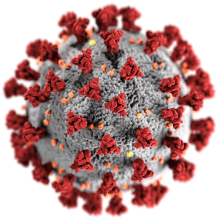
Modelling Covid-19

Data Engineering an Epidemiological Mathematical Model

Vlad-Stefan Tudor

2021

***Abstract*—The article aims to illustrate the essential concepts in mathematical modeling systems with Ordinary Differential Equations. A data-first approach is used in deploying Machine Learning techniques for fitting the model and simulating future behaviour of the pandemic.**



I. Credit and Source of Inspiration

The article is based on the research paper *Measuring and Preventing COVID-19 Using the SIR Model and Machine Learning in Smart Health Care*[3]. This material aims to provide a summary and to further analyze particular cases of interest derived from the original research.

Computations, visualizations, data imports are easily done by using the Python repository at:

https://github.com/lisphilar/covid19-sir[2]

Credit goes to the well developed and maintained code for greatly simplifying the research pipeline. Very insightful explanations for the code is also provided in [1]

II. The Pipeline: From Data to Prediction

Mathematical models can accurately describe a wide range of systems by encoding the overall *input-output* behaviour through equations. Such is the case for models that take as an input the state of the system at the current time and generate its expected evolution. There are two principal methodologies for mathematically modelling systems:

- *Analytically*, by using theoretical formulae and equations developed by researchers with background in the field
- *Empirically*, by feeding historical data into the system, containing both the input and the target, known output and gradually fitting the model to it.

While analytical models are the traditionally robust way-to-go, complex systems with hidden correlations between a large number of variables pose serious problems.

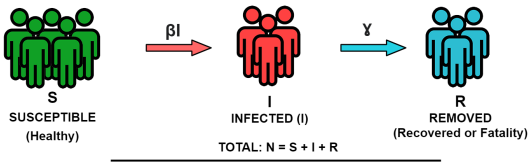
On the other hand, models that are treated like a *black box* (one does not care what is inside), if given sufficient data may produce unexpected results, identifying patterns and underlying mechanics of a system without an explicit theoretical framework of the subject. While this very well suits the current wide-availability of data, converging to a solution is not guaranteed.

In the presented method, a combination of the two methods is used. A general model for infectious diseases is used as a starting point: the *SIR Model* and data analysis is used for calibrating the parameters of the equations.

III. Infectious Disease Models

A. The SIR Model

The SIR (Susceptible-Infected-Recovered) model is a system of Ordinary Differential Equations that describe the transitions between 3 states that describe the population's affliction.



$$\frac{\mathrm{d}S}{\mathrm{d}T} = -N^{-1}\beta SI$$
$$\frac{\mathrm{d}I}{\mathrm{d}T} = N^{-1}\beta SI - \gamma I$$
$$\frac{\mathrm{d}R}{\mathrm{d}T} = \gamma I$$

Fig. 1: SIR Model

The system is defined by 2 parameters that assure generality for any kind of epidemic:

- β is the effective contact rate (contacts per person per time multiplied by the probability of transmission)
- γ models the recovery / fatality rate.

To obtain the model means, therefore, to obtain the values for these parameters. In practice, the values of the parameters are affected by factors such as governments' reactions against the disease, medication and mutations of the virus.

Non-dimensional SIR Model

The following substitutions are made in order to remove the units from the equations variables:

$$(S, I, R) \to N \times (x, y, z)$$
$$(T, \beta, \gamma) \to (\tau t, \tau^{-1}\rho, \tau^{-1}\sigma)$$

We obtain:

$dx/dt = -\rho xy$
$dy/dt = \rho xy - \sigma y$
$dz/dt = \sigma y$

The Basic Reproduction Ratio R_0

This value represents the expected number of new infections caused by a single infected host.

$$R_0 = \frac{\beta}{\gamma} = \frac{\rho}{\sigma}$$

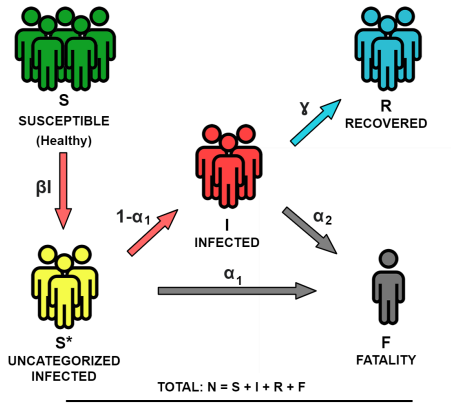
Solving the differential equations gives the following result which effectively quantifies the evolution of the disease:

$$S(t) = S(0)e^{-R_0 \frac{R(t)-R(0)}{N}}$$

R_0 not to be mistaken for $R(t)$ and $R(0)$ which is the removed population as a function of time t.

