

PROJ0016-1
BIG DATA PROJECT
Bertrand Cornélusse
Pierre Geurts
Gilles Louppe

Intermediate review 3

*Refinement of our model. Taking into account the
uncertainty*

Julien Hubar
Pierre Dumoulin
Andreas Duquenne
François Lievens

*Master 1 Data science Engineering
Faculty of Applied Sciences
Academic years 2020-2021*

1 Introduction

In the last review, we found a model that met our expectations and that corresponded well to the representation we have of the epidemic.

Find this model we named **SEIR+** below:

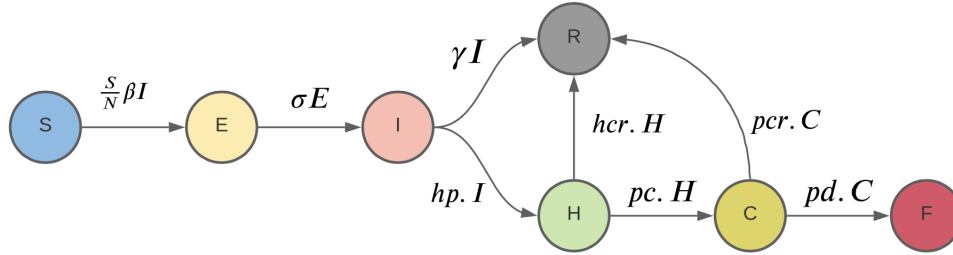


FIGURE 1: SEIR+(SEIRHCF) model

For this review, we have worked on improvements to our fitting methods, and introduced new concepts to be taken into account in order to allow the uncertainty in our model to be considered and to make our model more consistent with reality.

2 Update of the last review

In the previous review, we did not take into account the number of people tested as well as the sensitivity of the test. This therefore resulted in an underestimation of cases of infection for our curves.

In order to overcome this, we now consider the values of the positive cases(found among symptomatic people tested) to be a proportion of these real cases. The exact value therefore being included in a value interval determined with the aid of the possible sensitivity value interval(Find the value interval of positive cases in orange on the [fig.2](#)). From there we can therefore notice that among the tests carried out, "false-negatives" appear, these are symptomatic individuals tested negative while they are infected with the virus. We can now consider these individuals in our model thanks to the sensitivity and introduce a kind of uncertainty related to the sensitivity of the test.

Moreover, we know that the tests are carried out only on symptomatic individuals but that these tests are not carried out on all of these people. From the value of the number of daily tests carried out (which therefore gives us information on the number of symptomatic tested each day), we can determine an interval of value describing the true quantity of symptomatic individuals (knowing that we have between 50 and 100% of symptomatic people tested daily).

The value interval of symptomatic people can also be find in blue on the [fig.3](#)).

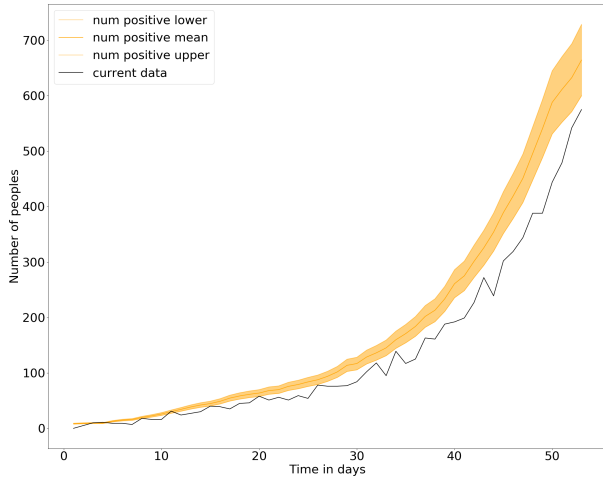


FIGURE 2: Positive tests and interval of positive cases including false negatives.

