Predicting SIRD model parameter for COVID-19 simulation using PSO

Tommaso Canova (240456), Mattia Franzin (239930), Thomas Trevisan (240458)
{*tommaso.canova, mattia.franzin, thomas.trevisan*}*@studenti.unitn.it*,

***Abstract*—This study employs the Particle Swarm Optimization (PSO) algorithm to predict the parameters of a SIRD (Susceptible, Infected, Recovered, Deceased) model that simulates the spread of COVID-19 virus. For this purpose Italian national data have been used, considering the time window from February 24th 2020 up to April 20th 2020 (8 weeks). Three methods are proposed: a simple baseline computing a single set of parameters for the entire simulation period, a time-varying approach with weekly parameter updates, and an extended time-varying approach utilizing an Long Short-Term Memory (LSTM) to refine the predicted parameters. As anticipated, the time-varying approach demonstrates superior predictive accuracy compared to the baseline due to its ability to correct errors over time. While the baseline approach performs adequately in the short term (28 days), it fails to accurately capture long-term trends (58 days). Unfortunately, the third method, incorporating the LSTM, does not yield coherent or useful predictions to enhance the time-varying approach. The project code is open source and available at the following repo: https://github.com/Thomas2710/Epidemic-modelling**

***Index Terms*—Particle Swarm Optimization, COVID-19, SIRD, pandemics, Long Short-Term Memory**

I. Introduction and Related Works

The most recent pandemic affecting humankind, known as SARS-CoV-2, has shown how the decisions of various governments must be absolutely data driven. Indeed, the consequences of any restrictive measures have had a threefold impact on people's health, the economy and the resulting social changes. In order to model these infectious diseases, mathematical models called compartmental models are used, the first of which were introduced in the early 20th century by Ronald Ross [1]. Typically, these tools simulate trends in the spread of a given disease through ordinary differential equations, which are deterministic given a defined set of parameters. There are several models in the literature, the simplest one that can be considered is SI (Susceptible and Infectious) which can be useful if one wants to track influenza without immunity properties. While, more complex models, such as SIRD (Susceptible, Infectious, Recovered, Deceased) can be used in a pandemic context. For example, Nisar et al estimated the basic reproduction number (R0) of the spread of COVID-19 in Wuhan city during the first 67 days using a fractional-order SIRD model [2]. Another study, carried out by Sebbag and Kechida revealed how the SIRD model can be more accurate, compared with real data, when the parameters used are predicted on a daily basis using an EKF (Extended Kalman Filter) [3]. In the context of Bio-Inspired algorithms, Haouari and Mhiri used only Particle Swarm Optimization (PSO) to predict pandemic trends using data offered by the State of Qatar, obtaining good accuracy, even though small dataset and mild assumptions were used. [4]. Another work, focused on using PSO and a more complex model such as the SEIQRD (Susceptible, Exposed, Infected, Quarantined, Recovered, Deceased) with Italian data, obtained an estimation of the pandemic trend over the 35 days following a 365-day training time window. With this study, Abdallash et al predicted with high accuracy the second-wave that struck Italy. [5]. A notable study is proposed by Godio et al., in which the parameters of an SEIR model are optimized using PSO, considering a fixed 30-day time window (medium term). The study analyzes data from Italian regions and compares them with those of Spain and South Korea, highlighting the main limitations of this approach [6].

Given these studies, we decided to try improving the estimation of the parameters of a relatively simple model like the SIRD (II-B) using Particle Swarm Optimization. We compared an approach with a single parameter estimation to one with updates at fixed intervals (time-varying) (Section III). Finally, to further tune the parameters, an additional experiment was conducted in which we adjusted the time-varying parameters using an LSTM. Unfortunately, this did not result in any improvement.

II. Problem Statement

This study focuses on the first 8 weeks of the pandemic in Italy, analyzing daily data on total positives, total recovered, and total deceased, and inferring the susceptible population. The goal is to determine the optimal parameters for the SIRD model using the Particle Swarm Optimization (PSO) algorithm provided by the Python library `inspyred`. By comparing a fixed time approach with a time-varying parameter approach, we aim to demonstrate the improved accuracy and relevance of the latter, addressing limitations of previous studies that rely on mild assumptions or short time windows.