B. *The SIR-F Model*

A major drawback of the original SIR model is that it does not take into account intermediary states between *susceptible* and confirmed *infected* individuals. In reality, a great deal of the dynamics of the system occurs exactly in-between these states, as individuals that are *infected but un-categorized* as such may travel, meet with their peers, thus easily spreading the virus. Additionally, distinct states are considered for *fatalities* and *recovered* individuals.



$$\frac{\mathrm{d}S}{\mathrm{d}T} = -N^{-1}\beta SI$$
$$\frac{\mathrm{d}I}{\mathrm{d}T} = N^{-1}(1-\alpha_1)\beta SI - (\gamma + \alpha_2)I$$
$$\frac{\mathrm{d}R}{\mathrm{d}T} = \gamma I$$
$$\frac{\mathrm{d}F}{\mathrm{d}T} = N^{-1}\alpha_1\beta SI + \alpha_2 I$$

Fig. 2: SIR-F Model

Non-dimensional SIR-F Model

$$(S, I, R, F) \to N \times (x, y, z, w)$$
$$(T, \alpha_1, \alpha_2, \beta, \gamma) \to (\tau t, \theta, \tau^{-1}\kappa\tau^{-1}\rho, \tau^{-1}\sigma)$$

We obtain:

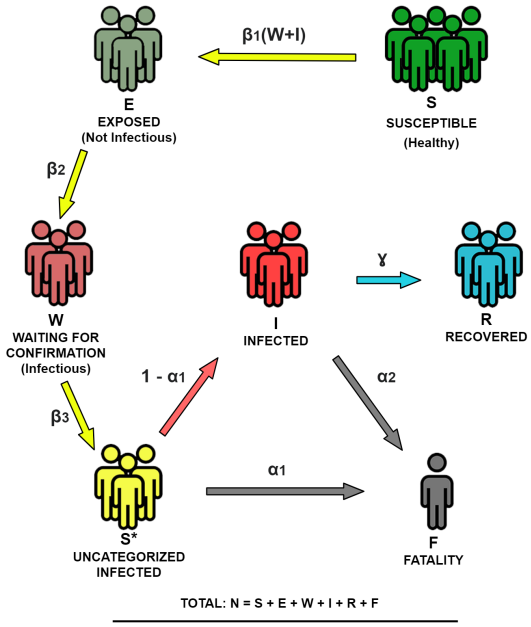
$dx/dt = -\rho xy$
$dy/dt = (1-\theta)\rho xy - (\sigma + \kappa)y$
$dz/dt = \sigma y$
$dw/dt = \theta\rho xy + \kappa y$

The Basic Reproduction Ratio R_0

$$R_0 = \beta\frac{1-\alpha_1}{\gamma + \alpha_2} = \rho\frac{1-\theta}{\sigma - \kappa}$$

C. *SEWIR-F Model*

In order to increase the accuracy, we have to model the latency of the virus spread. More states are needed in order to represent individuals that have *contracted the virus but are not infectious* and individuals who *are infectious but are not yet confirmed*, situations that account for a high level of uncertainty.



$$\frac{\mathrm{d}S}{\mathrm{d}T} = -N^{-1}\beta_1 S(W+I)$$
$$\frac{\mathrm{d}E}{\mathrm{d}T} = N^{-1}\beta_1 S(W+I) - \beta_2 E$$
$$\frac{\mathrm{d}W}{\mathrm{d}T} = \beta_2 E - \beta_3 W$$
$$\frac{\mathrm{d}I}{\mathrm{d}T} = (1-\alpha_1)\beta_3 W - (\gamma + \alpha_2)I$$
$$\frac{\mathrm{d}R}{\mathrm{d}T} = \gamma I$$
$$\frac{\mathrm{d}F}{\mathrm{d}T} = \alpha_1\beta_3 W + \alpha_2 I$$

Fig. 3: The SEWIR-F Model

Non-dimensional SEWIR-F Model

$$(S, E, W, I, R, F) \to N \times (x_1, x_2, x_3, , y, z, w)$$
$$(T, \alpha_1, \alpha_2, \beta, \gamma) \to (\tau t, \theta, \tau^{-1}\kappa\tau^{-1}\rho, \tau^{-1}\sigma)$$
$$(\alpha_2, \beta_i, \gamma) \to \tau^{-1} \times (\kappa, \rho_i, \sigma)$$

