LIÈGE université

FACULTÉ DES SCIENCES APPLIQUÉES
Ecole d'ingénieurs et d'informaticiens

PROJ0016-1
BIG DATA PROJECT

Bertrand Cornélusse
Pierre Geurts
Gilles Louppe

Final review

Modelling an epidemic: summary

Julien Hubar
Pierre Dumoulin
Andreas Duquenne
François Lievens

Master 1 Data science Engineering
Faculty of Applied Sciences

Academic years 2020-2021

1 Introduction

The purpose of this project is to model an epidemic (Covid20) in order to be able to make predictions and derive social measures. We decided to use a model that takes advantage of all the data we receive from the epidemic, in other words, a model taking into account the positive identified cases as well as hospitalizations or fatalities.

2 Model

Our model is composed of several states, each one representing a situation in which an individual could be. Each individual of the whole population must be in a given state such that the sum of all population states is the total population. The states and model parameters are given by:

- S: susceptible
- E: exposed
- I: infected
- R: recover
- H: hospitalized
- C: critical
- F: fatality

- β: Contamination rate
- σ: Incubation time (inverse)
- γ: Recovery rate
- hp: Hospitalization rate
- hcr: Recovery rate
- pc: ICU entry rate
- pdr: ICU death rate
- pcr: ICU recovery rate

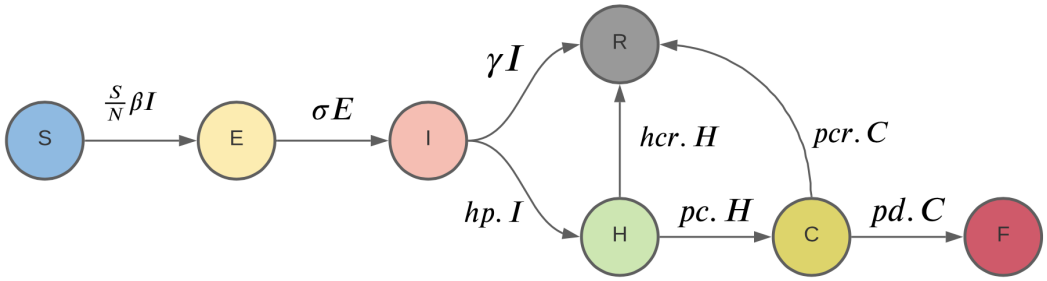

FIGURE 1: Epidemic model

3 Predictions

From parameters of the epidemic model, we are able to generate two kinds of predictions:

3.1 Deterministic

Our model can be described in a deterministic way through the following partial derivative equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta\frac{SI}{N} \\
\frac{dE}{dt} &= \beta\frac{SI}{N} - \frac{\sigma}{E} \\
\frac{dI}{dt} &= \sigma E - \gamma I - hp.I \\
\frac{dH}{dt} &= hp.I - hcr.H - pc.H \\
\frac{dC}{dt} &= pc.H - pd.C - pcr.C \\
\frac{dR}{dt} &= \gamma I + hcr.H + pcr.C \\
\frac{dF}{dt} &= pd.C
\end{aligned}$$

Each equation providing the evolution of a given state depending on the time and so describe moves of invidious of our population between each epidemic compartments. This model doesn't account for uncertainty and thus won't be used for prediction.

3.2 Stochastic

To add the stochastic effect of the true evolution of an epidemic in our model, we can model transitions between epidemic compartments by using multinomial laws (multivariate version of the binomial law). Therefore, computing a stochastic realization is achieved, at each time step, by computing the following transitions between epidemic states:

$$\begin{aligned}
S_{t+1} &= S_t - (S \to E)_t \\
E_{t+1} &= E_t - (E \to I)_t + (S \to E)_t \\
I_{t+1} &= I_t - (I \to R)_t - (I \to H)_t + (E \to I)_t \\
H_{t+1} &= H_t - (H \to R)_t - (H \to C)_t + (I \to H)_t \\
C_{t+1} &= C_t - (C \to F)_t - (C \to R)_t + (H \to C)_t \\
R_{t+1} &= R_t + (I \to R)_t + (H \to R)_t + (C \to R)_t \\
F_{t+1} &= F_t + (C \to F)_t
\end{aligned}$$

