Project coronavirus: Predicting ICU occupation over time

Noam ENCAOUA

I. Description of project and questions addressed (these will likely evolve over the course of the project).

My project first consists in predicting the close trend in ICU occupation due to COVID-19, which seems to be the key factor to handle any covid19 outbreak across the world.

I used the French hospital dataset https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/ containing hospitalized, ICU, returned home and dead people over time in every region.

Questions addressed: Predict ICU occupation over time. If possible, predict the population now immune to the covid19, and the number of deaths in hospitals.

II. Discussion of what methods were used, how they were used, and what motivated your choices.

I started using time delays method, multiple linear regression on the n recent days and a simple HIRD (Hospitalized, ICU, Returned home, Dead people).

All these models worked poorly on prediction and had a terrible convergence because they were based on the fact that H, I ,R and D were not predicted values or predicted from predicted values. The basic idea behind that is that as long as you're predicting with predicted H, I ,R or D, it starts diverging terribly.

III. You will likely discuss things like data preprocessing, feature selection, feature extraction and transformation, regularization, model optimization, model selection and cross validation, etc.

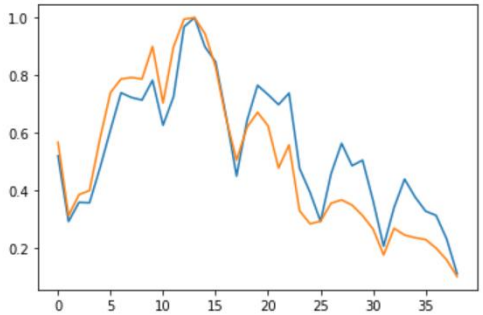
I also tried a simple model using time delays and their derivative, which gave me a matrix with this type of column (1-day delay)

[H(1) I(1) R(1) D(1) H(2) I(2) R(2) D(2) dH(1) dI(1) dR(1) dD(1) dH(2) dI(2) dR(2) dD(2)]

It worked quite well to predict each future value like H(3) or I(3) independently.

But, it started predicting really badly when none of the feature above was known but was only a predicted, and some convergence troubles appeared quite quickly.

I would say that I had also first assume that there was correlation or in some way an information in hospitalized people that will appear later in the ICU occupation, which appears to be untrue as shown in the graph on the right.



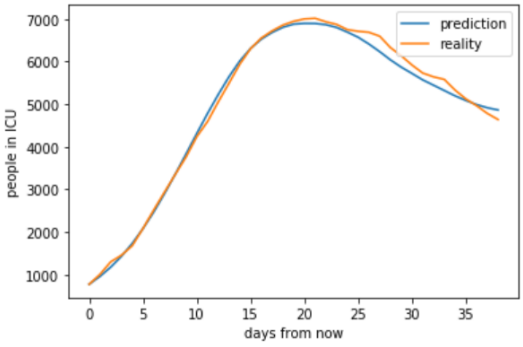
From all the previous methods above, I had 2 options:

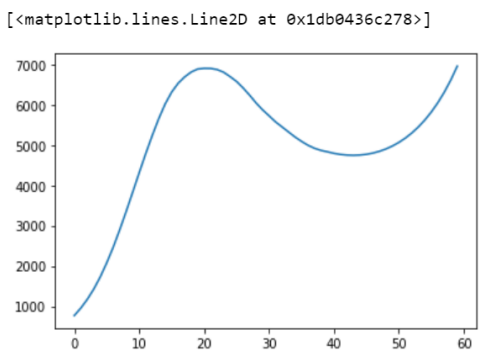
- I could consider ICU and Hospitalized data to be independent times Series and start predicting ICU or hospitalized people independently probably using neural networks …
- Pay more attention to the available epidemiological models and adapt one to the data I am using (what I chose because it made more sense to me and I knew it was going to give me more information concerning the epidemic such as immunity so far of the population …)

In order to refine my model, I decided to add the total number of available ICU units in France.

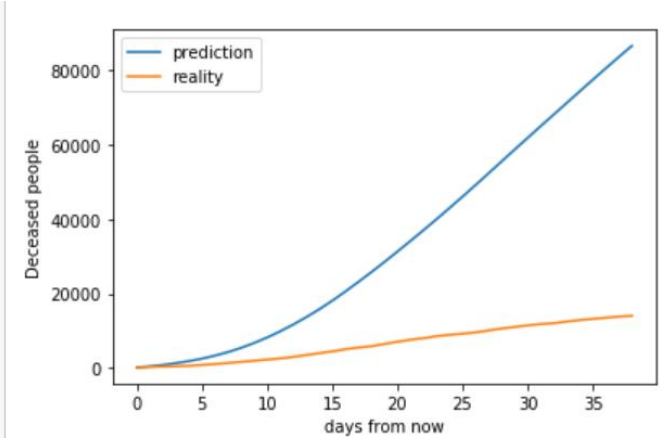
IV. Discussion of final model performance, including a full comparison to simple baseline methods. You will likely want to include plots, tables, or other figures, but do so judiciously. We don't need to see every experiment run.