We obtain:

$dx_1/dt = -\rho_1 x_1(x_3 + y)$
$dx_2/dt = \rho_1 x_1(x_3 + y) - \rho_2 x_2$
$dx_3/dt = \rho_2 x_2 - \rho_3 x_3$
$dy/dt = (1-\theta)\rho_3 x_3 - (\sigma + \kappa)y$
$dz/dt = \sigma y$
$dw/dt = \theta\rho_3 x_3 + \kappa y$

The Basic Reproduction Ratio R_0

$$R_0 = \frac{\rho_1}{\rho_2}\rho_3\frac{1-\theta}{\sigma+\kappa}$$

D. Modelling Vaccination

Vaccination through the population translates into a reduction of susceptible population. Therefore, we can model it by adding to the differential equation for the susceptible population S the term:

$$-\omega N$$

IV. Data

A wide variety of factual data is provided through open-source databases. All the presented information is available for any specific country. For demonstration, we chose *Italy*.

Making use of the available data, *Machine Learning* techniques can be deployed in order to:

- clean the data, filter out irregularities
- detect changes in trend
- decompose the system into principal components

Since we expect the system to be non-linear, it is highly improbable that a uniquely-determined model with constant parameters may fit adequately well the recorded data. Therefore, analysis is necessary in order to determine whether there are distinct phases in the evolution of the system. In these phases, the system may act locally linear, which would enable us to deploy the proposed models.

V. Susceptible/Recovered Trend

Following on the proposed SIR Model equations an analysis of the relationship between the number of susceptible population S and the recovered population R can be analysed:

$$\frac{dS}{dR} = -\frac{\beta}{N\gamma}R$$

This is equivalent to:

$$S(R) = Ne^{-\frac{\beta}{N\gamma}R}$$

We obtain:

$$logS(R) = -\frac{\beta}{N\gamma}R + logN$$

Recalling that N, β and γ are constants, we obtain that $logS(R)$ follows a straight line whose slope is given by the parameters of the model.