As an example, an individual in the class S can only end in the class E at the next time step. So, there are only two possible transitions for this people: to be infected and going to the state E, or staying in the state S. Thus, the probability distribution of individious's transitions from the S class can be view as follow:

$$P((S \to E) = k) = \frac{S!}{k!\,(S-k)!}p^k(1-p)^{S-k}$$
$$p = \frac{S}{N}\beta I$$

With S representing the number of people in the S state, and p the probability to be infected at this time step. A second example: To evaluate a transition from state I to others, we have extended our binomial law to tree possible outputs. Indeed, a person from the class I, during the next transition, can:

- Be cured and move to the R state with a probability γ
- Be hospitalized and move to the H state with a probability hp
- Staying infected with a probability $1 - \gamma - hp$

The probability distribution of transitions from I can be represented with the following multi-nomial law:

$$P((I \to R) = k, (I \to H) = l, (I \to I) = m) = \frac{I!}{k!\,l!\,m!}\gamma^k hp^l (1 - \gamma - hp)^m$$

With

$$I = k + l + m$$

The same principle is then used to determine the probability distributions for transitions between all states. In order to refine the precision of our predictions and so our interval of confidence, our stochastic predictor also take into account of the testing data, when there are available. Given these probabilities of transition of the distributions, all we need to do to generate stochastic realisations of our epidemic is a random number generator who respect the computed distributions.

Therefore, by using our stochastic predictor, we realized a large number of predictions. To generate *fig.*2, *fig.*3, *fig.*4 and *fig.*5, we made 200 stochastic realizations of our epidemic on 200 days. In *fig.*2 and *fig.*3, we have drawn 50 realizations from the 200, and the mean realization, for each type of curve. In *fig.*4 and *fig.*5, are drawn the average realization and the 97% confidence interval, again computed on the basis of our 200 simulations.

