EINDHOVEN UNIVERSITY OF TECHNOLOGY

COMPUTATIONAL BIOLOGY GROUP

Bachelor Eind Project SBMD Report Logs

Student

J.H. Cao, **studentid**: 1258540, *Department of Biomedical Engineering*

Supervised by

prof.dr.ir. N.A.W. van Riel, *Department of Biomedical Engineering*

dr. D. Bosnacki, *Department of Biomedical Engineering*

Contents

1 Outreach TU/e Innovation Space - Offering Help

My name is Hao Cao and I am a 3rd year Biomedical Engineering student. This is my last year and I have started with my Bachelor Eind Project (BEP). The topic of this BEP is **COVID-19**, since this is an ongoing world crisis, where the spread of COVID-19 is a serious threat to the global health. This BEP is part of the **Computational Biology Group of the Department of Biomedical Engineering**, under supervision of **prof.dr.ir N.A.W. van Riel** and **dr. D. Bosnacki**.

This BEP will focus on different models for the modelling and prediction of the spread of COVID-19 in the Netherlands. Many existing epidemic models are based on the SIR or SEIR models. Popular model for the prediction of daily cases, is to use a logistic growth curve, developed by Pierre François Verhulst. However, it is also possible to use a more dynamic model, such as the **Kalman filter**. Since the spread of COVID-19 in the Netherlands can be seen as a system that is continuously changing and as a real-time problem, using a Kalman filter might be an ideal approach.

The research of this BEP is to compare the three types of models:

1. SEIR (dynamic simulation model - can work independently from data)
2. logistic growth - data analytics / statistics
3. Kalman filter - combines system dynamics, (real-time) data and statistics/stochastics

, on how accurate and effective these models are on the prediction of

1. Daily disease counts in the Netherlands (to model this epidemic)
2. Furthermore, it is also interesting to apply these models, not only on the daily disease counts, but on the number of hospital admission as well as the Intensive Care Unit (ICU). The latter one is important, since it has a maximum capacity.

Lastly, with the SEIR model, an attempt is made to estimate the number of **"E"xposed** (to COVID-19) in the Netherlands, since this remains a big challenge and concern.

2 Preface

The purpose of this **Report Log** is to keep track of the notes, tasks and every other information provided during the Bachelor Eind Project (BEP) meetings. The table of contents contains a list of many different dates. For each of the given date, with the name ***Notulen***, there is a description, notes or information received during the meeting. The dates named ***SSA***, contain a summary or listings of all the findings/information and tasks performed on that day as a self-study assignment. Lastly, the section ***Results*** contain the results of the performed SSA of the given date.

3 Introductory meeting 19-march-2020

This section contains information regarding the introductory meeting of this BEP.

3.1 Notulen

The introductory meeting was mainly focused on the different kind of topics regarding the **BEP SBMD**.

The different **topics** were:

1. **Hospital patient's data/clinical data regarding**:
   (a) Obesitas after gastric bypass surgery
   (b) Heart failure, monitoring Troponin
   (c) Liver transplantion template model
2. **Dynamic, biological data**:
   (a) Diabetic patients
   (b) Metabolism/reaction on different meals
   (c) Parameter estimation
3. **Sars-Cov-2, aka COVID-19**:
   (a) Epidemic models
   (b) estimation of the infected/exposed

3.2 SSA

The tasks to be completed as self-study assignment before the next meeting are:

- Fill and sign BEP form
- Read the articles and give preferred topic

3.3 Results

After the given tasks, it was important to read and analyse the different articles and information, given by prof.dr.ir N.A.W. van Riel. All topics were interesting, but the topic of **COVID-19** is the most "recent and ongoing" topic and is a major concern for the global health.

The obtained results or the completed tasks from this SSA are:

1. The BEP form is signed and sent
2. Read all the given articles
3. The preferred topic is **Sars-Cov-2 aka COVID-19**

4 "Startbijeenkomst" 08-April-2020

The very first meeting after choosing the preferred subject.

4.1 Notulen

The chosen subject for this BEP is **COVID-19**.

Interesting topics to research regarding the subject of COVID-19:

1. Kalman filter to predict the number of daily infection
   (*https://towardsdatascience.com/using-kalman-filter-to-predict-corona-virus-spread-72d91b74cc8*)
2. The main programming language to be used is Python
3. Public Data as resource for model training
4. John Hopkins map and data is useful
   (*https://coronavirus.jhu.edu/map.html*)
5. Using the SIR/SEIR model
6. Making a blogpost and Github as "final report"
7. Should focus on the Netherlands
8. Would be good to compare 2 or 3 models (Kalman vs Logistic fit vs SEIR model)

Furhtermore, there is an outreach of the TU/e regarding research of the COVID-19.

4.2 SSA

The tasks to be completed as self-study assignment before the next meeting are:

- Study the Kalman filter and its applications
- Take a look at the outreach of TU/e regarding research of the COVID-19
- Make a Github repositories for the codes of this BEP

4.3 Results

It was important to read the medium article regarding the modelling of COVID-19, using the **Kalman Filter**. The Kalman filter is combines system dynamics, (real-time) data and statistics/stochastics. Furthermore, since the spread of COVID-19 in the Netherlands can be seen as a system that is continuously changing and as a real-time problem, using the Kalman filter might be an ideal approach.