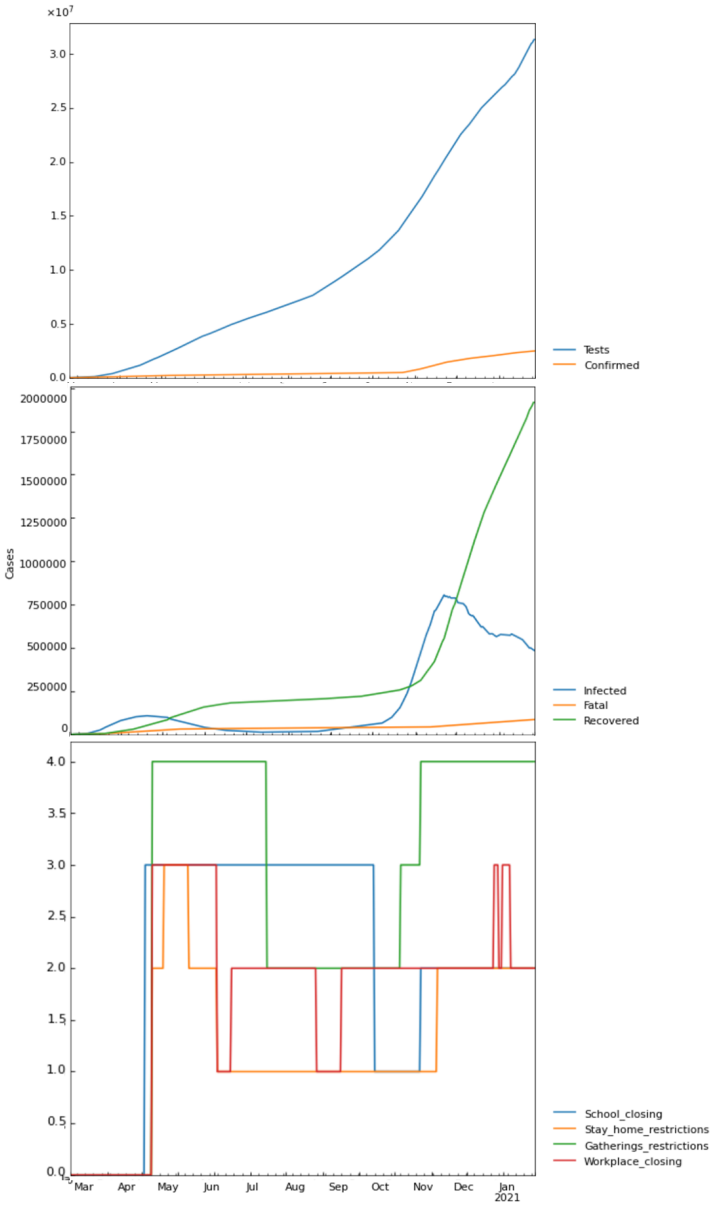


Fig. 4: Covid-19 Testing, Cases and Government measures in Italy

In reality, these parameters are not constant, since they are dictated by social dynamics and government measures against the pandemic. Therefore $logS(R)$ will *not* follow a straight line with a constant slope as the theory suggests. But the points where the slope changes can still be identified. This procedure is called *change point analysis*.

This analysis thus allows for segmenting the relationship between S and R into distinct *phases*. In the span of a phase, we consider the slope of $logS(R)$ and, implicitly, the parameters of the model to be all constant. Thus, we can estimate these parameters based on the data in the corresponding phase.

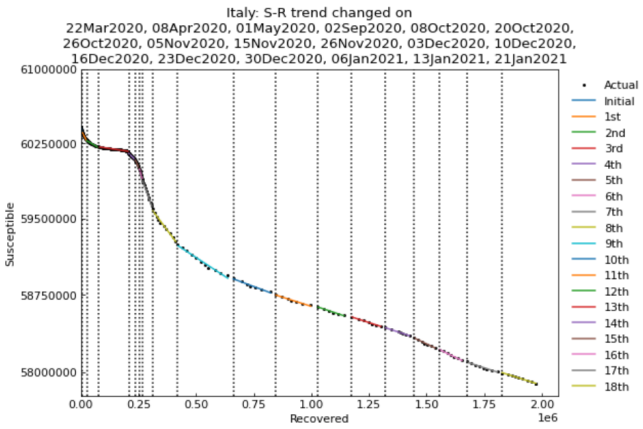


Fig. 5: Phases in Susceptible/Recovered Trend Analysis

VI. Parameter Optimization for the SIR models

The Susceptible/Recovered Trend Analysis served to identify the phases during which the parameters of the model can be considered constant. Now, optimization techniques can be applied in order to extract exact values for the parameters.

The Cost function

We take as objective for the parameter optimization minimizing the *Root Mean Squared Log Error (RMSLE)*:

$$\sqrt{\frac{1}{n}\sum_{param.}(log(true+1)-log(predicted+1))^2}$$

For each parameter of the model the observed, $true$ value is compared to the $predicted$ one. The lower the value of the RMSLE function, the more accurate the estimation. Minimizing the cost function therefore leads to the best approximation.

A. *Results: Parameters of the SIR-F Model*

By computing the parameter optimization, we obtain the values for the model parameters, segmented by the identified phases.

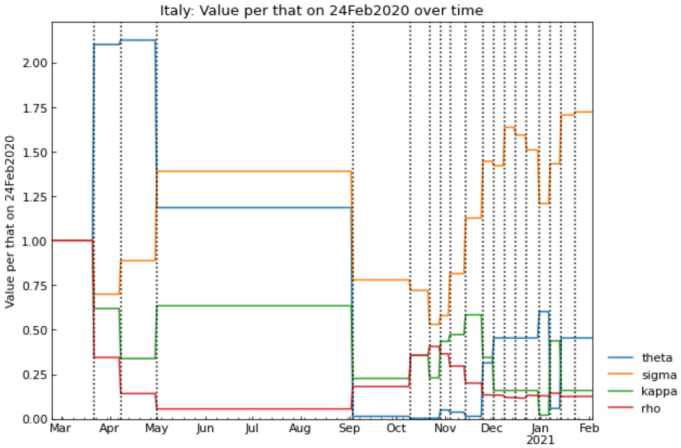


Fig. 6: Evolution of the SIR-F Model parameters for each phase

B. *Results: Model vs. Reality*

We now take a look at how the fitted model's estimations compare to the real world data. Consider that these are not predictions in the future. These are differences between the considered theoretical equations and the far more complex reality.