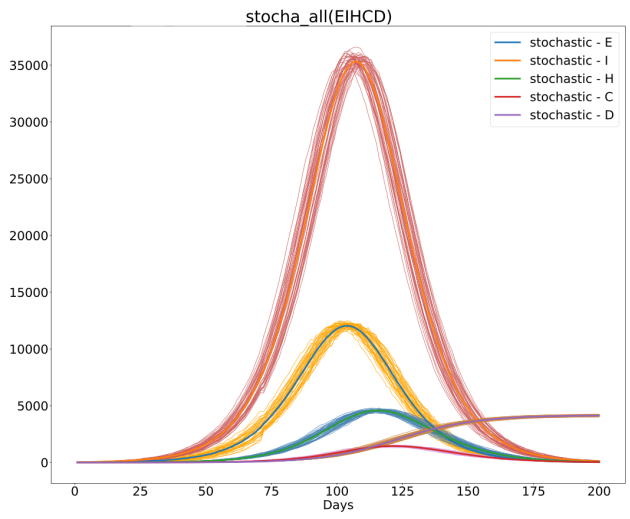

FIGURE 2: Stochastic modelling of I,E,H,C,F states

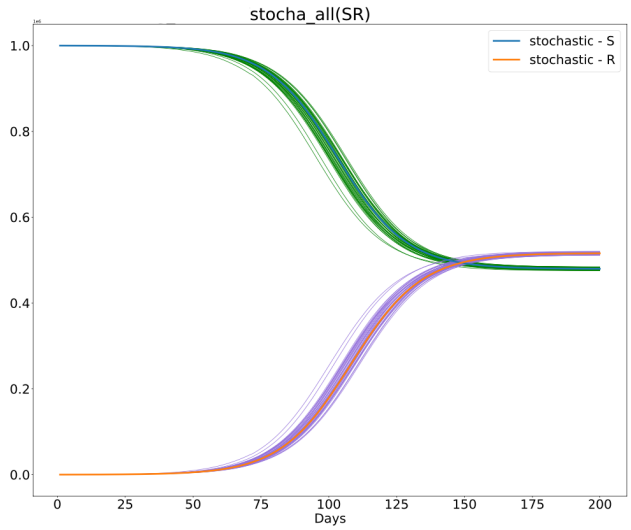

FIGURE 3: Stochastic modelling of S,R states

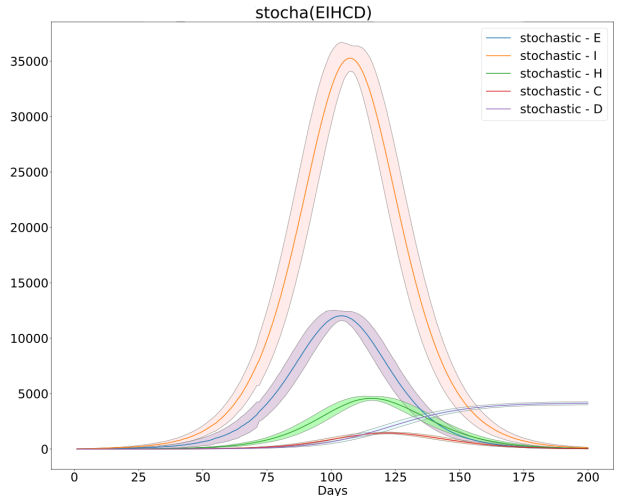

FIGURE 4: Confidence interval for I,E,H,C,F states

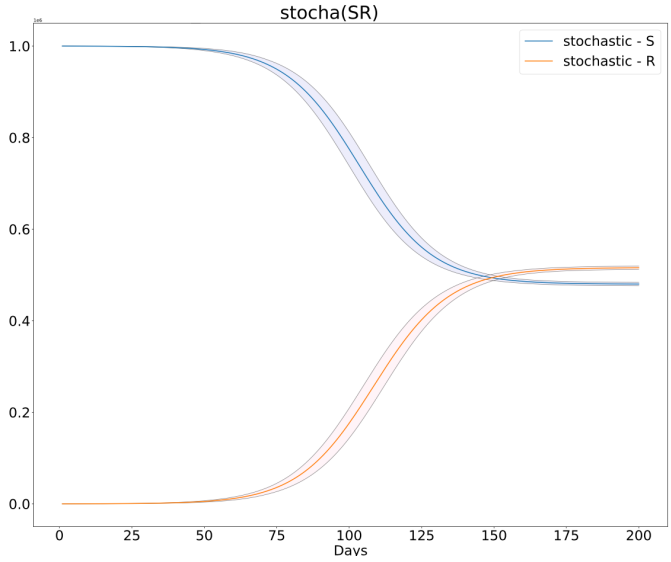

FIGURE 5: Confidence interval for S,R states

4 Fitting process

Now that we have modelled the evolution of an epidemic, we have to infer the parameters of our model to fit observed data. Since observed data evolution seems to have an exponential increasing shape, using least squares is impossible. To avoid this problem, the error between observations and comparable predictions are calculated by using a normal distribution, centered in zero and whose standard deviation is equal to the observed value. The error is given by the evaluation of the difference between predictions and observations on this Gaussian distribution.
Therefore we can define an objective function *obj* who return a real number between zero and infinity, and that we can minimize to find epidemic's parameters. This objective function is given by, where i is an integer $\in \{1, 2, 3, 4, 5\}$ which represent the five data comparison that we made (number of positive cases, hospitalizations,...), t is the time unit, and w represent the epidemic combinations of parameters that we are evaluating.

$$obj_w = \sum_{t=0}^{T} \sum_{i=1}^{5} -\log\left(\frac{1}{obs_{i,t}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{|obs_{i,t} - pred_{i,t}|}{obs_{i,t}}\right)^2\right]\right)$$

So, for each time unit of the fitting period, we are making the five following comparisons:

- The number of daily test is compared with the product of (E→I) transitions and the testing rate (t), divided by the positive rate. This latter is the proportion of positive test divided by the total number of test this day. This comparison is useful to optimize the testing rate.
- The product of (E→I), the testing rate and the sensitivity with the number of positive tests of the day.
- The number of predicted hospitalized people (H) with hospitalization's data.
- The number of predicted critical people (C) with the intensive care unit data.
- The number of fatality predictions (F) with fatality data.

For all these comparisons, the prediction data used are the mean predictions obtained from 1000 stochastic simulations of the epidemic over the time range of the available observations. It is also possible to use the deterministic predictor instead of the stochastic one during the fitting phase. By cumulating these five comparisons, our numerical optimizer is able to move in the search space so as to simultaneously optimize all our epidemic parameters.

5 Bruteforcing Strategy

During the fitting process, we were able to observe the non-convexity of the research space we are dealing with. Depending on the starting values used for our epidemic parameters, the optimizer could converge towards different solutions. To solve this problem, we implemented a kind of model selection process, which instantiates and fit models starting from uniformly random starting values for epidemic parameters. For each tested starting combinations of parameters, parameter's values before and after fitting are stored in a database, as well as the value returned by the objective function after fitting.

