Assessing COVID-19 Cases in Belgium‘s Provinces

Jonas De Boeck, Thibault Magnini
Department of Technology, UC Leuven-Limburg University of Applied Sciences
Geldenaaksebaan 335, 3001 Heverlee
jonas.deboeck@student.ucll.be, thibault.magnini@student.ucll.be

Cássio M. Oishi
Departamento de Matemática e Computação, Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista ‘Júlio de Mesquita Filho’,
19060-900 Presidente Prudente, SP, Brazil
cassio.oishi@unesp.br

Abstract

The coronavirus disease (COVID-19) has been rapidly spreading, leading to a severe health crisis all around the world. This is also the case in Belgium. With over 66.000 confirmed cases across the entire Belgian population, leading to negative economic and mental consequences. Therefore clustering of the provinces is helpful to enable an optimal vaccination strategy based on infection severeness per province. The provinces are clustered using the K-means algorithm according to their respective COVID-19 case numbers and testing numbers. Data obtained from the Belgian Institute for health on the 30th of March 2021, covering all 10 provinces adding the brussels region creating 11 separate regions. K-means lead to 4 optimal clusters. Later we separated Brussels and put it into its own cluster because of its high population density.

In addition a SIR-model study for each cluster was done. To provide insights on the current spread of the virus. Combining these results with the change in mobility data, the conclusion made was that outdoor activities have a much lower impact on the spread of the virus. Where mobilities to workplaces and schools have a significantly larger impact.

In addition to the SIR-model found in the paper a parameter estimation was done in 2 different ways. One way resulting in constants and the other way resulting in transient reevaluated parameters. Where the latter gave the better results for predicting.

Keywords

K-means, Covid-19, Belgium, Provinces, Clustering, SIR, Mobility, Neural Network

1. Introduction

The SARS-CoV-2 virus, or more known as COVID-19, that has been declared as a pandemic by WHO on the 11th of March 2020[1]. Since the global outbreak most countries around the world have been trying

[1] https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-mission-briefing-on-covid-19---12-march-2020

to control the spread of the virus by invoking different measures. A lot of research has been conducted regarding the COVID-19 pandemic on many different levels. Clustering is an often used technique to map cases to provide insights that assist the optimal handling of the pandemic.

In terms of Belgium, we find that it could be helpful clustering the 10 different provinces in addition to Brussels. Each having a significant part of the Belgian population. Conducting this research looks to provide insights on the general regions with higher risk of infection.

In this paper, data from March 2021 will be used to determine the different clusters and divide Belgium in general regions of risk amount. To obtain the correct number of regions, the algorithm will be run using different cluster counts. A visual representation of the regions on a Belgian map will be included in the paper.

The structure of this paper will continue as follows. Section 2 will discuss the different methodologies used to obtain results, Section 3 will contain greater detail on the K-Means algorithm, Section 4 describes the method used to calculate Cluster Validity, Section 5 Depicts the SIR-Model, Section 6 discusses the Nonlinear least squares methodology used. Section 7 Provides the structure of the used Neural Network to estimate the SIR-Model parameters. Section 8 Explains the used data and its estimations, Section 9 Summarizes the impact of different types of mobilities on the spread of the virus. Section 10, Provides the results of the different mathematical models used.

2. Methodologies

There are many different variants of clustering methods. In general these methods can be categorized in 2 main categories being hierarchical and non-hierarchical. The key difference being that hierarchical clustering methods use similarities to group different observations. The number of clusters is not determined in advance.

In the case of non-hierarchical clustering methods this number of clusters is predetermined. When pre-determining these, one must consider several pitfalls that will be discussed in the further sections.

In terms of Computer methodologies, the programming language Python was used to implement the clustering algorithm used to cluster the different observations. It was also used to manipulate the government data into a usable dataset with parameters fit for clustering.

The Neural Network used was made using Python and the scipy library.

3. K-Means Clustering

Partitioning is one of the most widely used clustering categories. K-means being the most popular choice within this category. K-means can be used in many different variations. Such as the K-medians, Fuzzy C-means, K-means++ and many more. In this paper the simplest variation was used, also referred to as Lloyd's algorithm.

The algorithm used proceeds as followed:

1. Chooses initial representatives from the given observations, equal to k. the predefined number of clusters.
2. Repeat the following process until convergence:

   a) Partition the given observations into k clusters, for each observation we assign it to the cluster with the nearest representative. This is done by calculating the squared Euclidean distance. We write

   $$min_{j=1,\ldots,k} \quad \left|\left| x_i - z_j \right|\right|^{2}$$

   where $z_j$is a representative, $x_i$is an observation.

   b) Update the representatives, we update the cluster's representative by setting it to the mean of its cluster, We write

   $$z_j = \left(1 / \left|G_j\right|\right) \sum_{i \epsilon G_j} \quad x_i$$

   Where $\left|G_j\right|$is the number of elements in the set G for a specific cluster.

The algorithm is run several times to find the optimal cluster representatives or also known as centroids. This way we minimize our J_{clust}which indicates improvement of or clustering results. This is further described in the next section.

4. Measuring Cluster Validity

The validity of the model gets measured using J_{clust} with J_{clust} being the sum of the mean squared distances for each cluster in the model. The following formula was used to calculate the cluster validity.

$$J^{clust} = J_1 + \ldots + J_k$$

Where,

$$J_j = (1/N) \sum_{i \epsilon G_j} \left|\left|x_i - z_j\right|\right|^2$$

With N being the number of features in the cluster, G being the set of elements in the cluster and z being the cluster's representative.

J_{clust} shows us the average distance of all the features in the model to their respective representatives. In other words, it shows us how accurately the model clustered the different elements in the k given clusters. By this our goal is to find the optimal number of clusters by minimizing the value of J_{clust}. To find the optimal number of clusters we made use of the elbow method, further described in the conclusion.

5. SIR-Model

The Susceptible (S), Infected (I), Removed (R) model is a well-known mathematical model used to predict the spread of diseases. Throughout the Covid-19 pandemic SIR has been frequently used to predict the way the pandemic is going to spread in certain countries. Often the SIR-model gets used together with different algorithms to get better results and a wider view of the current situation of the pandemic. There are various extensions to the SIR-model in which the developers decide to include different parameters such as deaths, exposures etc.

The SIR-model contains 3 differential equations who each depict a certain parameter included in the model. The differential equations used in this paper are the following:

$$\frac{dS}{dt} = -\beta \frac{SI}{N}$$

$$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Where S, I and R are the amount of susceptible, infected and removed people respectively at a current amount of time t and where $\beta$and γrespectively are the transmission rate and the recovery rate. The reproduction number R_O tells you the average amount of different people a single infected person infects. $R_O$can be found with the following formula $R_O = \frac{\beta}{\gamma}$.

6. Nonlinear Least Squares fitting model

The nonlinear least squares model is a form of least squares analysis to find the best fit for unknown parameters in a nonlinear equation. The model tries to fit the model by a linear one and then refines the parameters over several iterations, to then obtain the best possible fit of the parameters.

$$minimize \sum_{i=1}^{m} \quad f_i(x)^2 \;=\; ||\, f(x)\, ||^2$$

where $f_i(x), \ldots, f_m(x)$ are differentiable functions. In the case of this paper the equations found in the SIR-model.

The use of this model ensures that we can make reasonable predictions using the SIR-model. For more accurate results the model could be improved or the use of different models is required.

7. Neural network used for parameter estimation[2]

While using the SIR-model from the previous chapter we chose to use a constant $\beta$and γ. For the next step in this investigation we decided to use a neural network together with the SIR-model to estimate β and γas transient parameters for the model. This way we can compare the results between the use of constant and transient parameters in the SIR-model and compare which method is the most accurate to make epidemiological predictions.

The used model consists of the input layer, a hidden layer containing 10 neurons, an output layer with a single neuron. The model requires the following parameters to be fed to the input layer: Cumulative infected, active infections, cumulative recovered and cumulative deceases. Since in this case we used an SIR-model and not an SIRD-model the cumulative deaths fed to the model are just used to calculate the amount of removed people from the model. With the amount of removed people being the sum of the recovered people and the deceased.

The activation functions used in the model were the Sigmoid activation functions for the neurons in the hidden layer and the ReLU activation function for the single neuron found in the output layer.

Sigmoid,

$$f(x) \;=\; \frac{1}{1+e^{\;\;-x}}$$

ReLU,

$$f(x) \;=\; \{x \; if \; x > 0, 0.01x \; otherwise\}$$

As the loss function we minimize the following aggregated error measure, based on the model parameters, I and R.

[2] https://www.mdpi.com/1424-8220/21/2/540

$$\lambda(\beta_{net}(t), \gamma) \; = \; I_I \; + I_r$$

where,

$$I_I = \frac{1}{M}\sum_{n=0}^{M} \quad \left|\left| log(I_{data}(t_n) \; - \; log(\underline{I}(t_n)) \right|\right|_2^2 \qquad ,$$

$$I_R = \frac{1}{M}\sum_{n=0}^{M} \quad \left|\left| log(R_{data}(t_n) \; - \; log(\underline{R}(t_n)) \right|\right|_2^2$$

To assess the forecasts provided by the algorithm, the Mean Absolute Percentage Error (MAPE) evaluation metric was used. This can be noted as,

$$MAPE(\gamma_i, \widehat{\gamma}_i) \; = \; \frac{1}{n}\sum_{i=1}^{n} \quad \left|\frac{\gamma_i - \widehat{\gamma}_i}{\gamma_i}\right| \times 100,$$

where γ_i and $\widehat{\gamma_i}$ are the real and predicted values of any target variable as forecasted by the model. When assessing the results a threshold of 10 % is established to ensure a "satisfactory level" of accuracy from the predictive performance.