A. Dataset

Data at the national level was obtained from the official repository of the Department of Protezione Civile This repository has been publishing data related to the COVID-19 pandemic from February 24, 2020, to the present. The dataset includes 24 distinct columns that provide comprehensive information on the spread of COVID-19.

For the purpose of this study, we focused on three specific columns: the daily number of *infected* individuals, the total number of *recovered* individuals, and the total number of *deceased* individuals. Using this data, along with the total population of Italy, we calculated the number of susceptible individuals on a daily basis. These calculations were essential for providing the necessary inputs to the SIRD model.

B. SIRD model

As mentioned above, The SIRD model is an epidemiological framework that categorizes a population into four compartments: Susceptible (S), Infected (I), Recovered (R), and Deceased (D). Susceptible (S) represents individuals who can contract the disease, Infected (I) denotes those currently infected and capable of spreading the disease, Recovered (R) includes individuals who have recovered and gained immunity, and Deceased (D) accounts for those who have died from the disease. The model employs differential equations to describe the dynamics of disease transmission, recovery, and mortality, allowing for the estimation of how individuals transition between these states over time. In the context of the COVID-19 pandemic, the SIRD model has been instrumental in predicting the spread of the virus, assessing the effectiveness of public health interventions, and guiding policy decisions. Equations of the SIRD model are defined in Equation 1. The Transmission rate parameter (β) denotes the average frequency of person-to-person contacts per unit time, multiplied by the probability of disease transmission during each contact. The Recovery rate parameter (γ) signifies how swiftly infected individuals recuperate and transition to the recovered state. The Mortality rate parameter (δ) reflects the rate at which infected individuals succumb to the disease. Lastly, the total population size (N) aggregates susceptible (S), infected (I), recovered (R), and deceased (D) individuals, represented as $N = S + I + R + D$.
In formulas:

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta \frac{S(t)I(t)}{N} \\
\frac{dI(t)}{dt} &= \beta \frac{S(t)I(t)}{N} - \gamma I(t) - \delta I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t) \\
\frac{dD(t)}{dt} &= \delta I(t)
\end{aligned} \tag{1}$$

C. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a computational method used for optimizing a wide range of problems by iteratively improving candidate solutions. PSO simulates the social behavior of birds flocking or fish schooling, where each individual, referred to as a particle, adjusts its position in the search space based on its own experience and the experience of neighboring particles.

The position (x_i) and velocity (v_i) of each particle are updated according to the following formulas:

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1 \cdot (p_i - x_i(t)) + c_2 \cdot r_2 \cdot (g - x_i(t)) \tag{2}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{3}$$

where:

- w is the inertia weight,
- c_1 and c_2 are cognitive and social coefficients, respectively,
- r_1 and r_2 are random values between 0 and 1,
- p_i is the particle's best known position,
- g is the global best known position.

In our specific problem, an individual is a tuple:

$$(\beta, \gamma, \delta)$$

With every individual, we solve the differential equations of the SIRD model in order to get the new values of the epidemic spread. These values are compared with our ground truth in order to compute a loss that will guide our individuals towards the minima of the function we are optimizing. The fitness of the problem is the RMSE of the difference between the computed values of S,I,R and D and the true values of S,I,R and D. The 4 losses are turned into a fitness value using a Pareto Loss. By minimizing this fitness, we aim at finding an individual that represent the parameters of the SIRD model the best in this multiobjective setting. By doing so, we aim to identify the optimal parameters for the SIRD model, enhancing its accuracy and reliability.

In our search for optimal parameters using the Particle Swarm Optimization (PSO) algorithm, we chose a star topology over a ring topology. This choice was made to improve communication among all individuals toward a unified solution and to accelerate their convergence, since the best fitness outcome is identical in both topologies. The use of a Pareto loss function with constraints can help the Particle Swarm Optimization (PSO) algorithm to achieve better results. However, due to limited knowledge about the problem domain, we were not confident in adding such constraints.

