

Exploring Spatial-Temporal Patterns in COVID-19 Disease Data

A research in the Netherlands with the use of Self-Organizing Maps



Author: Bart van Liempd

Student number: 6000002

Supervisor: Ellen-Wien Augustijn

Responsible Professor: Menno-Jan Kraak



Source front page: (RTL, 2020)

Preface

Before you lies the master thesis: Exploring spatial-temporal patterns in corona disease data. This research has been written to fulfill the graduation requirement of the Geographical Information Management and Applications master's. It was written between September 2021 and March 2022, while the pandemic was still going on. This caused somewhat inconvenient circumstances to write the thesis.

The thesis was quite challenging not only due to the circumstances but also due to my inexperience with coding. At the end and extensive code was produced and nice results could be generated.

I would like to thank Ellen-Wien Augustijn for her guidance and support during the process of writing. I would also like to thank Menno-Jan Kraak for his feedback.

I hope you enjoy reading through this thesis.

Abstract

The world is currently facing the COVID-19 pandemic. To be able to understand the scale of the outbreak and to respond appropriately, it is required to track the spread of the virus. Currently, this tracking is done with the use of either temporal or spatial data. This research proposes a method to combine both dimensions to be able to track COVID-19 differently. This method is called the Self Organizing Map (SOM). With the use of SOM five datasets are compared to each other. These datasets are the positive percentage of tests, positive percentage of inhabitants, virus particles in sewage water, deceased cases, and hospitalized cases. For these datasets, the change of the spatial situation over time and the distribution of the local temporal variations over space are analyzed. Furthermore, the different waves of COVID-19 are compared to each other in the same way based on the virus particles in sewage water. In short, the positive percentage of tests and positive percentage of inhabitants showed nearly identical patterns. Hospitalized cases and deceased cases showed similar patterns, although not as similar as the datasets described above. The sewage dataset was the most similar to the hospitalized cases and deceased cases. To investigate this further, other methods should be used to evaluate the similarities. Primarily other clustering algorithms could provide a useful addition to the research.

Table of Contents

Abstract.....	4
1. Introduction	8
2. Theoretical framework.....	11
2.1 COVID-19.....	11
2.1.1 Confirmed cases.....	11
2.1.2 Deceased cases	12
2.1.3 Hospitalized cases	12
2.1.4 Sewage data	12
2.1 Diffusion patterns.....	13
2.4 Machine learning.....	14
2.5 Neural Network.....	15
2.6 Self-organizing maps	15
2.6.1 Sammon’s projection	16
3 Data framework	17
3.1 Data description.....	17
3.2.1 Confirmed cases.....	17
3.2.2 Number of tests	18
3.2.3 Daily Deaths	19
3.2.4 Hospitalized cases	20
3.2.5 Sewage treatment data	20
3.2 Data evaluation	22
3.2.1 Sewage treatment Data	23
3.2.2 Confirmed cases.....	24
3.2.3 Deceased.....	28
3.2.4 Hospital admissions	28
3.3 Data comparison	29
3.3.1 Comparison of different datasets Utrecht.....	31
3.4 COVID-19 in the spatial dimension	31
3.4.1 Hospitalized cases	31
3.4.2 Deceased cases	33
3.4.3 Positive percentage.....	35

3.4.5 Sewage data	36
3.5 Data conclusion	38
4 Methodology	39
4.1 Time period first sub-question	39
4.2 Time period second sub-question	41
4.3 Self-Organizing Maps	42
4.3.1 Input data	42
4.3.2 Training	43
4.3.3 Mapping	43
4.4 Preparation of Sewage Treatment locations	45
4.5 Visualization of results	46
Counts plot	46
Codes plot	46
Clusters plot	46
Mapping plot	46
Sammon's mapping & trajectories	46
4.6 Comparison of SOM results	46
4.7 Software	47
5 Results	48
5.1 Explanation of parameters	48
Grid size	48
Number of iterations	48
Number of clusters	49
5.2 Individual Results	50
Hospital	50
Positive percentage of tests	52
Deceased	53
Sewage	55
5.3 Comparison of SOM results with hotspot analysis	57
Hospital	57
Positive percentage of tests	58
Deceased	58
Sewage	59

Conclusion.....	60
5.4 Results without delayed datasets	61
5.5 Results with delayed datasets.....	67
5.6 Comparision of waves in COVID-19 based on sewage data	74
6. Discussion & Conclusion	81
7 Literature.....	84

1 Introduction

In December 2019, the COVID-19 virus emerged in Wuhan, China (Yang et al., 2020). This virus would grow out to be a pandemic that is still ongoing. COVID-19 arrived in the Netherlands in February 2020 (Ministerie van Algemene Zaken, 2020) and the country has experienced multiple so-called waves of infections in the last 18 months.

It is important to keep track of where and to what extent the coronavirus is present during the pandemic as it makes it easier to understand the outbreak's scale and respond appropriately (Roser et al., 2020). However, the total number of COVID-19 cases is not known. This is because the approximation of the cases is dependent on a variety of factors. Multiple types of data can be used to track COVID-19. The results derived from these data are reliant on the characteristics of the data used. Therefore, it is important to be aware of the data used to generate results.

It has become apparent that countries have varying success in tracking the number of COVID-19 cases. In third world countries, the World Health Organization (WHO) had to support in setting up test centers (World Health Organization, 2020a). Although the WHO has supported the set up of test centers, the reliability of the data still differs per country. Furthermore, not every data collection method is available in every country. When conducting research it is important to realize what data sources are available and how reliable they are.

There are a lot of different approaches to measuring the severeness of COVID-19 per country. A great example of a country that uses many of these different approaches is the Netherlands. The Netherlands is classified as a country that performs comprehensive tracing of COVID-19 (Ritchie, 2020), this means that the COVID-19 infections are fully traced. This is mostly done by tracking the number of positive tests conducted. However, more data is available to evaluate the severeness of the pandemic, these include the following: confirmed cases, hospitalized cases, intensive care admissions, deaths, reproduction number, disease burden, and the COVID -19 particles in the wastewater (RIVM, 2021).

Although there is a wide variety of data available, there is a lot of untapped potential. The Dutch COVID-19 data are solely published either as spatial data or temporal data (see Figures 1.1 and 1.2). This is not sufficient to analyze the spread of the virus. To properly analyze the diffusion of the virus it is necessary to display the data in a spatiotemporal manner.

The spread of a disease can be monitored via diffusion patterns. Diffusion is a spatial-temporal process, the spread of something between locations – a movement across space and a change through time (Schærström, 2009). The concept of diffusion can be applied to multiple phenomena. It has been applied in knowledge diffusion (J. Singh, 2005), violence diffusion (Schutte & Weidmann, 2011), housing price (Pollakowski & Ray, 1997), and most importantly disease diffusion (Schærström, 2009). Disease diffusion refers to the variable spatial occurrence of ill health, such as the epidemics of infections, caused by the spread of disease agents through space and time (Schærström, 2009).

The diffusion of diseases can be analyzed with the use of multiple methods. It has been done with the use of a hierarchical Bayesian approach (Assunção et al., 2001), a spatial clustering algorithm (Kuo et al., 2018), and self-organizing maps (Augustijn & Zurita-Milla, 2013; Melin et al., 2020). The self-organizing maps are a widely applied methodology and have proved to be suitable for the research of disease diffusion (Augustijn & Zurita-Milla, 2013; Melin et al., 2020). Therefore, it is used in this research to find patterns in the diffusion of COVID-19 in the Netherlands.

Self-organizing maps are described as follows:

'The self-organizing map has the property of effectively creating spatially organized internal representations of various features of input signals and their abstractions.' (Kohonen, 1998)

In short, SOM can find similar observations in the input data and create topological relationships between them.

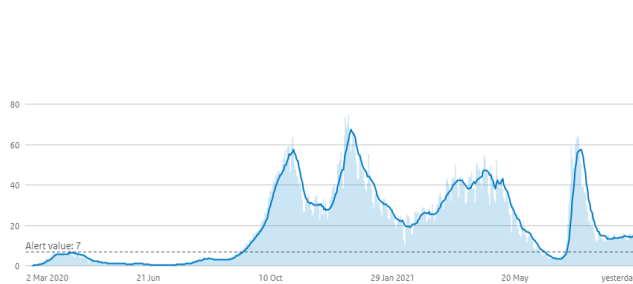


Figure 1.1: COVID-19 cases through time in the Netherlands (RIVM, 2021).

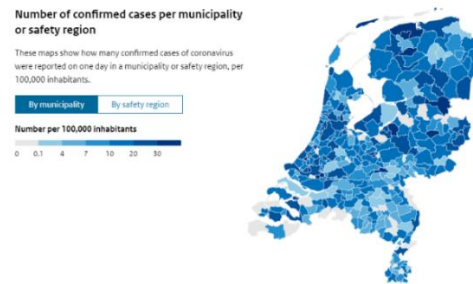


Figure 1.2: COVID-19 cases in space in the Netherlands (RIVM, 2021).

This research aims to find spatial-temporal patterns in the diffusion of COVID-19 in the Netherlands. This fills the gap in the current research which mostly investigates the temporal aspect of the coronavirus spread. To be able to understand the spatial diffusion of diseases is important for predicting and containing diseases (Schærström, 2009). This is more important than ever since researchers claim that we enter an age of pandemics (Joanna Roberts, 2021).

The number of research that takes both the temporal and spatial dimensions into account is limited (Melin et al., 2020)). A suitable methodology to investigate both dimensions is using self-organizing maps (SOM). Moreover, the papers that investigate both dimensions often take a global approach to diffusion (Melin et al., 2020). This research solely focuses on the Netherlands. The benefit of focusing on the Netherlands is that there is a lot of data available. Whereas most research mainly uses the number of confirmed cases (Burki, 2020; Covid et al., 2020; Stoecklin et al., 2020), there is much more choice in the Netherlands. This makes the Netherlands a suitable place for research. Thus, this research can be used to better deal with coronavirus or other virus outbreaks in the future.

The following question is used to explore the spatiotemporal spread of COVID-19 in the Netherlands:

Which spatiotemporal patterns can be found in the spread of COVID-19 in the Netherlands?

To answer this question, the first issue that needs to be addressed is to investigate which data is needed. Multiple data sources can be used to answer the main question. As stated in the introduction these consist of hospitalized cases, intensive care admissions, deaths, reproduction number, disease burden, and the COVID -19 particles in the wastewater (RIVM, n.d.-b).

The Institute for Health Metrics and Evaluation (Institute for Health Metrics and Evaluation, 2021) claims that the number of deaths is the best way to track the progression of COVID-19. However, other studies state that using other resources might be better. The reproduction number (R_0) is used in research (Ives & Bozzuto, 2021) and is described as a fundamental dataset in epidemiology. Hospitalized cases are also used in research (Kashyap et al., 2020). However, in this study, the main goal is not to track the rate of

the disease spread but to evaluate the pressure on hospitals. This can be used as an indication of the severeness of the pandemic. Another way to track the spread of COVID-19 is the number of confirmed cases. Although this is one of the most used methods (Franch-Pardo et al., 2020), the method is heavily dependent on the number of total tests.

All of the methods described above are to varying extents dependent on human effort to track COVID-19. There is a relatively new methodology that is not reliant as much on human effort. This is the tracking of the number of COVID-19 particles in wastewater (Mallapaty, 2020). As shown, there is much choice when researching the diffusion of diseases. To be able to conduct research properly, the most suitable datasets must be chosen. Moreover, the strengths and weaknesses of the datasets need to be addressed. Adjustments to the dataset must be made if the raw data is unfit for the goal of the research.

All of the datasets are not stable over the pandemic period. The stability of data is very important for analyzing spatial-temporal patterns. Data may depend on the number of tests conducted (this will vary over the pandemic period), but also medical expertise on treating COVID-19 cases, leading to higher recovery rates. To investigate which data sources are suitable and how much pre-processing is required to make the datasets useful for analysis, the following sub-question has been defined:

Which datasets are suitable to explore spatiotemporal COVID -19 patterns in the Netherlands and what is the implication of using these datasets.

This sub-question is answered for the most optimal period in time for all datasets. After identifying the most suitable datasets, their spatio-temporal patterns are analyzed. To be able to do this the following question is answered:

What patterns can be found in different datasets?

This question is answered in two different ways. Hotspot analysis and SOM analysis are conducted on the various datasets to detect clusters. These analyses are not performed to compare between datasets, but only to identify if clusters are found within single datasets.

When the suitable datasets have been identified and their spatiotemporal patterns have been identified, it is interesting to look further into the use of these datasets. The dataset concerning the samples of sewage water is used in further analysis. This is done by investigating if the patterns are the same over multiple waves in the Netherlands. The following sub-question is answered concerning this issue:

Which differences can be found in the spatio-temporal patterns of different waves of COVID-19 in the Netherlands with the use of samples of sewage water?

With the help of the sub-questions, the main question can be answered in multiple ways. The spatiotemporal patterns can be identified with the use of multiple datasets. The benefit of using multiple data sources is that the strength of all data sources is used, and the main question can be answered as extensively as possible.

Limitations

This research investigates spatial-temporal patterns for the coronavirus in the Netherlands. It aims to give an objective view of the diffusion of the virus. It does not try to look into the effects of the restrictions and rules introduced by the government. The rules can however be of great importance to the spread of the coronavirus and could be an interesting point of research for the future.

2. Theoretical framework

In this chapter, the theoretical background of the research is described. It starts at COVID-19, followed by the ways to track COVID-19. Then the theories behind the methodologies are described.

2.1 COVID-19

Covid-19, SARS-CoV-2, 2019-nCoV, and the coronavirus are all names for the disease that originated in the Hubei province in China, which has spread to other countries and caused a pandemic (Velavan & Meyer, 2020).

Coronaviruses exist for a long time and were first discovered in 1966, however the type of virus that is COVID-19 only existed in bats before 2019. The generally accepted theory is that the virus transition between bats and humans happened on the Huanan seafood market in Wuhan, China.

People infected with COVID-19 experience mainly pneumonia, which is described as follows (Mayo Clinic, n.d.):

Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing.

These symptoms result in droplets containing COVID-19 in the air, which are caused by coughing or sneezing. These droplets thus result in the transmission of COVID-19 (Galbadage et al., 2020). To be able to track the transmission and the spread of this virus, multiple datasets are used. These consist of confirmed cases, deceased cases, hospitalized cases, and the concentration of rDNA particles in sewage. COVID-19 tracking

COVID-19 can be tracked using multiple datasets. Commonly used datasets consist of confirmed cases, deceased cases, hospitalized cases, and COVID-19 particles in sewage water. In this section, these ways to track COVID-19 are defined and described.

2.1.1 Confirmed cases

One of the most used data to track the spread of COVID-19 is confirmed cases (Franch-Pardo et al., 2020). One of the reasons is that the number of confirmed cases is tracked by almost every country and is therefore widely available. A confirmed case of COVID-19 is defined by the World Health Organization as (World Health Organization, 2020c): A person with a positive Nucleic Acid Amplification Test (NAAT). This medical definition in simplified words is a person with laboratory confirmation of COVID-19 infection, irrespective of clinical signs and symptoms (Madabhavi et al., 2020). In the Netherlands, the same definition is used to construct the data. The COVID-19 tests conducted at the GGD, which have a positive result are used as the indicator for a confirmed case.

There is, however, a large issue with the tracking of COVID-19 through the number of confirmed cases. This data is dependent on the number of tests conducted. Africa is a continent with a low amount of cases, however, this is most likely due to the limited amount of testing (Chitungo et al., 2020). Because of this issue, it is better to use the percentage of tests with a positive result. This can be used to minimize the bias that is present when only using the number of confirmed cases (Pitzer et al., 2021). In the Netherlands, this number of tests is defined as the total number of conducted COVID-19 tests with a result by the GGD. This is the number of tests with a positive result and a negative result.

2.1.2 Deceased cases

The number of deaths is claimed to be the best indicator for the progression of the pandemic (Institute for Health Metrics and Evaluation, 2021). However, when critically analyzing this claim, it is safe to say that there are multiple difficulties when using the number of deaths to track COVID-19. There is a delay between the day that someone passes away and the day that the person's death is reported (Institute for Health Metrics and Evaluation, 2021; Ritchie et al., 2020). This delay is estimated to range between 17 and 21 days (Institute for Health Metrics and Evaluation, 2021). However, this is not the only issue with the data. The actual death toll from COVID-19 is most likely higher than the number of confirmed cases. This is again related to the issue of limited testing. Moreover, it can be difficult to determine whether the cause of death is directly linked to COVID-19 (Ritchie et al., 2020). The way the death figures are constructed differs between countries. In some countries, only hospital deaths are used as the death count of COVID-19. Other countries also include deaths in homes. The discrepancy can cause different numbers between countries, which results in non-comparable statistics. In the Netherlands, this RIVM reports the deaths daily. It concerns the number of deceased persons per day. However, they also state that the number of deaths is likely to be higher, since not every person is tested for COVID-19. This causes the deceased person to not count towards the number of deceased cases.

2.1.3 Hospitalized cases

Hospitalized cases are the best metric to use in research to evaluate the spread of COVID-19. It gives the best overview of the pressure on healthcare (Kashyap et al., 2020). The hospital admissions put more pressure on the healthcare than confirmed cases and deceased cases. This is due to people in the hospital needing care throughout the entire day. People are not directly taken into the hospital when they are diagnosed with COVID-19. The time between diagnosis and hospitalization ranges between 3 and 10 days (Faes et al., 2020). The range is large due to it being influenced by various factors such as age, living situation, etc. It is also important to mention that the length of the stay in a hospital also ranges between 3 and 14 days (Faes et al., 2020; Vekaria et al., 2021). This is also influenced by multiple factors such as age and health. The length of the stay has also decreased as time went on and the knowledge on treating COVID-19 expanded. In the Netherlands, the number of hospital admissions is described as the number of reported hospitalizations. The RIVM also states that the true number is probably higher since the GGD is not always notified immediately of hospital admissions.