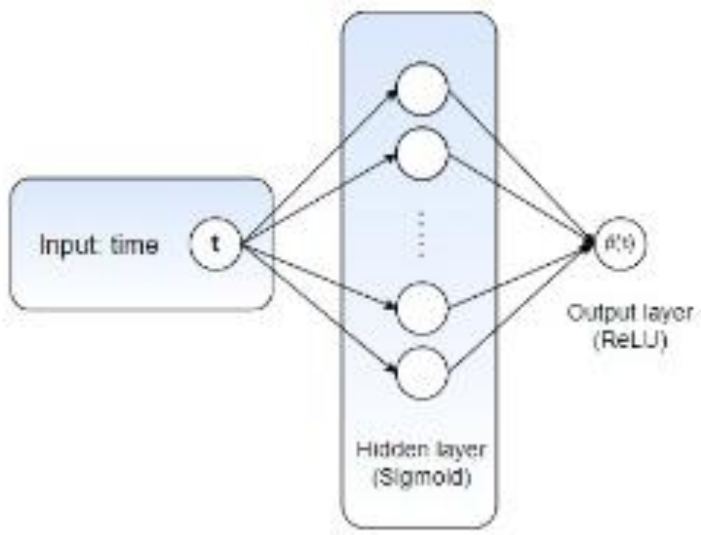


Figure 1 ANN layout for $\beta(t)$

8. Data and Parameters

8.1. Data K-means

The data used in this paper was obtained from the Belgian Institute for health. The 31st of March 2021. The clustering of the provinces was done based on 3 variables, Infection rate per province, ICU-rate per province, Positive test percentage per province. Infection rate and ICU rate were calculated using following formula:

$$\beta = \frac{Number\ (\#)\ of\ cases}{Population\ at\ risk} \cdot \alpha$$

Where β is the rate and $\alpha = 100.$
For the infection rate the population at risk being the entire province population and for the ICU rate the population at risk being the number of confirmed cases in the respective province. The Positive test percentage was calculated by the formula:

$$\gamma = \frac{Number\ of\ positive\ tests}{Total\ number\ of\ tests} \cdot \alpha$$

Where γ is the positive test percentage and α = 100.

These calculations were done for each of the Belgian provinces and Brussels. We count 10 provinces and Brussels bringing our number of regions to be clustered to N = 11.

Table 1 shows descriptive statistics results on the 11 pre-defined Belgian regions.

Rates	Minimum	Mean	Maximum	Standard Deviation
Infection rate	0.6489	1.0529	1.6834	0.2734
ICU rate	5.6917	13.3264	18.6458	3.9991
Positive Test Percentage	5.8291	7.7212	11.1391	1.6015

Figure 2 depicts multiple box-plots for three variables where N = 11. (Non-standard) in percentages. We identify that the ICU rates are much larger values compared to the other parameters. This leads to the ICU-rate having a larger influence on the final clustering result. To scale severity and determine which regions would benefit from vaccination priority we want to prioritize the regions with higher ICU-rates. When looking at the ICU-rate we see one outlier towards the lower side. The outlier is the province Flemish-Brabant with an ICU-rate percentage of 5.6917%.

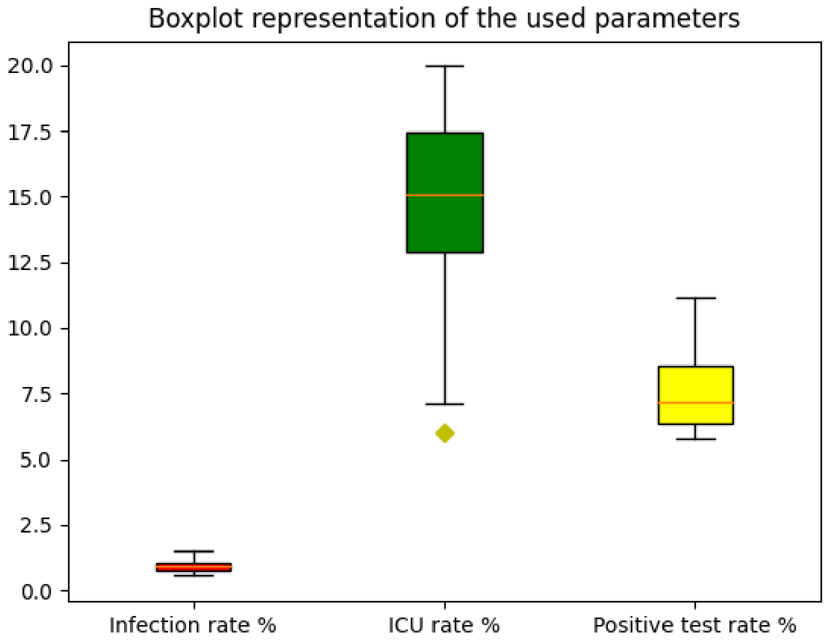


Figure 2 boxplots for the used parameters in the model

8.2. Parameters SIR-model

The SIR-model requires 3 different parameters to be known. Namely the active infections at a certain point in time, the cumulative amount of removed people at the same point in time and the amount of susceptible people. Where the amount of removed people is the sum of people who have recovered or are deceased. The Belgian government only provides specific data about the new cases on a specific day and the amount of deaths on a day. Because of this we had to estimate the current active infections and the cumulative amount of removed people. We estimated those using the following methods.

$$R(t) = C(t-T) - D(t-T)$$

Where $R(t)$ depicts the amount of recovered people on a day t. $C(t - T)$ is the amount of cumulative infections for Tdays before t. Tbeing a constant for the infection duration in our case we chose $T = 14$. $D(t - T)$ is the amount of people who died from the virus T days before t. Using this method of estimation we state that any infected person will only recover from their infection after 14 days of getting infected. Now we can calculate the estimated removed people using the following formula.

$$R_{SIR}(t) = R(t) + D(t)$$

We can calculate the active infections using the following method.

$$A(t) = C(t) - R(t) - D(t)$$

With $A(t)$ being the active amount of infection on day t.

Because of the Belgian government not providing information about deceases per province we had to use an alternative way of estimating recoveries for each province. This was the following method.

$$R(t) = C(t - T)$$

Where T is the infection duration, a constant which we chose to be 14 days. C(t) is the amount of cumulative confirmed cases on a specific day t. This method considers that people recover or die after exactly 14 days of infection.

9. Mobility Difference Analysis

An additional analysis was done using new data. This data contains the difference in mobility from the population during the COVID-19 Pandemic. Daily data gets compared to the daily date from February 6th - 13th 2020 (Baseline).

The mobility difference numbers get divided in 6 subcategories, being, Recreation and retail, Groceries and pharmacies, Parks, Transit stations, Workplaces, Residential.

The data was used in percentages, these percentages being the difference from the baseline numbers measured from February 6th - 13th 2020.

This data was also analyzed per province. So a relation between mobility differences and our clustering results could be looked at.

For first hand analysis the mean of all differences for march was calculated for each province.

10. Results and Discussion

10.1. K-Means results

Figure 3 depicts the elbow graph used to determine the optimal number of clusters. In this paper the algorithm was run several times using a different predefined number of clusters (k). In this paper k = 4 is preferred due to small differences within the cluster sum of mean squared distances for k > 4. The sum of the mean squared distances or J_{clust} in this case is equal to 4.4934.

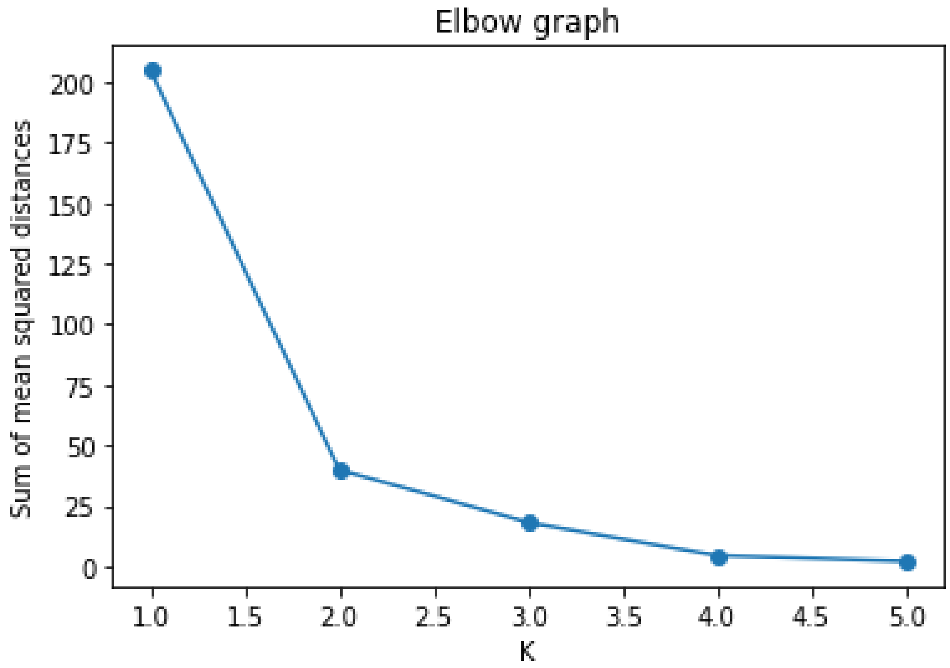


Figure 3 Within cluster sum of mean squared distance for every pre-defined number of clusters, K

Table 2 shows statistical information based on the ICU rate for each cluster, we identify 4 different clusters.

Cluster	N	min	mean	max	Standard deviation
1	5	13.4287	15.0823	16.3656	1.3162
2	2	16.8529	17.7493	18.6458	0.8964
3	2	5.6917	6.0403	6.3889	0.3486
4	2	11.6689	11.7998	11.9307	0.1309

Table 3 shows the results after running the algorithm until convergence of the representatives.