The obtained results or the completed tasks from this SSA are:

1. Kalman fitler and its application is studied. Background information is obtained
2. Wrote a proposal for the application regarding the TU/e - outreach for the research of COVID-19.
3. Public repository, containing the codes and progress of the work, is available on Github at : **https://github.com/jiehaocao/BEP-SBMD-COVID19**.

5 Meeting on 16-April-2020

This time, dr. D. Bosnacki joined the team as a supervisor.

5.1 Notulen

It is difficult to use the SEIR model, since the are currently no available data regarding the **Recovered** patients of COVID-19 in the Netherlands. dr. Bosnacki suggested to use a model without the recovered, since the recovery at the beginning is almost negligible. Futhermore, it is difficult to know the **Exposed**, since it is difficult to track the amount of people who are actually exposed to this virus, due to lack of test kits.

It is still advisable to study the Kalman fitler, and applying this to the data of RIVM, made available to the public. This way it is possible to check the performance of the Kalman filter on the data of the Netherlands. Another suggestion from dr. Bosnacki, was to use the data regarding the amount of patients taken in the hospital or ICU. This is more trustworthy and this data is also available. Furthermore, it would be possible to use the Kalman filter to predict the daily hospitalized patients due to COVID-19. Lastly, it is advisable to write a proposal for the TU/e innovation space, regarding the outreach and offering help to the research of COVID-19.

5.2 SSA

The tasks to be completed as self-study assignment before the next meeting are:

- Try to model and find a parameter for **Exposed**
- Model the hospital intake
- Try to implement the codes of Kalman filter, specifically on the data of the Netherlands (predicting the daily infected people)
- Proposal for TU/e innovation space (Offer help)
- Study the potential SEIR model

5.3 Results

The number of **Exposed** can be estimated from the SEIR model, if and only if the estimated parameters are accurate. Since the SEIR model, is a dynamic model and consists of differential equations, it is really sensitive to outliers, thus the different parameters. By changing of the parameters, the graph and model would have a complete different output.It would also be possible to add "Hospital" as an extra compartment to the SEIR model, since these models are compartmental models.
The implementation of the Kalman filter was unsuccessful. There were many errors when converting the R script to Python. Therefore, it was better to focus on the SEIR model and implement this in Python for to try to model the epidemic in the Netherlands.

The obtained results or the completed tasks from this SSA are:

1. **Exposed** can be estimated with the SEIR model, but is highly dependent on the estimated parameters (for the differential equations of the different compartments).

2. Extra compartment con be added to model the hospital intake.

3. Proposal is written and sent.

4. The SEIR model is implemented and available on Github repository.

5. Attempt was made to implement the Kalman filter, but is was not successful. z

6 Meeting on 23-April-2020

6.1 Notulen

It is quite difficult to estimate the parameters using the SIR or SEIR models, since this is a dynamic model. Therefore, one would have to estimate the parameters and solving the differential equations, simultaneously.

It is interesting to apply the SEIR model on the hospital admissions or ICU admissions. Furthermore, it is also possible to add extra compartments to this model,taking hospital and ICU admissions into account.

Instead of estimating the parameters, since these are sensitive (sensitivity can be tested by taking the derivative of these parameters function as time). One can also use **parameter sweep**. However, it might be better to use values obtained from literature or the ongoing research on COVID-19.

It is recommneded to focus on the **Logistic Growth Curve (Verhulst)** and using this function to estimate the parameters (less complicated), with the help of *Maximum Likelihood Estimation*. With these estimations, the SEIR model can be tweaked and a simulation can be performed on the SEIR model. Finally, the combination of the information obtained from the SEIR and Verhulst model, can be applied on the Kalman Filter.

From now on, prepare the work and questions on a Powerpoint Presentation, same as OGO computational biology (recap, work done, upcoming, questions for supervisors)

6.2 SSA

The tasks to be completed as self-study assignment before the next meeting are:

- Implement Logistic Growth Curve on Python
- Use Logistic Growth Curve to estimate the parameters using MLE
- Tweak and improve the Logistic Growth Curve
- (Optional) Implement these estimated parameters in the SEIR model

6.3 Results

The **logistic growth curves** are successfully implemented in Python and on the data of COVID-19 in the Netherlands. This logistic growth is applied on:

1. Number of confirmed cases (daily) in the Netherlands
2. Hospital Admissions (daily) in the Netherlands
3. ICU Admissions (daily) in the Netherlands

Furthermore, a **exponential growth trajectory** is also implemented in Python, regarding the data of COVID-19 in the Netherlands.

The detailed results and explanations are shown in the README.md on Github.

6.3.1 Logistic Growth

After estimating the parameters and fitting the model, it is possible to get an idea of the predicted maximum number of cases in the Netherlands. These models can also be used to simulate and check whether a lockdown had any effect, by fitting a model with data before the strict measurements and a model with data during these strict measurements. A difference in the **predicted maximum** can be a sign that there the measurements are effective.

Furthermore, this is important, because it can be predicted what the **maximum intake** is and when is "peak" is achieved. With this information, it is possible to take preventive measurements (such as necessary lockdowns, making sure that there are enough medical supplies and/or that the hospitals and ICU are not overloaded).

Lastly, it is possible that the actual data passes the predicted maximum, but these models are still valid, meaning that there exist a chance that the actual admissions can pass the predicted maximum. Therefore, it is important use these models to achieve insights when taking preventive measurements.