2.1.4 Sewage data

The use of data generated by sewage treatment plants can reveal the true scale of the COVID-19 outbreak (Mallapaty, 2020). It provides a better estimate of the extent of the coronavirus than the other datasets. This is due to sewage being able to account for people who have not been tested and those who have only minor symptoms (Mallapaty, 2020). It has shown to be a leading indicator of community infection ahead of COVID-19 testing data and hospital admissions (Peccia et al., 2020). A large benefit of this data source is the fact that it is not reliant on testing. This is the case for most other data sources. Moreover, there is no room for human judgment which is the case for deceased cases. The only human interaction required is the sample taking. In the Netherlands, the sample taking is as follows. For every sewage treatment plant, the sample is taken for 24 hours. These samples are then investigated on the presence of virus particles by researchers of the RIVM. The data is displayed as the average number of virus particles

in sewage, corrected for the volume of sewage water, and is calculated per 100.000 inhabitants of the service area (RIVM, n.d.-a).

2.1 Diffusion patterns

Diffusions patterns are used to investigate the patterns that are shown by the datasets described in the sections above.

Diffusion is the ability to spread outwards or to disperse from one or more limited centers to a wider geographical area. It can be seen as a spatial pattern with a temporal element added to it. A spatial pattern is the arrangement of individual entities in space and the geographic relationships among them (Chou, 1995). Thus adding a time period to this definition nears the definition of diffusion. Diffusion can be classified into four categories which represent the characteristics of the spread. The four categories are expansion diffusion, relocation diffusion, contagious diffusion, and spatial diffusion (Cliff et al., 1981). These are all forms of specific spatial patterns. In Figure 2.2.1, the different diffusion classes are visualized (Cliff et al., 1981).

Expansion diffusion is a spread where the spreading phenomena start at the source and diffuse outwards into new areas. Common examples of this type of diffusion are the spread of wildfires or the diffusion of innovation.

Relocation diffusion is quite similar to expansion diffusion. It also spreads outwards from its source into new areas, however, it leaves behind its origin. This means that the phenomenon is no longer present at the source. An example of relocation diffusion is migration. A person lives in a certain country (the source) and leaves to a different country, thus the person is no longer present at the source.

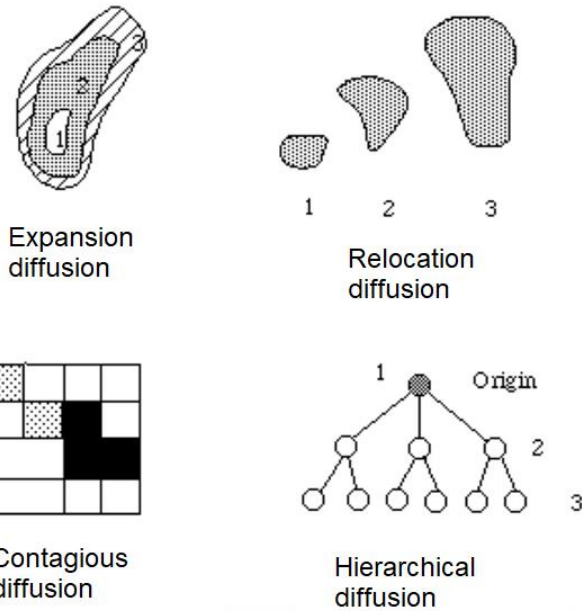
Contagious diffusion is the spread of a phenomenon to neighboring objects in space. These objects can vary in form. An example of contagious diffusion can be the spread of terrorism to countries, the neighboring objects, in this case, are the countries that share a border (LaFree et al., 2018). A different example is the spread of an infectious disease that requires direct contact between individuals for an infection to occur.

Lastly, hierarchical diffusion is the spread of a phenomenon through an orderly sequence of classes or places. In other words, diffusion follows a certain hierarchy. Diseases can also be used as an example of hierarchical diffusion. Studies describe the spread of AIDS as hierarchical diffusion because they observed a diffusion pattern from large urban centers to smaller towns in the U.S (Gould et al., 1991).

While a distinction can be made between these four classes, a diffusion pattern can meet the requirements of multiple classes. Thus, one diffusion pattern class does not exclude the presence of another class.

Spatial patterns are widely used in machine learning algorithms to process and visualize large amounts of data (Ngoc Thach et al., 2018).

Figure 2.2.1: Multiple ways of diffusion visualized (Cliff et al., 1981)



2.4 Machine learning

Machine learning has been around since 1959 (Samuel, 1959). He describes machine learning as the programming of a digital computer to behave in a way that would be described as involving the process of learning. The study concludes with the conclusion that in the future it is possible to write learning schemes that will outperform an average person and become economically feasible. This prediction turned out to be correct, as it is one of today's most rapidly growing technical fields (Jordan & Mitchell, 2015). They state that machine learning nowadays is focused on the question: How can one construct computer systems that automatically improve through experience? Thus the emphasis on machine learning has shifted from learning to improving. The first study (Samuel, 1959) also predicted that machine learning would become economically feasible. This is confirmed by the statement that machine learning is now applied in several fields including health care, manufacturing, education, financial modeling, policing, and marketing (Jordan & Mitchell, 2015). The strength of machine learning is in tasks involving intelligence or pattern recognition which are difficult to automate but are easily done by humans (Dongare et al., 2008).

There are two types of machine learning, supervised- and unsupervised machine learning (el Naqa & Murphy, 2015).

Supervised machine learning is the construction of algorithms that can produce general patterns using labels to predict the classification of input data (el Naqa & Murphy, 2015; A. Singh et al., 2016). These classifications are defined beforehand and the algorithm thereafter places the input data into the most suitable category.

Unsupervised machine learning is different from its supervised counterpart because the categories or classifications are not defined beforehand. The machine has a variety of degrees of freedom in the way it shapes these classifications based on the input data (el Naqa & Murphy, 2015). The longer the algorithm

runs, the better the machine can construct classes based on the input data. Thus the use of larger datasets can aid in constructing better classes.

Neural networks are a popular framework to perform machine learning. They are a specific type of machine learning and the workings are described in the next section(Keijsers, 2010).

2.5 Neural Network

Neural networks are mathematical models that use learning algorithms inspired by the functioning of the human brain to store information(Keijsers, 2010; Marini, 2009). When the neural networks are used in machines they should be referred to as 'artificial neural networks'. The similarity between the human brain and neural networks is that both are built up of many neurons with connections between them (Keijsers, 2010). Neural networks are used to emulate features of biological neural networks to address a range of difficult information processing, analysis, and modeling problems (de Smith et al., 2007).

The neural network consists of nodes (neurons) and the connections between them. These connections are weighted, meaning that the strength of the relationship between nodes varies (Dongare et al., 2008). As neural networks are a form of machine learning, they are also focused on gradually improving. This improvement is done by training. The training of a neural network consists of feeding a dataset into the algorithm and letting the algorithm adjust to this dataset. For supervised networks, the training is complete when the algorithm provides the desired output. For unsupervised networks, the desired outcome is not exactly known, therefore the end of the training is mostly decided by the algorithm itself (Parisi et al., 2003).

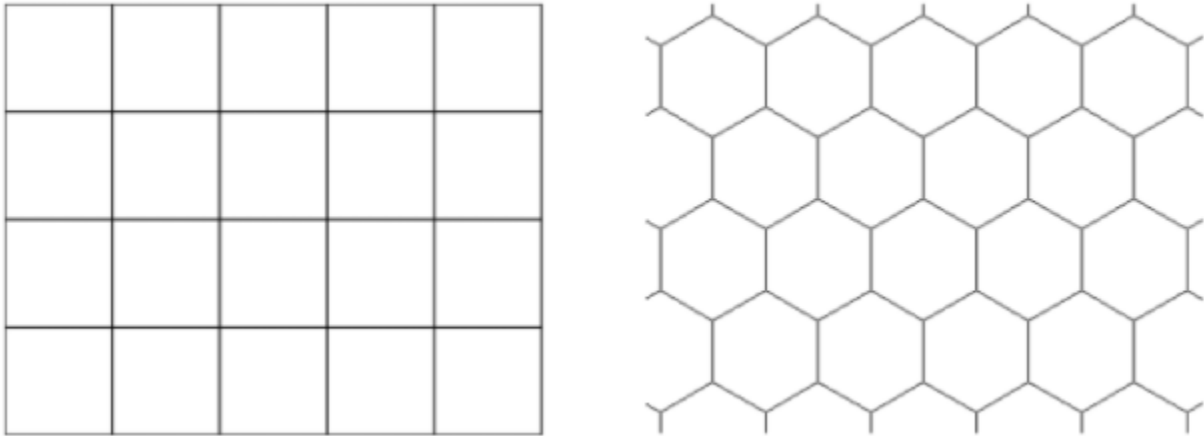
A self-organizing network is a specific type of neural network, which uses a grid of neurons instead of neurons with connections between them. In the next section, the self-organizing network is described.

2.6 Self-organizing maps

A self-organizing map is a software tool for the visualization of high-dimensional data (Kohonen, 1998). It is a neural network tool and can be categorized as an unsupervised machine learning technique (de Smith et al., 2007). In broad terms, the SOM method attempts to find a set of attributes, which represent a large set of input attributes. This means that the method tries to find similar attributes in the input data and groups these attributes in a cluster. The unique aspect of SOM is that the output vectors (the clusters) do have a topological relationship, meaning that they are spatially connected (de Smith et al., 2007).

The starting point of a self-organizing map is an empty grid, this can either be a hexagonal grid or a rectangular grid (see Figure 2.6.1). This empty grid is filled at the end of the processing of the data. Every grid cell is then filled with observations that are similar in their attributes. The principle that is used to determine how the data is organized is based on Tobler's first law of geography, which is the following: 'Everything is related to everything else, but near things are more related than distant things' (Tobler, 1970). For SOM this means that similar values are mapped close together and dissimilar values are mapped further away from each other. It is important to understand that the output map is not a geographic map, but a two-dimensional representation of the similarity of the models applied (de Smith et al., 2007).

Figure 2.6.1: SOM grids (Algobeans, 2017)

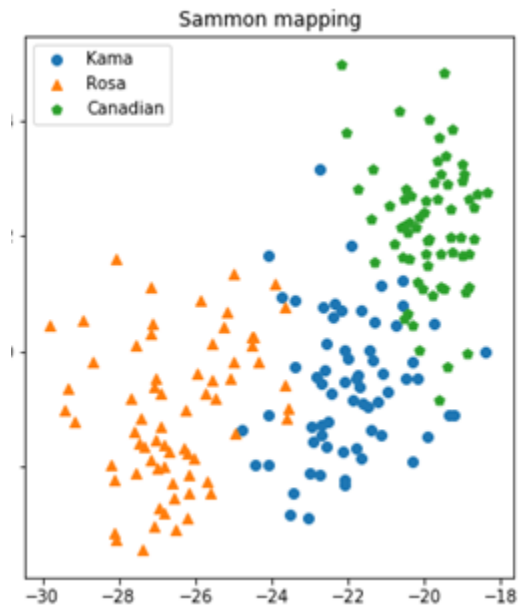


2.6.1 Sammon's projection

An addition to the SOM is Sammon's mapping (Sammon, 1969), which is a useful tool in pattern recognition practice (de Ridder & Duin, 1997). Sammon's projection allows reducing the dimensionality of the data even further. Whereas a SOM is dependent on a grid, Sammon's mapping is not. A SOM uses a grid to map the neurons, which can be visually deceiving. It looks like every cell is evenly similar to its surrounding cells, however, this is not the case. Sammon's projection does not have this issue due to not depending on a grid. Sammon's projection is visualized in a 2d-space (Figure 2.6.1.1). In this space, the distance between neurons visualizes the differences between them (de Ridder & Duin, 1997). In other words, the inter-point distance is preserved.

Sammon's projection is also capable of mapping trajectories, which can be used to evaluate the direction of disease spread (Augustijn & Zurita-Milla, 2013). This can be done by creating Sammon's projection and thereafter connecting the neurons in chronological order.

Figure 2.6.1.1: Example of Sammon mapping (Data farmers, 2019)



3 Data framework

In this chapter, the datasets that can be used in research are explored, compared, and analyzed in multiple dimensions. This is a critical chapter in the study since the studies towards COVID-19 data are quite limited at the time of writing. COVID-19 is still a relatively new phenomenon and therefore research is still in its infancy. Data is the foundation of this research and therefore, a large chapter is dedicated to it.

3.1 Data description

In this section, the datasets suitable for research are described and explored individually.

3.2.1 Confirmed cases

Widely used is the number of confirmed cases, which is described as follows by the WHO: ‘a person with laboratory confirmation of COVID-19 infection’ (World Health Organization, 2020c).

The confirmed cases in the Netherlands are published every day by the RIVM, which is the National Institute for Health and Environment in the Netherlands. This is done daily and contains the number of positive tests which are communicated with the RIVM before 10:00 the same day. The number of positive tests is communicated by instances of the GGD, which are the municipal health services. The GGD conducts positive tests itself, however, there are also commercial tests available. These commercial institutes have to communicate the number of positive tests to the GGD. The results are published per municipality. In Figure 3.2.1.1 a visualization of the municipalities is displayed.

Figure 3.2.1.1: Municipalities in the Netherlands 2021 (Esri Nederland, 2021)



Weaknesses

This methodology is heavily dependent on the number of tests performed. Research (World Health Organization, 2020b) suggests a positive test rate lower than 10% as a general benchmark of adequate testing. When this suggested percentage is compared to the positive rate in the Netherlands, the conclusion can be drawn that there are not enough tests conducted in the Netherlands. Therefore, using this data needs to be done very carefully.

The second weakness of the data is that communication is partially done manually by the person responsible. This can cause the communication to be delayed, which results in an inaccurate number of cases. Due to this issue, the RIVM suggests using the weekly average. This average minimizes the influence of the daily fluctuations.

Strengths

One of the strong points of this data source is the frequency. The number of confirmed cases per municipality is uploaded daily by the government. Due to this frequency, the spread of the coronavirus can be tracked precisely.

Table 3.2.1.1: Properties of confirmed cases

Title	Confirmed cases
Source	https://data.rivm.nl/covid-19/COVID-19_aantallen_gemeente_per_dag.csv
Area	Municipalities
N	352
Start	28-02-2020

3.2.2 Number of tests

The number of tests is not one of the main data sources, however, the positive tests are heavily reliant on the tests. Therefore it is important to analyze the number of tests with the strengths and weaknesses. As stated earlier, multiple instances conduct Covid-19 tests in the Netherlands consisting of the GGD and commercial facilities.

Weaknesses

As stated in the confirmed cases subsection, the GGD communicates the positive tests to the RIVM. However, there are multiple difficulties with the total number of tests reported. The commercial testing instances only communicate the number of positive tests. These instances do not have to communicate the number of tests conducted. This is a weakness of the system since the percentage of positive tests cannot be calculated due to the incompleteness of the data. The RIVM calculates the percentage of positive tests only based on the GGD tests conducted. However, this influences the percentage since only a part of the tests is included.

The second weakness of this dataset is the limited amount of time this is tracked. Although Covid-19 appeared in February 2020 in the Netherlands (Ministerie van Algemene Zaken, 2020), the number of tests conducted are only tracked since the first of June 2020.

Strengths

The number of tests is communicated daily and therefore there is a lot of data available. Moreover, the data is very detailed due to every municipality having its test facilities.

Table 3.2.2.1: Properties of the number of tests

Title	Number of tests
Source	https://data.rivm.nl/covid-19/COVID-19_uitgevoerde_testen.csv
Area	Municipalities
N	352
Start	01-06-2020

3.2.3 Daily Deaths

Research states that the number of deaths is the best way to track the progression of COVID-19 (Institute for Health Metrics and Evaluation, 2021). The GGD publishes the number of deaths daily (RIVM, 2021). This contains any death that is caused by COVID-19 that is reported to the GGD. This means that not only deaths in hospitals are reported, but also the deaths that happen at home for example. For this data source again the RIVM suggests that using the weekly average is better due to some deaths being communicated later than the day they happen. Although the number of deaths is reported often, still some cases are not included in the data. This can happen when the person is not tested for COVID-19. This shows that the reported deaths are also dependent on the number of tests conducted, however not to the same extent as the number of confirmed cases.

Weaknesses

The report of deaths can be delayed, however, this is repaired afterward. This can still cause problems in the short term and can play a role in this research due to the last days not being complete. Moreover, the number of deaths can also be reliant on two other data sources (vaccinations and tests) as described above. However, the vaccination's influence is only limited for this research due to the vaccinations happening after the third wave. The influence of tests is also small, especially when compared to the influence they have on the number of positive cases.

Strengths

The data is published daily, which allows for precise tracking. Due to this, the spread of COVID-19 can be done precisely. Research also states that the data is very robust (Ives & Bozzuto, 2021), which means that the errors in the results are limited.

Table 3.2.3.1: Properties of the number of deaths

Title	Number of deaths
Source	https://data.rivm.nl/covid-19/COVID-19_aantallen_gemeente_per_dag.csv
Area	Municipalities
N	352
Start	27-02-2020

3.2.4 Hospitalized cases

The number of hospital admissions has been used before in studies to evaluate the pressure in hospitals due to COVID-19 (Kashyap et al., 2020). The benefit that hospital admissions have compared to the number of deceased people is that the number is generally bigger. However, there are still 0 values present. It is important to note that the example used is again the municipality of Utrecht, which is one of the largest in the Netherlands. Up Until the first of June only severely ill people were tested, a large portion of these people was taken into the hospital right thereafter. Due to this, the registered hospital admissions were more complete during the first wave. From the first of June, everyone can get tested and this results in earlier diagnosis. The GGD is not always notified or with a delay anymore due to this reason. Since the 6th of October, the RIVM uses the registered hospital admissions from the National Intensive Care Evaluation (NICE).

