depending on whether we allowed the algorithm more or less freedom to explore the search space. The fitness functions for this experiment are the RMSD of the

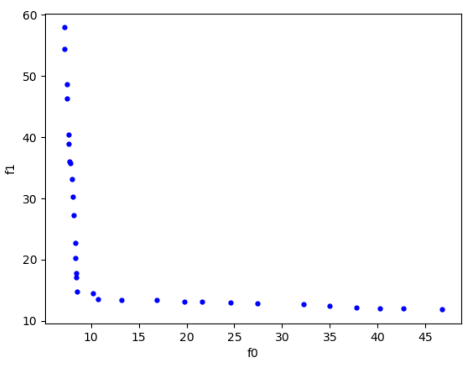

Fig. 3. Pareto frontier referring to the solution in Figure 2. The frontier seems to be defined and with equally spaced solutions.

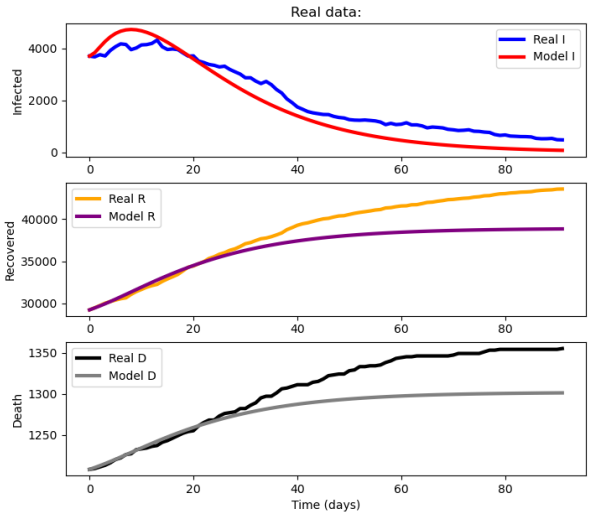

Fig. 4. Data predicted versus real data in the period from 01-03-2021 to 01-06-2021.

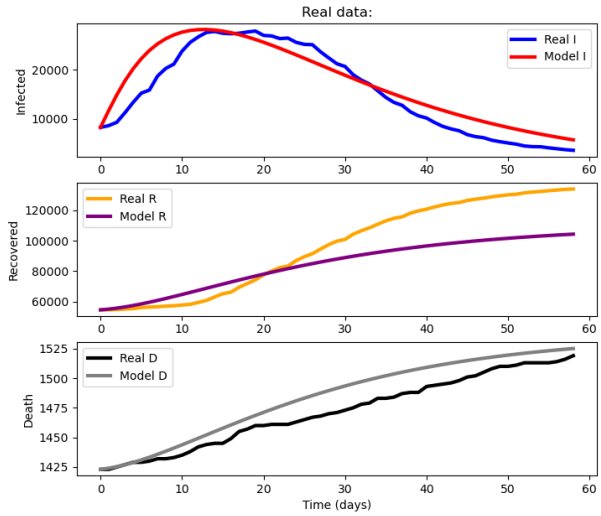

Fig. 5. Data predicted versus real data in the period from 01-01-2022 to 01-03-2022.

infected, recovered and dead people each with respect to their ground truth values; the addition of another fitness function was part of the strategy for seeking a more reliable model. The GA used was NSGA-II in a multi-objective optimization setting with the enforcement of some simple constraints. The results presented in Figure 4 and Figure 5 show that, despite the simplicity of the model, the curves of the predictions very much pick up the behaviour of the real curve. As another experiment we tried to use a single-objective optimization strategy by combining the fitness functions in a weighted sum, but we obtain a worse performance than the previous experiment.

V. Discussion and conclusions

This project is a clear example of validity of the Occam's razor principle: in spite of the extreme simplicity of the models used, most of the times results showed that they were enough for obtaining reasonable predictions. We investigated the possibility of predicting the parameters of different compartmentalized models of the COVID-19 disease by using a number of evolutionary computation strategies. The populations of parameters to estimate the number of susceptible, exposed, infectious, recovered and dead people in P.A. Trento were initialized, evolved and mutated according to the pipeline of a standard Genetic Algorithm. Both mutation and crossover were applied in each experiment performed, proving better performances together than alone. Simple constraints were enforced in order to guide the predictions of the model towards the curve of real data while maintaining meaning *per se*, while single-objective optimization and multi-objective optimization strategies were tested and results analysed. We focused on building from scratch a set of assumptions and guessings that were both tied to reality and worked well for our ends, but not every time we have been successful. Indeed, some settings performed much better than others, up to reaching good results that closely match the reality of the COVID-19 situation in different time frames throughout 2020 to 2022. Even though through simple means and several assumptions, we have empirically demonstrated that genetic algorithms, but more in general evolutionary computation, is capable of handling a wide variety of applications with successful outcomes, thanks to its adaptability and sound theoretical foundations.

Contributions

Luna contributed to the implementation, testing and analysis of results of the SEIRD model, to the conceptualization, writing of the initial draft, final drafting of the document, retrieval and organization of bibliography and citations.

Raffaele contributed to the implementation, testing and analysis of results of the modified SEIRD model, to the conceptualization and writing of the initial draft.

References

[1] S. Ghamizi, R. Rwemalika, M. Cordy, L. Veiber, T. F. Bissyandé, M. Papadakis, J. Klein, and Y. L. Traon, "Data-driven simulation and optimization for covid-19 exit strategies," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3434–3442, 8 2020.

[2] R. Miikkulainen, O. Francon, E. Meyerson, X. Qiu, E. Canzani, and B. Hodjat, "From prediction to prescription: Evolutionary optimization of non-pharmaceutical interventions in the covid-19 pandemic," 5 2020. [Online]. Available: http://arxiv.org/abs/2005.13766

[3] P. Yarsky, "Using a genetic algorithm to fit parameters of a covid-19 seir model for us states," *Mathematics and Computers in Simulation*, vol. 185, pp. 687–695, 7 2021.

[4] Z. Qiu, Y. Sun, X. He, J. Wei, R. Zhou, J. Bai, and S. Du, "Application of genetic algorithm combined with improved seir model in predicting the epidemic trend of covid-19, china," *Scientific Reports*, vol. 12, 12 2022.

[5] A. Godio, F. Pace, and A. Vergnano, "Seir modeling of the italian epidemic of sars-cov-2 using computational swarm intelligence," *International Journal of Environmental Research and Public Health*, vol. 17, 5 2020.

[6] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proc. R. Soc. Lond. A*, vol. 115, 1927.

[7] E. Loli Piccolomini and F. Zama, "Monitoring italian covid-19 spread by a forced seird model," *PLOS ONE*, vol. 15, no. 8, pp. 1–17, 08 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0237417