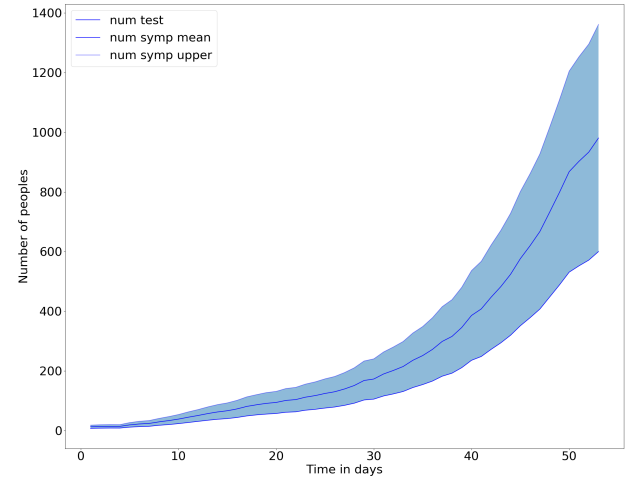


FIGURE 3: Curves of the number of test and interval of symptomatic people.

By combining this new information, we can therefore improve our estimates of positive cases for our model.

3 Fitting

Since the beginning of the project, we have been using the minimum numerical optimizer of the *scipy* package in order to fit the values of our parameters to the received data. To do so, the optimizer implemented will generate the epidemic curves of the model for each combination of tested parameters. These curves are then compared to the set of data containing the measures observed by the objective function.

From the prior beliefs on the distribution of the observations each day according to the real values, it is possible to calculate the joint probability of all the values observed in our dataset given our predictions. So, the probability to be tested for a symptomatic people follow a binomial distribution.

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where,

- n is the real number of symptomatic people.
- k is the observed values of daily test,
- p the testing rate.

The probability to be positively tested follow the same logical but with a probability the product of the testing rate and the sensitivity to take account the fact that each infected people can't return a positive test. The same process can also be used to compare hospitalization, intensive care and death curves with their respective forecasts.

However, the data for the infectious, these observations can be considered accurate and reliable, albeit with some noise. In order to allow our optimizer to have a value to optimize that remains stable with the exponential growth of epidemic curves, we will once again use a probability mass function

of a binomial distribution. But this time, we will use an hyper parameter '*binomial-smoother*', whose inverse will be the probability used and who will multiply the predicted daily value to obtain the n (number of symptomatic people) of the binomial.

We can so consider the product of theses probability values over each days data like a likelihood probability between our predictions and observed data that we can maximize. Optimizing the variables of our model is therefore equivalent to maximizing the sum of the logarithms of these probabilities. Given the complexity of our model, hyperparameters have been added in order to influence the behaviour of the objective function and to allow the optimizer to converge more easily towards a solution. In addition, weights are added in order to vary the relative importance of each curve compared during the fitting process.

We have also extended the concept of '*binomial-smoother*' with an hyperparameter that can be applied to each part of the objective function. This hyperparameter thus makes it possible to manage the enlargement of the probability mass function(see *fig.4*) of each of the binomials. Thus making it possible to linearize more the evolution of the probability returned according to the gap between the predictions and the observed values. The different combinations of these hyper parameters can then be tested in our model selection process in order to determine which ones provide the best convergence towards a solution.

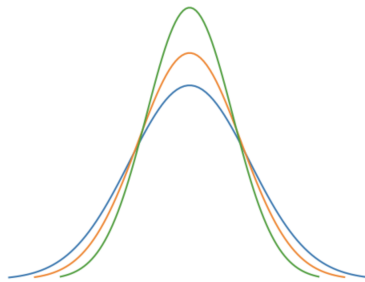


FIGURE 4: Illustration: enlargement of a probability

4 Model selection

Since the search space seemt to be not convex, there are several combinations of the values of the model state transition parameters (**MSTPs**). Therefore, one of the acceptable solutions to solve this problem is to define a set of values for the **MSTPs**. Once these parameters have been chosen, it is necessary to improve the values of the **MSTPs**. To do this, a brute force optimization has been carried out. This was necessary to achieve a convergence of the *scipy.minimize* algorithm. This is very sensitive to the initial value.