Province	Cluster	Severity	Infection rate (%)	ICU rate (%)	Positive test rate (%)
Antwerp	1	Medium	0.8416	16.2631	5.8787
Liège	1	Medium	0.6489	13.4287	7.2704
Limburg	1	Medium	0.8498	13.5461	6.4798
East-Flanders	1	Medium	1.1302	15.8081	7.006
West-Flanders	1	Medium	0.9448	16.3656	6.416
Brussels	2	High	1.2256	18.6458	8.539
Hainaut	2	High	1.1697	16.8529	9.7003
Walloon-Brabant	3	Low	1.604	6.3889	7.9821
Flemish-Brabant	3	Low	0.7767	5.6917	5.8291
Luxembourg	4	Medium-High	1.2481	11.9307	8.6927
Namur	4	Medium-High	1.6834	11.6689	11.1391

We can classify the 4 different clusters in degrees of severity. Cluster 2 being the provinces with the highest ICU rate and fairly high infection rate. This implies higher severity.

Cluster 4 being a cluster containing 2 provinces with high infection rates, but ICU rates on the rather low side around 12%. The number of positive tests in regard to the number of total tests taken is fairly high.

Cluster 3 is to be considered a ‘safer’ region, both infection and ICU rates are on the low side, as well as the positive test percentage. This could be viewed as a region with lower severity.

Cluster 1 contains provinces with a rather low to average infection rate and positive test percentage but show some rather high ICU-rates.

Figure 4 shows a visualization from Table 2 on a map from Belgium.
We can see here that the southern provinces, seen in the color green and purple, are the provinces that show the worst rates. This being cluster 2 and 4. The more northern part is made up by cluster 1 and 3, meaning less severe numbers.

We can conclude here that according to the numbers from March 2021, southern regions could see priority for vaccination, according to the current COVID-19 numbers. These regions have increased ICU rates, vaccinating here would be necessary as soon as possible to prevent hospitals from reaching maximum capacity.

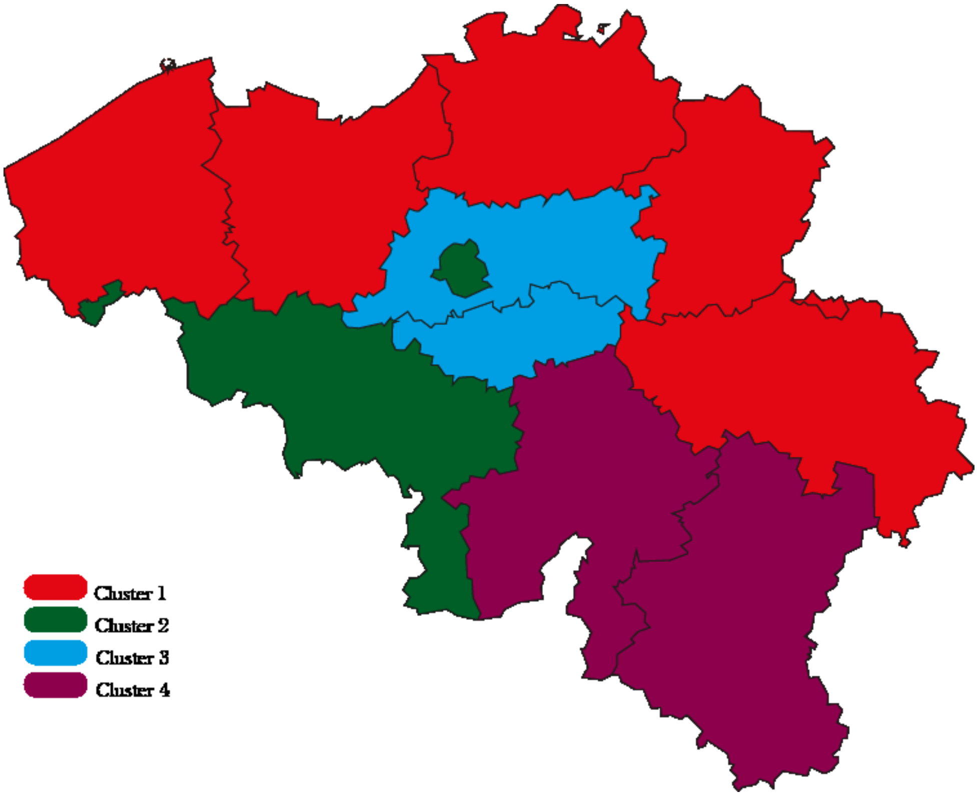


Figure 4 map from Belgium showing the clusters

Figure 5 shows the clusters in a 3D plane. With the infection rate on the X-axis, the ICU-rate on the Y-axis and the positive test rate on the Z-axis. The representatives are also displayed. Each element of each cluster is connected with its respective representative.

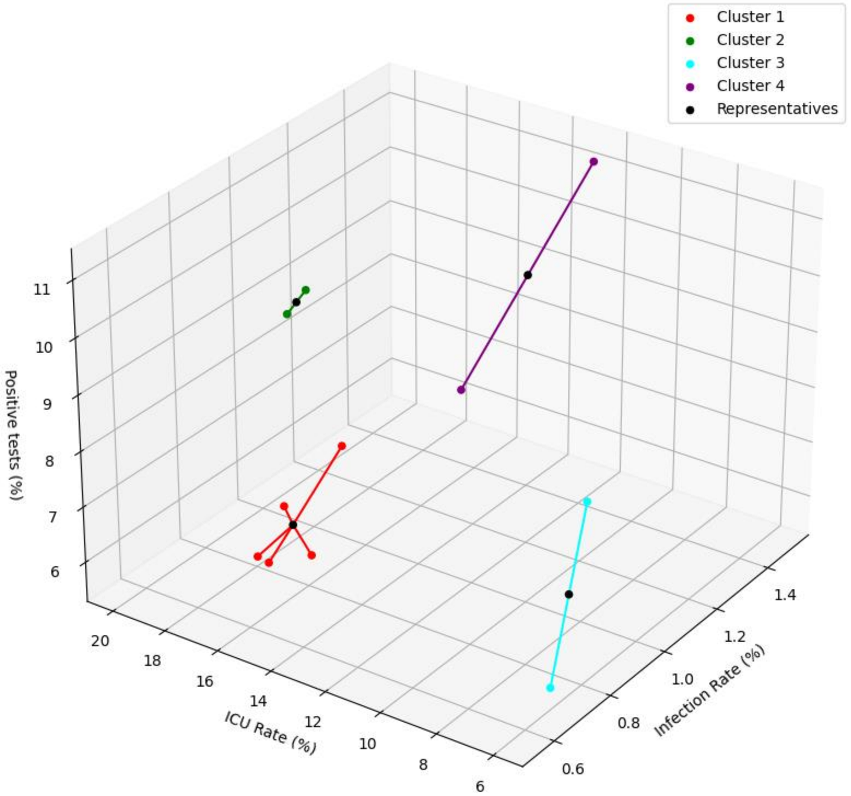


Figure 5 the clusters visualized on a 3D-plane

10.2. Discussion Mobility differences

Looking at the mobility data and comparing it to the clustering result. Some interesting results can be found.

Table 4 shows the mean mobility difference for each category, for each province.

Province	Cluster	Recreation	Grocery	parks	transit	workplaces	Residential
Flemish-Brabant	3	-33.2258	2.129	51.9032	-42.6452	-27.7742	13.4516
Walloon-Brabant	3	-27.2083	6.2083	60.7917	-18.625	-22.4167	11.75
Antwerp	1	-38.0968	5.0323	35.7097	-33.129	-19.7097	11.1613
Liege	1	-28.3226	2.4516	21.8065	-18.8387	-19	8.871
Limburg	1	-32.7419	5.2903	55.0645	-24.3226	-16.129	10.1613
East-Flanders	1	-33.129	10.9032	18.4839	-23.7419	-19.4516	11.5161
West-Flanders	1	-31.7419	15.7097	17.2581	-16.129	-14.7742	10.129
Luxemburg	4	-26.129	11.3548	55.1613	-18.0323	-18.871	9.0323
Namur	4	-25.2903	8.4194	46.7419	-21.6774	-19.9032	9.6774
Hainaut	2	-25.3226	2.4839	21.0323	-31.2903	-16.1935	8.5484
Brussels	2	-50.3871	-5	35.0968	-39.129	-37	12.3548

The first conclusion to be made based on these numbers is that provinces that differ the least in the recreation category from the baseline, are also clustered in the more severe clusters. This is the case for Namur, Luxemburg and Hainaut. With the exception of Brussels, Brussels being the city with the highest population density in Belgium. This leads to higher ICU-rates and infection rates.

But there is a fairly big difference of around 5% between the provinces in severe clusters 3 and four and the provinces in cluster 1 and 2.

This could help us conclude that recreational and retail mobilities have a larger impact on the spread of the COVID-19 virus.

Secondly we can also see that outdoor activities concerning park visits have increased immensely, but have a significantly smaller impact on the spread of the virus. Provinces where the increase of park visits is above 50% still remain in lower severity clusters, with one exception being Luxembourg. This could indicate outside activities having a less significant impact on the spread of the COVID-19 virus.

Thirdly we find that workplace mobility has an impact on the severity of the COVID-19 pandemic in the Belgian regions. We can see that the clustered regions having the lowest COVID-19 numbers also have the largest decrease in workplace mobility, for example, Flemish-Brabant and Walloon Brabant. Regions clustered in worse clusters show a smaller decrease in workplace mobility. With the exception of Brussels. This is because Brussels is the largest city in Belgium, providing the most employment opportunities and being the international center of Belgium.

Figure 6 depicts for each province the difference in recreational and retail mobilities from the baseline, being the daily averages before the COVID-19 pandemic.