Strengths

Hospital admissions do not only give a good overview of the spread of COVID-19, but also the pressure on hospitals.

Weaknesses

The way data of hospitalized cases are collected differs over time, which results in small fluctuations. Therefore it might be hard to compare hospitalized cases over time. Moreover, this dataset is delayed similar to the number of deaths, however, the delay is not as long as the number of deaths.

Table 3.2.4.1: Properties of the dataset hospital admissions

Title	Hospital admissions
Source	https://data.rivm.nl/covid-19/COVID-19_ziekenhuisopnames.csv
Area	Municipalities
N	352
Start	27-02-2020

3.2.5 Sewage treatment data

Research suggests that the Covid-19 outbreak can be measured through the analysis of sewage treatment data (Peccia et al., 2020). More than a dozen research groups worldwide have started analyzing wastewater to estimate the total number of infections (Mallapaty, 2020). This method supposedly does not give the same results as the confirmed cases data.

In the Netherlands, the sewage treatment data is collected by taking a sample that contains sewage from 24 hours of a sewage treatment plant. This sample is corrected according to the flow of sewage that comes through. This is done to assure that an increase in sewage does not necessarily lead to more covid particles in the sewage. The data is thereafter corrected for the number of inhabitants per sewage treatment (RIVM, n.d.-a). In Figure 3.2.5.1 a visualization of the locations of sewage treatment plants is displayed.

Figure 3.2.5.1: Locations of sewage treatment plants in the Netherlands, 2018 (Over Morgen, 2018)



Strengths

The wastewater data does not rely on testing, therefore it can be a suitable method for regions where testing facilities are not set up adequately yet or where people are unwilling to get tested. The Netherlands is one of the leading countries on this methodology and is therefore suitable for research. (Mallapaty, 2020) states that research groups in the Netherlands detected traces of SARS-CoV-2 in wastewater at Schiphol airport only four days after the country confirmed its first case of COVID-19 using clinical testing. This indicates that it is a reliable method to gain information on COVID-19 outbreaks.

Weaknesses

The difficulty in using this data is the low number of times that the data is collected. In smaller treatment plants only a few samples have been taken, whereas larger treatment plants have over a hundred samples taken since the beginning of the method. This is due to the method being relatively new and therefore not being implemented as widely as possible. This issue is likely to be influential on the results and therefore needs to be considered when analyzing those results.

Table 3.2.5.1: Properties of sewage treatment data

Title	Sewage treatment data
Source	https://data.rivm.nl/covid-19/COVID-19_rioolwaterdata.csv
Area	Sewage treatment installation
N	326
Start	30-03-2020

3.2 Data evaluation

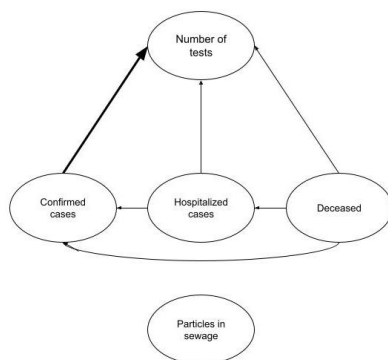
In this chapter, the data is described and analyzed.

Before analyzing the data, it is important to know how the datasets relate to each other. When measuring the spread of COVID-19, multiple datasets can be used. These datasets are all related to each other to some extent. In Figure 3.2.1, the relation between these datasets is displayed. The number of tests is displayed at the top because it is one of the most influential datasets. Confirmed-, deceased, and hospitalized cases are all to some extent related to the number of tests. Confirmed cases are the most heavily reliant on the number of tests. In general, when more tests are conducted, more confirmed cases are found. Hospitalized are also to some extent reliant on the number of tests. When people are hospitalized they are sent to a specific COVID-19 department of the hospital, to be able to determine if a patient is infected with COVID-19, a test must be conducted. Deceased cases are dependent because when people pass away a cause for death is determined, to be able to determine if this is COVID-19 related a test must be conducted.

Confirmed cases, hospitalized cases, and deceased cases are also all related to each other. Before people pass away, they are hospitalized most of the time. Before they are hospitalized in the COVID-19 department, there must be confirmed whether the patient is infected with COVID-19. However, this was not possible at the beginning of the pandemic due to limited testing capacity.

At the bottom, the concentration of RDNA particles is displayed on their own. This is due to this dataset not being reliant on any other data source. This dataset is collected by taking samples of sewage in sewage treatment plants.

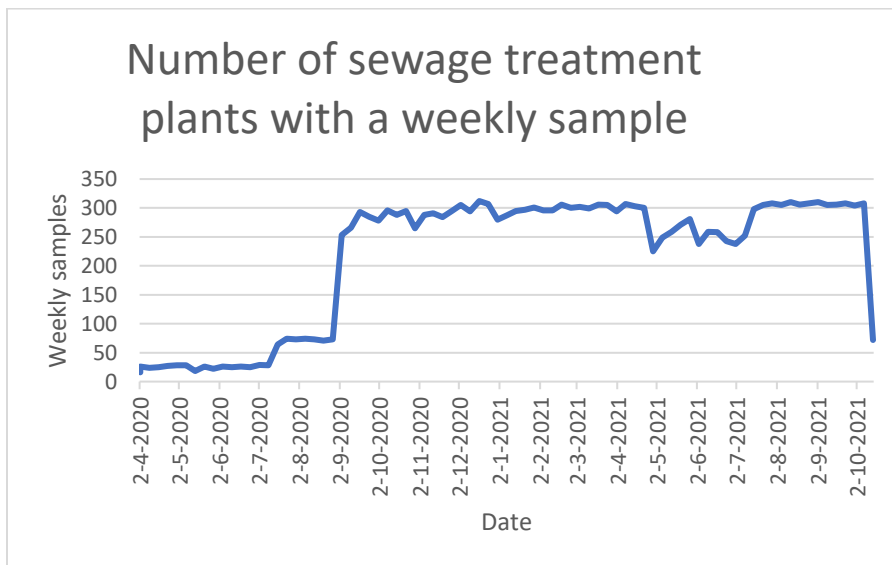
Figure 3.2.1: Relationship between the datasets



3.2.1 Sewage treatment Data

Sewage treatment data can be used to trace COVID-19 particles in the wastewater. However, to be able to do this the data needs to be of adequate level. The collection of this type of data is relatively new, which results in missing values in the dataset. In Figure 3.2.1.1, the number of observations of sewage data can be seen. The figure shows an increase in observations as time progresses. In the beginning, there are as little as roughly 20 observations per week. The majority of these observations are located at the larger treatment plants. From the end of April till mid-July, a remarkably steady drop is observed. The RIVM does not provide a reason for this occurrence. Moreover, when looking at the data, the lack of observations is not present at specific plants. The missing observations are spread equally throughout all of the treatment plants. As an example of the data, a treatment plant in Utrecht is shown in Figure 2. It is important to realize this is one of the bigger plants and therefore it has a lot of observations relative to the smaller plants.

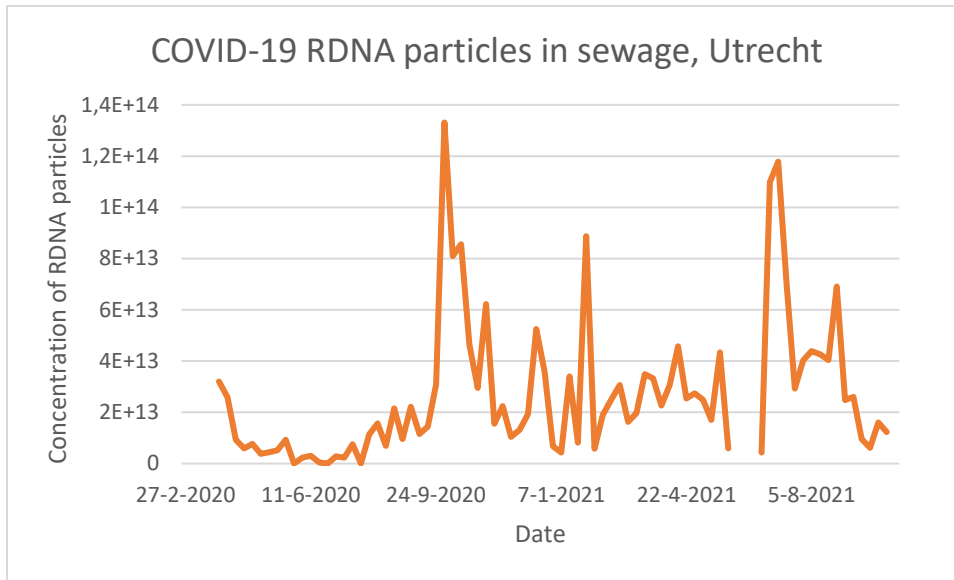
Figure 3.2.1.1: Amount of sewage values collected per week



As can be seen in Figure 3.2.1.2, there are a lot of missing values per week. However, a pattern can still be distinguished from the graph. The pattern is the most important part of this study since a SOM can handle missing values and therefore they do not cause an issue.

The weekly average of the concentration of rDNA particles is the best measure to use since the number of observations varies per week. Using the weekly average reduces the influence of the varying number of observations.

Figure 3.2.1.2: Weekly Average of COVID-19 particles in the sewage treatment plant in Utrecht



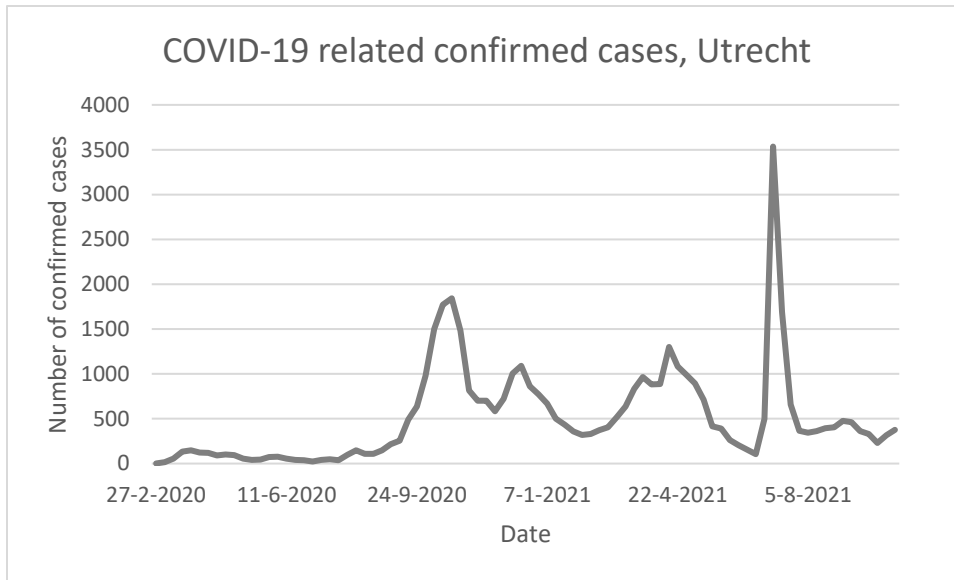
Suitability for research

When the average of the observations per treatment plant per week is used the data can be used for this research. The lack of observations, in the beginning, will influence the results. However, this can be interesting to analyze as well because the smaller plants will have the same patterns at the beginning (missing data). The difference in the way they develop can then be seen in waves 2 and 3. The missing values are likely to have a very limited influence on waves 2 and 3. This data source is important to look into since it is not dependent on human interpretation.

3.2.2 Confirmed cases

The number of confirmed cases is a widely applied data source to track the number of COVID-10 cases. It is one of the most tracked data sources in the Netherlands. However, it is heavily dependent on the number of tests. As can be seen in Figure 3.2.2.1, the number of confirmed cases is very low in the beginning. This is a large difference from the other data sources. This is likely due to the number of tests conducted. To be able to track confirmed cases properly the percentage of positive tests needs to be used. However, there are two issues while trying to calculate this percentage. The number of tests is only tracked from the first of June. The biggest issue is that the number of tests is only tracked per safety region in the Netherlands. This means that the municipalities cannot be used directly as a unit of measurement. This is a disadvantage compared to the other data sources, which can be tracked on a smaller scale.

Figure 3.2.2.1: Number of confirmed cases per week for the Municipality of Utrecht



Although the number of tests is not available per municipality, there is an alternative way to calculate this. This calculation is based on the assumption that the number of tests is correlated to the number of inhabitants per municipality. The first part of the calculation consists of determining how big of a portion of the population each municipality represents within a safety region.

This is done by the following formula:

$$(1) \quad P_m = \frac{I_m}{I_s} * 100$$

Where:

P_m = Portion P of municipality m

I_m = the number of Inhabitants I per municipality m

I_s = the number of Inhabitants I per safety region s

When this is known, the number of tests per municipality can be calculated according to the formula:

$$(2) \quad T_{mw} = \frac{P_m}{100} * T_{sw}$$

Where:

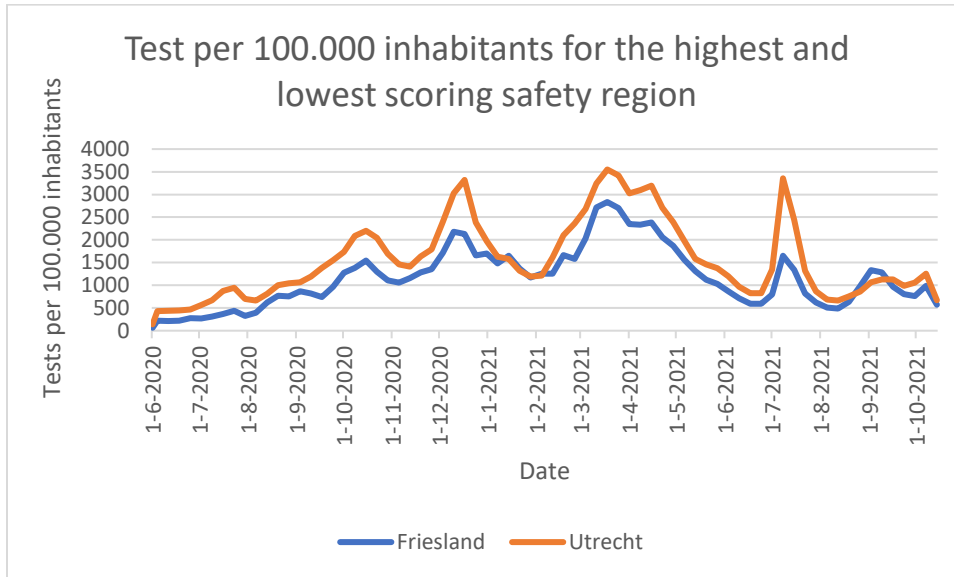
T_{mw} = the tests T in municipality m in week w

P_m = Portion P of municipality m

T_{sw} = Tests in safety region s in week w

The number of tests per safety region is used instead of the total number of tests to calculate the tests per municipality. This is done due to the number of tests varying throughout the safety regions as can be seen in Figure 3.2.2.2. The use of safety regions, therefore, gives a better estimation of the number of tests per municipality.

Figure 3.2.2.2: Tests per 100.000 inhabitants for the highest and lowest scoring safety region, Netherlands



The next step is to calculate the percentage of positive tests per municipality, this can be done by using the tests per municipality and the number of positive tests per municipality. The following formula is used to calculate this:

$$(3) \quad PPTmw = \frac{PTmw}{Tmw} * 100$$

Where:

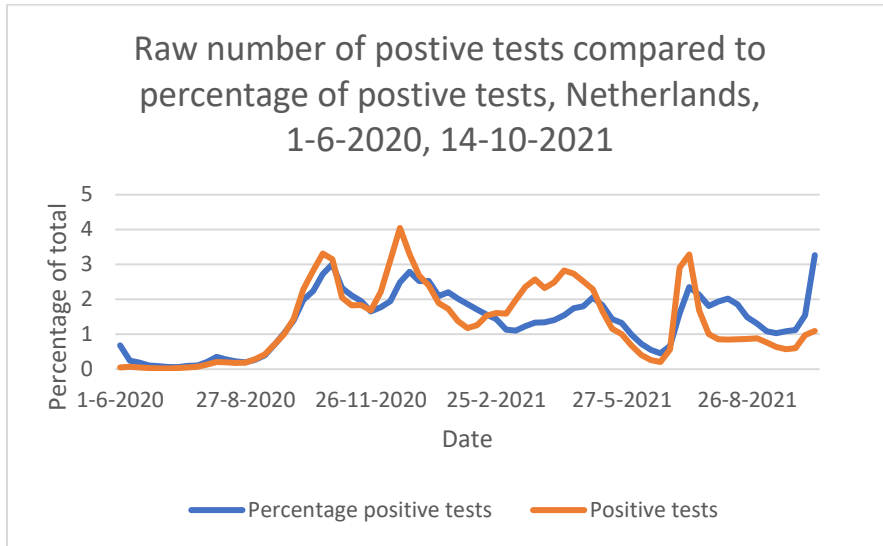
PPTmw = Percentage of Positive Tests PPT in municipality m in week w

PTmw = Positive Tests PT in municipality m in week w

Tmw = Tests T per municipality m in week w

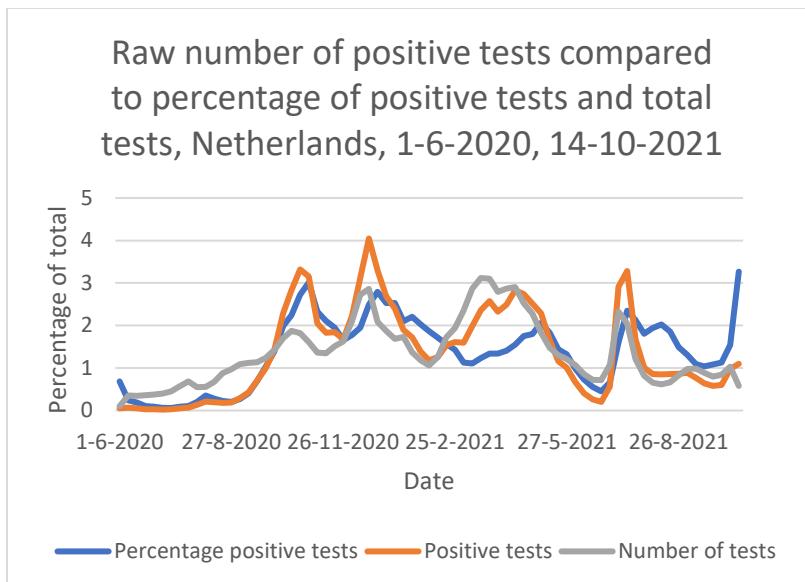
This results in the percentage of positive tests per municipality. However, it is important to realize that this is a calculated percentage based on assumptions and therefore not as reliable as other sources. In Figure 3.2.2.3, the raw number of positive tests is compared to the percentage of positive tests. In this figure can be seen that in general the peaks in the percentage of positive tests are lower than for the raw number.