Note: detailed explanation and results are available on Github

6.3.2 Exponential growth trajectory

Eventhough exponential growth function are quite abstract and difficult to tell when the peak is reached, by plotting the **log-scale**, a straight line is obtained. This represents the trajectory of confirmed COVID-19 cases. By plotting the actual data (in the Netherlands), it can be seen whether the actual data deviates from the trajectory.

A downward deviation, would mean that things are getting better, but a upward deviation is a sign that more preventive measurements should be taken.

7 Meeting on 30-April-2020

7.1 Notulen

Today everything went as planned. A Powerpoint presentation with the work done and the results were given. Throughout the presentation, some important remarks and feedbacks were mentioned. One of the important feedback is to make sure that the source of the data that were used for the research, are mentioned.This is mentioned on **Github**.

The data of the daily confirmed cases and the hospital admissions are obtained from the website of **RIVM (National Institute for Public Health and the Environment)**, whereas the data of ICU admission is obtained from **Stichting Nice : National Intensive Care Evaluatie: https://www.stichting-nice.nl/**.

Furthermore, most other important remarks regarding the logistic growth model were:

1. Logistic growth is not 'suitable' for hospital admissions, because the logistic growth curve is only taking the admissions into account, but not the discharge of the patients (recovered or deceased). Thus, the section of **Hospital Admission** under Logistic Growth, is ***deprecated (on Github)***.

2. Moreover, it is not 'suitable' for ICU admissions either, because the logistic growth curve is only taking the admissions into account, but not the discharge of the patients (recovered or deceased). Thus, the section of **ICU Admissions** under Logistic Growth, is also ***deprecated (on Github)***.

For the above applications (hospital & ICU admissions), a SEIR model would be much better (since this deals with both admissions and discharge). Lastly, more the SEIR and the Parameters of the logistic growth (as well as the caveats of a logistic growth curve) should be studied.

7.2 SSA

The tasks to be completed as self-study assignment before the next meeting are:

1. Implementation of Kalman filter (on Python) and:
   - Applied on the daily confirmed cases
   - The hospital admissions
   - The ICU admissions
2. Improve SEIR model
   - Study the article and see what extra compartments are parameters are used in for the model. Furthermore, these parameters are not a constant, but rather a function of time.
   - Ultimately, it is interesting to add extra compartments to this model and try to mimic the course of the spread of COVID-19 in the Netherlands (our own parameters for each compartment).
3. Improve or analyse the Logistic growth curve carefully
   - Try to use less data (first two weeks), so the model does not use all available data for the training, but rather using the first two weeks and making a prediction. This predicted value is then compared to the data *"that have never been seen"*. Therefore, it is possible to calculate the error margin/deviation.
   - By applying the above method, more insights regarding the parameters are obtained. The parameters are not constants, as the model estimates and returns, but are rather a function of time.
   - The predicted maximum will be different than the predicted maximum when using all available data.

7.3 Results

The Kalman filter is a mathematical model that can be used as a predictor or filter. This means that this Kalman Filter can also be applied on the SEIR models or the Logistic Growth Curve, as a filter. But in this section, Kalman Filter is used as a predictor, to forecast the spread of COVID-19 in the Netherlands. Since, the Kalman Filter is ideal for systems that are continuously changing, the spread of COVID-9, which is time-dependent can also be seen as a continuously changing system.

The advantage is that Kalman Filter uses the observed measurements, which are measured over time (confirmed daily cases). These measurements can contain noise and inaccuracies (for example in this case: lack of testing, social distancing etc.), which the Kalman Filter will take into account and therefore produce values that are closer to the actual data (the prediction would be closer to the actual confirmed cases of that day).

7.3.1 Kalman Filter

The **Kalman Filter** is applied on:

1. The daily confirmed cases (**After further evaluation, this approach might be incorrect**)
2. It is chosen to apply it on the SIR/SEIR model, as a filter but also an estimator or predictor. This would mean, that the **"state-space"** would consist of the differential equations from the corresponding models.

Confirmed daily cases, can be used to update the Kalman Filter, which would enhance the next prediction.

7.3.2 SIR/SEIR model

After reading different articles, it is clear that these compartment models can have a high complexity. More compartments, such as **quarantine** or **hospital admission** can also be added. These models would not be dependent on the available data, but the available data is necessary to determine the different parameters in the differential equations found in these models. By applying the **Kalman Filter**, real-time data can be used to provide a more accurate estimation (as a filter on the SIR/SEIR models) and also for the forecast of new cases in the future.

7.3.3 Logistic Growth Curve

By using less data (20 days of the actual data) on the **Logistic Growth Curve**, it was remarkable that the model did **not** perform well, which was not unexpected.
In Figure 1, it can be seen what would happen if only 20 days of data was used to fit a model and using this (without updating the model with new available data) to predict the number of daily cases over a longer period of time. It is clear that the model's predictions are off-track.