III. Methods

Three incremental approaches were tested to obtain predictions that were as consistent as possible with the dataset used

A. Baseline
B. Time varying
C. Time varying with LSTM

A. Baseline

Our baseline approach estimates the parameters that best fit the considered timespan. In this baseline scenario, the SIRD model parameters are assumed to be constant and do not vary over time, remaining unique throughout the entire period. The parameters were estimated using PSO and considering as loss the average RMSE of the entire period of time. With the obtained parameters we simulated the trend of the curve using the *solve_ivp* function provided from *scipy.integrate* on the SIRD equations. Given that the progression of a pandemic varies over time, alternating between periods of high transmission and periods of low transmission (thanks also to the use of vaccines), this model is perfectly suited to be considered a baseline approach.

B. Time Varying

The second approach we considered was to try to estimate the change in the model behaviour, by optimizing the SIRD parameters on a weekly basis. The optimization process is therefore looped over the weeks of the pandemics, starting from the previously estimated SIRD values. This process tries to optimize with respect to the ground truth in the dataset the parameters, and solves the differential equations in order to get the new estimate of the spread values, to be used as starting point for the new iteration. This approach works as long as the weekly optimizations give reasonable results. When this condition is not met, then a substantial error in the parameters will spread to the subsequent weeks, resulting in a curve that is not very representative of the data. The final curve is computed utilizing the computed parameters per week, starting each considered week from the SIRD values obtained running the ODE solver with the parameters of the previous week. While this approach represents an evolution from the baseline method, it remains insufficient for accurately predicting future pandemic numbers. This limitation arises because, through parameter estimation, model values are computed under the assumption that the pandemic trend will continue as observed in the current week. Such an assumption may provide a reasonable approximation during certain phases of the pandemic but fails to accommodate situations where the transmission trend reverses. Especially for the COVID-19 pandemic, this has been observed to be often the case, due to:

- the high mutability of the virus, which has led to numerous variants
- the usage of several rounds of vaccines

Then, in order to predict a pandemic with such a high variability, we tried a deep neural network approach, utilizing a Long Short Term Memory (LSTM) network to try and predict the behaviour of the model.

C. Time varying with LSTM

As mentioned above, we implemented an LSTM-based approach to predict the time series of model parameters. This approach was inspired by the study "Epidemic Modeling using Hybrid of Time-varying SIRD, Particle Swarm Optimization, and Deep Learning" [7]. Our goal was to determine if there was a recurring pattern in the evolution of a pandemic. To train the LSTM, we used parameters obtained from Particle Swarm Optimization (PSO) estimation in a time-varying setting, resulting in a set of parameters for each of the considered weeks. The training process involved predicting the parameters for the next week based on the input of either 1 or 3 previous weeks' parameters. However, simply comparing the predicted parameters with the actual parameters is insufficient, since the actual parameters are only estimates derived from the PSO. The optimal approach is to use the predicted parameters to simulate the 7-day evolution of the initial states of *S, I, R, D* for the week in question, and then compare the calculated values with the ground truth on a daily basis. This process requires the ability to back propagate the error from the *S, I, R, D* values to the β, γ, and δ parameters. To achieve this, we used a differentiable implementation of the statistical model obtained with torchdiffeq.

IV. Experiments and Results

We decided to work with weekly data that was extrapolated from the daily data described in II-A, as it is more robust and less susceptible to noise.

We ran experiments focusing on detecting the parameters for the first wave of the pandemic, going from February to April 2020.

Unfortunately, contrary to the findings in the reference paper [7], our use of the LSTM model did not yield the desired results, regardless of the number of weeks used for prediction. When only one week was used for input, the model failed to detect upward or downward trends, leading to exponential growth in infection predictions. Conversely, when three weeks were used as input, the predicted values showed excessive variation, indicating that increased context does not necessarily improve prediction accuracy. These results demonstrate that the LSTM model is inadequate for capturing time series that are highly variable and influenced by multiple factors, particularly when data availability is limited.

A. Experimental findings

Simulating the baseline within a short time frame (28 days) shows fairly satisfactory results, as the model slightly underestimates the actual data (Figure 1). However, when considering a larger time frame (56 days), the SIRD model drastically undershoots the real trend (Figure 2). From Figure 3, it can be observed that the time-varying approach may appear less accurate than the baseline, as there is a slight overshooting compared to the actual data starting from the second week. Nevertheless, in the 56-day simulation shown in the Figure 4, the overshooting is corrected in the fifth and eighth weeks, thereby providing greater flexibility. The single plot comparing the amount of Susceptibles, Infected, Recovered and Deceased for the best approach (time-varying in 56 days) are reported in the Appendix A.