In order to perform the optimization of **MSTPs** by brute force the model has been slightly reduced (see *fig.5*) in order to make the computation time reasonable.

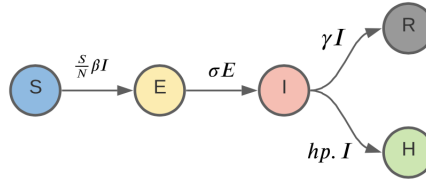


FIGURE 5: Caption

This led to the identification of the following values in *fig.6*:







MSTP		
β		[0.375]
σ		[0.313]
γ		[0.318]
h_p		[0.017]
s		[0.74]
t		[0.37]

FIGURE 6: First range of MSTPs values

5 Uncertainty

We previously introduced the uncertainty related to the interval of values given for the sensitivity of the tests as well as for the variable proportion of symptomatic people tested (**see part 2.Update of the last review**). We can now, when moving to a stochastic model, introduce a new kind of uncertainty.

A stochastic model is defined by a set of random variables characterizing the evolution of the system over time. This kind of model provides a dynamic dimension to our model, adding a temporal variable. Therefore this variable allows us to take into account modifications of the evolution of the epidemic such as a lockdown. In the case of this epidemic, the random variables will be defined as the transition from one state to another (several other) state(s).

This transition is a quantity given by a multinomial random variable. A multinomial random variable takes as input, a quantity of experiences (here the number of people in a given state) and a set of probabilities (here the probabilities of exits from one state to one (several) other (s) state(s)). The number of people in a given state is a variable whose dynamics depend on equations describing the inputs/outputs relation from one state to another.

An illustration of this idea can be done through a walk as in the figure below (*fig.7*).

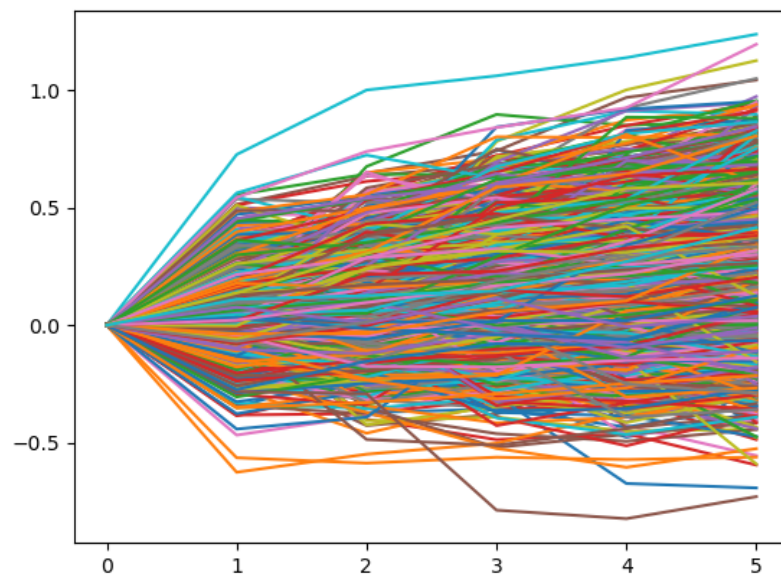


FIGURE 7: stochastic model defined by a set of random variables characterizing the evolution of the system over five time steps

The usefulness of a stochastic process in the monitoring of an epidemic is to be able to obtain predictions of the behavior of the curves by means of certain uncertainty (these curves can be found in *fig.8*). Being able to introduce uncertainty is essential in models like this in order to be able to give weight to our forecasts and therefore to be able to quantify the legitimacy of our potential actions.

