Danmarks Tekninske Universitet

Advanced Time Series Analysis
Course number: 02427

Computer Exercise 4

Authors:
Mads Esben Hansen, s174434
Karl Emil Takeuchi-Storm, s130377

December 23, 2020

1 Introduction

In this report we will try to look into how the spread of covid-19 can be modelled. We will have a strong focus on whether weather data can be used to improve local forecasts. It is the answer to "Computer Exercise 4" in the course in advanced times series analysis at DTU. Whenever there is a referral to *the lecture notes*, the notes in mention is *Modelling Non-Linear and Non-Stationary Time Series*, by Henrik Madsen and Jan Holst, December 2006.

2 Methodology

2.1 SIR

The most used model for modelling of infectious decease is to use a Susceptible-Infected-Removed (SIR) model. The essential idea is to divide a group into 3 parts, people who are susceptible to the decease, people who are infected, and people who have been removed due to immunity, death etc. It classical formulation is as follows:

$$\begin{aligned}\frac{dS}{dt} &= -\beta \cdot S \cdot I \\ \frac{dI}{dt} &= \beta \cdot S \cdot I - I \cdot \gamma \\ \frac{dR}{dt} &= I \cdot \gamma\end{aligned}$$

2.2 SIHR

We have measurements from positive tests and hospitalizations, in turn we will expand the model slightly to a Susceptible-Infected-Hospitalized-Removed (SIHR) model.

$$\begin{aligned}\frac{dS}{dt} &= -\beta \cdot S \cdot I \\ \frac{dI}{dt} &= \beta \cdot S \cdot I - I \cdot \gamma_I \\ \frac{dH}{dt} &= H_r \cdot I \cdot \gamma_I - H \cdot \gamma_H \\ \frac{dR}{dt} &= (1 - H_r) \cdot I \cdot \gamma_I + H \cdot \gamma_H\end{aligned}$$

This model is still quite simplistic, in fact it is too simplistic for real-world use, which is why much research has been done regarding how to improve it. In this regard many avenues can be explored, we will however limit ourselves to a fairly simple model and instead add diffusion terms to the equation, generating a system of stochastic differential equation. In turn we obtain the follow system:

$$\begin{aligned}dS &= (-\beta \cdot S \cdot I)dt + p_1 dW_t \\ dI &= (\beta \cdot S \cdot I - I \cdot \gamma_I)dt + p_2 dW_t \\ dH &= (H_r \cdot I \cdot \gamma_I - H \cdot \gamma_H)dt + p_3 dW_t \\ dR &= ((1 - H_r) \cdot I \cdot \gamma_I + H \cdot \gamma_H)dt + p_4 dW_t\end{aligned}$$

Where dW_t describes the Wiener process.
It can be quite tedious to solve SDE, therefore we will use the software package CTSMR[1] developed at DTU for solving SDEs.

[1] For documentation please refer to: http://ctsm.info

2.3 Model comparison

For model comparison, we will use 2 metrics, root mean squared error (RMSE), and Bayesian information criterion (BIC). Which are given by:

$$RMSE = \sqrt{\frac{r^T r}{n}}$$
$$BIC = k \cdot log(n) - 2 \cdot log(\hat{L})$$

Where r is the residual vector, n is the number of observations, k is the number of parameters in the model, and $\hat{L}$ is the likelihood of the model. We will further divide the the data into two parts, one we use for parameter estimating, and one for testing how well the model generalizes.

When evaluating the performance of the model we are interested in how well the model interpolates, i.e. one step predictions within the training data. In this regard we will refer to both the BIC score and RMSE. However, since our primary focus is forecasting, we want to evaluate how well the model generalizes, i.e. extrapolates. In this regard, the BIC is meaningless, since the likelihood is computed on the model. We will therefore use the RMSE of the testing data to compare the models.