B. Experiments parameters

After running several experiments, we ended up with an optimal configuration for our problem's parameters, such as:

- Losses weights: [S:0.8, I:5, R:2, D:3]
- Population: 300
- Generations: 50
- Cognitive rate: 2.1
- Social rate: 1.2
- Inertia: 0.65

From experiments, we observed that our fitness landscape is made of several local minima. As shown in figure 5, the algorithm converges rather quickly (20 to 40 generations).

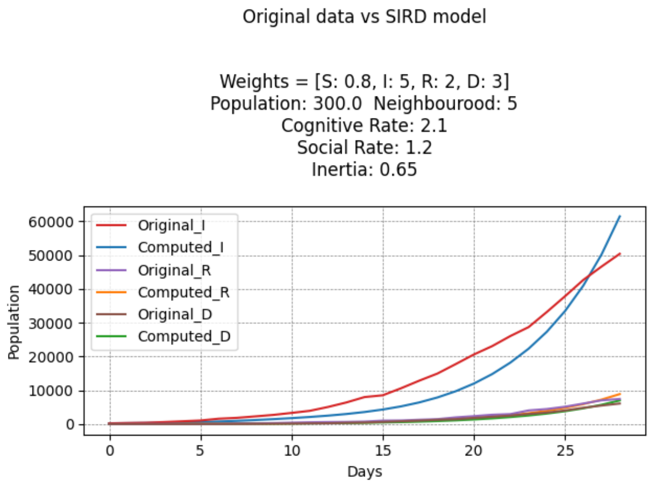

Fig. 1: **Baseline** simulation over 28 days

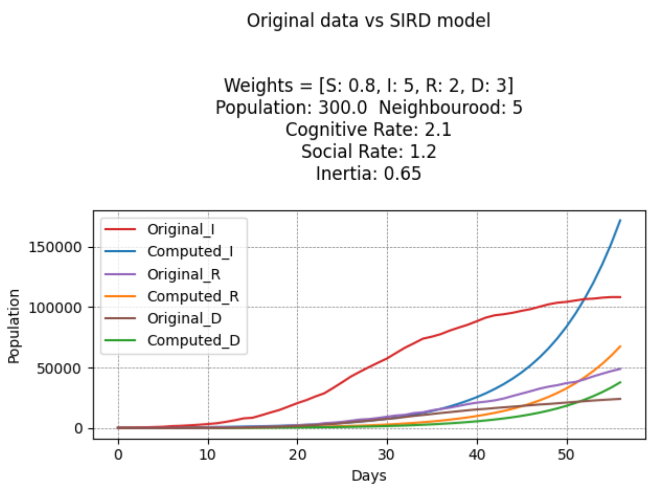

Fig. 2: **Baseline** simulation over 56 days

Fig. 3: **Time varying** simulation over 28 days

Fig. 4: **Time varying** simulation over 56 days

Since the convergence is clearly not to a global minima, we adjusted the PSO settings to enhance exploration as much as possible. This means:

- Increasing the population size
- Reducing the social rate
- Increasing the inertia rate

In fact, having a lower social rate encourages the swarm to explore the search space more extensively rather than relying solely on the best-known solution (position). At the same time, increasing the inertia places greater emphasis on velocity, thereby accelerating exploration.
For the first 8 weeks of pandemic, the best parameters we found with our time varying PSO are:

week	β	γ	δ
1	0.34300	0.02437	0.00786
2	0.23362	0.02290	0.01335
3	0.23943	0.02344	0.02052
4	0.13470	0.01878	0.01460
5	0.13584	0.02152	0.01811
6	0.00409	0.00810	0.00298
7	0.11882	0.03488	0.01639
8	0.001	0.04412	0.001

TABLE I: Best optimized parameters

C. Fitness Landscape

From our experiments, we deduced our fitness landscape might:

- have a lot of local minimas
- be non-convex, leading to unreachable points in the pareto front

Additionally, since it was observed that the algorithm quickly converges to a local minimum, employing an extension of the PSO, such as the *Guaranteed Convergence PSO* (*GCPSO*), could have been beneficial. This variant helps avoid stagnation in suboptimal solutions.