Figure 6 difference in recreational and retail mobilities

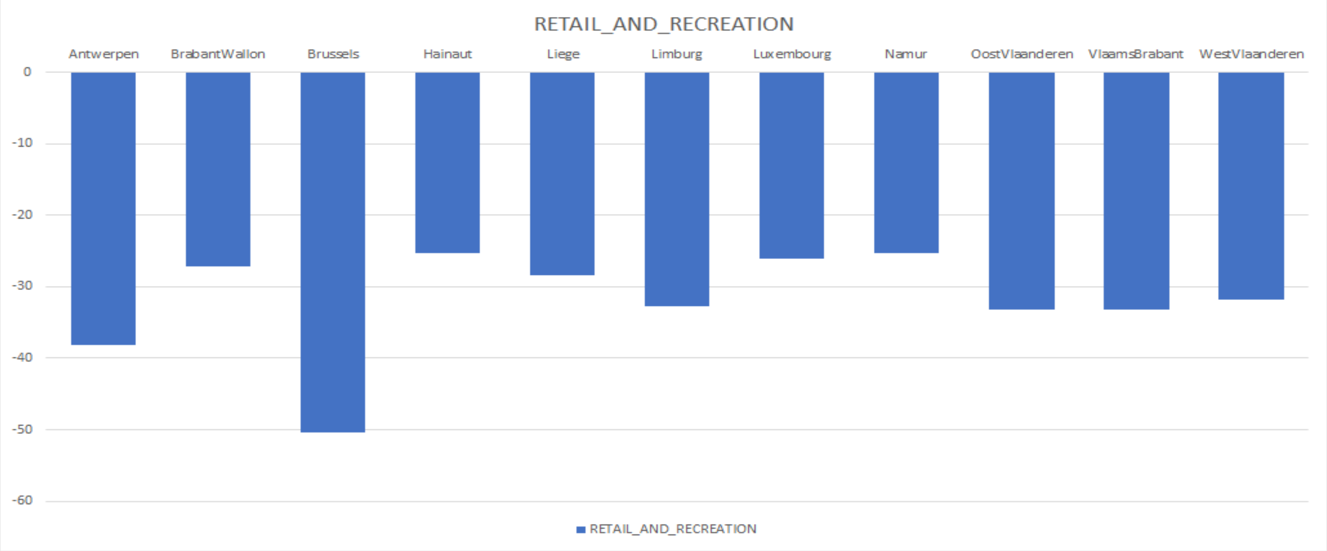


Figure 7 depicts the increase of park visits per province during the month of March 2021.

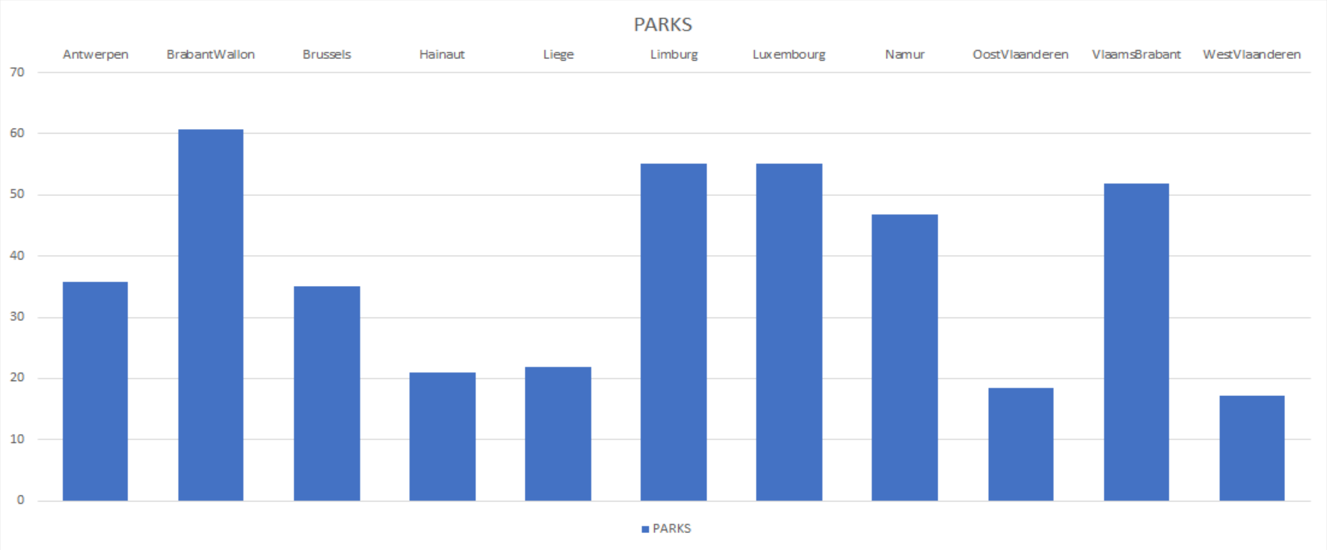


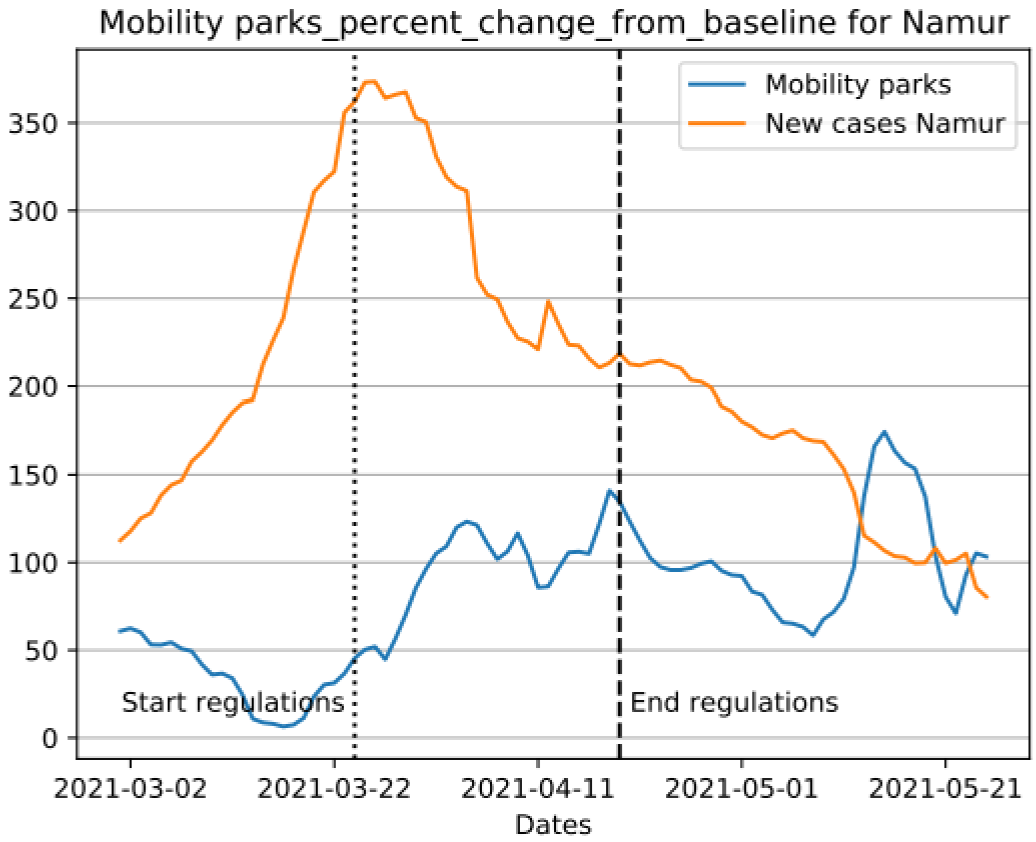
Figure 7 increase of park visits per province

When looking at the trends provided by the Neural Network graphs, several conclusions can be made. A first one to look at is the opening and closing of the schools and their impact on the increase of active infections and reproduction number.

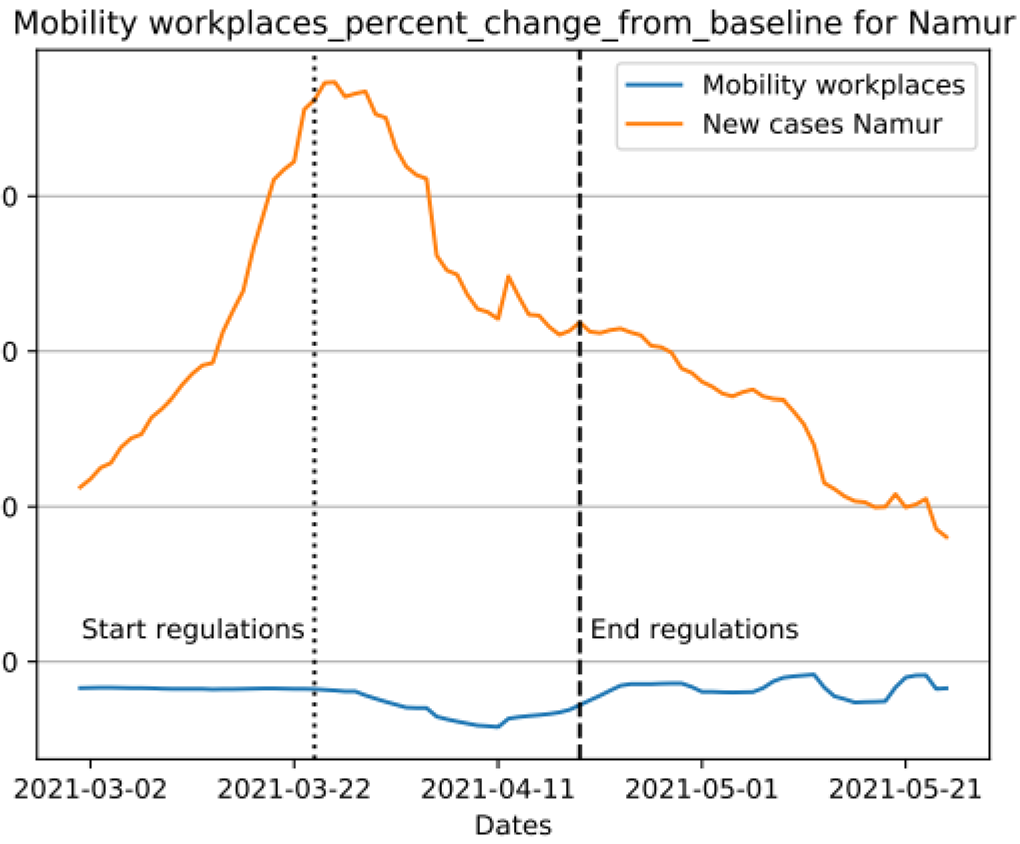
When analyzing the decrease in COVID-19 cases measured by the end of march, we can see the government taking additional regulations, more specifically closing the schools. But also non-medical contact professions had to close. It is clear the impact these regulations had. When looking at the results and conclusions previously made in this paper we can see alignment with the government's choices. Both recognize the larger impact of indoor activities than outdoor activities.