3 Data

3.1 Background

The primary driver for this project has been the notion that the temperature effects the spread of the decease. This notion has been stated in some variant by many domain experts. The idea is not that the actual temperature increases the spread, but rather that peoples behaviour changes. When it is colder more people meet at indoor activities, rooms are less adequately ventilated as windows are opened less and many types of activities shift indoors (sports, social gatherings and work). Few assumptions has been made, based on knowledge from domain experts. The mean time from infection to hospitalization is 13 days. It takes 4.7 days on average to reach a infected state where symptoms appear. Many factors play into the model of COVID-19, some are very difficult to model.

3.1.1 Data from Statens Serum Institut

The modelling is carried out using the publicly available data from SSI. The data has both raw and processed elements. The main interest is of course in the raw data, which contains data for total number of tested, the number of new positives, hospitalized and deceased for each day since the beginning of the pandemic. On the processed data we have used the R_t (Reproduction number), as an initial data to test our correlations on.

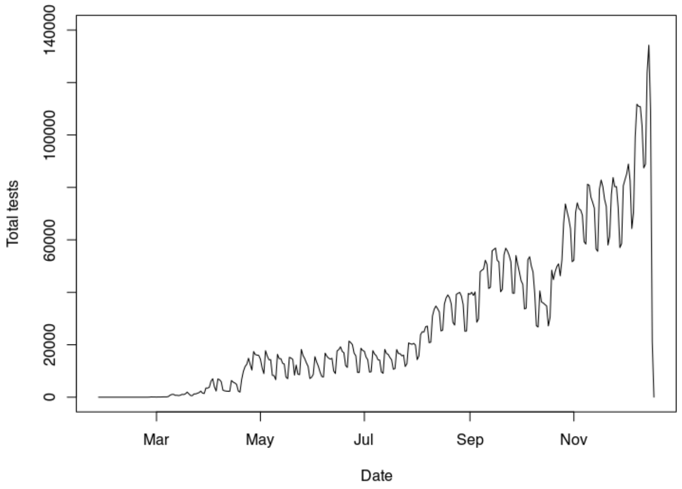

Figure 1: Number of daily tests

The testing strategy has changed over time, and so has the testing capacity. This of course has a great effect on the number of positives found in a test setting. The best indicator for the number of people who are infected, is the number going to hospital, with COVID-19. The problem with using hospitalized is the time it takes to get hospitalized after infection. In the remaining of this report, it is assumed to take exactly 13 days. In figure 2 both positive tests and hospitalized are plotted over time. It is clearly visible that the severity of the disease has changed, or (which seems more plausible) the large amount of testing currently, gives a better picture of the number of infected, than the low number of tests earlier in the year.

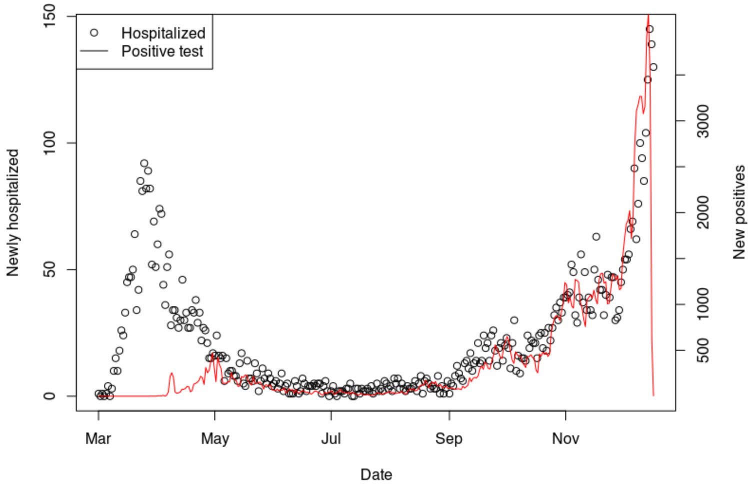

Figure 2: Hospitalized

In order to look for these differences, the percentage of positive tested, that goes to hospital is plotted as a function of total tests. The problem being that early on, test was only conducted on people with clear symptoms, and further a person is more likely to get tested with even mild symptoms than no symptoms. Thus testing few people will give a higher positive percentage, which in turn gives a higher number of people going to hospital.