Figure 3.2.2.3: Raw number of positive tests compared to the percentage of positive tests in the Netherlands



To further analyze the correlation between the number of tests and the number of confirmed cases, both datasets are visualized in Figure 3.2.2.4. In general, the number of tests and the number of positive tests experience the same peaks. However, the percentage of the positive test has different peaks. The assumption can be made that the percentage of positive tests is less reliant on the total number of tests. Therefore it is a better alternative to use than the raw number of positive tests

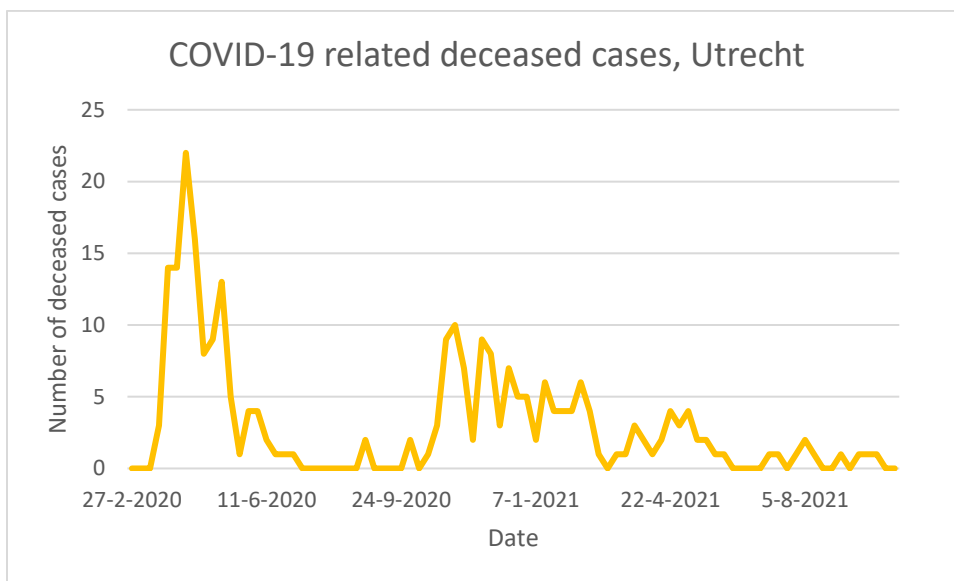
Figure 3.2.2.4: Raw number of positive tests compared to the percentage of positive tests and the total number of tests in the Netherlands



3.2.3 Deceased

The number of deceased people due to COVID-19 has been used before in studies. However, when using municipalities as a research unit, there is a chance that there are a lot of 0 values. In Figure 3.2.3.1, the number of deceased people in the municipality of Utrecht is displayed. As can be seen, there are multiple weeks where there are 0 deaths. This can be an issue as all of the municipalities have the same values in these weeks. Utrecht is one of the largest cities in the Netherlands, which means that smaller municipalities are going to have even more 0 values. This can cause patterns to look very similar and thus it might be hard to draw a proper conclusion. It is also clear that the number of deaths slowly decreases towards the end of the graph, this is likely due to the number of people that are vaccinated. However, this is not possible to prove with data, because the percentage of vaccinated people is not available through time.

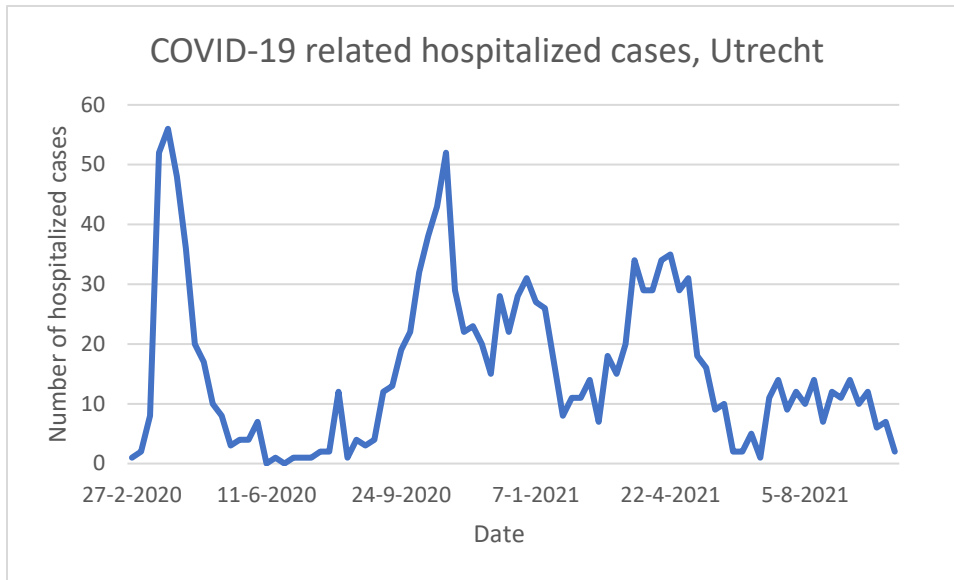
Figure 3.2.3.1: Number of deceased people due to COVID-19 in Utrecht.



3.2.4 Hospital admissions

The number of hospital admissions has been used before in studies to evaluate the pressure in hospitals due to COVID-19. In Figure 3.2.4.1, the number of hospital admissions in the municipality of Utrecht per week is displayed. As can be seen, it follows a sort of similar pattern in the beginning as the deceased graph. However, it looks like the number of vaccinations does not lower the number of hospital admissions. Again, this is not possible to prove with data, because the percentage of vaccinated people is not available through time. The benefit that hospital admissions have compared to the number of deceased people is that the number is generally bigger. However, there are still 0 values present. It is important to note that the example used is again the municipality of Utrecht, which is one of the largest in the Netherlands.

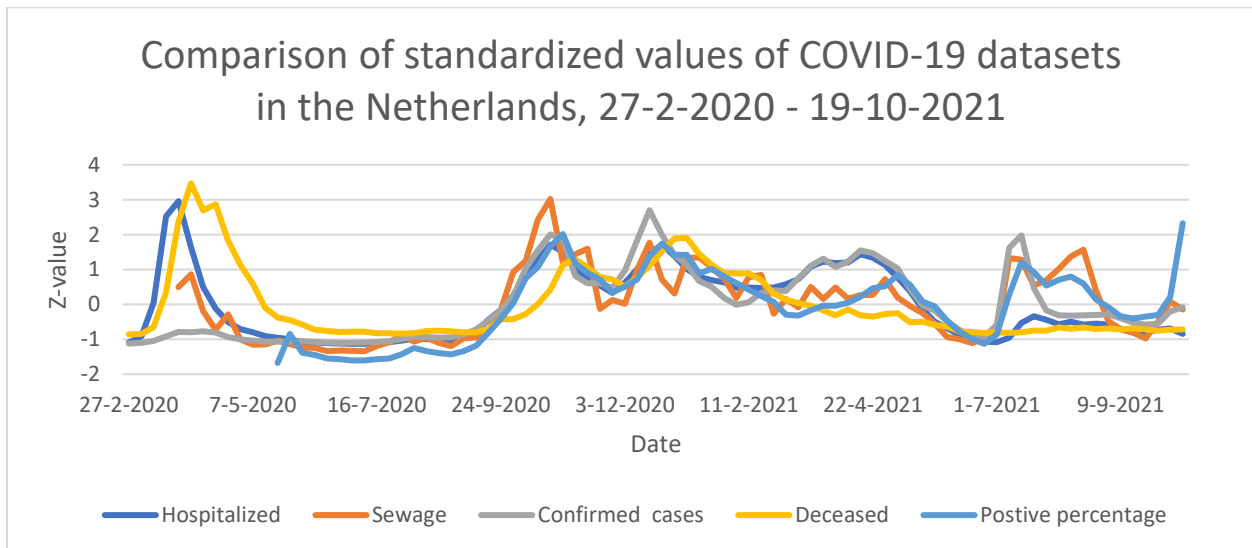
Figure 3.2.4.1: Number of hospital admissions per week for the Municipality of Utrecht.



3.3 Data comparison

After analyzing the datasets independently it is important to compare them to each other. This allows to evaluate the differences and provides insight into what steps need to be taken to be able to compare them properly. In Figure 3.3.1, the different datasets are compared through time. The datasets are standardized based on the Z-value to be able to compare them. It is important to realize that this shows the peaks of the standardized scores. The graph does not show the absolute numbers.

Figure 3.3.1 Comparison of different COVID-19 datasets (RIVM, 2021)



Based on Figure 3.3.1, can be concluded that the peaks of the different datasets do not align properly. This stands out better when the dates of the peaks are presented. In Table 3.3.1, the first three peaks are compared. These are compared because they are minimally influenced by the number of vaccinations. The first vaccination in the Netherlands took place on the 6th of January 2021 (Huisman, 2021), meaning

that from this point onwards the number of vaccinations influences all datasets used. As can be read in Table 3.3.1, the last used date is 6-1-2021 which is the same day as the first vaccination.

Table 3.3.1: National Peaks of different datasets compared, the peak is within the week of the displayed dates.

Dataset	Peak 1	Peak 2	Peak 3
Sewage	2-4-2020, 8-4-2020	22-10-2020, 28-10-2020	17-12-2020, 23-12-2020
Hospital	26-3-2020, 1-4-2020	29-10-2020, 4-11-2020	24-12-2020, 30-12-2020
Deceased	2-4-2020, 8-4-2020	5-11-2020, 11-11-2020	31-12-2020, 6-1-2021
Confirmed cases	2-4-2020, 8-4-2020	22-10-2020, 28-10-2020	17-12-2020, 23-12-2020
Confirmed cases percentage	No data available	29-10-2020, 4-11-2020	31-12-2020, 6-1-2021

In the first wave, it is noticeable that every dataset experiences its peak at the same time except for the hospitalized cases, which has its peak a week earlier. This is a strange occurrence and not in line with the literature. This can be caused by the lack of proper data infrastructure in the first wave. In the second wave, the confirmed cases and RDNA concentration in sewage are the first datasets to experience a peak. These are followed consecutively by hospitalized cases and confirmed cases. In the third peak, the same pattern occurs. Confirmed cases and RDNA concentration in sewage are first, followed by hospitalized cases and confirmed cases. Based on this the conclusion can be drawn that hospitalized cases and deceased cases experience delays compared to other confirmed cases and RDNA concentration in sewage in the second and third peak. These delays can have various reasons.

The reported deceased cases can have a delay due to multiple reasons. The first kind of delay in death counts occurs due to the lag between the data a person dies and the date the authorities take note of the death. This lag is usually a week (Schechtman K, 2021). The second reason for the delay is caused by the reporting capacity of the authorities. When deaths surged, the daily counts were too large for the authorities to process. This resulted in a lag of deaths (Schechtman K, 2021). It is also important to know that people that die in hospitals due to COVID-19 have an average stay of 6-7 days (Faes et al., 2020). Therefore the peak of deceased cases can lag behind hospitalized cases. However, there are a lot of patients that recover from COVID-19 and therefore do not contribute to this lag. Therefore can be concluded that the influence on this delay is expected to be minimal.

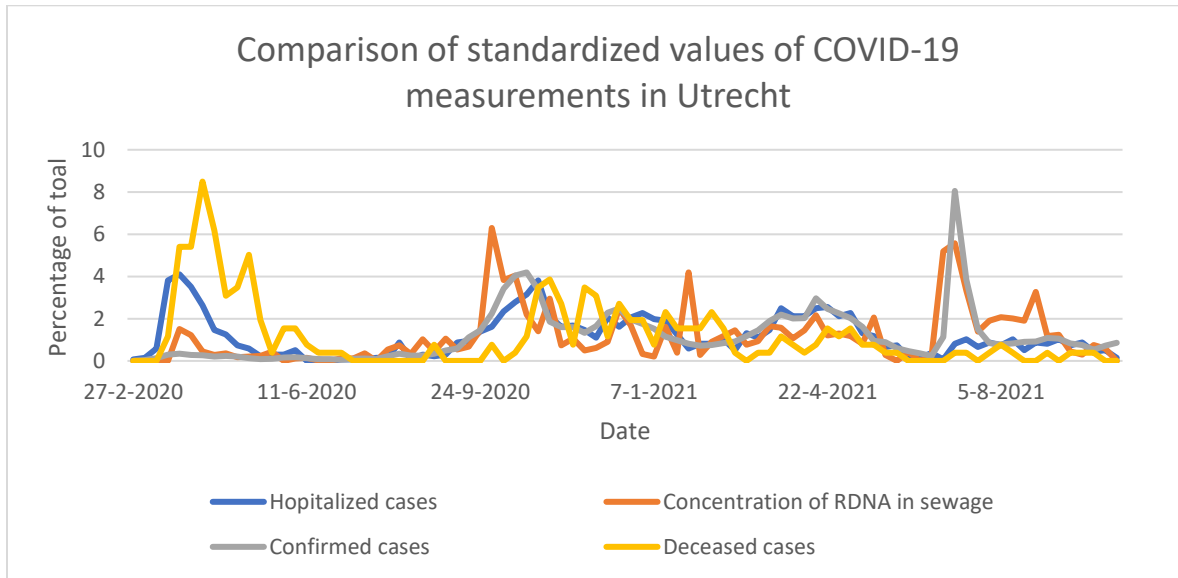
The hospitalized cases experience lag as well. This is quite logical because there is a delay between diagnosis and hospitalization ranges between 3 and 10 days (Faes et al., 2020). The range is large due to it being influenced by various factors such as age, living situation, etc. It is also important to mention that the length of the stay in a hospital also ranges between 3 and 14 days (Faes et al., 2020; Vekaria et al., 2021). This is also influenced by multiple factors such as age and health. The length of the stay has also decreased as time went on and the knowledge on treating COVID-19 expanded.

Studies also suggest that there is a lag between symptom onset and diagnosis (Faes et al., 2020). This means that there should be a difference between RDNA concentration in sewage and confirmed cases. However, it is hard to decide whether or not this is the case based on the current comparison. A more in-depth comparison through time and space can provide better answers.

3.3.1 Comparison of different datasets Utrecht

The development of COVID-19 in Utrecht is different than that of the national levels. As can be seen in Figure 3.3.1.1, the development is less smooth and more importantly, the peaks are different from the national level. Some of the peaks are delayed compared to the national level, this can be caused by COVID-19 being present in other areas first. It is unlikely that most of the municipalities follow the same development as the national level. To be able to study the differences in these developments, SOM is used. SOM is capable of processing high-dimensional data and can simplify those.

Figure 3.3.1.1: Comparison of different COVID-19 datasets (RIVM, 2021)



3.4 COVID-19 in the spatial dimension

3.4.1 Hospitalized cases

Now the data is compared through time and certain differences have been found, it is interesting to investigate whether differences can also be found in space. This can be done by mapping the datasets onto a geographical map. This is done in Figures 3.4.1.1, 3.4.1.2, and 3.4.1.3. In these Figures, the hospitalized cases per municipality per 100.000 inhabitants are mapped for every peak as displayed in table 3.4.1.1, which is displayed here again for clarity purposes. In Figures 3.4.1.1, 3.4.1.2, and 3.4.1.3, the data is displayed in quantiles, meaning that every class is 20% of the total observations. Although there are patterns visible, it is better to compare these maps differently. In Figures 3.4.1.4, 3.4.1.5, and 3.4.1.6. clusters are visualized. In red the high-value clusters are displayed and in blue the low-value clusters are displayed.