6 Recommendation

6.1 Implementation of social data

Population distribution

In order to be able to study the different scenarios a study of the data provided has been made. The first thing was to divide the population according to their age. For this purpose four categories were created, the "young" aged between 0 and 6 years old, the "junior" aged between 6 and 23 years old, the "*medior*" aged between 23 and 65 years old and the "*senior*" aged over 65 years old. This cutout was made under the assumption that a "*junior*" goes to school(school being compulsory above 6 years old), a "*medior*" goes to work and that the "*young*" and the "*senior*" don't go to work or to school. Then the same kind of idea was used to determine the size of households according to the categories (see the households section 6.1 below).

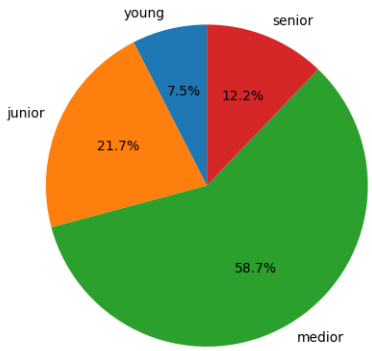

FIGURE 6: Distribution of the population according to the dataset stats according to age

Schools and workplaces

In order to define the number of people a person was likely to meet each day in schools or in workplaces. The weighted average has been computed.

- **Schools**: In a school the juniors are likely to meet **298.85** people.
- **workplace**: In a workplace the medior are likely to meet **19,85** people.

Households

Starting from the household division provided by the stats file. In order to calculate the average household size for each age group, the following assumptions were made. An illustration of the cut-out has been made *fig*8.

- **Junior**: They can be found in 3,4,5,6 and 7 household. They didn't end up in 1 because it was considered that a junior doesn't live alone and still live with both parents.
- **Young**: They can be found in 2,3,4,5,6 and 7 household. They didn't end up in 1 because it was considered that a junior doesn't live alone.
- **Medior**: They can be found in all kinds of household.
- **Senior**: They can be found in 1 and 2 household because they live either alone or in couples

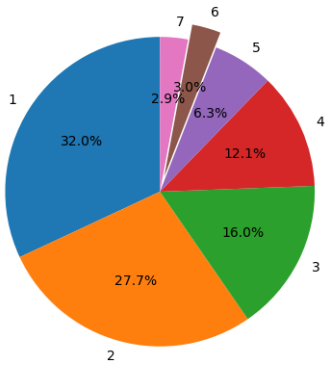

FIGURE 7: Distribution of the population according to the dataset stats according to households

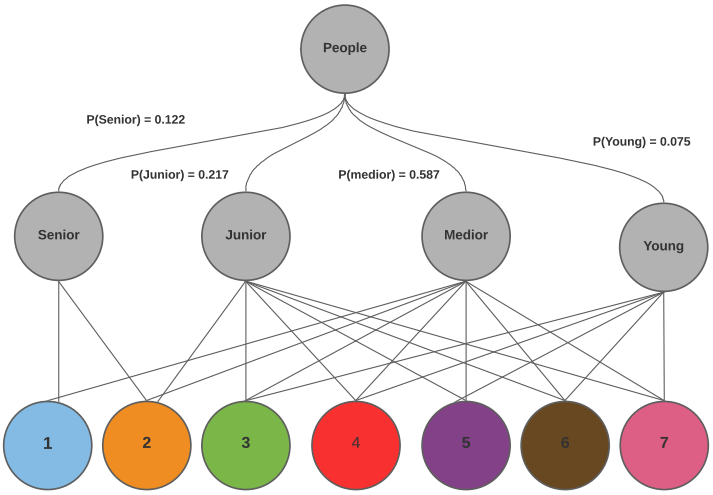

FIGURE 8: Illustration of the grouping of individuals by household according to their age.

households/Category	Young	Junior	Medior	Senior
mean people met	4.124	3.258	2.536	1.46

Communities

Finally, the community distribution provided allows us to identify the number of individuals met on average. Therefore, it was calculated that an individual met 93 people on average.

6.2 Stochastic predictor adaptation

Firstly, we have to remember the formula of the daily number of contamination's:

$$\text{Contamination}_t = \beta_{t-1} S_{t-1} \frac{I_{t-1}}{N}$$

Thus, on one hand we have the β parameter, which represents the probability that a contact between an infectious individual and a healthy individual will transmit the disease, on the other hand, we have $\frac{SI}{N}$, which represents the number of daily contacts between healthy and infectious individuals. We apply a product factor to the parameter β in order to modify the contamination rate of the day to health measures.