Baseline Model: HIRD model



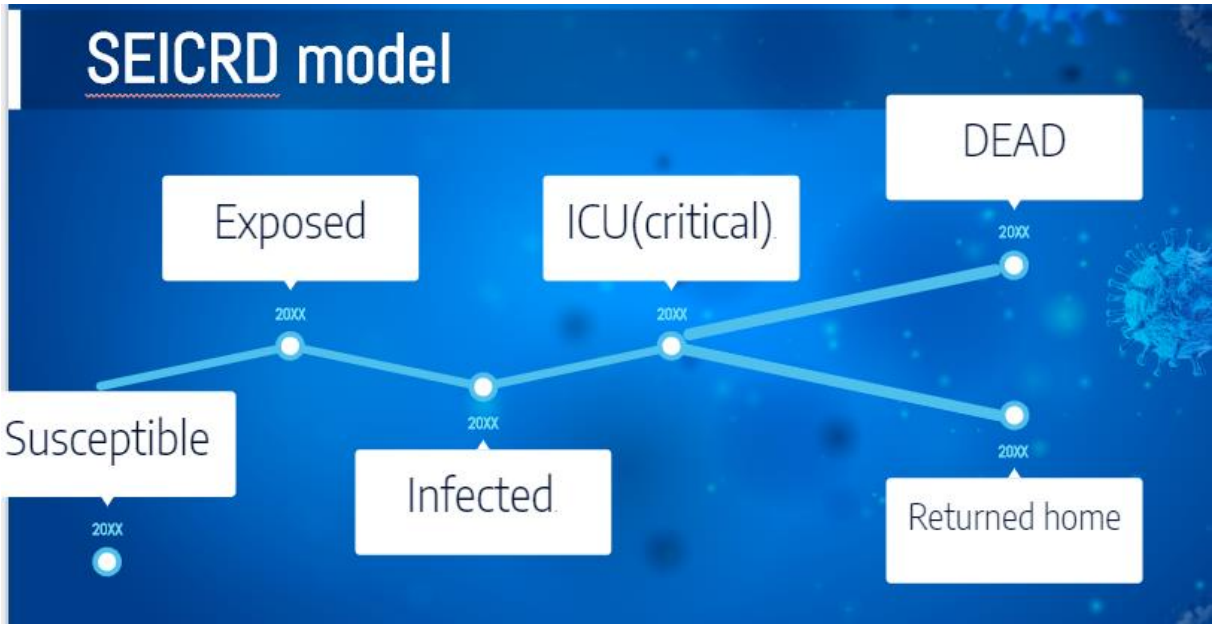


Trouble: Convergence, prediction of dead people …



Consequence: Need for a more complex model and a more refined optimization.

I opted for this model:



Using the following ODE (Ordinary Differential Equation)

```
def deriv(y, t, beta, gamma, sigma, N, p_I_to_C, p_C_to_D, Beds):
    S, E, I, C, R, D = y

    dSdt = -beta(t) * I * S / N
    dEdt = beta(t) * I * S / N - sigma * E
    dIdt = sigma * E - 1/12.0 * p_I_to_C(t) * I - gamma * (1 - p_I_to_C(t)) * I
    dCdt = 1/12.0 * p_I_to_C(t) * I - 1/7.5 * p_C_to_D(t) * min(Beds(t), C) - max(0, C-Beds(t)) -
    (1 - p_C_to_D(t)) * 1/6.5 * min(Beds(t), C)
    dRdt = gamma * (1 - p_I_to_C(t)) * I + (1 - p_C_to_D(t)) * 1/6.5 * min(Beds(t), C)
    dDdt = 1/7.5 * p_C_to_D(t) * min(Beds(t), C) + max(0, C-Beds(t))
    return dSdt, dEdt, dIdt, dCdt, dRdt, dDdt
```

```
def logistic_R_0(t, R_0_start, k, x0, R_0_end):
    return (R_0_start-R_0_end) / (1 + np.exp(-k*(-t+x0))) + R_0_end

def double_log(t, R_0_start, k_start, x0_start, R_0_mid, k_end, x0_end, R_0_end):
    return logistic_R_0(t, R_0_start, k_start, x0_start, R_0_mid) + logistic_R_0(t,0, k_end, x0_end, R_0_end-R_0_mid)
```