Table 3.4.1.1: Peaks of different datasets compared

Dataset	Peak 1	Peak 2	Peak 3
Sewage	2-4-2020, 8-4-2020	22-10-2020, 28-10-2020	17-12-2020, 23-12-2020
Hospital	26-3-2020, 1-4-2020	29-10-2020, 4-11-2020	24-12-2020, 30-12-2020
Deceased	2-4-2020, 8-4-2020	5-11-2020, 11-11-2020	31-12-2020, 6-1-2021
Confirmed cases	2-4-2020, 8-4-2020	22-10-2020, 28-10-2020	17-12-2020, 23-12-2020

Figure 3.4.1.1: The spatial spread of COVID-19 related hospitalized cases in the first peak

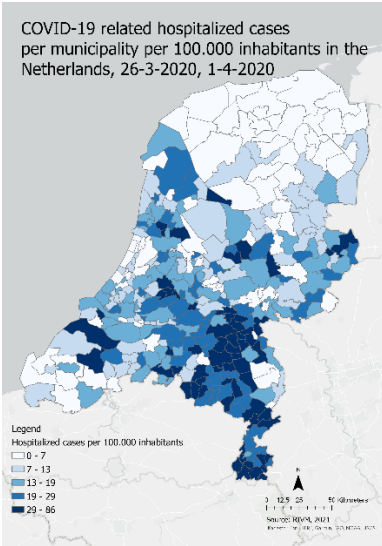


Figure 3.4.1.2: The spatial spread of COVID-19 related hospitalized cases in the second peak

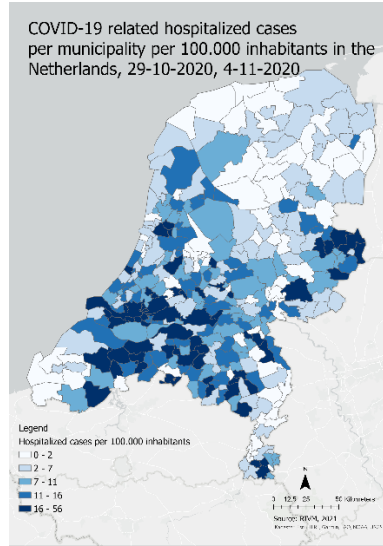
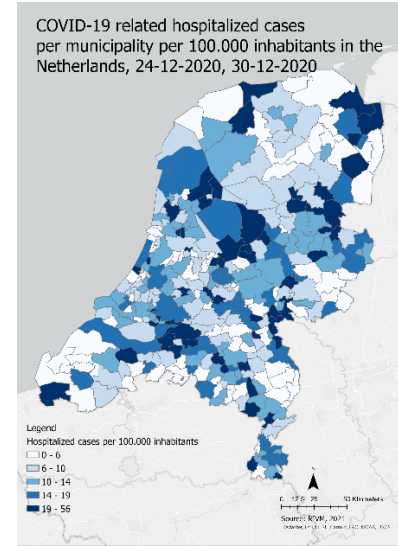


Figure 3.4.1.2: The spatial spread of COVID-19 related hospitalized cases in the third peak



In Figures 3.4.1.4, 3.4.1.5, and 3.4.1.6., remarkably, the three peaks have three different cluster patterns. The first wave is the most outstanding one because it is the only one with a high concentration in the south of the Netherlands. This is likely due to a multiple-day festival called 'Carnaval' which is a tradition in the southern part of the Netherlands (Chen et al., 2020). Moreover, the North-eastern part of the Netherlands seems to be a cold spot throughout the waves.

Figure 3.4.1.4: Hotspots of the spatial spread of COVID-19 related hospitalized cases in the first peak

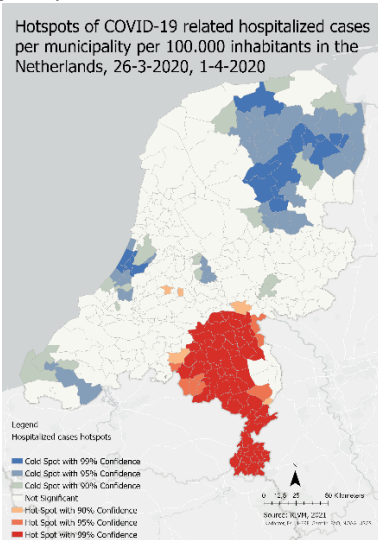


Figure 3.4.1.4: Hotspots of the spatial spread of COVID-19 related hospitalized cases in the second peak

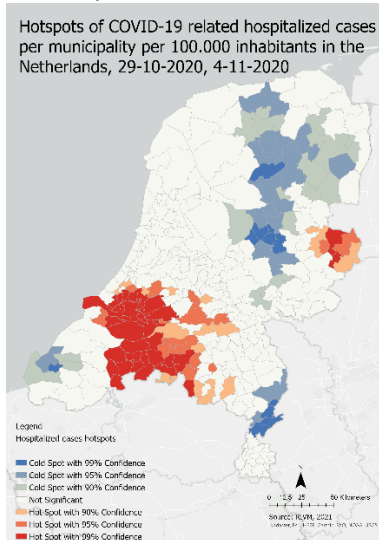
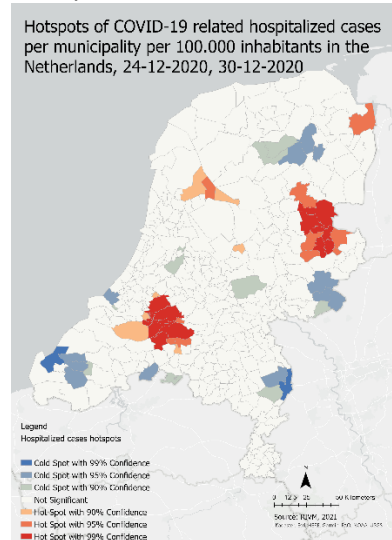


Figure 3.4.1.4: Hotspots of the spatial spread of COVID-19 related hospitalized cases in the third peak



3.4.2 Deceased cases

In this section, the spatial pattern of deceased cases during the three peaks as shown in table 3.4.1.1 are compared. In Figures 3.4.2.1, 3.4.2.2, and 3.4.2.3, the data is displayed in quantiles, meaning that every class is 25% of the total observations. In Figures 3.4.2.4, 3.4.2.5, and 3.4.2.6 the generated hotspots are displayed. It is noticeable that again the south of the Netherlands has a darker color during the first peak and the northern part of the Netherlands is lighter. This pattern shows similarities to the first peak of hospitalizations. However, the second peak still shows that most cases are within the southern part of the Netherlands. This is not in line with the pattern of hospitalizations. It is important to realize that in the second peak there are far fewer cases. The maximum number of cases in a municipality in the second peak is 4, whereas this is 44 in the first peak. A variety of factors can influence this pattern such as the age of the population, the average health, and other health-related factors. Due to this, it is hard to determine the reason behind this discrepancy. In the third peak, the pattern is similar to the peak of hospitalizations. In the middle and the northeastern part, the cases are high. However, when looking at the hotspots the southern part of Limburg is again a hotspot, while this is not the case in hospitalizations. Another aspect that stands out is that the most northern part of the Netherlands, which is the Eemsdelta is dark in the choropleth map and not significant in the hotspots maps. This is due to the Eemsdelta being a large municipality. This means that the largest part of the dark spot is just one municipality. Therefore, there are not many neighboring municipalities with equally high cases. This results in a non-significant spot.

Figure 3.4.2.1: The spatial spread of COVID-19 related deceased cases in the first peak

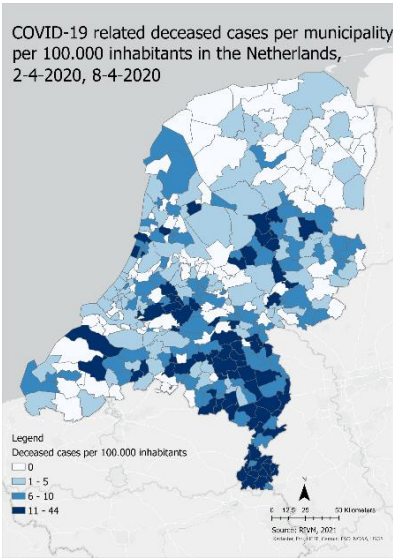


Figure 3.4.2.2: The spatial spread of COVID-19 related deceased cases in the second peak

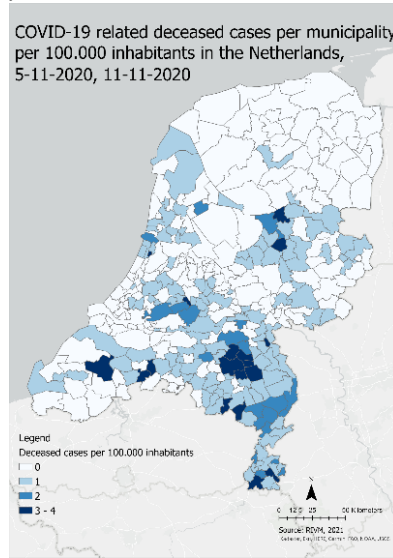


Figure 3.4.2.2: The spatial spread of COVID-19 related deceased cases in the third peak

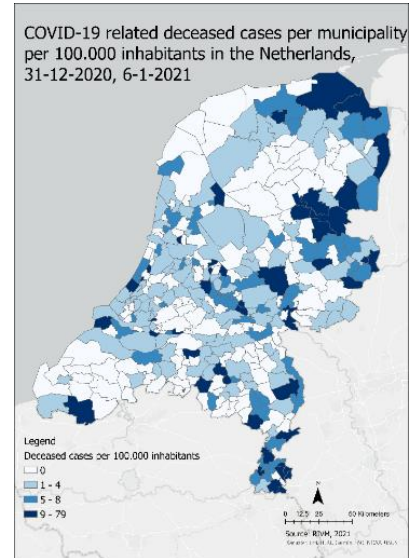


Figure 3.4.2.4: Hotspots of the spatial spread of COVID-19 related deceased cases in the first peak

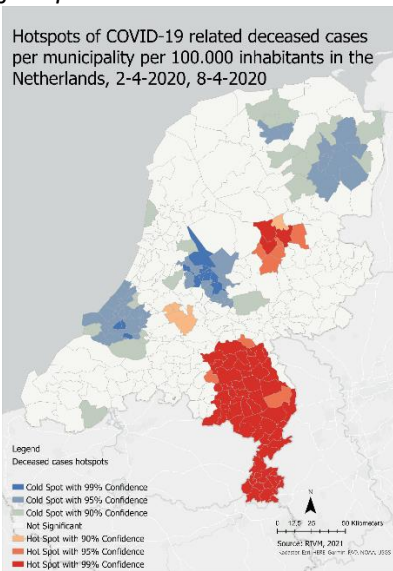


Figure 3.4.2.4: Hotspots of the spatial spread of COVID-19 related deceased cases in the second peak

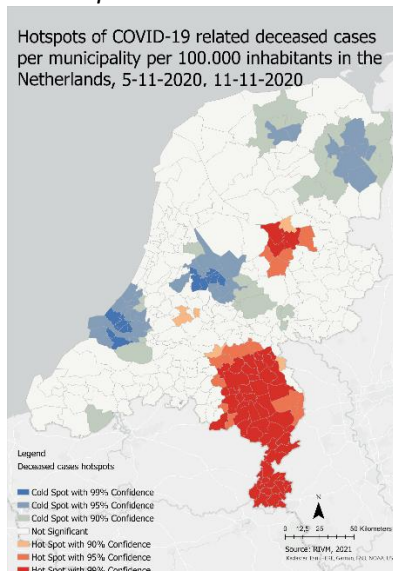
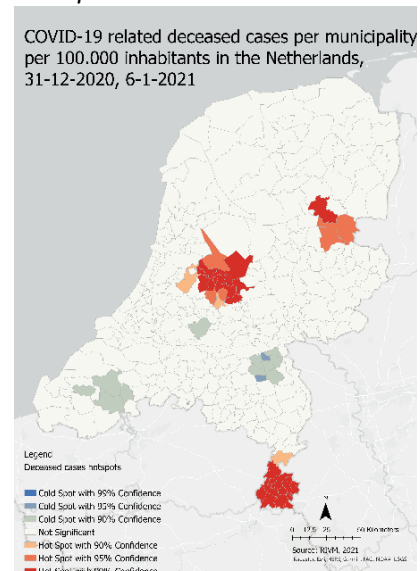


Figure 3.4.2.4: Hotspots of the spatial spread of COVID-19 related deceased cases in the third peak



3.4.3 Positive percentage

In this section, the spatial pattern of the percentage of tests that are positive during the three peaks as shown in table 3.4.1.1 is compared. In Figures 3.4.3.1 and 3.4.3.2, the data is displayed in quantiles, meaning that every class is 20% of the total observations. In Figures 3.4.3.3 and 3.4.3.4, the generated hotspots are displayed. In this section, only two peaks are analyzed, because the required data is not available during the first peaks. When comparing the maps to the other datasets, it stands out that the pattern of the peak at the beginning of November 2020 is similar to the patterns of the hospitalized cases. For both datasets, there is a large presence of cases in the southern part of the 'Randstad' and the eastern part of the Netherlands. However, during the wave at the end of December 2020, the percentage of tests that have positive results looks like a combination of the pattern of deceased and hospitalized cases. The hotspot of cases in the northern part of the 'Randstad' is more similar to the pattern of the deceased cases. However, the other two largest hotspots located in the eastern part of the Netherlands and a bit south of the northern part of the 'Randstad' are more similar to the pattern of deceased cases. Another interesting occurrence is the small hotspot in the northern part of the Netherlands. This hotspot is quite unexpected, due to these municipalities being islands. Therefore, it would make sense that COVID-19 does not spread as well between these municipalities. However, at the end of the year, most people have holidays. This can indicate that the hotspot can be caused by tourists. This is, however, speculation and the real reason can not be retrieved from the data.

Figure 3.4.3.1: Spatial spread of COVID-19 percentage of tests with a positive result in the first peak

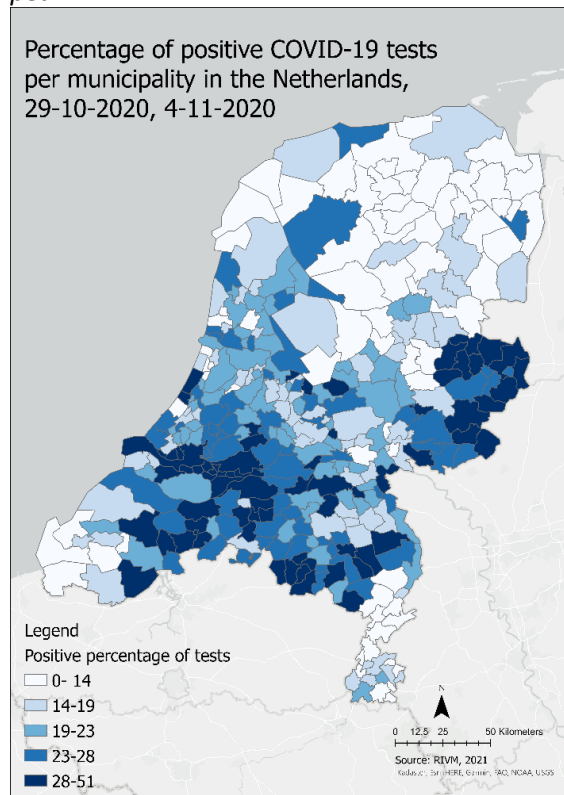


Figure 3.4.3.2: Spatial spread of COVID-19 percentage of tests with a positive result in the second peak

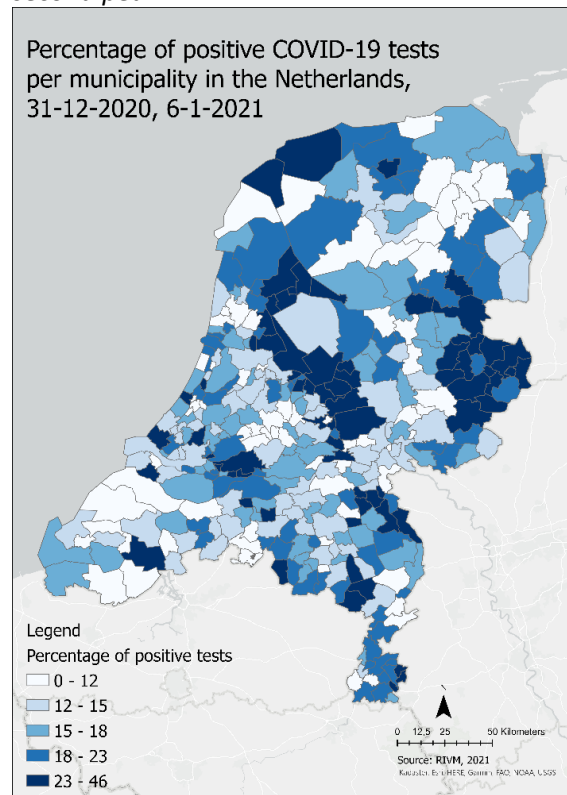


Figure 3.4.3.3: Hot spots of spatial spread of COVID-19 percentage of tests with a positive result in the first peak

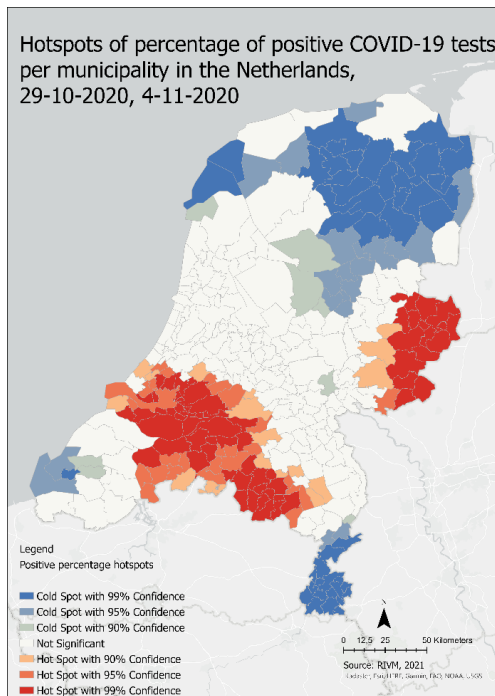
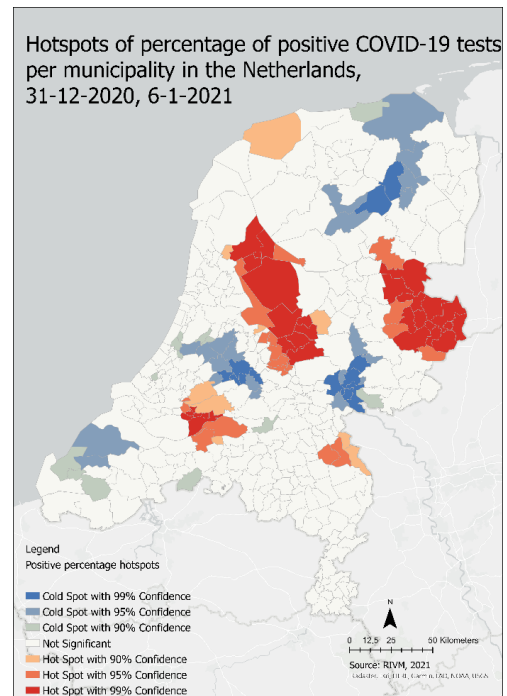