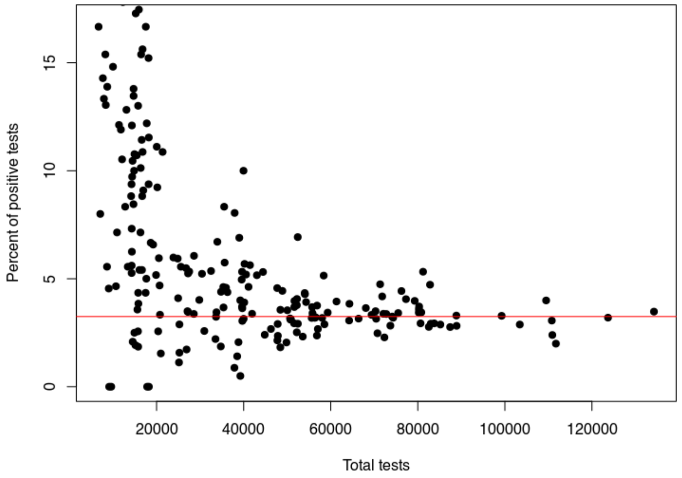

Figure 3: Hospitalized

This short analysis, could give raise to the thought that approximately 3.25 percent of infected people goes to hospital.

3.1.2 Data from Danish Meteorological Institute

Data for the weather input is obtained from DMI, and contains the mean temperature for each full day.

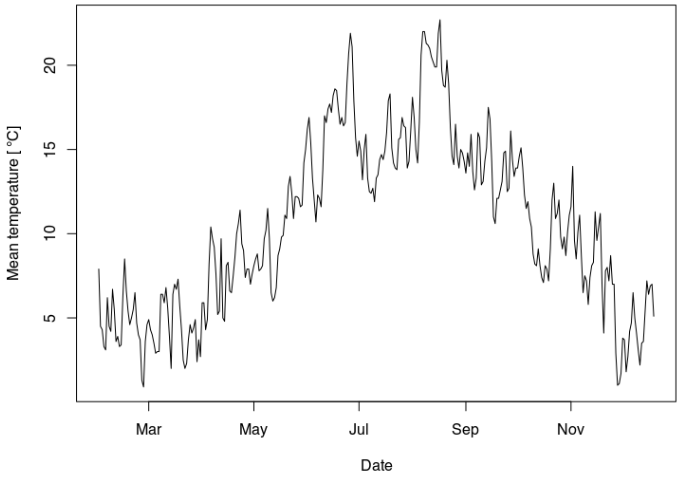

Figure 4: Hospitalized

3.1.3 Further data acquired

In addition to the above data, the number of registered outbreaks on mink farms for a given week has been sources, as well as the dates for changes to restrictions related to corona.

4 Results

4.1 Base Model

As mentioned, we started off using the very simple SIHR model given by:

$$\begin{aligned}
dS &= (-\beta \cdot S \cdot I)dt + p_1 dW_t \\
dI &= (\beta \cdot S \cdot I - I \cdot \gamma_I)dt + p_2 dW_t \\
dH &= (H_r \cdot I \cdot \gamma_I - H \cdot \gamma_H)dt + p_3 dW_t \\
dR &= ((1 - H_r) \cdot I \cdot \gamma_I + H \cdot \gamma_H)dt + p_4 dW_t
\end{aligned}$$

Where dW_t describes the Wiener process. Fitting the model yielded the following results:

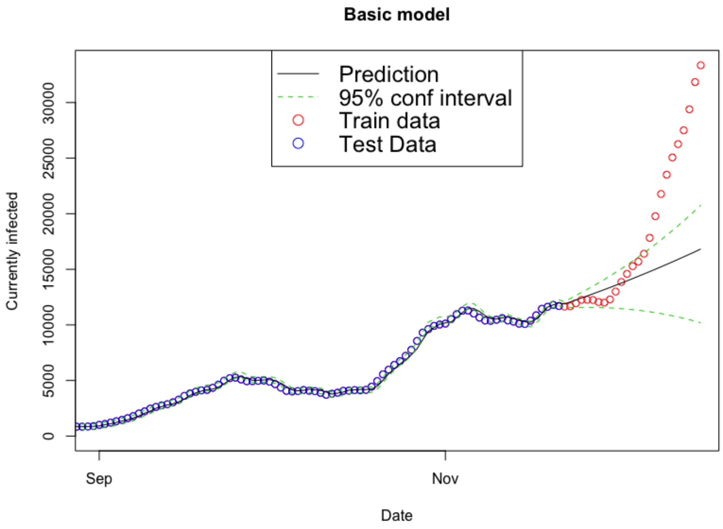

Figure 5: Base model

Using this approach implicitly assumes that the infections coefficient β is constant over time. This means that the predictions of infections will follow a standard non-stochastic SIHR model, i.e. logistic growth. From the figure, it is evident that the growth rate cannot be assumed constant, and in turn this method is too simple for our purpose.

4.2 β as state

The easiest way to handle a time varying parameter is simply to include β as a new state, and model it as a random walk; i.e. using the model given by:

$$\begin{aligned}
dS &= (-B \cdot S \cdot I)dt + p_1 dW_t \\
dI &= (B \cdot S \cdot I - I \cdot \gamma_I)dt + p_2 dW_t \\
dH &= (H_r \cdot I \cdot \gamma_I - H \cdot \gamma_H)dt + p_3 dW_t \\
dR &= ((1 - H_r) \cdot I \cdot \gamma_I + H \cdot \gamma_H)dt + p_4 dW_t \\
dB &= p_5 dW_t
\end{aligned}$$

This allows β (which we now represent as a state denoted by B) to change over time to follow the behaviour of the data. We notice that when predicting using this approach, the B state will maintain the same value, and hence does not necessarily generalize very well. A good way of managing this is manually tuning p_5, i.e. the variance of B. Decreasing p_5 forces the B value to change slower and vice versa. This fact can be used to tune the model to allow for change over time, while still generalizing fairly well. To investigate these properties, we tried fitting the model using different values for p_5.

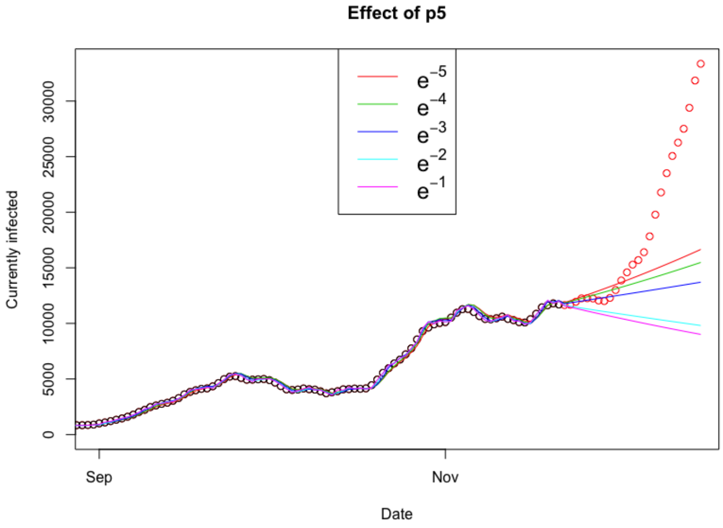

Figure 6: Predictions using different values for p_5.

In this case, the general trend is increasing, but the very local trend is decreasing. We can see that for $p_5 < e^{-3}$, the model predicts that the number of infected people will increase. It is our opinion that $p_5 = e^{-3}$ is a good value for p_5, since it captures some local trends, without compromising the general trend too much. Fitting this model with $p_5 = e^{-3}$ yielded the following results:

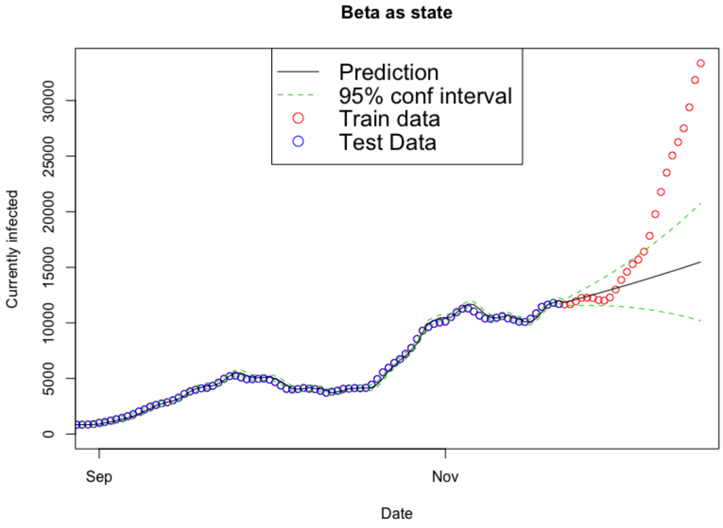

Figure 7: Model using β as new state.

We can see that while the first few steps in the forecast are relatively good, this method greatly underestimates the general trend.

4.3 Using weather

Another approach is to account for a time varying parameter is to model the parameter directly, i.e. model the parameter using some related data and some fit. We want to try to use temperature data to improve the predictions. We choose to use the difference between the temperature of the day and the average temperature over the past 10 days.

$$\tau_i = \frac{T_i}{\frac{1}{10}\sum_{k=i-10}^{i} T_i}$$

We wanted to investigate, if there was any dependency between $/tau$ and β. To solve this we looked at the estimated value of B in the previous model as a function of τ, as they described the time varying β parameter quite well (while not extrapolating). We plotted this using different off-sets, i.e. off-setting the τ value by eg 10 days to account for delay in development of the sickness. We found that using an off-set of 13 days produced the strongest dependency between the value of B and τ.

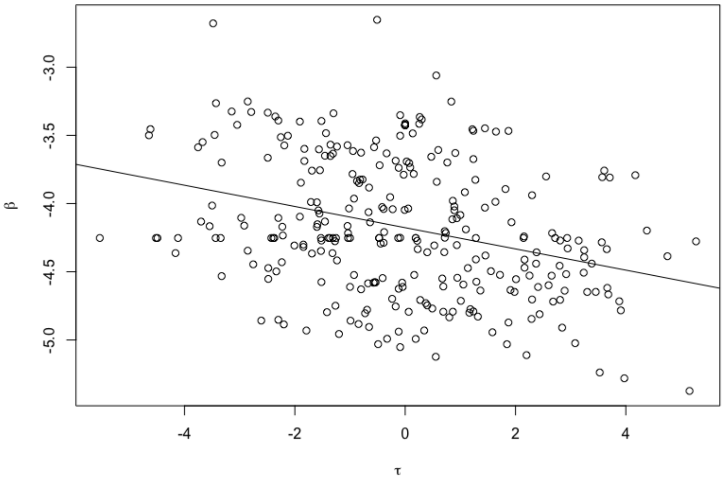

Figure 8: β dependence of τ with a 13 day off-set.

The plot suggests that the β value over time can be described using a first order linear function, i.e. $\beta = c_0 + c_\beta \cdot \tau$. Exploiting this yields the following model:

$$\begin{aligned}
dS &= (-(c_0 + c_\beta \cdot \tau) \cdot S \cdot I)dt + p_1 dW_t \\
dI &= ((c_0 + c_\beta \cdot \tau) \cdot S \cdot I - I \cdot \gamma_I)dt + p_2 dW_t \\
dH &= (H_r \cdot I \cdot \gamma_I - H \cdot \gamma_H)dt + p_3 dW_t \\
dR &= ((1 - H_r) \cdot I \cdot \gamma_I + H \cdot \gamma_H)dt + p_4 dW_t
\end{aligned}$$

Note that since there is a off-set of τ the true value is known for future values, reducing variance in the predictions. Utilizing this and fitting using CTSMR yielded the following results:

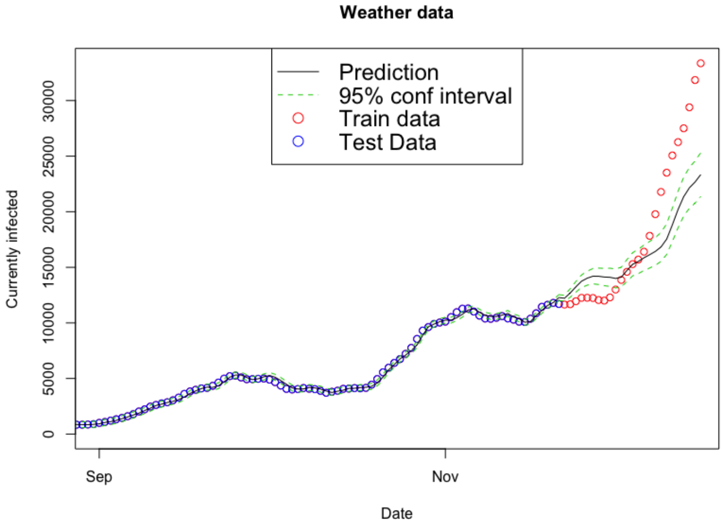

Figure 9: Results when including τ in the model.

4.4 β as state, and weather data.

Finally we wanted combine the last two models together and model the β value as a first order linear model with a slowly changing c_0 (remembering this can be obtained by lowering the variance of the state).

$$\begin{aligned} dS &= (-(B + c_B \cdot \tau) \cdot S \cdot I)dt + p_1 dW_t \\ dI &= ((B + c_B \cdot \tau) \cdot S \cdot I - I \cdot \gamma_I)dt + p_2 dW_t \\ dH &= (H_r \cdot I \cdot \gamma_I - H \cdot \gamma_H)dt + p_3 dW_t \\ dR &= ((1 - H_r) \cdot I \cdot \gamma_I + H \cdot \gamma_H)dt + p_4 dW_t \\ dB &= p_5 dW_t \end{aligned}$$

This method allows the mean of the β value to change over time, while still exploiting the dependence of τ. As before, we want to investigate the properties and a suitable of p_5, since we have change the sense in which it is used.

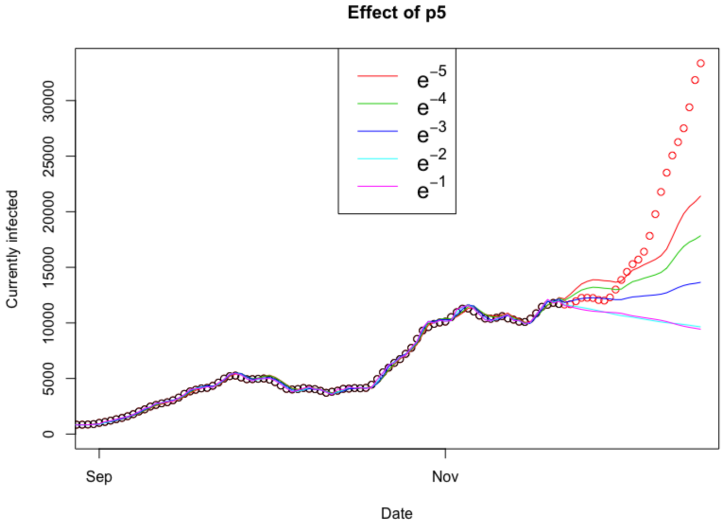

Figure 10: Predictions using different values for p_5.