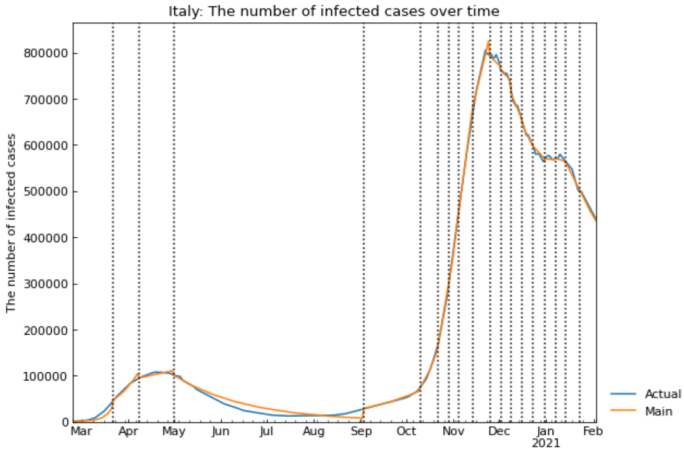


Fig. 7: SIR-F Model against real data

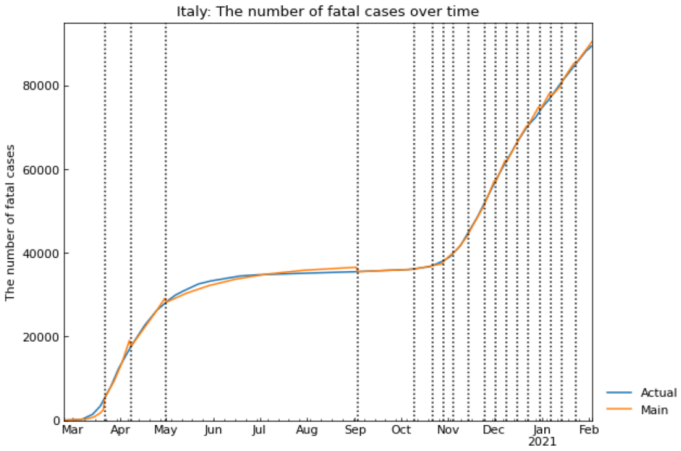


Fig. 8: SIR-F Model against real data

C. *Results: Predictions with SIR-F Model*

Having estimated the model's parameters on the available data, we now can deploy the model so as to predict the future evolution of the pandemic. Of course, the parameters are considered constant since changes in the parameters are caused by external factors which are not determinable by the model itself. This leaves a rather short window of certainty for the prediction but being able to foreshadow the underlying pattern still proves valuable.

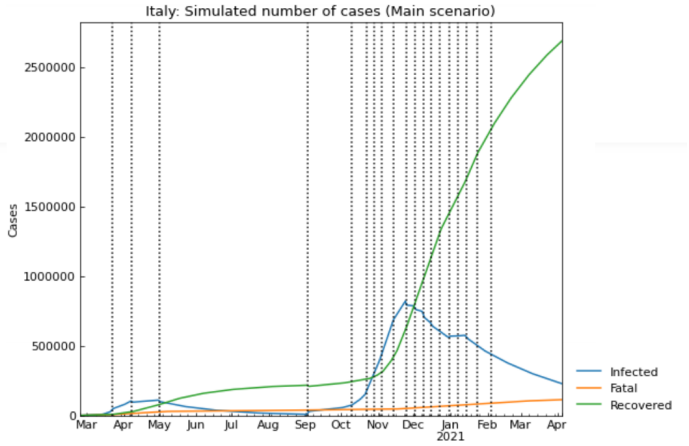


Fig. 9: SIR-F Model prediction 60 days in advance from February 3rd

RUNNING THE CODE

All the code is in *.ipynb* format, a type of file that can be opened with *python notebooks*. Notebooks can be run online on platforms such as *google Colab*, without the need of installing any dependency.

In order to be able to run the experiments presented in the paper, access the *Covsirphy Repository*[2], and upload the example notebooks on *Google Drive*. Then run them with *Google Colab*

REFERENCES

[1] Covid-19 Data with SIR Model. https://www.kaggle.com/lisphilar/covid-19-data-with-sir-model.
[2] Covsirphy Repository. https://github.com/lisphilar/covid19-sir.
[3] Madallah Alruwaili Nasser Alshammari Salman Ali Alqahtani Ali Karime Saad Awadh Alanazi, M. M. Kamruzzaman. Measuring and preventing covid-19 using the sir model and machine learning in smart health care. 2020.