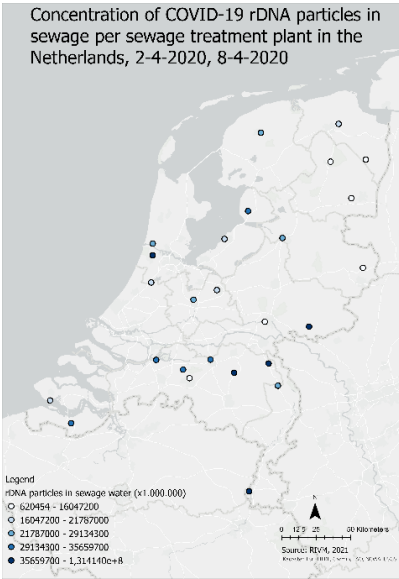
Figure 3.4.3.4: Hot spots of spatial spread of COVID-19 percentage of tests with a positive result in the second peak



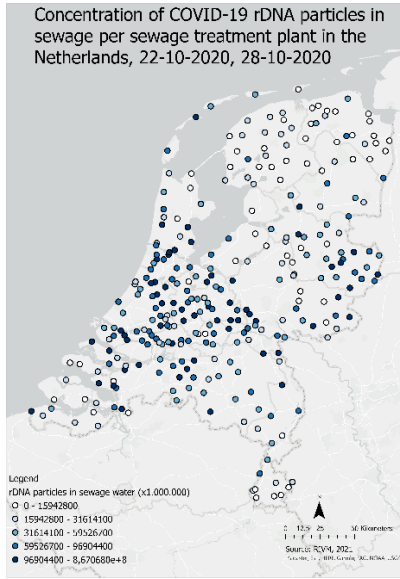
3.4.5 Sewage data

In this section, the spatial pattern of the concentration of COVID-19 rDNA during the three peaks as shown in table 3.4.1.1 are compared. In Figures 3.4.5.1, 3.4.5.2, and 3.4.5.3, the data is displayed in quantiles, meaning that every class is 20% of the total observations. In Figures 3.4.5.4, 3.4.5.5, and 3.4.5.6, the generated hotspots are displayed. The first important point to mention is that during the first peak, the number of treatment plants with data is very limited. Therefore, it is impossible to draw conclusions based on these maps. The other two peaks are quite similar to each other. Both have their main hotspots located in the Randstad. The second peak has another small hotspot in the province of Zeeland. This is quite surprising because the other datasets have not yet shown a hotspot only located in Zeeland. The underlying reason is unknown, but it might have to do with tourism again. Zeeland is a popular province for tourism for Germans. The third wave shows a hotspot in the east of the Netherlands, which corresponds to the findings in the other datasets. Moreover, similar to the percentage of positive tests there is also a small hotspot on the Wadden Eilanden. This dataset is different from the other datasets because it does not require testing or other human interaction. It does still show similar patterns compared to the other datasets. However, this can be further investigated with the result of the SOM.

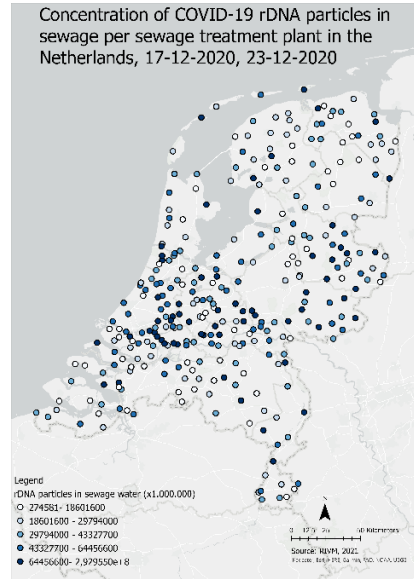
Figures 3.4.5.1: Spatial spread of COVID-19 particles in sewage water in the first peak



Figures 3.4.5.2: Spatial spread of COVID-19 particles in sewage water in the second peak



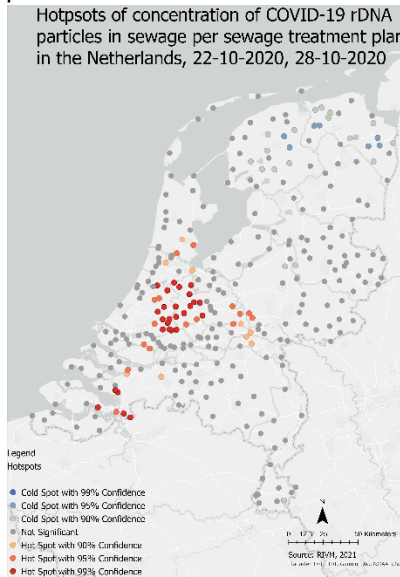
Figures 3.4.5.3: Spatial spread of COVID-19 particles in sewage water in the third peak



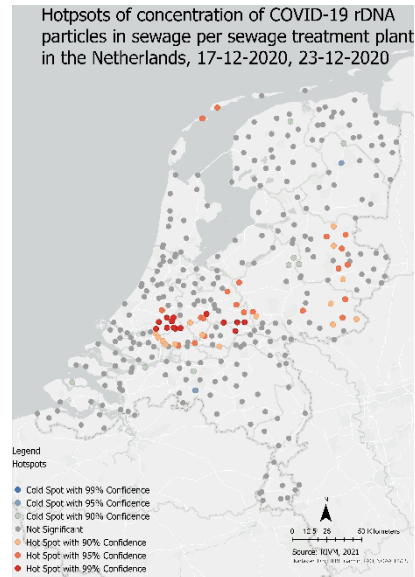
Figures 3.4.5.4: Hot spots of spatial spread of COVID-19 percentage of tests with a positive result in the first peak



Figures 3.4.5.5: Hot spots of spatial spread of COVID-19 percentage of tests with a positive result in the second peak



Figures 3.4.5.6: Hot spots of spatial spread of COVID-19 percentage of tests with a positive result in the third peak



3.5 Data conclusion

To summarize, there is a noticeable difference when comparing the datasets in the temporal dimension. The confirmed cases and sewage data peak earlier than the hospitalized cases and deceased cases. This is in line with the literature, which states that both experience a varying degree of lag. The comparison of the datasets in the spatial dimension is harder because only snapshots of time can be displayed. However, there are still noticeable differences between the datasets. There are still a lot of similarities between them. There seem to be hotspots in the south, the Randstad, and the eastern part of the Netherlands throughout all datasets. The cold spots are primarily located in the northern part of the Netherlands. The SOM can give more insight into whether these first observations are correct. The way this is explored further is described in the next chapter.

4 Methodology

4.1 Time period first sub-question

To be able to determine how the diffusion patterns of COVID-19 can be researched effectively, multiple variables need to be considered. First, it is important to realize what the goal of the research is. In this case, this goal is phrased as *Which spatiotemporal patterns can be found in the spread of COVID-19 in the Netherlands?* Thus, an effort is made to find patterns in certain datasets. These datasets are evaluated and compared in the previous chapter. To effectively find patterns the data must be stable through time. The stability is dependent on data completeness, medical care, and vaccination grade.

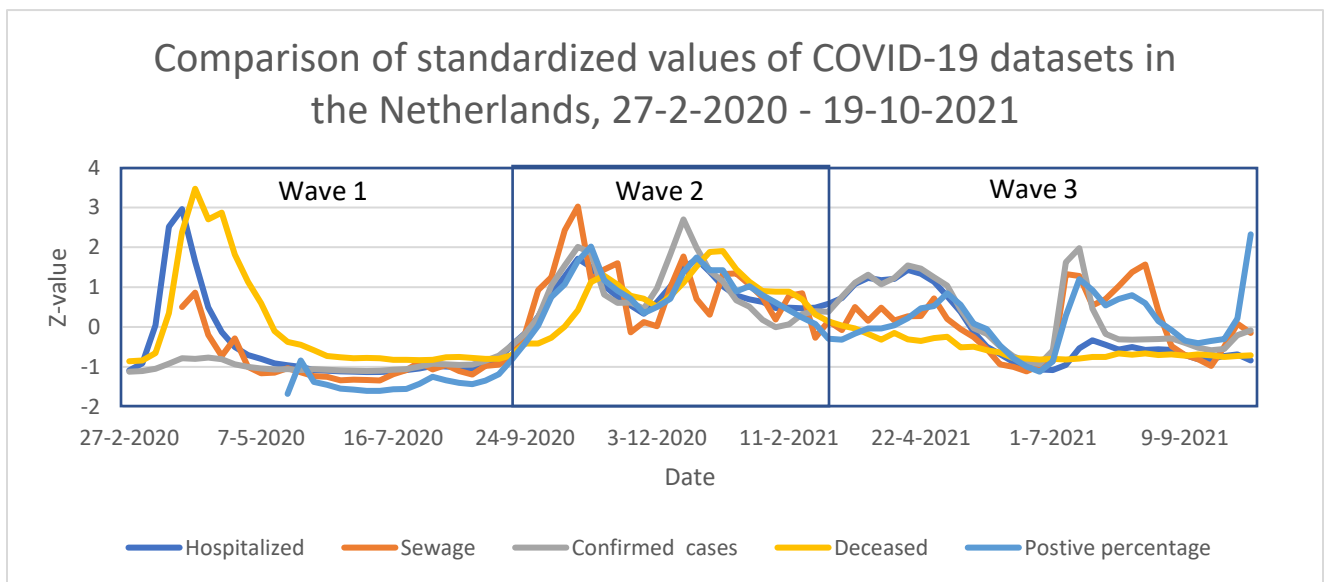
The government has defined multiple waves in the Netherlands (Ministerie van Algemene Zaken, 2020) and this provides a good structure to investigate the periods throughout the pandemic. Unfortunately, they only provide the starting month of the wave. Therefore, the assumption is made that the end of the wave is the start of the next wave. The defined waves are shown in Table 4.1.1:

Table 4.1.1: definition of COVID-19 waves (Ministerie van Algemene Zaken, 2020)

Wave	Start date	Assumed end date	Total duration
1	February 2020	October 2020	8 months
2	October 2020	March 2021	5 months
3	March 2021	To be determined	To be determined

In Figure 4.1.1, the waves are visualized on the peaks, which are used earlier in the research.

Figure 4.1.1: Different datasets and waves visualized



Each wave has its characteristics. Throughout the pandemic, the knowledge on how to deal with COVID-19 has developed. Which affects the mortality rates of COVID-19. Moreover, vaccinations play a big role in the mortality rate. To be able to do research effectively, a period must be established where these numbers are stable through time. Below, the different waves and their influential variables are described.

In the first wave, data incompleteness is an issue. This can be caused by the inexperience of data tracking at the beginning of the pandemic. The major issue here is that the number of tests performed is quite limited and the data of the number of tests before 1-06-2020 is not published. Only in June the capacity for public testing was scaled up and the data was tracked. Moreover, the sampling in sewage water treatment plants was also limited to a small selection treatment plant. Due to these limitations, the beginning of the pandemic is not very suitable for research.

Another characteristic of the first wave is the high mortality. At the beginning of the pandemic, a higher percentage (as high as 48.6% in some countries) of hospitalized cases lead to death in patients (Consuegra et al., 2021). This is due to the inexperience in treating COVID-19 and the absence of the required materials to treat COVID-19 properly.

In the second wave, it stands out that there is a continuous number of COVID-19 cases throughout every dataset. In this wave, the tracking of data is more advanced than in the second wave and the same as in the third wave. Due to the steady stream of COVID-19 cases, this wave is quite interesting to investigate. In Figure 4.1.1, only the cases through time can be seen, which vary quite a lot. When using this time period in a SOM, the spatial dimension can also be explored.

In the third wave, a steep decline can be seen in the deceased- and hospitalized cases. The rising number of vaccinations is likely to be responsible for this trend. This is confirmed when looking at the number of confirmed cases and the concentration of RDNA in sewage, which both experience less of a decline. This is due to vaccination not preventing a person from getting ill, but preventing the occurrence of severe symptoms, thus making the person less ill.

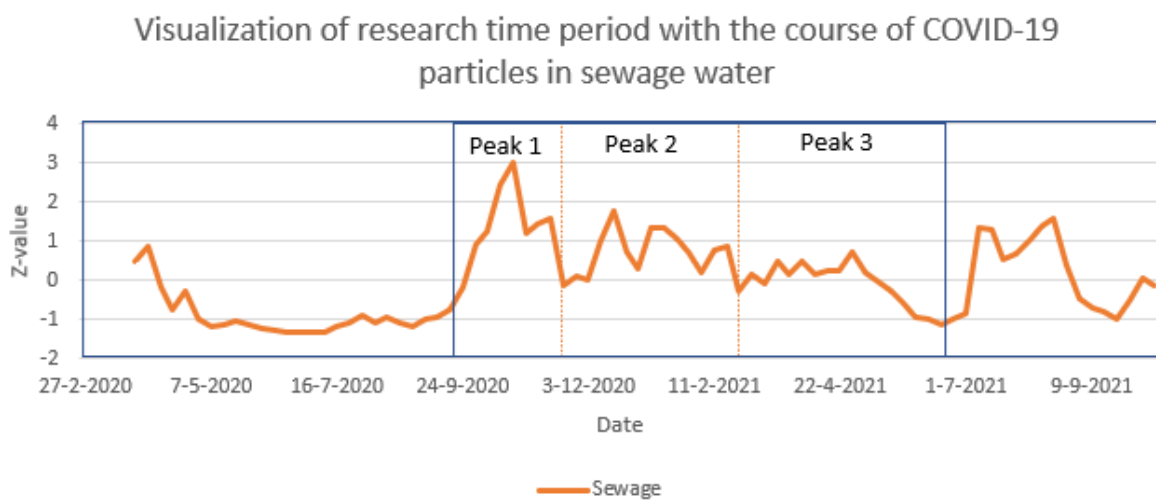
In short, in the first wave issues regarding data completeness are present and in the third wave, the number of vaccinations plays a significant role. The absence of both problems in the second wave makes this wave an excellent period to use for research. This period is interesting to investigate because as can be derived from Figure 4.1.1, there are still some fluctuations in the degree of COVID-19 presence. Therefore, the second wave is chosen as the time period of research. Since only this limited time is chosen as the period of research the goal of the research is more focused on the data side. This short time period makes it difficult to analyze the spread of COVID-19 in general. However, within this period the different datasets can be explored very well. Moreover, the influence of the different datasets on the resulting patterns can be analyzed within this period. Most of the limitations of the datasets do not play a role within this time period and therefore the datasets can be used in their best form. The result of this focus is that finding spatiotemporal patterns in COVID-19 is a secondary goal. However, this research can be used as a basis for later research to further explore COVID-19 patterns.

4.2 Time period second sub-question

The second aim of the research is to find differences in the spread of COVID-19 throughout multiple waves. As described before, the research throughout multiple waves is difficult due to the shortcomings of each respective wave. The first wave lacks data and in the third wave, vaccinations are too big of an influence at first sight. However, vaccinations only play a major role in Hospitalized cases, confirmed cases (and a positive percentage of those), and deceased cases. Vaccinations lower mortality (Gupta et al., 2021), reduce hospital admissions (Rosenberg et al., 2021), but asymptomatic breakthrough infections are still possible when vaccinated (Bergwerk et al., 2021; Birhane et al., 2021). This means that there are still cases of COVID-19 under-vaccinated people, but can remain undetected. This is due to people not testing for COVID-19 when they do not suffer from any symptoms. However, the COVID-19 particles can still be discovered by sewage water samples. This means that the sewage sampling methodology can be suitable to investigate further into the spread of COVID-19 even when the vaccination grade is high.

Thus to investigate the spread past wave 2, only sewage water samples can be used. For this sub-question, the start is at the same time as in the first sub-question, which is the end of September 2020. This is because the sewage data is not present during the majority of the first wave. Moreover, the last part of the first wave is not interesting to look at, since the infections are low compared to the rest of the time. The beginning of the official second wave marks the start of an interesting time period to investigate. This time the end of the wave has been moved to the beginning of July, this can be seen in Figure 4.2.1. The reason behind this move is that the infections only really hit zero at the beginning of July. This movement provides a longer period of research, which is needed when trying to compare multiple time periods. The orange lines indicate different phases within this period. When looking carefully it can be observed that the lines have been drawn where infections are relatively low compared to the weeks around it. The periods between these orange lines can be seen as small waves within a larger wave. These smaller waves can be compared to evaluate if the spread of COVID-19 occurred in the same pattern throughout the Netherlands in space and time.

Figure 4.2.1: Different waves visualized



4.3 Self-Organizing Maps

The data has been analyzed in great detail, however remains useless if not utilized properly. To use the data effectively SOM is used. In the following section, the process of generating results from the data is described.

4.3.1 Input data

The organization of data influences the structure of the output of a SOM. Therefore, it is important to organize data in such a way that the required output is created. To obtain the goals of this study, the data needs to be organized in two different ways. These are spatial- and temporal organization, which is in line with the goal of the research, which is to identify spatiotemporal patterns. Andrienko et al. (Andrienko et al., 2010) propose to structure data dually:

- A temporally ordered sequence of spatial situations (Space over time)
- A set of spatially arranged places where each place is characterized by its particular temporal variation of attribute values (Time over space)

Those two ways result in two different ways of analyzing spatiotemporal data:

- Analyze the change of the spatial situation over time
- Analyze the distribution of the local temporal variations over space.

This translates into structuring the data according to the space over time structure as shown in table 4.3.1.1

Table 4.3.1.1: Space over time structure

Space/Time	Timestep 1	Timestep 2	Timestep 3
Municipality A			
Municipality B			
Municipality C			

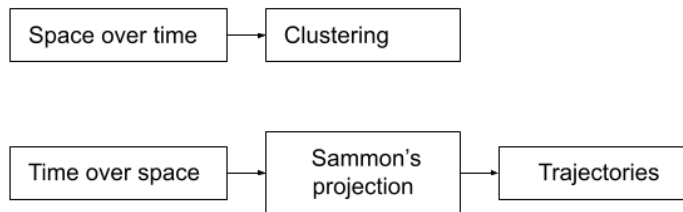
And for the time over space structure, the data structure as shown in table 4.3.1.2 is used.