We can see that this time around there seem to be no real difference between $p_5 = e^{-1}$ and $p_5 = e^{-2}$, since this is simply a too large value (forgetting factor is too great). Since we primarily care about how well our generalized, we choose to use $p_5 = e^{-4}$ this time. The results were the following:

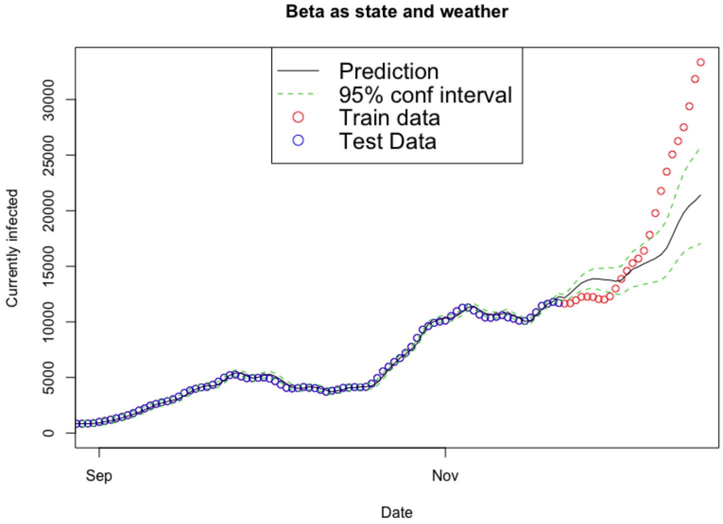

Figure 11: Using both new state and weather data.

By the looks of it, the predictions are not quite as good as the previous model. This is because the local trend and general trend are in opposite directions.

4.5 Performance of the 4 models.

As mentioned, we will use RMSE and BIC scores to compare the models.

Table 1: Performance of the 4 models.

Model	$RMSE_{in}$	$RMSE_{out}$	BIC
1	160	6792	620
2	110	10525	621
3	144	4248	644
4	145	5031	666

As can be seen in the table, it seems that model 1 and 2 performs best on data used to estimate the parameters (they have the lowest BIC score). However, model 3 and 4 have the lowest RMSE on the training data, indicating that they generalizes better than the other 2 models. In general, model 3, i.e. including weather data but not β as a state, generally performs the best of the four model.

5 Discussion

When fitting model 2 and 4 using CTSMR we ran into issues where the solution that optimized the likelihood resulted a B state with a extremely high variance. Effectively most of the variance in the data could be described with a B value that varies very much. However, such a model generalizes extremely poorly, which is why we had to manipulate the parameter estimating. We found the manually setting the value that controls the variance in a given state is an effective way of tuning the equivalent to a forgetting factor, when dealing with time varying parameters. We were in this way able to use the variance in B as a knob, between letting the value change very rapidly (high variance), and change very slowly (low variance).

In our model we choose to use the difference between the temperature a given day and the moving average until that day. We did this because we wanted to capture the sense of a *cold* day or a *hot* day. There are endless possibilities when choosing how to manipulate this data, and this could obviously have been done differently with unknown results. We also choose to focus on the temperature of a day; however, it can easily be speculated that other weather data, e.g. rain, could also be used to improve the predictions of covid-19. We simply choose temperature because we did not have an infinite amount of time to finish the report.

As discussed earlier, the total number of test carried out, has increased continuously through the duration of the pandemic. The meaning of this is problematic to model, as more test will cause more people to test positive. As the given model uses positive test as an input for I (infected), this cause some problems to the model. A prediction will then predict that more people are going to test positive, which cannot be interpreted as a direct indicator of the disease spreading further in society. This is due to the fact, that the model will react to increased testing as well.

When comparing the models we only looked at one possible division of training/testing data, and compared based on that. In general it would be a good idea test many possible division and compare based on all of these. In our case the situation was that the local trend was decreasing whereas the global was increasing. This meant that the models using β as a new state performed poorly on unknown data, but this could partly be as a cause of this, and might not be the case in general.

6 Conclusion

We found that a good way of modelling the spread of covid-19 is to use a quite simplistic compartment model, and extend it by turning them into a system of stochastic SDEs. We found that including some key parameters as new states in the model, effectively generating time varying parameters, reduced the errors in the one-step predictions within the training data, however, we found no evidence to suggest that it improved the predictive properties of the model. Conversely, we found some evidence to suggest that modelling key parameters as a function of temperature can improve the predictive value of the model, however, it also increases the error in the one-step predictions within the training data slightly.