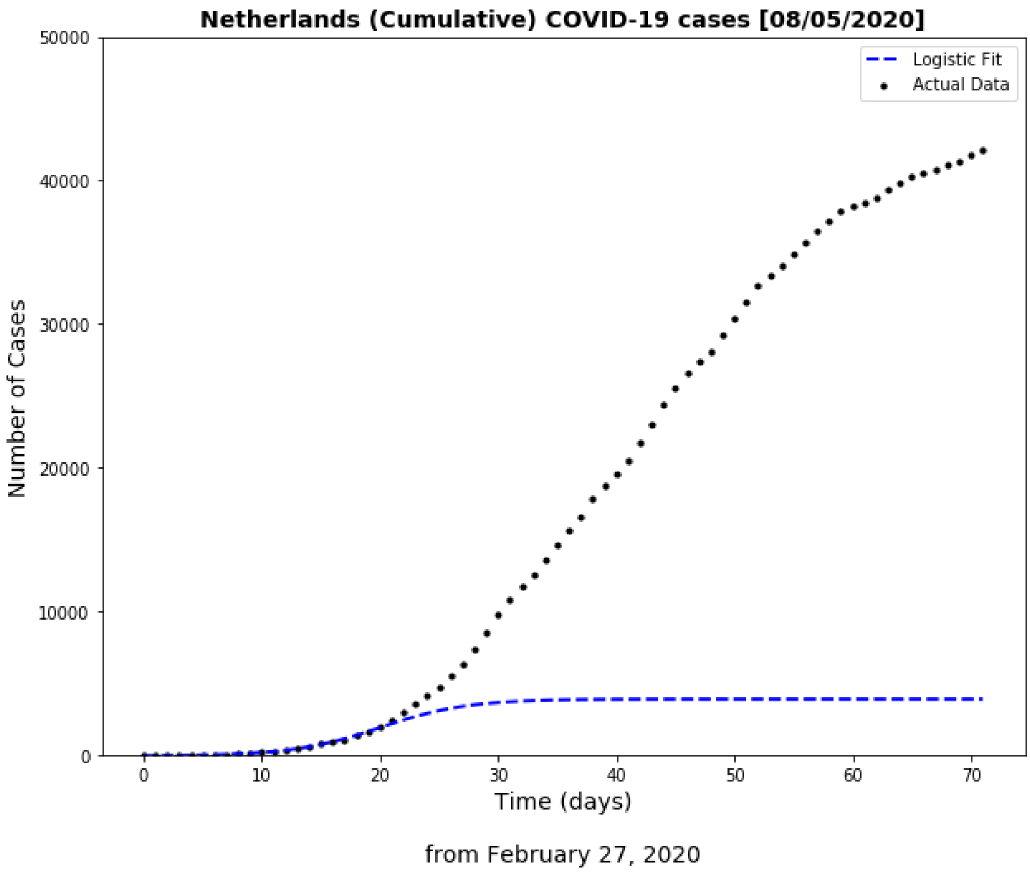

Figure 1: *Logistic Growth Curve model (only 20 days of available data and no further feeding of new data)*

Thus, the **Logistic Growth Curve** model, is a powerful model on predicting the number of cases accurately, but only for a short period of time. Furthermore, this model is highly dependent on the available data and has to be fed with new data continuously for an accurate model. Only with large data and continuously updating this model with new data, the predictions are accurate. However, the latter one is only accurate for prediction of short period of time (such as the cases in the upcoming 2-5 days).

Lastly, depending on the purpose or what question has to be answered, different models should be used. These models are only as good as the data they rely on. Such **Logistic Growth Curve** can be used to accurately predict the cases of the coming days (when sufficient data is available), but for a longer period of time, the SIR/SEIR (compartmental models) should be used.

8 Meeting on 07-May-2020

8.1 Notulen

8.1.1 Logistic Growth Curve

As mentioned before, the **Logistic Growth Curve** does have caveats. These caveats were mentioned in the previous results and also on the Github page. While this type of model does a really good job in predicting the new cases in the upcoming days, it still remains a challenge to have a *"constant"* **predicted maximum**. Each time, this model is fed with new data, the predicted maximum changes, thus it is really difficult to have a definite estimated **predicted maximum**. One challenge would be to use the available data and statistics to achieve a rather accurate estimation of the **predicted maximum** (in the coming weeks or days). An approach for this latter problem is to vary the predicted maximum with time, this way the predicted maximum would be in a certain range.

At last, it depends on the type of question that has to be answered in order to choose the proper model. No models are "good" or "wrong", but it depends on its applications and are only as good as the data they rely on.

8.1.2 SIR/SEIR model

These model, as mentioned before, can get really complex. For the Netherlands, it would be interesting to add **"Verzorgingshuizen"** also known as **Nursing home** as an extra compartment to this model. The reason is because there are many elderly who passed away during this COVID-19 pandemic, but is not certain that the cause was COVID-19, since these elderly at the nursing homes do not get tested (whether the death is indeed caused by COVID-19). This new compartment model can be seen in Figure 2 below:

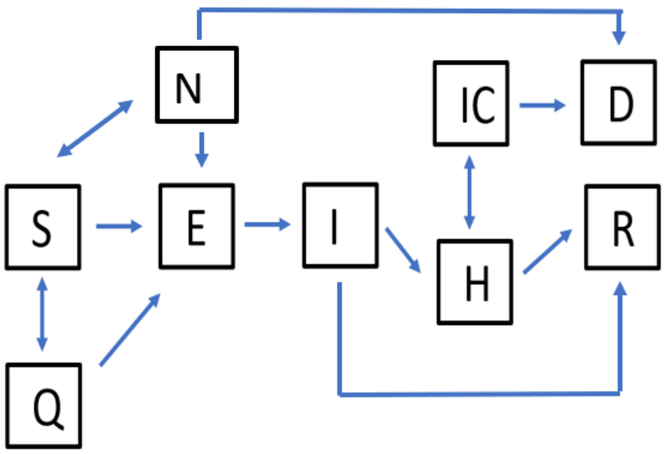

Figure 2: *Extended SEIR model : S = Susceptibles, Q = Quarantined, E= Exposed, N = Nursing Home, I = Infected, H = Hospital, IC = Intensive Care, D = Deceased, R = Recovered*

Thus, if there are data available regarding the nursing homes in the Netherlands, it would be possible to make an assumption for the in-and-out flow of this compartment (assumption or estimation of the parameters for the differential equation).

Furthermore, with these models, it is possible to make the parameters time dependent. This way the effect of the "lockdown" can be simulated and might give insights on how these parameters develop before, during and after lockdown.

Lastly, the Kalman Filter can be applied to these compartment models, as this was already mentioned.

8.1.3 Kalman Filter

There are many applications of the Kalman Filter. Some possible applications can be:

- Using the Kalman Filter to estimate the parameters after lockdown. This is to check if the rates or parameters have changed (to lower values/rates).
- Using the Kalman Filter to estimate the spread of COVID-19 after lockdown was announced.

- Adjusting the Kalman Filter to predict new cases over a longer period of time, since it is accurate for predicting the cases of the upcoming day, but can have high uncertainty when predicting the upcoming week (just like weather forecasting).
- This Kalman Filter approach might even be more accurate than the Logistic Growth Curve.
- Kalman Filter applied on the SIR/SEIR might actually be a discretization of the SIR/SEIR models' differential equations.

Thus, there are many applications where one can make use of the Kalman Filter, whether it is for estimation/prediction, smoothing or filtering. It can even be used for estimating the parameters, thus constantly updating these parameters over time. Depending on the type of application, one can consider using a Kalman Filter.

8.2 SSA

The tasks to be completed or interesting to look into as self-study assignment before the next meeting are:

1. Try to find an algorithm or method to have an accurate estimation of the so called ***predicted maximum*** for the Logistic Growth Curve, since this remains a big challenge with such model.
2. Add Nursing Home compartment to SEIR model and play around with the parameters.
3. Play around with the parameters, by making them time-dependent, to check whether lockdown actually has any effect on the spread of COVID-19.
4. Read articles or information regarding the update of parameters by using a Kalman Filter. Thus, using the Kalman Filter to estimate and update the parameters over time.
5. Read the blogpost and try to figure out which **"state-space"** the author used, or was it just using ***time*** as a "state".
6. **Logistic Growth** VS **Kalman Filter**: Which is best for forecasting and when should it be used? Furthermore, which is better for long period or short period forecasting?
7. Kalman Filter applied on the SIR/SEIR might actually be a discretization of the SIR/SEIR models' differential equations. Take a look at this phenomena.

8.3 Results

A **General Introduction** is added to the Github page, giving a short description on the work done. Furthermore, the **TU/e - Outreach** for this BEP is attached in this **Report Logs**.

8.3.1 Logistic Growth Curve

It remains a challenge to have an accurate estimation of the ***predicted maximum***. If the amount of tests performed are lower or vice versa, this can have a great impact and increase uncertainties in these logistic growth models. Nevertheless, with these models, you can get an impression to see if lockdown measurements are indeed working, by analysing the growth model and its parameters. Recalling that α is the number of days at which the expected number of cases is half the maximum **C**, whereas β is the growth factor.

From analysis it shows:

1. That $\alpha = 18.4$ before lockdown (***"intelligent lockdown was announced and imposed on March 16, 2020 by the Dutch Government***), is much smaller than the $\alpha = 41.7$ after lockdown.
2. Whereas, the $\beta = 0.30$ before lockdown was also higher than the $\beta = 0.12$ after lockdown.

The above analysis is expected, because it is expected that the growth parameter would be smaller and decrease when a lockdown is imposed. This would mean that the number of days at which the expected number of cases is half the maximum, would be larger. Thus, both phenomena shows the effect of *"flattening the curve"*, where the it takes longer to reach the peak and the peak is smaller because the growth parameter is smaller.

However, it still remains a challenge to estimate an accurate maximum for the Logistic Growth Model. The previously estimated maximum was **C** = **43136**, but the **confirmed cases of 13-05-2020** = **43211**. Thus, it is clear that there is under modelling and underestimation of the severity, because as of today's (*13-05-2020*) daily cases have already exceeded the estimated maximum.

The main reason for this under modelling or underestimation is because the model contains a parameter C, which is the maximum capacity that should be time-dependent but unfortunately it is remains as a constant. Therefore, underestimation can be seen and this will eventually affect the model's performance.

A proposed solution is to make all 3 parameters, found in the *Logistic Growth Model*, time-dependent.

8.3.2 Kalman Filter

As mentioned before, there are two main applications (regarding the epidemic modelling of COVID-19 in the Netherlands) of the Kalman Filter that are being analyzed and discussed in this BEP.