To provide a clear graph to analyze trends in mobility and COVID-19 cases the paper uses a 7 day moving average to plot the different graphs.

The figure above also shows an interesting shift of behavior. We can see a clear increase in Park visits starting around the 24th of march 2021. Yet a significant decrease in new COVID-19 infections. This again indicates that outdoor activities have little to no negative impact on the evolution of the virus.



To identify what might have a negative impact on the spread of the virus we analyze the next graph.



This figure depicts the change of work place mobility for the province of Namur. When looking at the graph we can see a significant dip in workplace mobility. This can point to the fact that this type of mobility has a negative impact on the spread of the virus. The new government regulations advised working from home during this period. The decrease in workplace mobility has also led to a decrease in new cases. This can be an indicator of a correlation between these trends.

10.3. SIR-model results

To create a better idea of the impact of different mobility changes on the spread of the COVID-19 virus, SIR-model predictions were made for each of the clustered regions. To get to our results we first had to undertake some steps. Those steps are described in the chapters below.

10.3.1. Fitted parameter results

The model contains 2 constants who have to be estimated, namely $\beta$and γ. They were both estimated using nonlinear least square fitting on data from March 1st until the 30th of April.

Estimating both parameters is challenging, these parameters depend on many different factors. To provide several insights the paper looks at 2 different implementations of these estimations one based on a constant βparameter and one with a transient βestimated with the ANN.

We combined the clustering results with the SIR-Model fitting and prediction. With one exception being Brussels, We consider Brussels a separate cluster, since its population density is way higher than the remaining clustered provinces. Separating Brussels from its original cluster also means Hainaut will be a cluster on its own.

The paper proceeds to analyze these clusters. So for each cluster SIR-Model estimations were made and fitted. Leading to the following results.

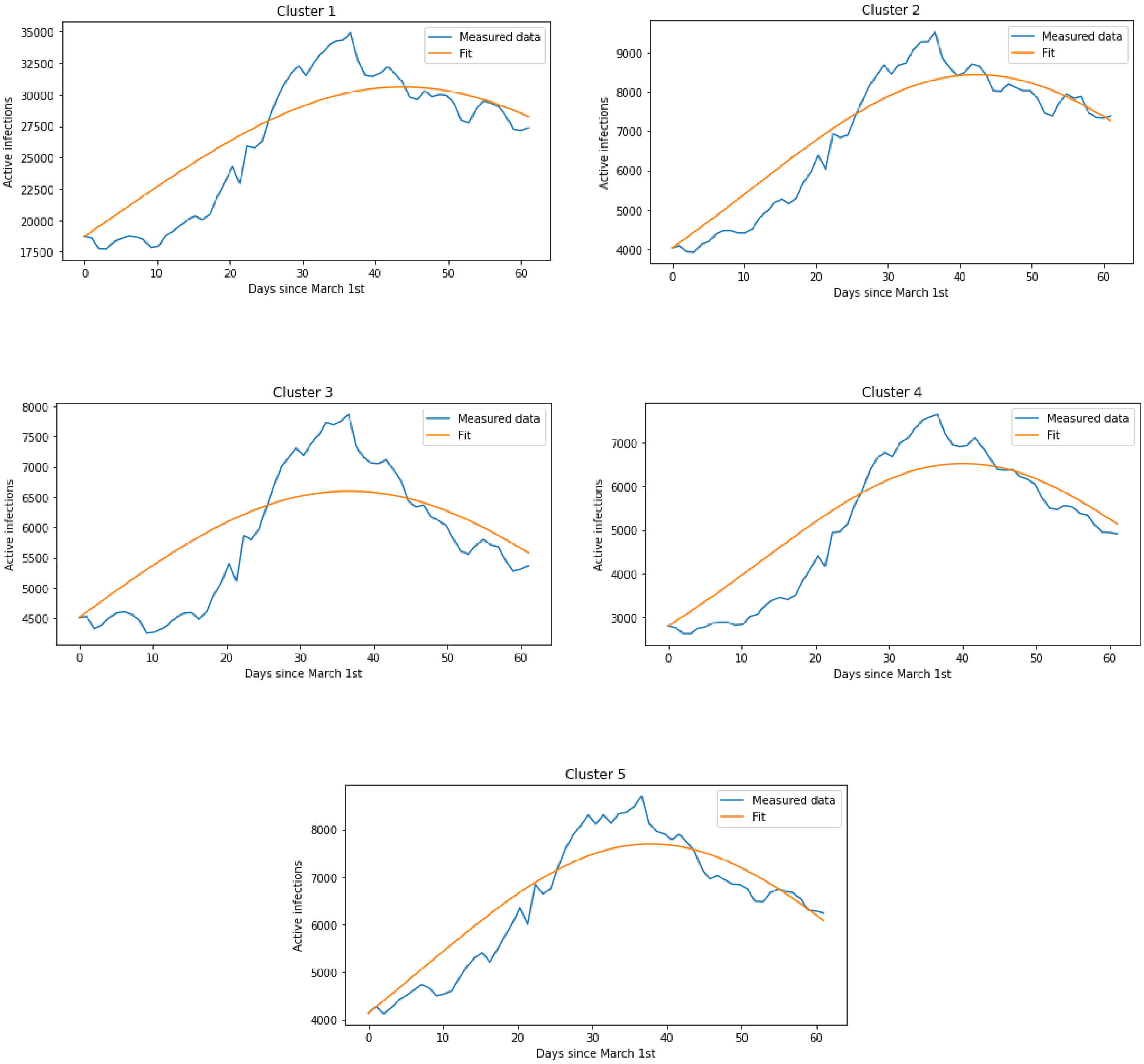


Figure 8 depicting the found best-fit to the data for each cluster

After fitting the SIR-model for each cluster we found the following $\beta\ and\ \gamma$parameter values.

Table 5 shows the found values for $\beta$and γ and the calculated value for R_0.

Cluster	β	γ	R_0
1	0.3645	0.3217	1.1329
2	0.4139	0.3466	1.1940
3	0.3957	0.3534	1.1197
4	0.4100	0.3397	1.2069
5	0.4158	0.3533	1.1769

According to the best fitted $R_0$the clustering results get confirmed. We see that cluster 2 and 4 are the clusters with the highest R_0 values. Cluster 5 originally belonged to cluster 2, but were split up because of the significantly higher population density. We can see that the regions clustered in the "best" cluster group, also have the lowest R_0value.

This is also to be compared with the mobility results. Confirming that regions where recreation and retail mobilities have decreased the least, also have the highest R_0value.

The same conclusion can be made for the park visits. We still see the regions having significantly higher park visit percentages, having a lower R_0value.

10.3.2. Predictions

Using these parameters we made predictions for the next 10 coming days. The graphs below show the results for those 10 days.