Because our actual model make robust fit of our parameters (including β), we have so chosen to consider that this value of beta that we have find is valid for a population whose contact has not been restricted, and so, that this parameter represent the contamination rate multiply by a "contact restriction factor" equal to 1.

With the help of the statistics mentioned above, we can divide our population into different age groups and represent the different contacts that this population maintains daily in a two-dimensional matrix like shown in the figure 9. From this matrix, we can compute the total number of daily contacts in our population and divide it by the total number of daily contacts without health measures. By this way, we obtain the "contact restriction factor" that we can apply by multiplying the original value of β and we obtain the vector represented in the figure 10.

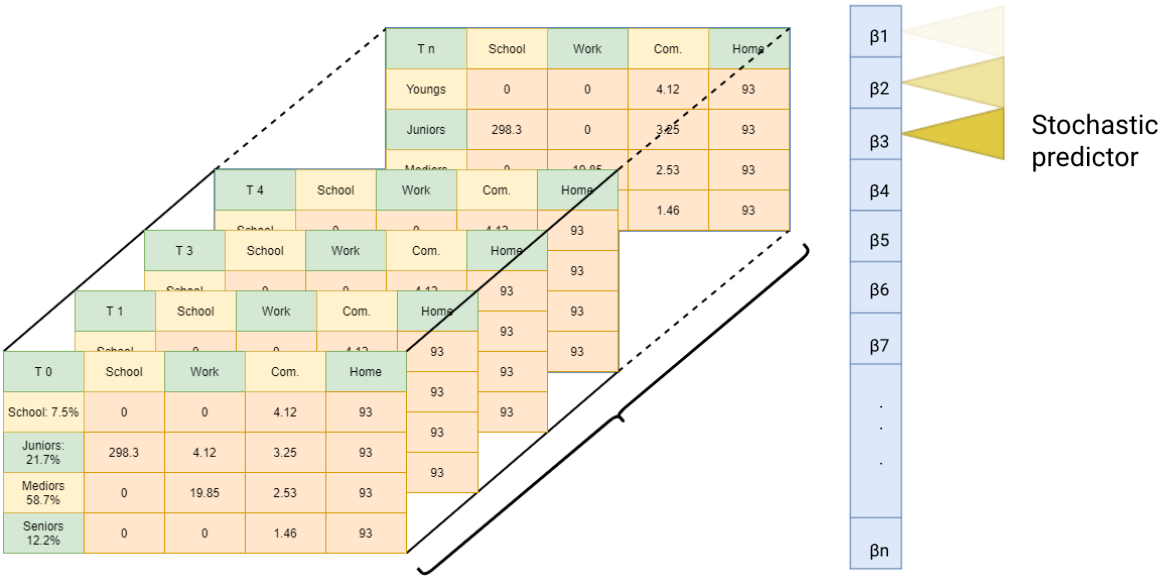

FIGURE 9: Evolution of contact distribution according to health measures

FIGURE 10: β time line

From the vector of the figure 10, our stochastic predictor can now apply simulation taking account of programmed health measures.

7 Measure effects and recommendation

In order to evaluate the effects of the measurements on the evolution of the curves, we simulated, individually and by combination, the effects of the different available measurements.

We then discovered that the measures taken individually were not satisfactory or too restrictive(lockdown) and we therefore opted for a combination of measures.

From the *fig.*11 we concluded that our recommendation would be to combine "mask-wearing" along with "social distancing".

Note that these curves were computed considering measures starting on the 15 December(day 73) on the x-axis. Moreover, the red zone represent the period between day 73 and 128 which is the period during the one we had to keep critical cases and hospitalizations bellow the given thresholds(number of peoples on the y-axis).