Note: For in-depth analysis and the results, please check the Github Page

The two main applications are:

1. A self-defined **exponential growth system with randomization on the growth parameter** (based on the available data - daily confirmed cases)
2. The **SIR/SEIR** model, to take uncertainties into account and use this as a filter on the dynamic systems of SIR/SEIR models

In this result, the focus is on **the first application**. An explanation on how the Kalman Filter (model) was developed and used, for the simulations and predictions regarding this application, is given in the following steps:

1. First of all, in this model, the standard Kalman Filter is used meaning that this Kalman Filter application is quite simple.
2. Calculate the most recent growth factor that can be obtained from the available data (*daily confirmed cases*). This is calculated with the following equation, where **N = number of cases, and x = day**:

$$growthfactor = \frac{N_{x+1} - N_x}{N_x - N_{x-1}} \tag{1}$$

   Thus, the growth factor is the **ratio between two successive changes**.

3. The growth factor is most likely "not constant" (this also depends on the amount of tests performed each day, but this is neglected for now), thus we add some noise to this growth factor that will account for the uncertainties. The prediction is calculated with the following equation:

$$\text{Predicted number of cases} = (growthfactor + noise) * N_{x+1} \tag{2}$$

4. Since the growth factor and the number of cases are developing over time, this can be seen as a time-series system. Thus, each day there are new number of cases.

5. The Kalman Filter is used to take all the uncertainties into account (noise, potential lack of test-kits, sudden rise in testing or poor administrative reporting etc.). Since the above predictions are in the *a priori state*, the new available data of each day (*daily confirmed cases* are used for the update of the Kalman filter, where the *a priori state* will transition to the *a posteriori state.*

   However, the simulation and modelling with the Kalman Filter does not give or represent the true number of infections in the country, but it is giving a prediction based on the available data and information

By implementing those steps into a Kalman Filter model, it was possible to run simulations and generate numbers for **long-term** and **short-term predictions**.

The results for the **long-term predictions -** ***when running the simulation on 11/05/2020*** and the **short-term predictions -** ***running the simulation when new data is available*** are given in Figure 3 & 4 , respectively.

It is important to note that the Kalman Filter Prediction for 14-05-2020 = 43480.0 found in Figure 4, is different than the Kalman Filter Prediction for 14-05-2020 = 43499.0 found in the Figure 3 that shows the results of the long-term predictions. The main reason is because new available data (daily confirmed cases) is being fed to the model, where the model retrains itself (also taking uncertainties into account) and produces a new prediction. While this is a simple model that uses the standard Kalman Filter, it is clear that the predictions are not too far off the confirmed cases.

Thus, when more complex components are added to the model and switching from standard Kalman Filter to the *extended Kalman Filter*, the Kalman Filter can be used for epidemic modelling and can produce promising results.

	Kalman Filter Prediction
2020-05-13	43223.0
2020-05-14	43499.0
2020-05-15	43818.0
2020-05-16	44189.0
2020-05-17	44638.0
2020-05-18	45169.0
2020-05-19	45785.0
2020-05-20	46517.0
2020-05-21	47394.0

Figure 3: *Results regarding the long-term predictions of the Kalman Filter after running the simulation on* ***11/05/2020***

f

	Confirmed Cases	Kalman Filter Prediction
2020-05-12	42984	42980.0
2020-05-13	43211	43223.0
2020-05-14	?	43480.0

Figure 4: *Results regarding the short-term predictions of the Kalman Filter after running the simulation on* ***13/05/2020***

9 Meeting on 14-May-2020

9.1 Notulen

It was a productive meeting where many results, regarding the **Logistic Growth Model** and **Kalman Filter**, were discussed. Furthermore, there is an ongoing research at the **Catharina Ziekenhuis in Eindhoven** which includes mathematical modelling using the data of the COVID-19 patients. The main goal or purpose is to model "the fate" of the patients and hospital environments (for example: the likelihood that a patient will transfer to the IC).

9.1.1 Logistic Growth Model

Both the short-term and long-term predictions were not too accurate. The model was not performing well and underestimating the severity of the spread of the infectious disease. Moreover, for *long-term predictions* the model was becoming very unstable. This is due to the parameter **"C = Maximum Capacity"**. A possible solution or fix for a more accurate estimation of this parameter, is to use the **Von Bertalanffy function**. However, this is still uncertain thus it would be best to send an e-mail to the research team of *prof. Edwin van den Heuvel* to ask how they used that function to make adjustments to the Logistic Growth Model.

Another possible solution is to use **Bootstrapping** for a more accurate estimation of the parameters (confidence interval / sampling covariance) in the Logistic Growth Model.

9.1.2 Kalman Filter

From the last results, it can be seen that the Kalman Filter does a good job in its short-term predictions. However, without new updates, the long-term predictions of the Kalman Filter Model becomes much more unstable which was expected. The **Kalman Filter Model**, mentioned in the last Results section, was based on a simple model. Thus, by implementing more factors to the model the complexity can be increased and might result in more accurate predictions.

In summary, the Kalman filter can even be applied on epidemic models and of course many other biomedical applications. Even if the Kalman filter predictions does not say much about the course of COVID-19 in the Netherlands (except for showing that the predictions are quite accurate), it is clear that the Kalman filter is a powerful filter that can be used when dealing with large uncertainties. If this Kalman filter is applied on more complex epidemic models, an even more accurate result can be obtained.