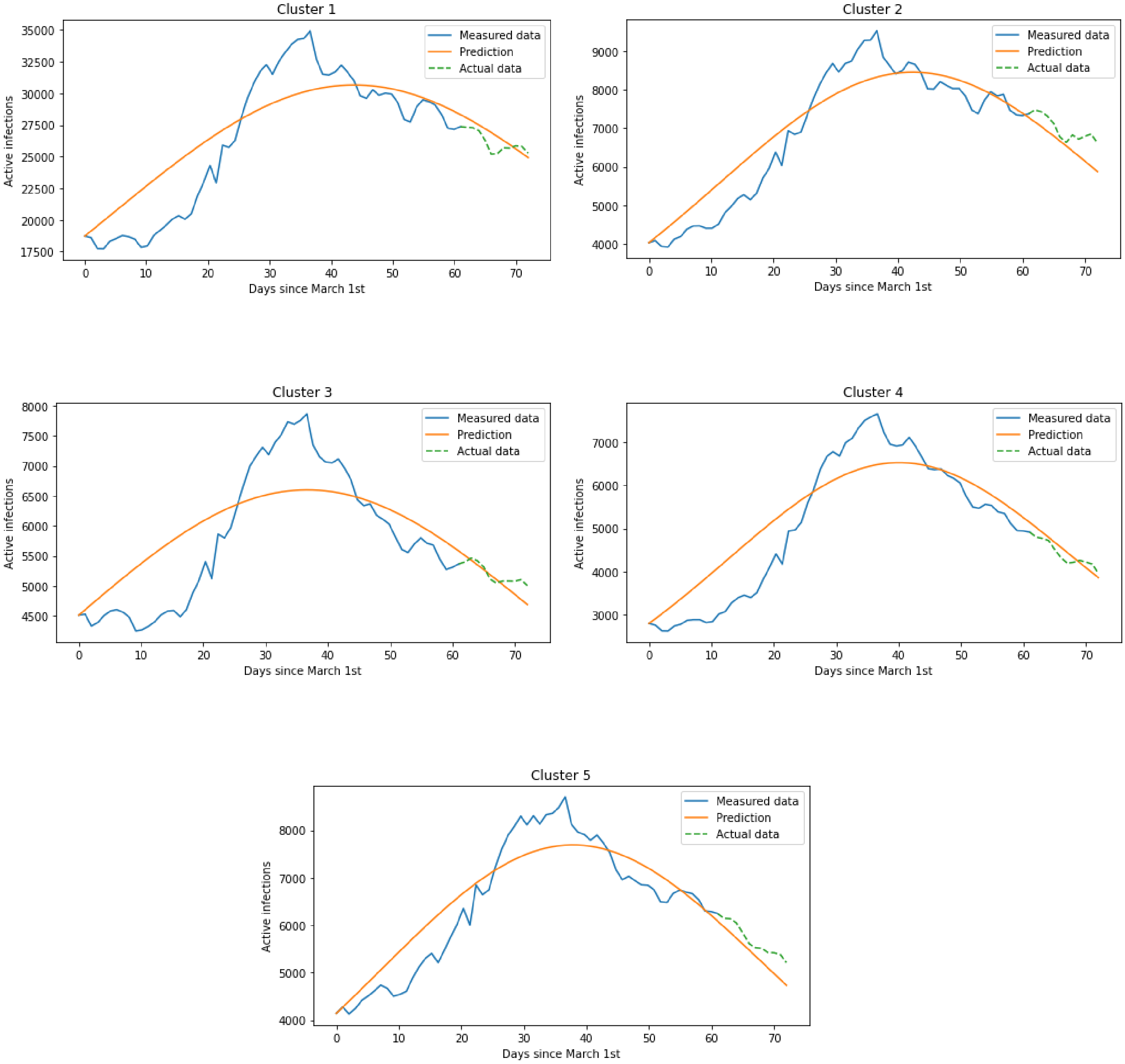


Figure 9 Predictions for the first 10 days of May for each cluster

Comparing the actual data to the prediction the model made we can see that the prediction was fairly close to the actual data measured for the first 10 days of May. We can observe a downwards going trend going further into May.

When combining these results with the mobility reports from March and April, several conclusions can be made. Firstly we can see that increased park visits have no influence on the number of COVID-19 infections. When comparing the mobility numbers from March and April, a large increase of park visits can be observed. When looking at the SIR-Model predictions, We see that the number of infected people started decreasing in the month of april.

Secondly we see the largest decrease in workplace mobility over the month of April. This points to the impact of indoor work related activities. Similarly for the use of public transport.

Combined with the decrease in workplace mobilities we see some change in the Belgian COVID-19 regulations, The main change being the Schools having to close. Having several schools closing around early April, and on the 10th of april all schools having to close obligatory, indicates the large impact of the school environment on the COVID-19 numbers. Seeing as how most mobility categories in the reports stay fairly stable. with the exception of workplace and public transport mobilities.

10.4. SIR-Model prediction using transient parameters

To increase accuracy of the model predictions we used ANN to determine the $\beta$and γ as transient parameters. Both $\beta$and γget reevaluated for each day in the training data. This allowed us to have a more accurate evaluation compared to using a constant $\beta$and γ. Training the model for Belgium resulted in these forecasts provided by the model.

10.4.1. Predictions of the development of active cases

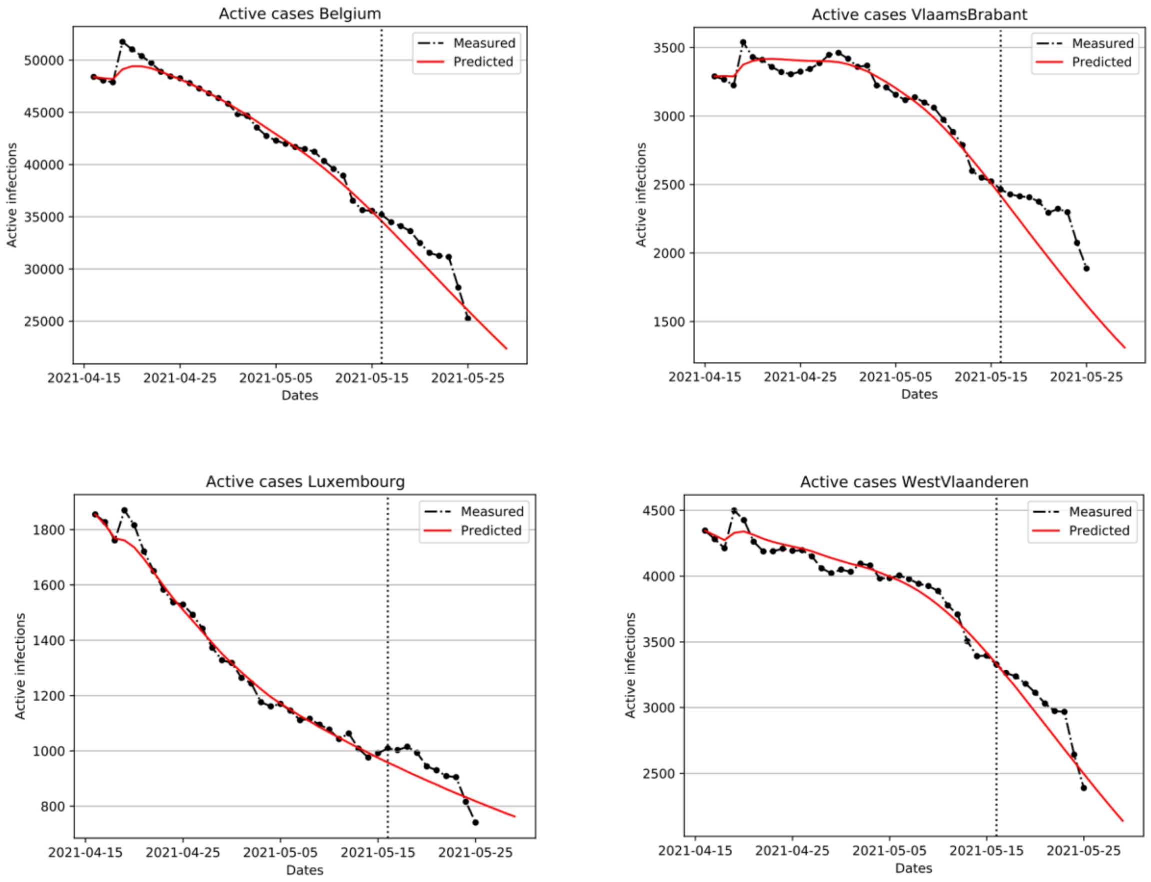


Figure 10 the predictions for Belgium, Flemish Brabant, Luxembourg and West Flanders respectively

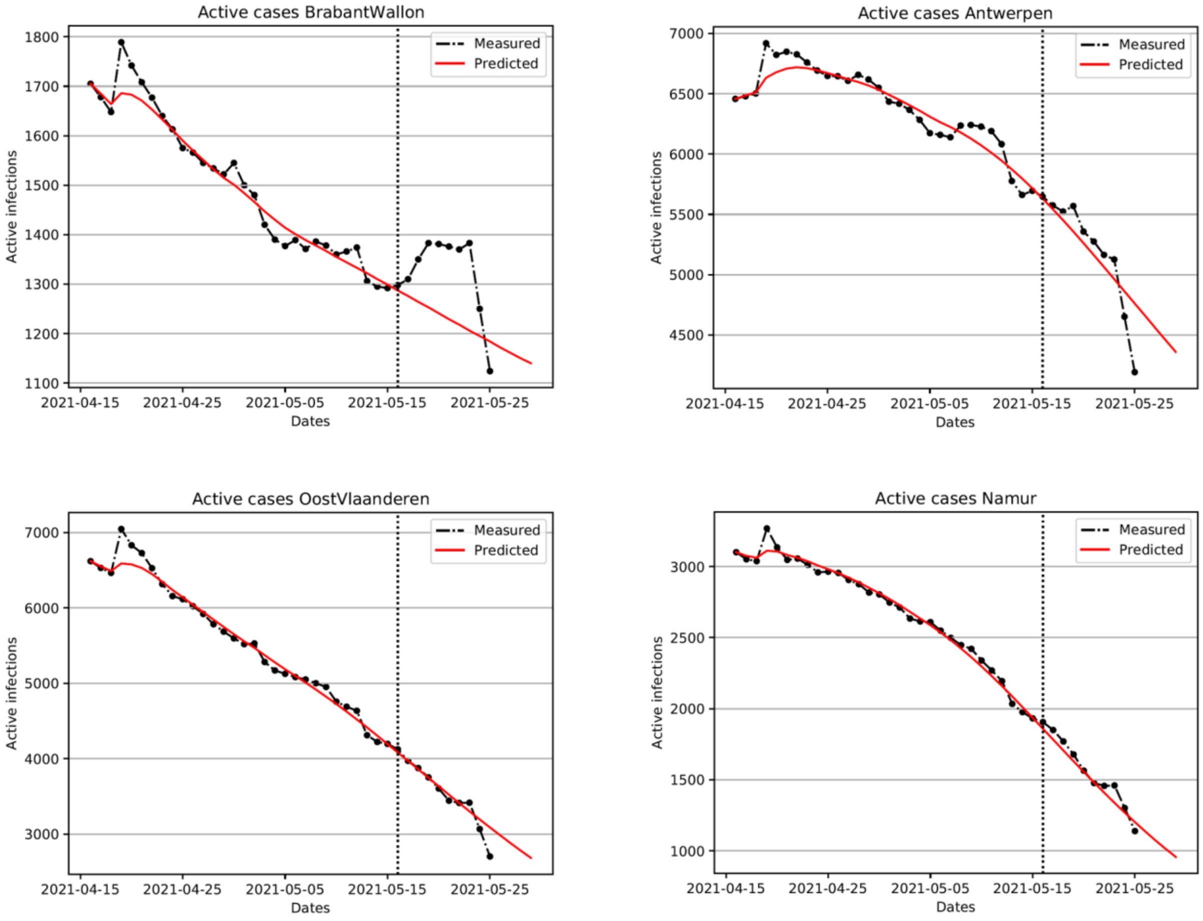


Figure 11 the predictions for Walloon Brabant, Antwerp, East Flanders and Namur respectively