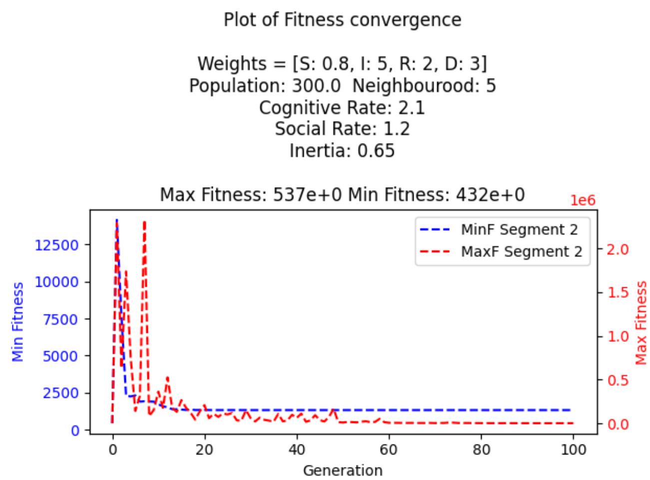

Fig. 5: Fitness values for 2nd week

V. Conclusion

Our project aimed at providing a forecast of the Susceptible-Infected-Recovered-Deceased (SIRD) model values used to describe the spread of an epidemic. We tried achieving this result by estimating the parameters of the model with a Particle Swarm Optimization algorithm on a weekly basis, and then training a time series forecast model such as LSTM to predict the parameters for the incoming week.
The problem we considered carried some intristic complications.
Firstly, the spread of an epidemic is very area dependent, considering that the applied countermeasures were different between each region/state, and not the same for the whole country. Therefore, trying to estimate the parameters on a national level might result being too unpredictable.
Secondly, the deterministic model employed to characterize pandemic dynamics serves as an approximation of the actual scenario, missing the real situation behind the data and increasing the unpredictability of the parameters.

Lastly, our experiments with the time varying configuration achieves good results in the long term, enhancing our baseline. On the other hand, the LSTM doesn't provide the expected prediction capabilities.

Thanks to this project, we were able to delve into the world of computational epidemiology by studying the SIRD model. Additionally, we learned to optimize a bio-inspired algorithm in a multi-objective scenario through a hands-on approach.

Appendix A
Predicted SIRD values

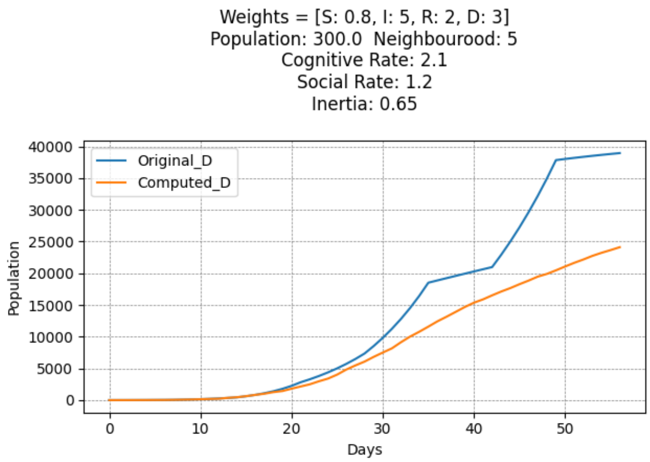

Fig. 6: Comparison of computed vs real deaths using Time-varying simulation over 56 days

Fig. 7: Comparison of computed vs real susceptibles using Time-varying simulation over 56 days