Compare_stocha_and_deter.pdf(smoothed=True)

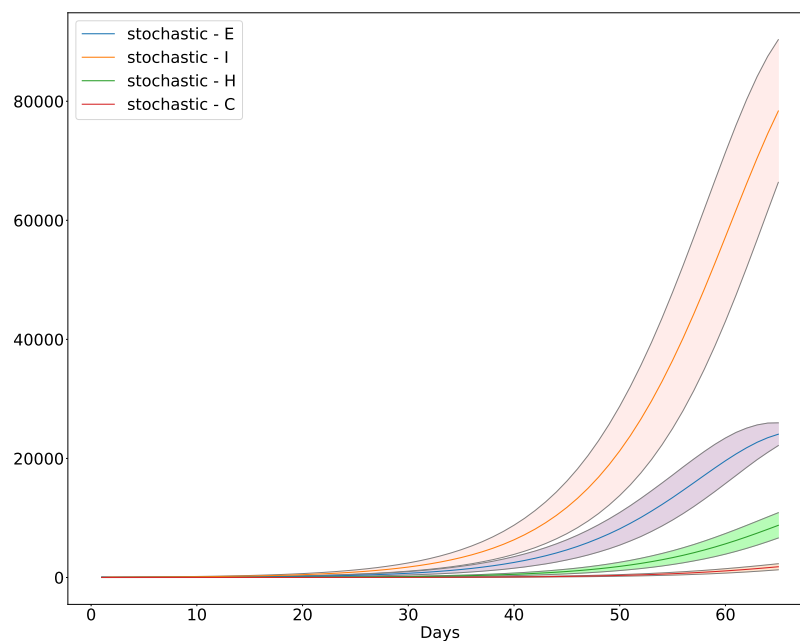


FIGURE 8: Prediction of the curves considering the uncertainty

As can be seen on *fig.9* and *fig.11*, we have been able to generate stochastic predictions of the

epidemic. Using the mean and the standard deviation of those predictions we computed a 95% confidence interval. Thus, *fig.10* and *fig.12* shows the comparison between both models and the confidence range.