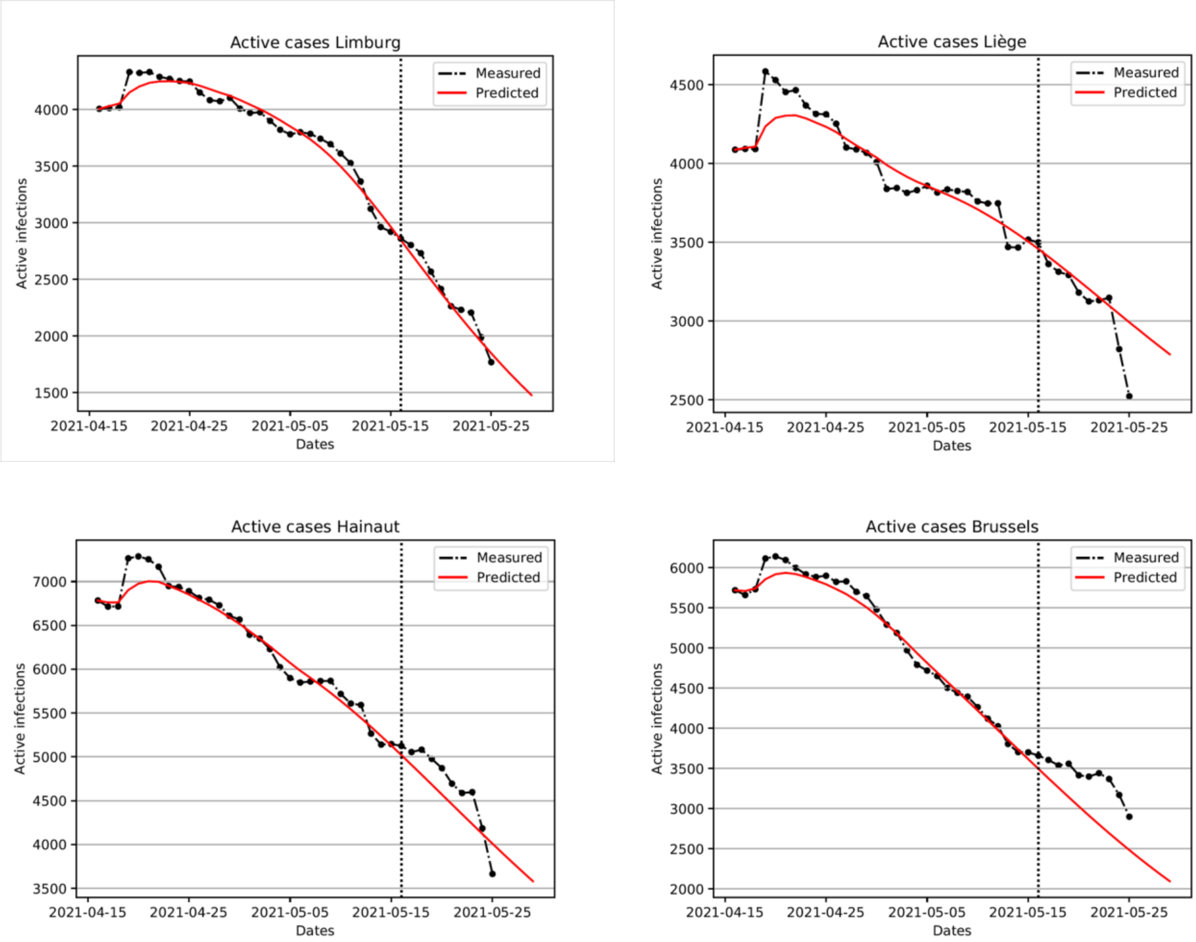


Figure 12 the predictions for Limburg, Liège, Hainaut and Brussels respectively

In general we see a decrease in the active amount of cases. It's also clear when looking at the left side of the vertical drawn line in the graphs, which shows the training period, that the model has been fit way more accurately in comparison to the predictions with constant parameters. Sometimes we can see that the prediction isn't as accurate as we would hope for. For example this is the case in Walloon Brabant one of the reasons for this slight inaccuracy in the prediction would probably be the sudden small spike in active cases right after the end of the training period.

10.4.2. Predictions of the reproduction number

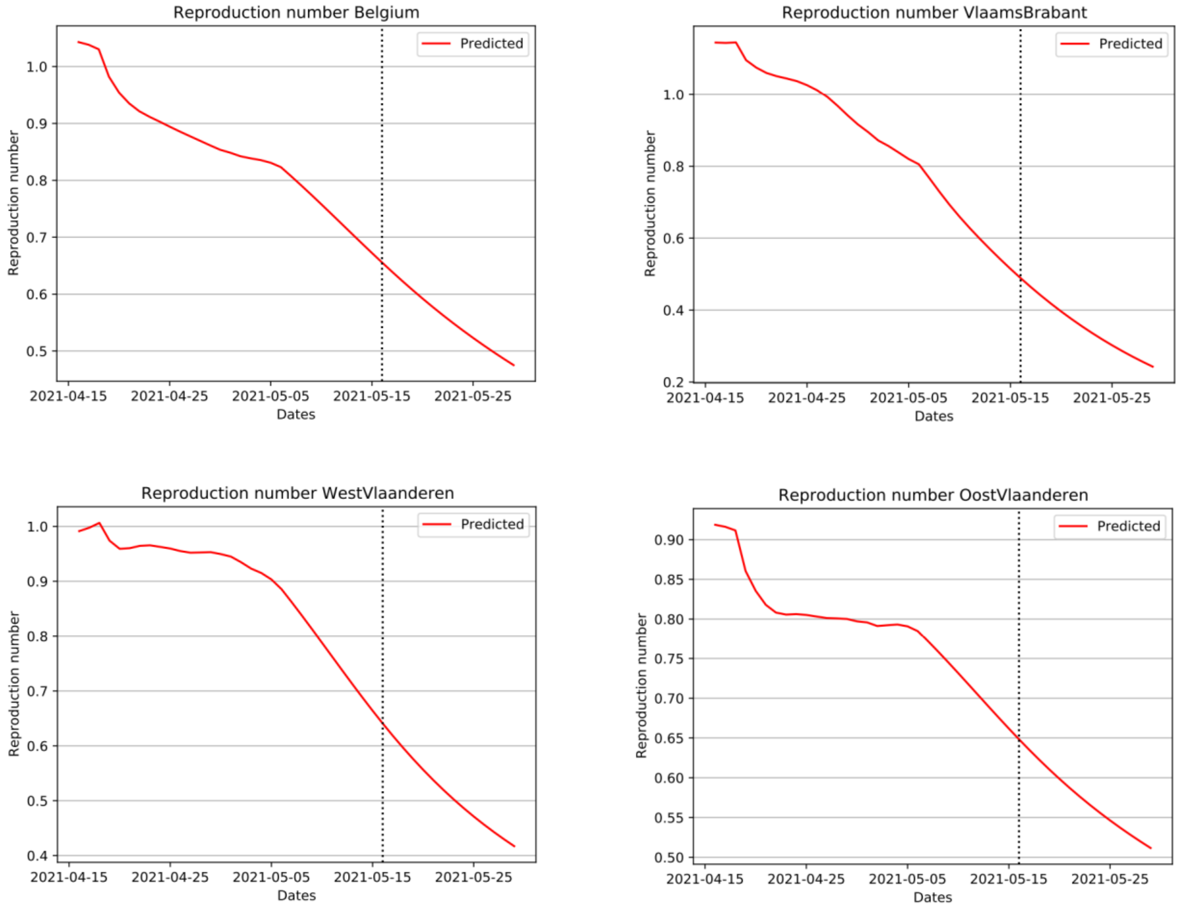


Figure 13 the development of the reproduction number for Belgium, Flemish Brabant, West Flanders and East Flanders respectively