*****Update: The prediction for 14/05/2020 was 43480 and the total confirmed cases of that day was 43481. This is only a difference of 1 case!*****

9.2 SSA

The tasks to be completed or interesting to look into as self-study assignment before the next meeting are:

1. Finish the miscellaneous tasks from last week
2. E-mail the research team of *prof. Edwin van den Heuvel* for more information regarding the adjustments made to the Logistic Growth Model, using the **Von Bertalanffy function**
3. Try the **Bootstrapping method** for a more accurate estimation of the parameters in the Logistic Growth Model
4. Start with tranferring the content on Github README to an official thesis' report.
5. Finish up the Kalman Filter applied on the (extended) SIR/SEIR model
6. Use more accurate paramteres (from the Netherlands) or other research paper, to get an estimation of the number of **Exposed** in the Netherlands.

10 Results

10.1 Report

Started with the **Report** by transferring the information that is on the Github README to a bachelor thesis template. The difference between the Report and the current README on Github were the changes made to the different sections in the Github README, by extending these sections with extra information and references.

10.2 Parameter estimation of the SIR/SEIRD model

The estimated parameters, derived from articles, are different per article or conducted research. The models and parameters are adjusted for a sepcific region of the world or a specific country. Using these parameters to model the spread of COVID-19 in the Netherlands, is inaccurate and results in large deviations.

Thus, an attempt was made to use the available data of COVID-19 in the Netherlands to estimate the models' (*SIR* / *SEIRD*) parameters. This way, the estimated parameters and output would have a closer resemblance to the spread of COVID-19 in the Netherlands.

However, these models do not show the true number of infected or exposed, but is using the available data to try and model the spread of COVID-19 in the Netherlands.

The first attempt was to use the **Cumulative confirmed cases**, to fit the SIR model where the β (the number of people an infected person infects on average) parameter can be estimated. As the **Infected** compartment contains this parameter, an attempt was made to use the **Cumulative confirmed cases (number of Infected)** to fit the model and estimate the parameters using **least squares estimation**, while solving the differential equations simultaneously. For the parameter γ, is was assumed that this remains constant and the assumed value was derived from literature.

The result after fitting the model and plotting the *best fit (Infected compartment)* versus the *actual data (infected)* can be seen in Figure 5 :

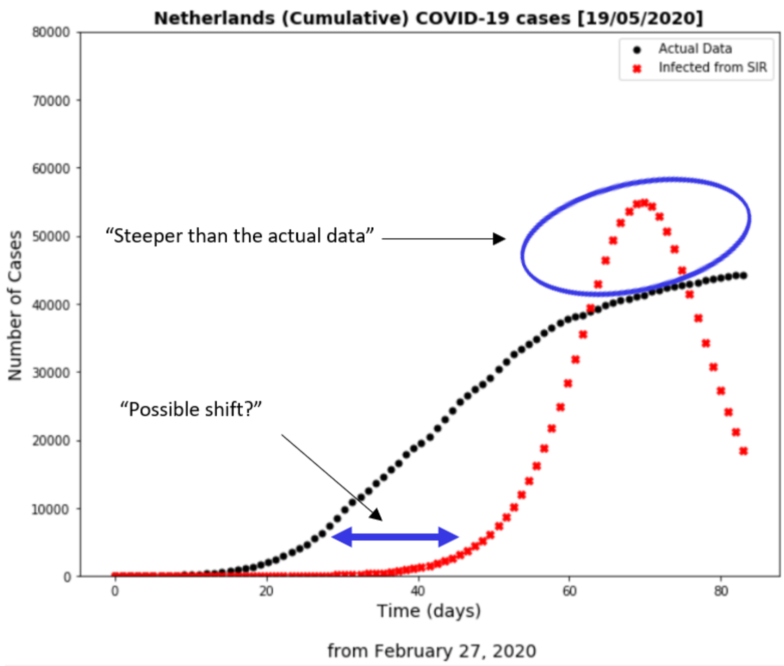

Figure 5: ***Best fit of the Infected compartment (SIR model)*** *versus* ***the Actual Data***

It can be seen that the *Best fit* is not the best at all. It is clear that underfitting is taking place. It seems that the data is shifted to the right (red curve) and at interval [60, 80] there is a huge peak. Furthermore, the estimated parameter $\beta = \mathbf{2.24} \pm \mathbf{0.12}$ (95% CI), which is most likely too large, considering that strict measurements took place at March 16, 2020. Thus, the underfitting might be caused by the assumption that β is a constant.

The **Cumulative number of Cases (Infected)** is highly dependent on the amount of tests performed. Furthermore, not everyone is tested thus this data might not be the most "trustworthy" data for parameter estimation. On the other hand, data such as **Cumulative number of deaths (Deceased)** is much more accurate or trustworthy (assuming that all deaths due to COVID-19 are registered and there are no "unknown deaths). Therefore, a second attempt was made to fit a model and estimate the parameters using **least squares estimation**, but for the SEIRD model instead.