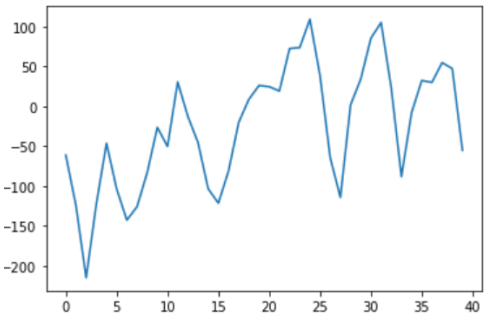
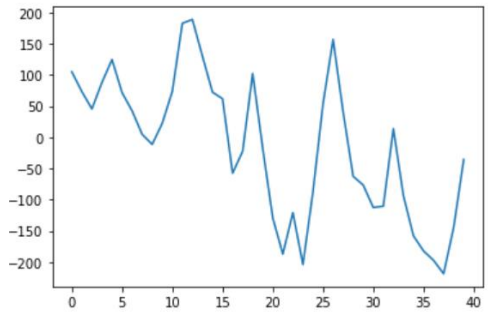
```
def Model(days, agegroups, beds_per_100k,R_0_start, k_start, x0_start, R_0_mid, k_end, x0_end, R_0_end,pc_0_start,
          kc_start, x1_start, pc_0_mid, kc_end, x1_end, pc_0_end,pd_0_start, kd_start, x2_start, pd_0_mid, kd_end,
          x2_end, pd_0_end,s):

    def beta(t):
        return double_log(t, R_0_start, k_start, x0_start, R_0_mid, k_end, x0_end, R_0_end) * gamma
    # agegroups is list with number of people per age group -> sum to get population
    N = sum(agegroups)
    def Beds(t):
        # the table stores beds per 100 k -> get total number
        beds_0 = beds_per_100k / 100_000 * N
        return beds_0 + s*beds_0*t  # 0.003
    def p_I_to_C(t):
        pc=double_log(t, pc_0_start, kc_start, x1_start, pc_0_mid, kc_end, x1_end, pc_0_end)
        return pc   # 0.003
    def p_C_to_D(t):
        pd=double_log(t, pd_0_start, kd_start, x2_start, pd_0_mid, kd_end, x2_end, pd_0_end)
        return pd   # 0.003
    y0 = N-1.0, 1.0, 0.0, 0.0, 0.0, 0.0  # one exposed, everyone else susceptible
    t = np.linspace(0, days, days)
    ret = odeint(deriv, y0, t, args=(beta, gamma, sigma, N, p_I_to_C, p_C_to_D, Beds))
    S, E, I, C, R, D = ret.T
    R_0_over_time = [beta(i)/gamma for i in range(len(t))]  # get R0 over time for plotting
    pc=[p_I_to_C(i) for i in range(len(t))]
    pd=[p_C_to_D(i) for i in range(len(t))]
    return t, S, E, I, C, R, D, R_0_over_time, Beds, p_I_to_C, p_C_to_D,pc,pd
```

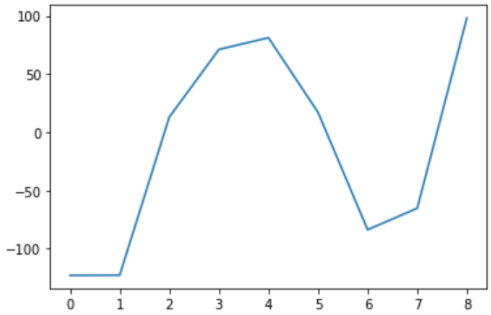
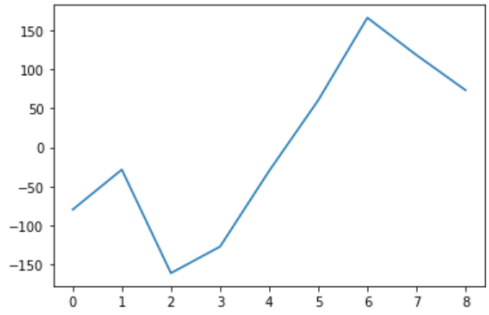
This model takes into account the number of ICU beds available.

I choose to change twice the values of Probability to go to ICU and probability to die because I wanted my model not only to fit the ICU curve but also the deaths curve. I had to change for example, the ICU admission probability when the outbreak was at its peak because even if officially ICU were not overwhelmed in France, we clearly observed much younger patients in the ICU when the epidemic was peaking (this only means all of the older patient couldn't enter in the ICU). So I had to lower the ICU admission ratio in order to state this fact. If I don't do that, I have a model where ICU are overwhelmed and thus I obtain dead people who didn't attend an ICU (data which no one really knows) .This is also why, the death rate decreased over time as younger people were in ICU.

About the error in prediction, on the training part, it looks that way (ICU error on the left, Death error on the right)

On the testing part, it looks like that:

V. A conclusion describing what you would have done with more time: are there modifications of your approach worth trying? Other questions to address? Different data sets or features

The government started to release a dataset with age categories but this dataset starts lately after the beginning of the outbreak (just before the peak I think). If the government releases in order to refine the probability to die over time, to change the duration of your ICU occupation according to the age.

I could also complexify my model, apply it to other areas in the world, it will essentially depends on the new data available…

(I could also try the approach of the neural networks but I will probably use them on another project soon anyway)