Table 4.3.1.2: Time over space structure

Time/Space	Municipality A	Municipality B	Municipality C
Timestep 1			
Timestep 2			
Timestep 3			

Both data structures are needed, as they can be used to generate different sorts of results. The space over time structure is used to detect the municipalities that show similar behavior over time. The time over space structure is used to study the diffusion of COVID-19 over time. The diffusion needs one extra step in the process. To map the diffusion, a Sammon projection is used. As an overview in Figure 4.3.1.1 the data structuring and the use of the data is shown.

Figure 4.3.1.1: Overview of data structures and their use



4.3.2 Training

It is important to establish the way a SOM is trained because the result is dependent on the training. Thus, the training needs to be done with the goal of the study in mind. For the first sub-question, this goal is to compare different datasets both in space and time. For the second sub-question, the goal is to compare different waves for the sewage treatment data.

For sub-question one, the SOM needs to be trained on all datasets simultaneously to be able to compare them all to each other. To achieve the training of all the datasets simultaneously, a combination of all the data into one file must be constructed.

For the second sub-question, the SOM is only trained on the sewage treatment data. Although the different waves are compared, this is not relevant for the training of the SOM. The separation into waves only applies to the mapping part of the SOM. The mapping of the SOM is the next step and is described in the next section.

4.3.3 Mapping

To visualize the results several different approaches can be used, these are a U-matrix (counts plot), SOM lattice, codebook vector (codes plot), Sammon's trajectories, and GIS maps. A codebook vector is a list of numbers that have the same input and output attributes as your training data (Brownlee, 2016). In this case, they are generalizations of the input data and consist of the neurons in a lattice (Augustijn-Beckers, 2018). A U-matrix (unified distance matrix) visualizes the distance between the codebook vectors of neighboring neurons (Utsch, 1990). In this research, the U-matrix is also used to visualize the counts within a neuron. In this case, it is referred to as a counts plot. Sammon's projection is an algorithm that constructs a trajectory that allows for the tracking of the diffusion of a phenomenon, in this case, COVID-19. The GIS maps are used as the final tool. With the use of GIS, the results of a SOM can be displayed on a geographic map, thus the spatial pattern can be further explored with this use of GIS. In Figures 4.3.3.1 and 4.3.3.2, the process of the SOM and the corresponding visualization methodologies are displayed. The main difference between the two is the input data. The process that follows is more or less the same. The difference is in comparing the datasets in sub-question 1 and the comparison of waves in sub-question 2.

Figure 4.3.3.1: Process of the result generation for sub-question 1

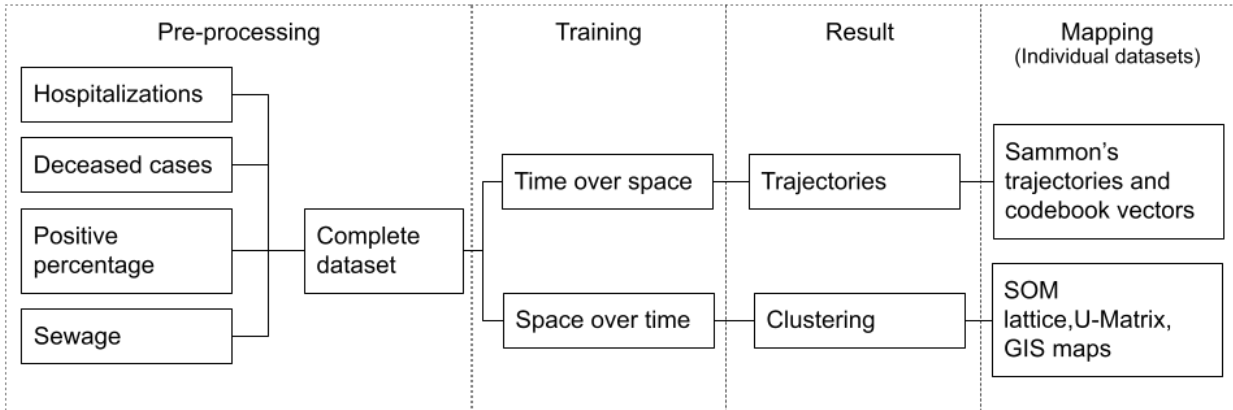
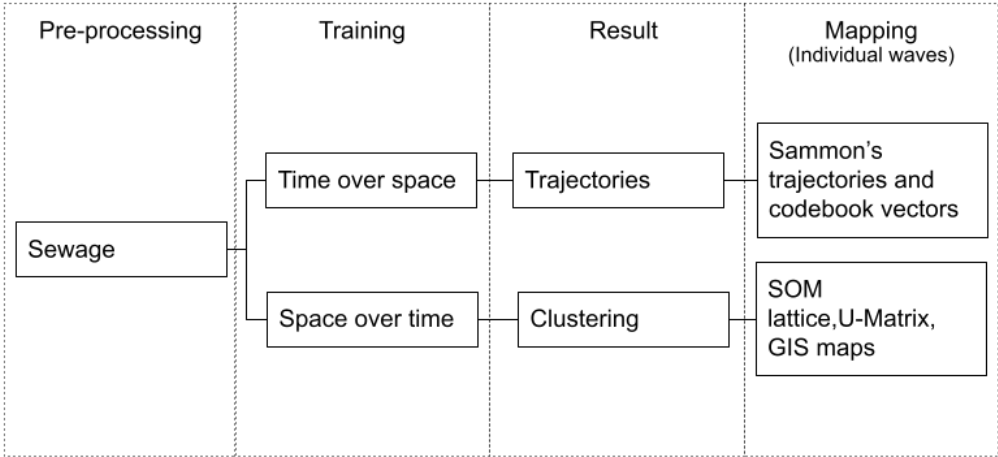


Figure 4.3.3.2: Process of the result generation for sub-question 2



4.4 Preparation of Sewage Treatment locations

To be able to generate the results of the SOM, all datasets must be comparable. Since the sewage treatment data is collected per treatment plant it is not suitable for comparison with the data per municipality. Therefore, the data must be edited in such a way that it can be compared to the municipality data. Every sewage treatment plant has its service area. However, the locations of these service areas are not available. Therefore, it is harder to determine which municipality is with the service area of a sewage treatment plant. Thus, this requires extra analysis to construct data in a way that sewage treatment plants are joined with municipalities. However, the issue here is that not every municipality has its own single treatment plant. There are areas where there are fewer sewage treatment plants than municipalities and areas where the contrary is true. Examples of these situations are displayed in Figures 4.4.1 and 4.4.2. These situations occur everywhere in the Netherlands to varying extents.

To solve the issue, every municipality is assigned the three closest sewage treatment plants. The value of the COVID-19 particles in the sewage per week is calculated based on the average of those three sewage treatment plants. This construct undoubtedly influences the results and therefore is something to keep in mind during the processing of the results. To be able to use the data of the sewage treatment plants in their original form, the second sub-question is solely dedicated to the use of this dataset. This provides insight

Figure 4.4.1: The distribution of sewage treatment plants in the region of Eindhoven (Watersector, 2021; Esri, 2021).

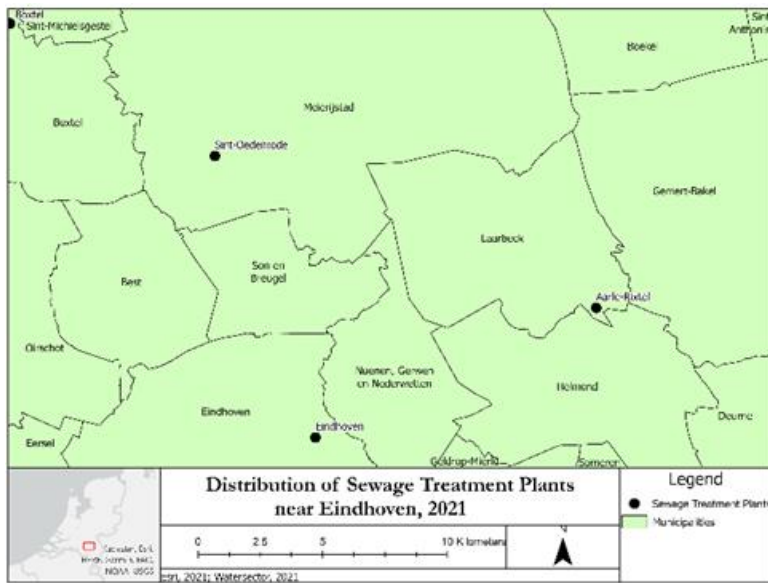
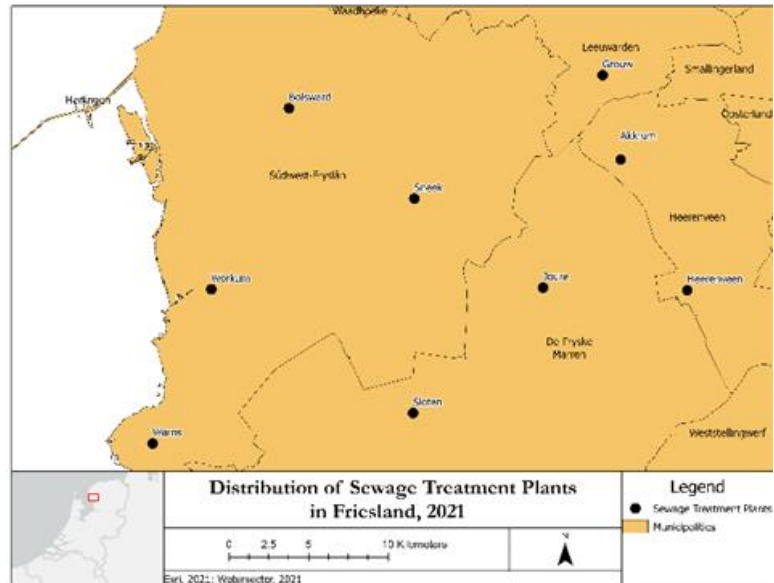


Figure 4.4.2: The distribution of sewage treatment plants in the regions of Eindhoven (Watersector, 2021; Esri, 2021).



4.5 Visualization of results

The results of a SOM can be presented with the use of multiple so-called plots. In the sections below the use of these plots is described.

Counts plot

The first result that is plotted is the counts plot. This can be used to analyze how municipalities are spread over the neurons. Based on this plot, the decision can be made whether the SOM algorithm was run with the right parameters. This is the case when almost all of the neurons have several municipalities assigned. This plot can also be used in combination with the plots described later to analyze how common certain graphs in the codes plot are.

Codes plot

The codes plot can be used to visualize the graph of the course of the presence of COVID-19 through time. After the training, the plot should be organized in a way that similar graphs are grouped. This plot can also be used to draw the border of the clusters based on the hierarchical clustering. However, this can be confusing therefore a clusters plot can be used to improve the clustering visualization.

Clusters plot

The clusters plot provides a visualization of the hierarchical clustering. This plot is needed because just drawing the borders of the clusters over a plot can be confusing. When the clustering function finds clusters that are located at the edge of the grid, it is impossible to tell whether they belong to the same cluster with just the cluster borders. The clusters plot gives every cluster a unique color, which makes it much easier to understand. Furthermore, it is possible to also to visualize the number of values within the neurons on top of the cluster plot. This provides an overview of the size of a cluster.

Mapping plot

The mapping plot is relevant when comparing different datasets with each other. Through this plot, it is possible to see where the values of the datasets are assigned. It is similar to a counts plot because it also visualizes the number of values per neuron. However, with the mapping plot, it is easier to visualize the differences between the different datasets. This is because the color can be changed for each dataset accordingly.

Sammon's mapping & trajectories

Sammon's mapping can be used to visualize the similarity between neurons based on distance. The closer neurons are mapped together, the more similar the neurons are.

Sammon's projection can be used to map the diffusion trajectories of a time series (Augustijn-Beckers, 2018).

4.6 Comparison of SOM results

Besides comparing the SOM results visually, they can be compared with concrete numbers. This can be done by using the figure of merit (Augustijn & Zurita-Milla, 2013). A figure of merit is a quantity used to characterize the performance of a method relative to its alternatives (Olivieri & Escandar, 2014). Thus in this study, the figure of merit can be used to compare different datasets and waves to each other. Research (Augustijn & Zurita-Milla, 2013) has shown that it is possible to measure the extent to which the clustering for the subsamples corresponds to the clustering of other subsamples.

The datasets can be compared by calculating the number of municipalities that are in the same neurons or clusters between datasets. This can be done using the following formula:

$$M(V) = \frac{mn}{tm}$$

In which:

M(V) = The figure of merit (M) for datasets ran with the same parameters (V)

Mn = municipalities within the same neuron

Tm = total number of municipalities

This formula is a simplified version used in other research (Levine & Domany, 2001) to make it more suitable for this research.

For clarity purposes, an example of the resulting matrices is given in table 4.6.1.

Table 4.6.1: Crosstab of the figure of merit to compare datasets

Trained /Mapped	Deceased	Hospitalized	Percentage confirmed	Sewage
Deceased	1			
Hospitalized		1		
Percentage confirmed			1	
Sewage				1

The comparison of waves for the second sub-question is the same as the comparison of datasets. The figure of merit, in this case, tells how different the waves progress through time and space compared to each other. The results can then be displayed in a cross table as presented in table 4.2.2.

Table 4.2.2: Crosstab of the figure of merit to compare waves

Trained/Mapped	Wave 1	Wave 2	Wave 3
Wave 1	1		
Wave 2		1	
Wave 3			1

4.7 Software

The software used in this research is 'R', which is a free software environment for statistical computing and graphics (R project, n.d.). This software is suitable for organizing maps as there are multiple so-called packages available. The most important one is the 'Kohonen' package which refers to the creator of self-organizing maps. Moreover, previous research (Augustijn & Zurita-Milla, 2013) has already been done on self-organizing maps and disease diffusion, and the R code behind that research is available and can therefore be used as a guideline. The output of the self-organizing map is however not a geographic map. To visualize the output of the SOM in space other software is required. ArcGIS Pro is used to do this since it is one of the most used GIS software across universities. Moreover, it is suitable for mapping data in space.

5 Results

In this chapter, the results of the research are described and analyzed. Firstly, the datasets are analyzed individually to detect clusters. After this, the findings of the individual results are compared to the findings in the hotspot maps earlier in the research. This is done to evaluate the added value of the SOM. Secondly, the datasets that do not have a delayed method of tracking are compared to each other by mapping them back on the same SOM. These datasets are the percentage of positive tests, positive percentage of the population, and virus particles in sewage. Lastly, all datasets are compared by mapping them back on the same SOM. These datasets consist of the three named above with the addition of hospitalized cases and deceased cases.

5.1 Explanation of parameters

Before results can be generated, it is important to determine the parameters that need to be set. In the case of a SOM, these parameters are the grid size, the number of iterations, and the number of clusters in the case of clustering.

Grid size

A simple formula can be used to determine the right size for the grid. The proposed methodology (Tian et al., 2014) is shown below.

$$M = 5\sqrt{N} \quad (1)$$

Where:

M = Number of neurons

N = Number of municipalities

This results in the following formula being applied to the datasets.

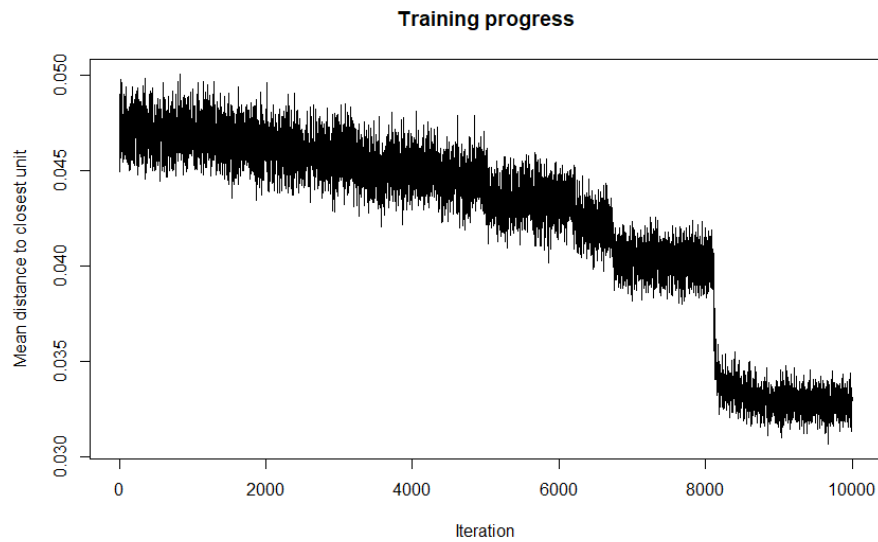
$$93,8 = 5\sqrt{352}$$

The grid size closest to 94 neurons is a 10x10 grid size. This grid size can be reduced if there are neurons that remain empty. Thereafter it is checked whether there are empty neurons within the grid. If this is the case, the grid is generated again with a smaller size until there are only neurons with at least one value. The grid is also referred to as the map of neurons in this research.

Number of iterations

The number of iterations of the SOM can be determined by checking whether the training progress changes over time. The training progress becomes steady and will no longer change after more iterations. At this point, the training can be stopped (Kohonen, 2013). The training progress of the SOM of all the datasets looks like Figure 1. Therefore can be concluded that 10.000 iterations are enough since the training progress remains steady after 8000 iterations.

Figure 5.1.1: Example of training progress of the SOM networks



Number of clusters

The number of clusters for the datasets is determined through the elbow method. The Elbow method is a method that looks at the percentage of variance explained as a function of the number of clusters. The optimal number of clusters is reached when adding another cluster does not give much better modeling of the data. The percentage of variance explained by the clusters is plotted against the clusters. The first clusters will add a lot of information and therefore give a relatively big drop in the graph. At some point, the gain will drop dramatically and give an angle in the graph. The ideal number of clusters is chosen at this point. This point vaguely resembles an elbow, which explains the name of the method (Bholowalia & Kumar, 2014).