Recall the functions of the SEIR(D) compartment:

$$\frac{dS}{dt} = -\frac{\beta(t)IS}{N}$$

$$\frac{dE}{dt} = \frac{\beta(t)IS}{N} - \delta E$$

$$\frac{dI}{dt} = \delta E - \gamma \cdot (1 - P[I \longrightarrow D]) \cdot I - \alpha \cdot P[I \longrightarrow D] \cdot I$$

$$\frac{dR}{dt} = \gamma \cdot (1 - P[I \longrightarrow D]) \cdot I$$

$$\frac{dD}{dt} = \alpha \cdot P[I \longrightarrow D] \cdot I$$

$$\delta = \text{rate that an exposed person becomes infectious } \frac{1}{\text{incubation time}}$$

$$\alpha = \text{fatality rate, } \frac{1}{\text{days after critical patient passes away}}$$

$$P[I \longrightarrow D] = \text{Probability that an infected person passes away}$$

The reproductive number R_0 is chosen to be a logistic function:

$$R_0(t) = \frac{R_{0_{start}} - R_{0_{end}}}{1 + e^(- k \cdot (-t + x_0))} + R_{0_{end}}$$

,whereas the parameter $\beta(t)$ is calculated with the function:

$$\beta(t) = R_0(t) \cdot \gamma$$

Thus, the parameters that need to be estimated are $R_{0_{start}}$, $R_{0_{end}}$, x_0 and $P[I \longrightarrow D]$.

The parameter x_0 is the inflection point, the day where the value of R_0 is half the maximum (steepest decline in this case).

After fitting the model using the data of the **Cumulative number of deaths (Deceased)** , it was possible to estimate the parameters and fit a curve. The values of α, γ and δ are assumed to be constant and are derived from literature.

The result after fitting the model and plotting the *best fit (death compartment)* versus the *actual data (number of deaths)* can be seen in Figure 6 :

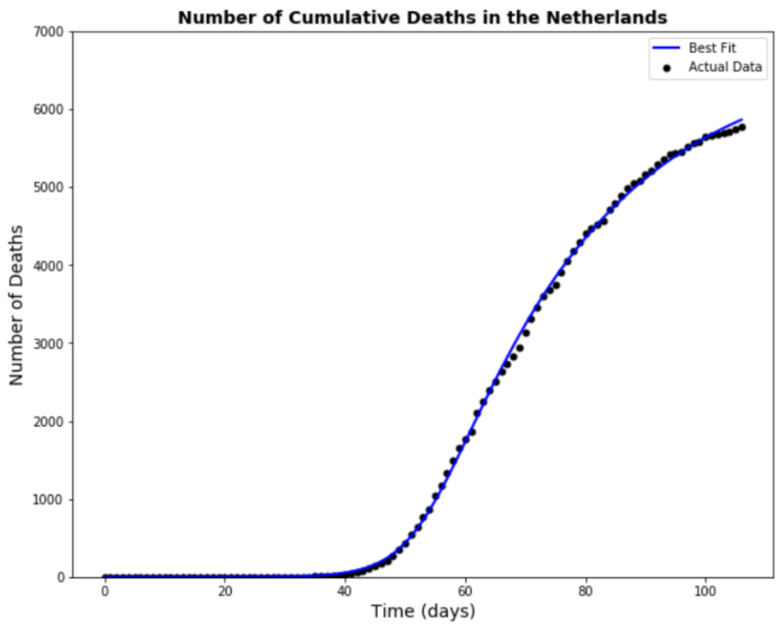

Figure 6: ***Best fit of the Death compartment (SEIRD model)*** *versus* ***the Actual Data***

And the estimated parameters are given in Table 1:

Parameter	**Value**	**Standard Error**
$R_{0_{start}}$	5.00	0.86
$R_{0_{end}}$	0.55	0.048
x_0	50.1	1.73
k	0.28	0.05
$P[I \longrightarrow D]$	0.08	0.09

Table 1: Estimated parameters and its Standard Error

An important remark is that there is always a trade-off between model's complexity and model's performance. Having more compartments added to the standard SIR model, will increase complexity. However, having more compartments means that there are also many more parameters that has to be estimated thus more data is needed. When there is a lack of data (especially in the case of COVID-19), it is difficult to obtain accurate parameters for the extra compartments. Therefore, having a more complex model does result in a closer resemblance to biological systems but the uncertainty of the model also increases.

10.2.1 Kalman Filter Model & Logistic Growth Model

The 1 week predictions of the Kalman Filter Model, applied to the self-defined exponential growth system, is finalized. The results are given in Table 2:

Date	Confirmed Cases	KF Prediction	LGM Prediction
12-05-2020	42984	42980.0	42124.0
13-05-2020	43211	43223.0	42222.0
14-05-2020	43481	43480.0	42398.0
15-05-2020	43681	43787.0	42572.0
16-05-2020	43870	43904.0	42738.0
17-05-2020	43995	44078.0	42896.0
18-05-2020	44141	44107.0	43042.0
19-05-2020	44249	44301.0	43180.0
20-05-2020	44447	44351.0	43308.0
21-05-2020	44700	44671.0	43434.0

Table 2: *The actual confirmed cases (Confirmed Cases), the Kalman Filter (KF) predictions and the Logistic Growth Model (LGM) Predictions*

It can be seen that the Kalman Filter Model is still performing quite well, but there are deviations, sometimes overestimation and other times underestimation is taking place. Although it might not be too interesting too analyse the Kalman Filter Model's predictions, it is important to note that the Kalman Fitler is a strong mathematical model that can be used for estimations of the past, present or future states. The Kalman Filter is widely used in navigation system, but it is clear that this can also be applied to biological systems and other biomedical models.