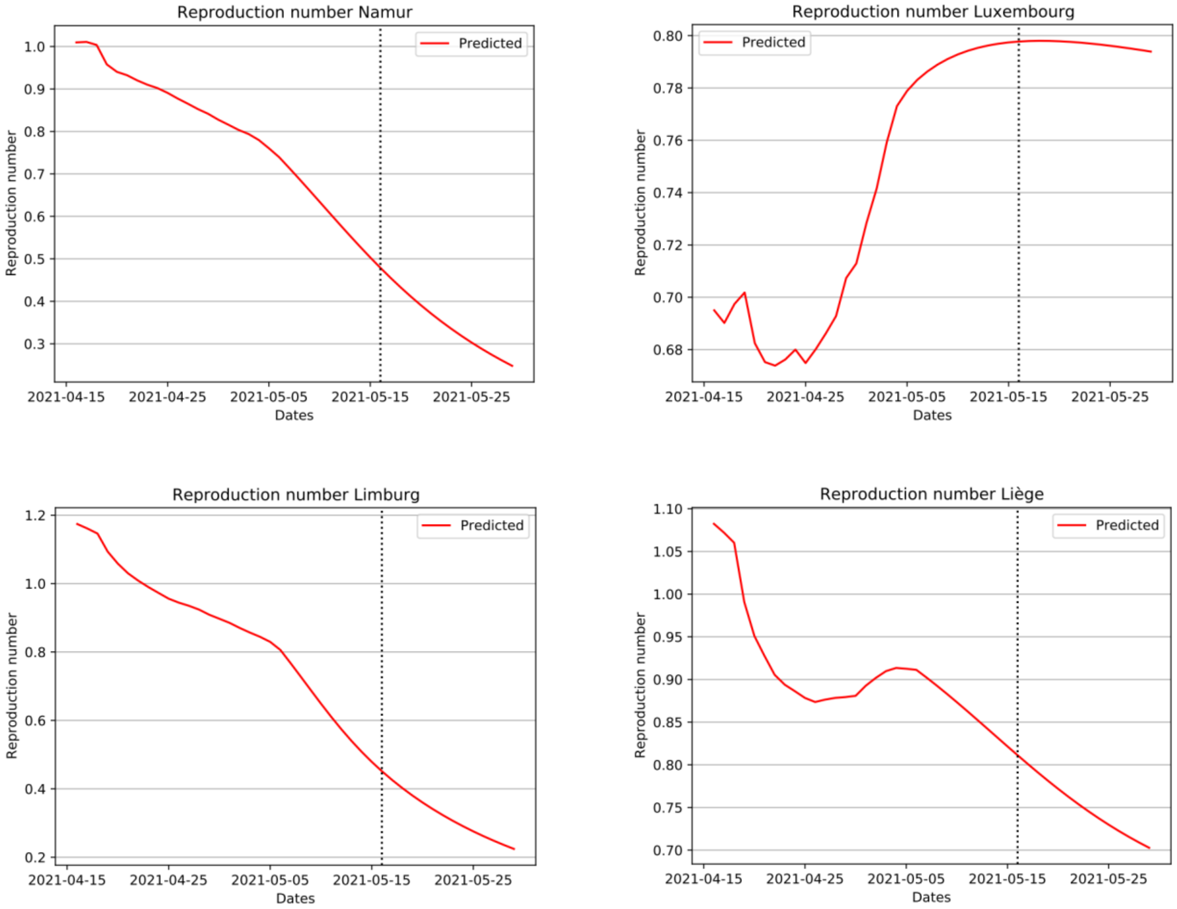


Figure 14 the development of the reproduction number for Namur, Luxembourg, Limburg and Liège respectively

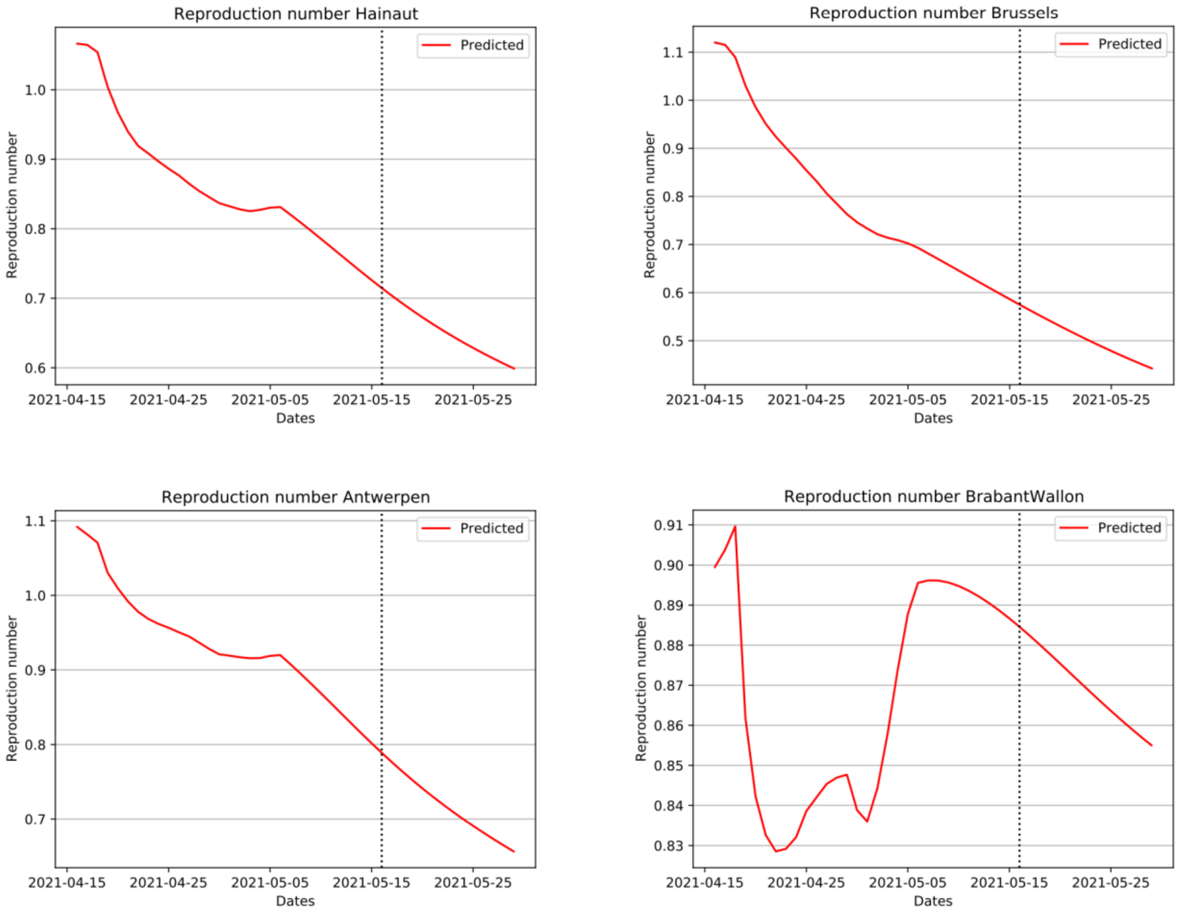


Figure 15 the development of the reproduction number for Hainaut, Brussels, Antwerp and Walloon Brabant respectively

In most provinces we can see a decrease of the reproduction number. A decrease in the reproduction number will also result in a decrease of the total active cases over time. We do see a slight increase in the provinces Brabant Walloon and Luxembourg. Nevertheless the reproduction number stays below 1 which leads to a decrease in active cases.

11. Conclusion

We found that when using the infection rate, ICU rates and the positive test percentage Belgium's provinces can be clustered into 4 main clusters. Later we decided to put Brussels in it's own fifth cluster, this because of its high population density in comparison to the other provinces.

When comparing the clustering results with the mobility data provided by Google we found that park visits hardly have any negative impact on the current active cases. We found that obligating people to work from their homes and closing schools had a relatively high impact on not only the amount of infections but also the amount of ICU beds used.

When using both versions of the SIR-model we could see a downward going trend in the amount of active infections and in most cases for the reproduction number. But it was obvious looking at the different graphs that the SIR-model together with the neural network gave better and more accurate results. The SIR-model with constant parameters can be used to make rough predictions but it's definitely advised to use it together with a neural network to get best results. We saw some inaccuracies in the active infections from the neural network, this because of sudden spikes in infections right after the end of the training period.

References

Boyd, S. and Vandenberghe, L., 2018. *Introduction to applied linear algebra.* Cambridge: Cambridge University Press, pp.45-87**.**

Virgantari, F. and Erika Faridhan, Y., 2020. *K-Means Clustering of COVID-19 Cases in Indonesia's Provinces.* [online] Available at: https://repository.unpak.ac.id/tukangna/repo/file/files-20201228112135.pdf [Accessed 22 March 2021].

Fakhreddine, K. and Hammouda, K., 2000. *A Comparative Study of Data Clustering Techniques.* Waterloo: University of Waterloo.

Md. Zubair, Iqbal, A., Shil, A., Haque, E., Moshiul Hoque, M. and Sarker, H., 2020. *An Efficient K-means Clustering Algorithm for Analysing COVID-19*.

Aydin, N. and Yurdakul, G., 2020. *Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms.* Applied Soft Computing Journal, 97 (106792).

P. Sinaga, K. and Yang, M., 2020. *Unsupervised K-Means Clustering Algorithm.* IEEE Access, [online] 8(19582086), pp.80716 - 80727. Available at: https://ieeexplore.ieee.org/abstract/document/9072123 [Accessed 27 March 2021].

Google, L. L. C. (2021) "Google COVID-19 Community Mobility Reports." Available at: https://www.google.com/covid19/mobility/ (Accessed: April 12, 2021).

Sciensano (2021) "COVID19BE_CASES_AGESEX." Available at: https://epistat.wiv-isp.be/covid/ (Accessed: May 28, 2021).

Sciensano (2021) "COVID19BE_HOSP." Available at: https://epistat.wiv-isp.be/covid/ (Accessed: April 19, 2021).

Sciensano (2021) "COVID19BE_MORT." Available at: https://epistat.wiv-isp.be/covid/ (Accessed: May 28, 2021).

Sciensano (2021) "COVID19BE_tests." Available at: https://epistat.wiv-isp.be/covid/ (Accessed: April 19, 2021).

Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2001. *On Clustering Validation Techniques*. Journal of Intelligent Information Systems, 17:2/3, pp.107–145.

Amaral, F., Casaca, W., Oishi, C. and Cuminato, J., 2021. *Towards Providing Effective Data-Driven Responses to Predict the Covid-19 in São Paulo and Brazil*. Sensors, 21(540).

Newville, M., Stensitzki, T. and Otten, R., 2021. *Modeling Data and Curve Fitting — Non-Linear Least-Squares Minimization and Curve-Fitting for Python*. [online] Lmfit.github.io. Available at: <https://lmfit.github.io/lmfit-py/model.html> [Accessed 29 April 2021].