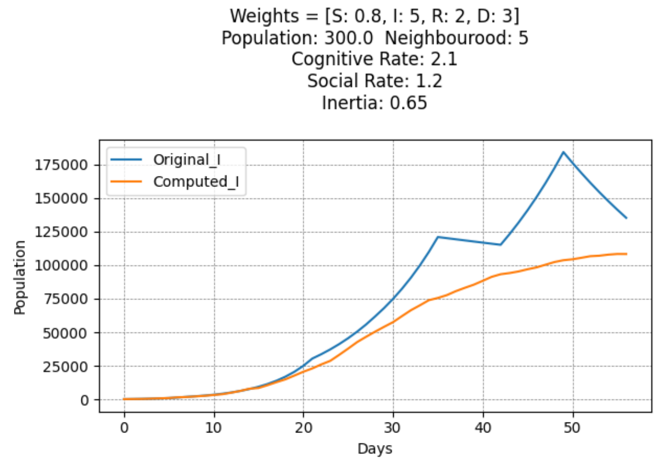

Fig. 8: Comparison of computed vs real infected using Time-varying simulation over 56 days

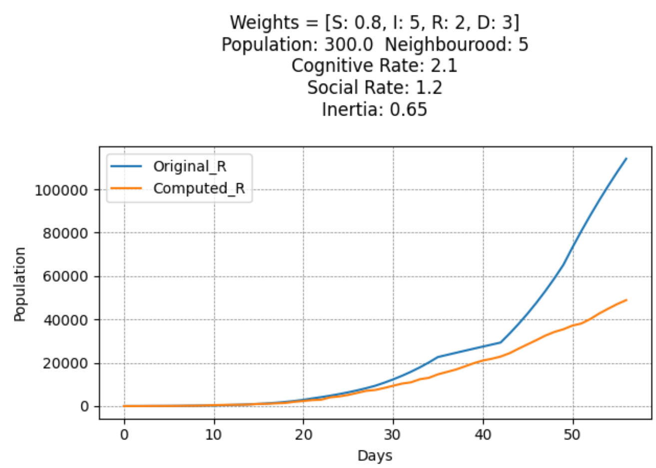

Fig. 9: Comparison of computed vs real recovered using Time-varying simulation over 56 days