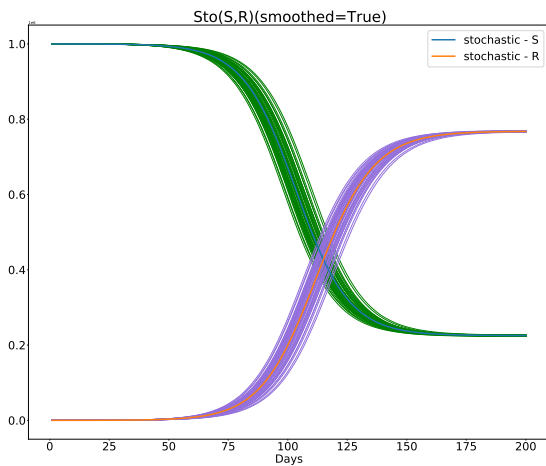


FIGURE 9: Stochastic model for S and R states

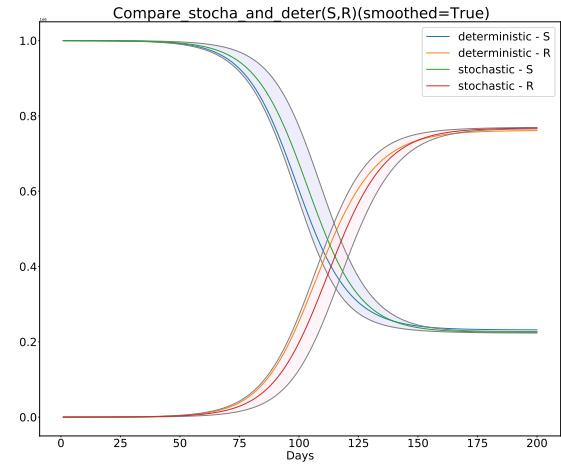


FIGURE 10: Comparison of deterministic and stochastic models for S and R states

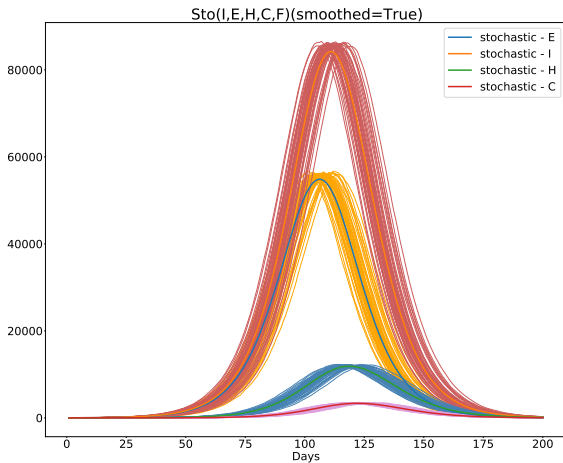


FIGURE 11: Stochastic model for E, I, H and C states

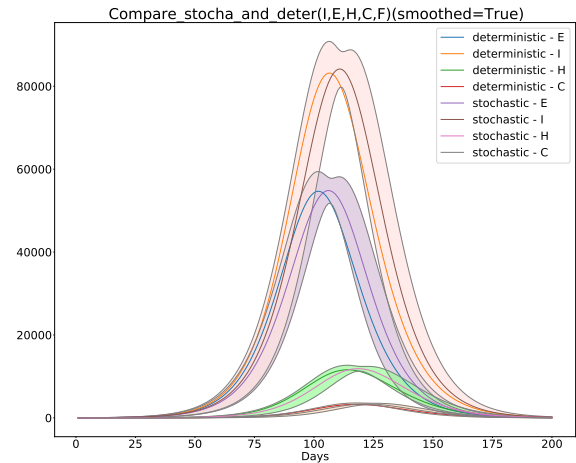


FIGURE 12: Comparison of deterministic and stochastic models for E, I, H and C states

In order not to create new people in our population. We had to be careful that the individuals could not go in two compartments at the same time (see *fig.13*). In fact, because of the multiple transitions from one state to many other, it should have been possible to have more people coming in those states than leaving the initial one. This is why multinomial distribution was used to avoid this issue.

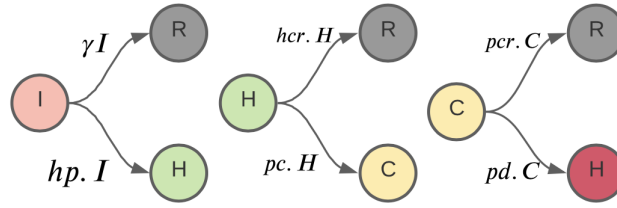


FIGURE 13: Steps in which individuals could find themselves in two states

6 Improvements

In order to be able to deliver a recommendation in an optimal way for the next review, we will then implement, on the bases of our stochastic model, a first-order Markov process(see [fig.14](#)). This process is a stochastic model that will provide us decent predictions considering the consequences of potential actions relating to our recommendations.

A Markov process fit well with the study of the epidemic because we can work with a first-order Markov process that consider that the element X_t and the elements $X_{0:t-2}$ are conditionally independent given the element X_{t-1} .

This is indeed representative of our epidemic. The number of people in each compartment of our model is computed at each time step considering the value of this same compartment at the time step $t - 1$.

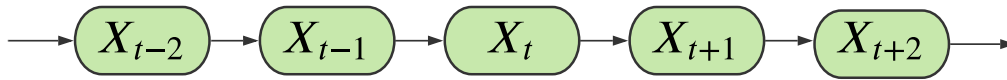


FIGURE 14: First order Markov process

This process is made up of 4 inference tasks:

1. Prediction: prediction are already manage in our stochastic model by the use of the objective function explain in [3.Fitting](#) where we use the data find as described in [5.Uncertainty](#).
2. Filtering: filtering inference task will be add for the next review in order to re-weight our beliefs and to decreases the uncertainty at each time step. This will help us to search after the optimal recommendation to do.
3. Smoothing: smoothing task will help us to reconsider our predictions with sufficient hind-sight(given by the new evidences).
4. Most likely explanation(MLE): MLE will helps us to refine our estimates at each step to choose the best way.