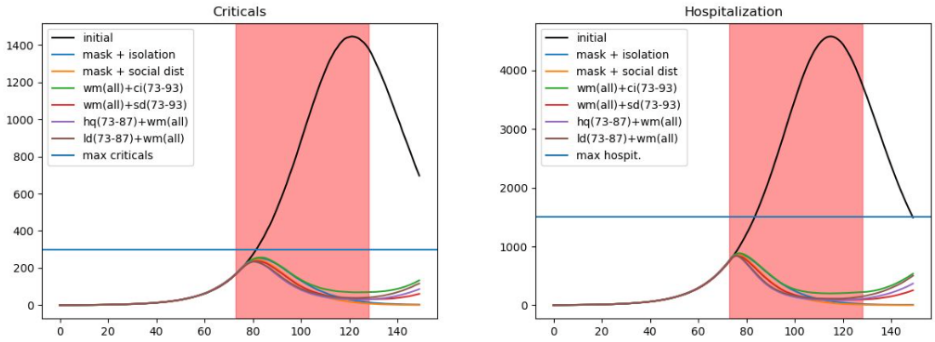

FIGURE 11: Evolution of the critical cases and the hospitalizations with regards to the combinations of measures.

Our recommendation respects the thresholds below which we must stay for critical cases and hospitalizations. Moreover, our recommendation is not very restrictive.
Note that in order to be able to easily compare the different curves, we used the means of our stochastic predictions in order to have a fine curve for each combination of measures.
Our recommendation obviously respects the thresholds for every simulation of the stochastic predictions.

8 Comparison to received data and improvements

In *fig.*12 and *fig.*13, we can observe respectively the comparison between our prediction distribution and the given epidemic evolution, for hospitalization and critical curves. The measures taken to generate these data are the wearing of masks and social distancing from day 73 to 128. As we can see, the data received does not fall within our confidence interval (threshold 97.5%). Our predictions seem to show a more rapid effect of the measures on the growth of the hospitalisation curve. As for the critical curve, we can observe a gap of a few days between predictions and observations. In both situations, the measures taken seem to meet the given target of less than 1,500 hospital admissions.

This difference between our predictions and observations could be explained in several ways:

- During the writing of this report, some small errors were discovered that could induce bias. Unfortunately, the fitting process is very long due to the brute force strategy
- As we have seen, many combinations of parameters can give very similar results during the fitting procedure. It is therefore possible that our model converged to a different combination of parameters than the one that best represents the initial model. We can notice that the effect

of taking measurements is faster for our predictions than for observations. This may mean that the actual incubation period of the disease is longer than the one we have estimated.

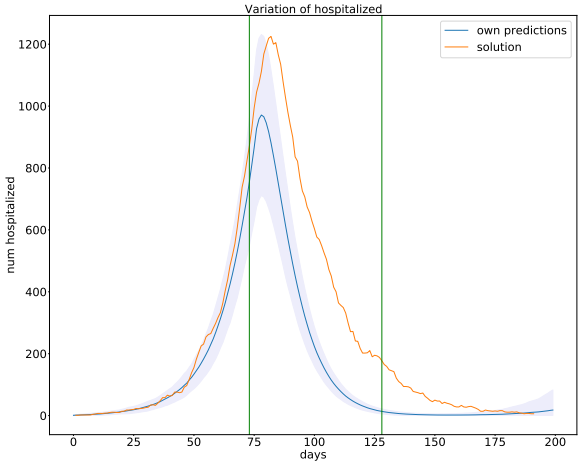

FIGURE 12: Difference of the H state between our predictions (95% confidence range) and the real world

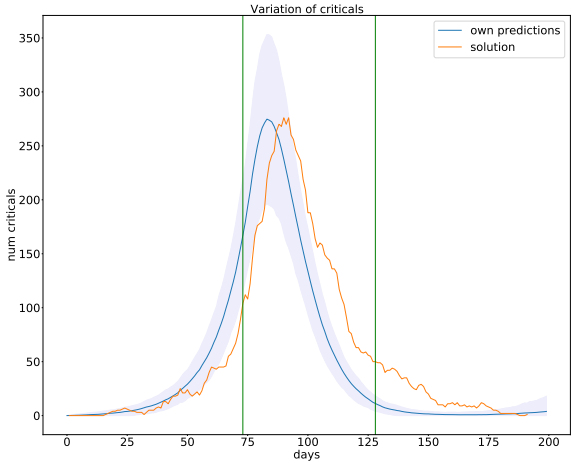

FIGURE 13: Difference of the c state between our predictions (95% confidence range) and the real world

8.1 Improvements

In order to improve the performance of our model, we will restart our bruteforcing procedure with the corrections of the bugs we have identified. This should allow us to find combinations of parameters to better approach the observations. We also think it would be interesting, in the future, to test our implementation on real covid-19 data, in order to see whether or not our model could be an indicator to determine actions to take in a more complex real world.