APPENDIX B
FITNESS VALUES ACROSS WEEKS

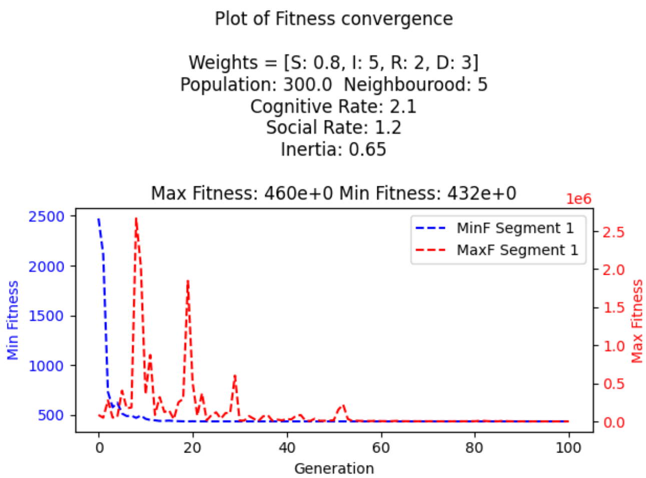

Fig. 10: Fitness values for 1st week

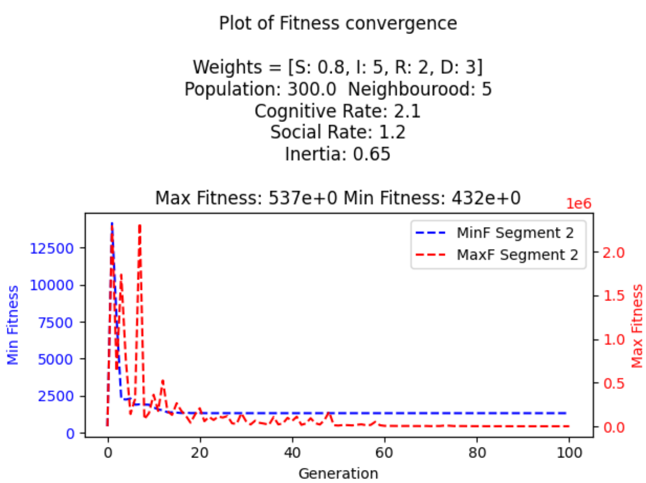

Fig. 11: Fitness values for 2nd week

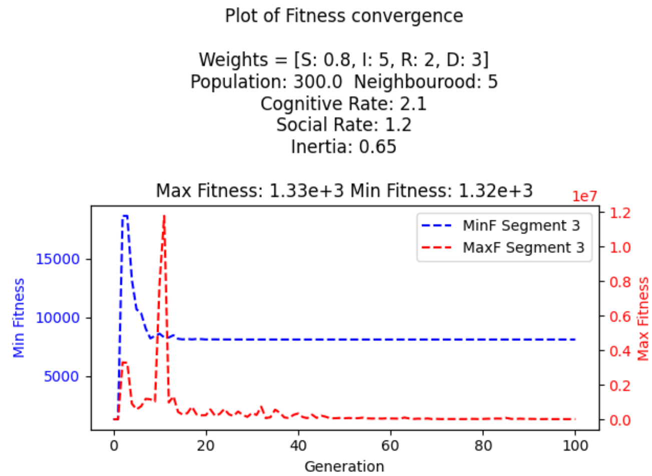

Fig. 12: Fitness values for 3rd week

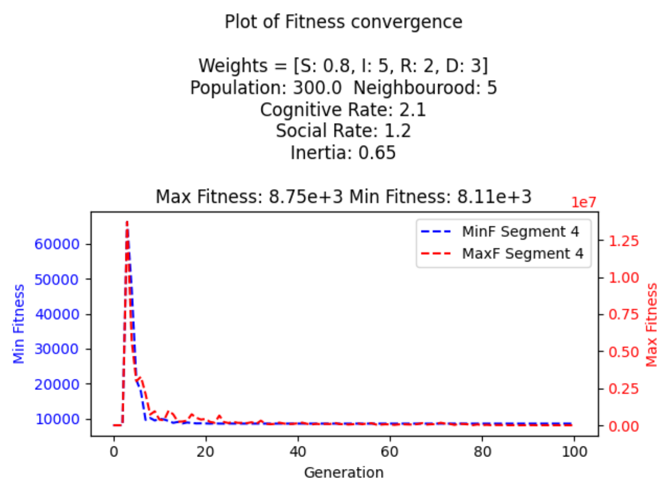

Fig. 13: Fitness values for 4th week

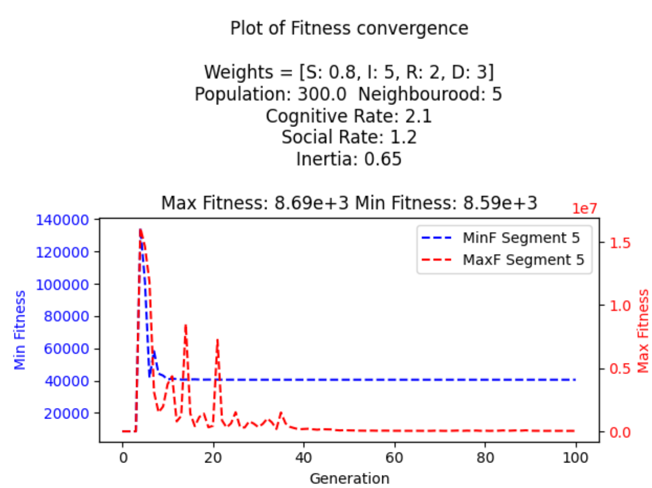

Fig. 14: Fitness values for 5th week

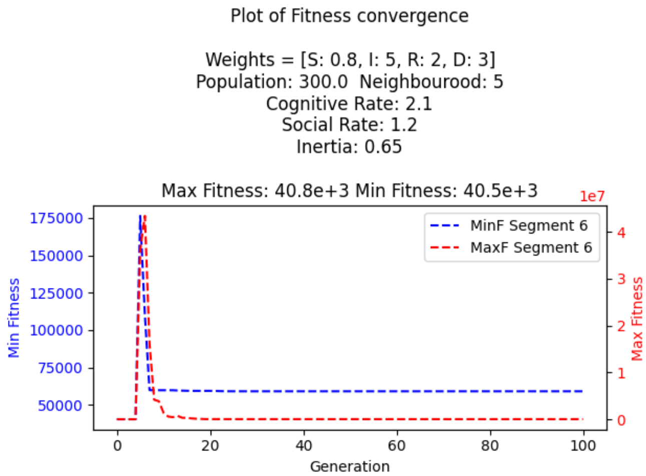

Fig. 15: Fitness values for 6th week

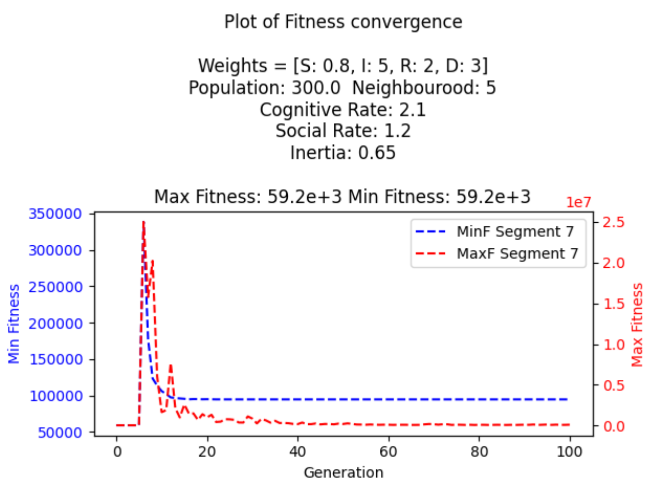

Fig. 16: Fitness values for 7th week

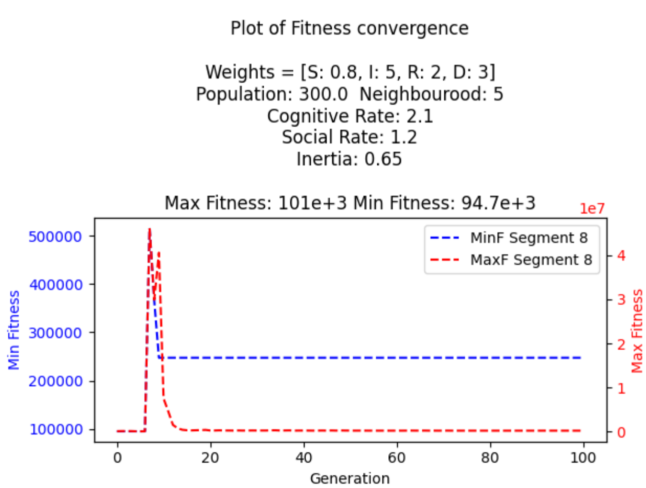

Fig. 17: Fitness values for 8th week

CONTRIBUTIONS

All members contributed equally to the topic research, implementation, analysis of the results and to the writing of the document.

REFERENCES

[1] R. Ross, "An application of the theory of probabilities to the study of a priori pathometry.—part i." *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character 92.638 (1916): 204-230*, 1916.

[2] K. S. e. a. Nisar, "Mathematical analysis of sird model of covid-19 with caputo fractional derivative based on real data. results in physics, 21, 103772." *Results in physics vol. 21*, 2021.

[3] A. Sebbagh and S. Kechida, "Ekf-sird model algorithm for predicting the coronavirus (covid-19) spreading dynamics." *Scientific Reports 12.1: 13415.*, 2022.

[4] M. Haouari and M. Mhiri, "A particle swarm optimization approach for predicting the number of covid-19 deaths." *Scientific reports 11.1: 16587.*, 2021.

[5] M. A. Abdallah and M. Nafea, "Pso-based seiqrd modeling and forecasting of covid-19 spread in italy," *2021 IEEE 11th IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE), Penang, Malaysia, 2021, pp. 71-76, doi: 10.1109/ISCAIE51753.2021.9431836.*, 2021.

[6] A. V. Alberto Godio, Francesca Pace, "Seir modeling of the italian epidemic of sars-cov-2 using computational swarm intelligence," *International Journal of Environmental Research and Public Health. 2020; 17(10):3535.*, 2020.

[7] N. Kumar and S. Susan, "Epidemic modeling using hybrid of time-varying sird, particle swarm optimization, and deep learning," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, Jul. 2023. [Online]. Available: http://dx.doi.org/10.1109/ICCCNT56998.2023.10308066