5.2 Individual Results

In this section, the results of the individually plotted SOMs are presented and discussed. The goal of this is to explore the datasets individually. The goal is not to compare the datasets, because this is not possible when they are trained on separate training data. Later in the research, the datasets are compared.

Hospitalized cases

For this dataset, a 9x9 size map of neuron is used, because it is the largest map of neuron (with a maximum of 10x10) where there are no neurons with zero municipalities in them (See Figure 5.2.1). However, when using a 10x10 size map of neuron size there were some neurons with zero municipalities. This resulted in the need to make the map of neuron smaller. To decide how many clusters should be assigned, the elbow method is used. This results in an ideal number of clusters that is 3, as can be seen in Figure 5.2.2. However, due to unsatisfying results with the use of 3 clusters, 8 clusters are used. In Figure 5.2.2 can be seen that there is a slight increase up until 8 clusters. Therefore, 8 clusters is a good alternative.

Figure 5.2.1: The number of municipalities within a neuron

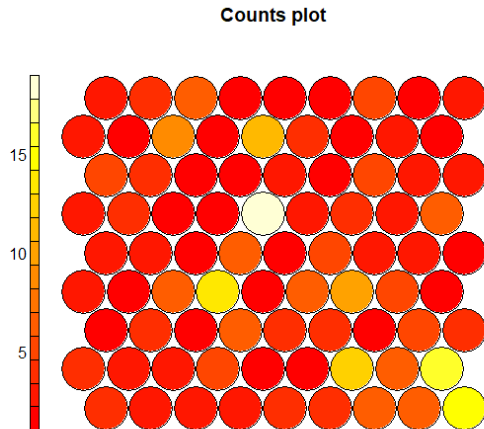
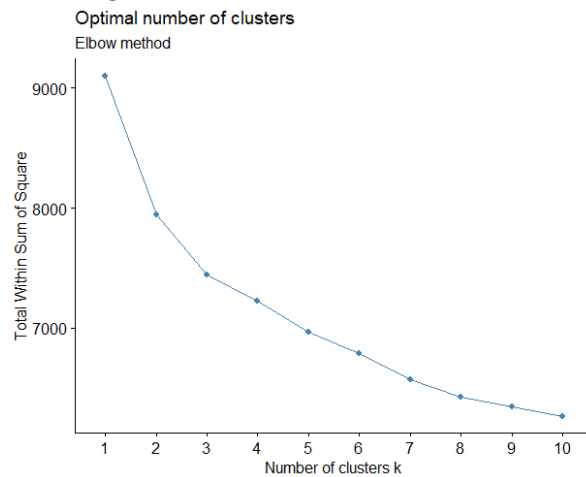


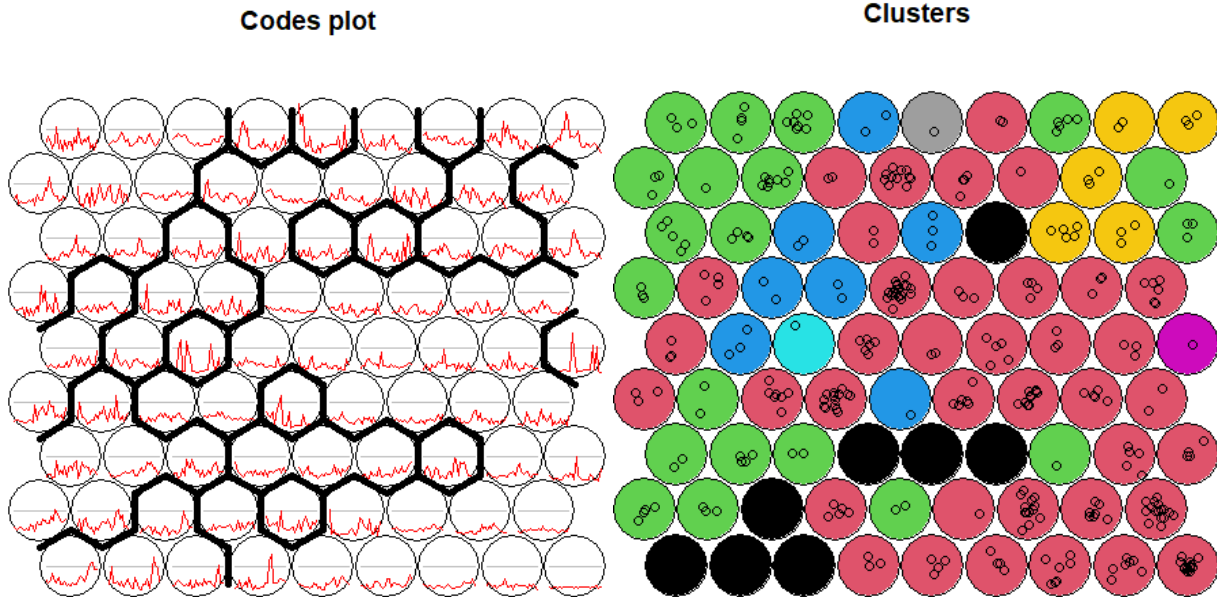
Figure 5.2.2: the ideal number of clusters according to the elbow method



These clusters are then visualized on the codes of the SOM (see Figure 5.2.3). Because it can be unclear which neurons belong to a certain cluster, in Figure 4 the clusters are visualized using colors. The pattern that can be seen in the codebook vector is that in the bottom left corner the biggest peak is at the end of the timespan. In the top left corner, the peaks are still towards the end but appear earlier than in the bottom left. In the bottom right corner, there is are low values through time without distinct peaks.

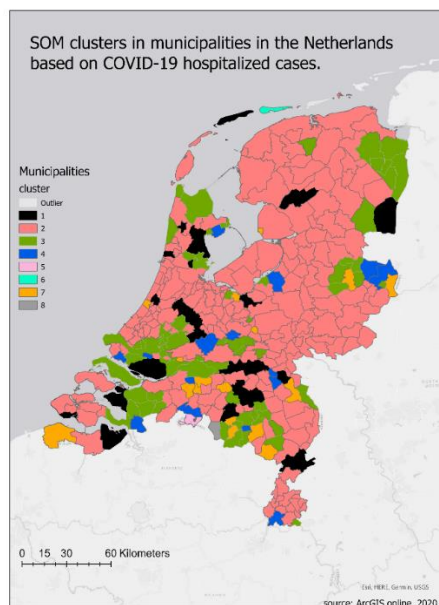
Figure 5.2.3: Codes plot with hierarchical cluster boundaries

Figure 5.2.4: Color visualization of the location of clusters



In the map in Figure 5.2.5, the spatial pattern of the clusters is shown. Cluster 3 is spread out over the entire country but shows clustering where it occurs. Most of the municipalities in this cluster are medium-sized cities and not the most important city in the region. The exception here is Rotterdam, which is a large city within the same cluster. In general, in this cluster, the codebook vectors (Figure 5.2.3) show continuous medium values without any high peaks. This pattern can be an indication that the SOM works well for medium to larger-sized cities. All of the other clusters do not show any distinct pattern. They are all spread out and do not show up in large cities either.

Figure 5.2.5: SOM clusters in the Netherlands based on hospital admissions



Positive percentage of tests

For this dataset, a 5x5 size map of neurons is used. This is due to it containing neurons with zero values if the map of neurons is a larger size. To decide how many clusters should be assigned, the elbow method is used. This results in an ideal number of clusters that is 4, as can be seen in Figure 5.2.7.

Figure 5.2.6: The number of municipalities within a neuron

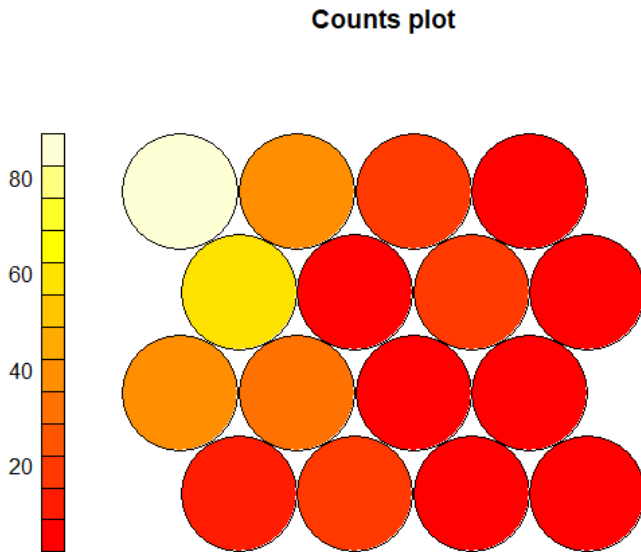
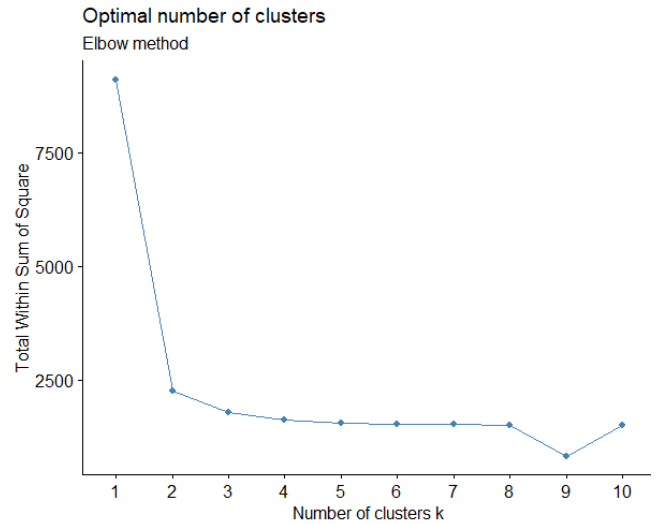


Figure 5.2.7: the ideal number of clusters according to the elbow method



These clusters are then visualized on the codes of the SOM (see Figure 5.2.8). Because it can be unclear which neurons belong to a certain cluster, in Figure 5.2.9 the clusters are visualized using colors. As can be seen in the codes plot (Figure 5.2.8), most of the neurons have a flat development through time. Only on the right side of the plot, do the neurons have high values. This results in the neurons with higher values being assigned to different clusters.

Figure 5.2.8: Codes plot with hierarchical cluster boundaries

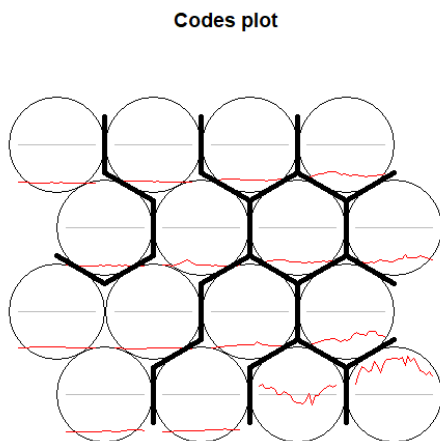
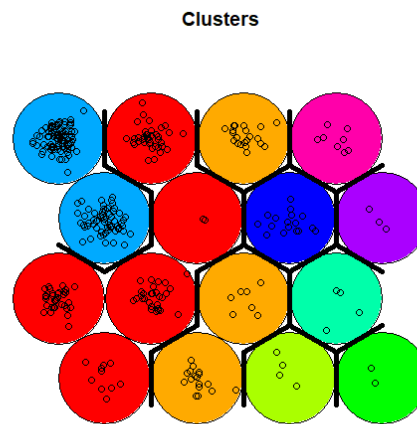
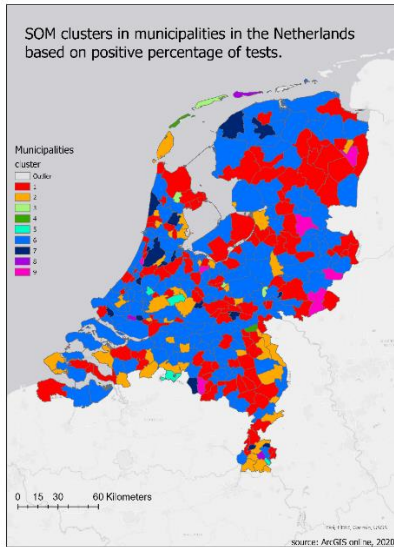


Figure 5.2.9: Color visualization of the location of clusters



The spatial spread of the clusters can be seen in Figure 5.2.10. The most interesting cluster to look at is clusters 3,4 and 8 as they have a codebook vector that is vastly different from the other codebook vectors. However, these clusters only show up in a very limited number of municipalities. These high values occur due to the low number of inhabitants for these municipalities. This results in the peaks being higher.

Figure 5.2.10: SOM clusters in the Netherlands based on the positive percentage of tests



Deceased cases

For this dataset, it was possible to use a 10x10 size for the map of neurons, because there are no neurons with zero municipalities in them (See Figure 5.2.11). To decide how many clusters should be assigned, the elbow method is used. This results in an ideal number of clusters that is 3, as can be seen in Figure 5.2.12. However, there is also a decrease in benefit after eight clusters. Therefore, eight clusters are used because this provided better results than three clusters.

Figure 5.2.11: The number of municipalities within a neuron

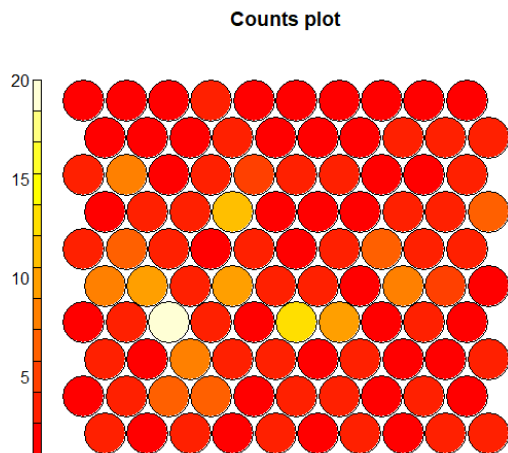
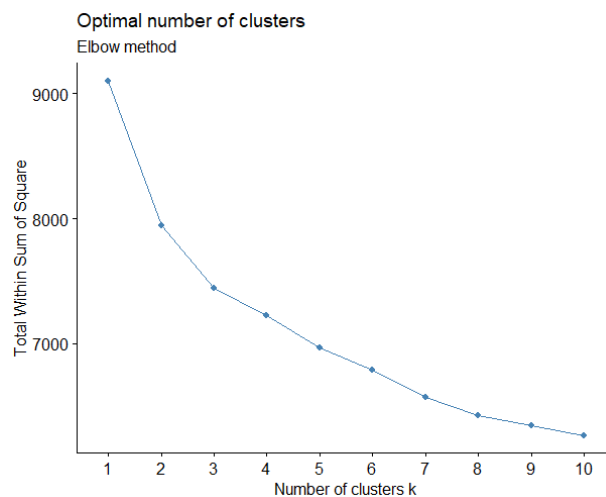


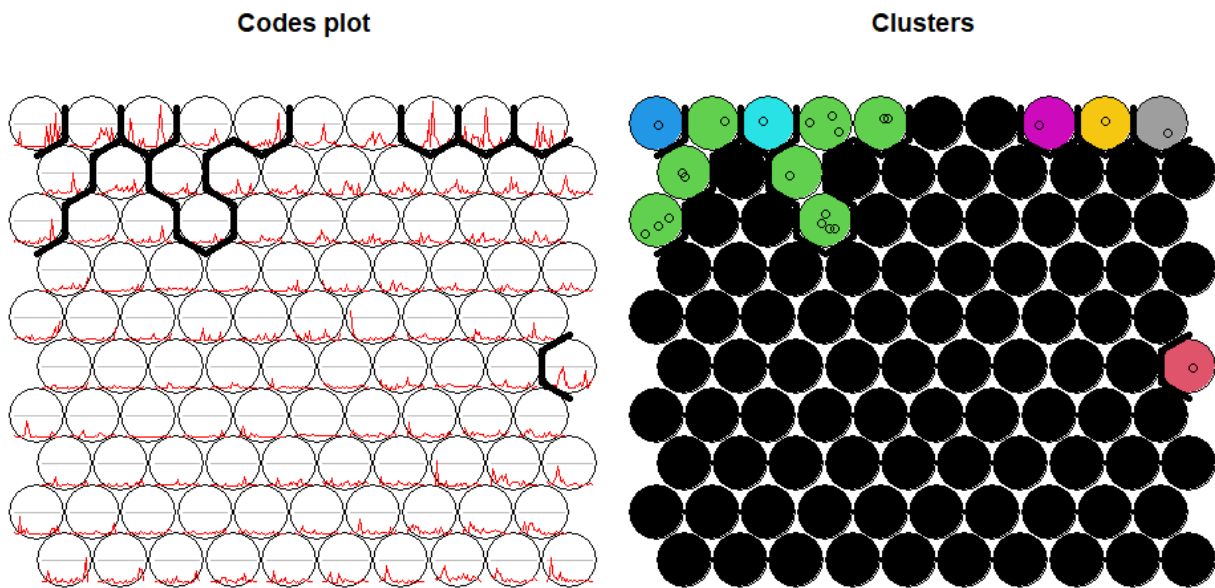
Figure 5.2.12: the ideal number of clusters according to the elbow method



These clusters are then visualized on the codes of the SOM (see Figure 5.2.13). Because it can be unclear which neurons belong to a certain cluster, in Figure 5.2.14 the clusters are visualized using colors. What stands out in these codes is that most of the displayed graphs are relatively flat. This is due to the number of deaths being close to 0 most of the time. This results in municipalities with higher deceased cases per 100.000 people being assigned to a separate cluster. Another pattern that can be distinguished is that the graphs in the right bottom corner have their peak early and the graphs in the top left corner have their peaks all quite late.

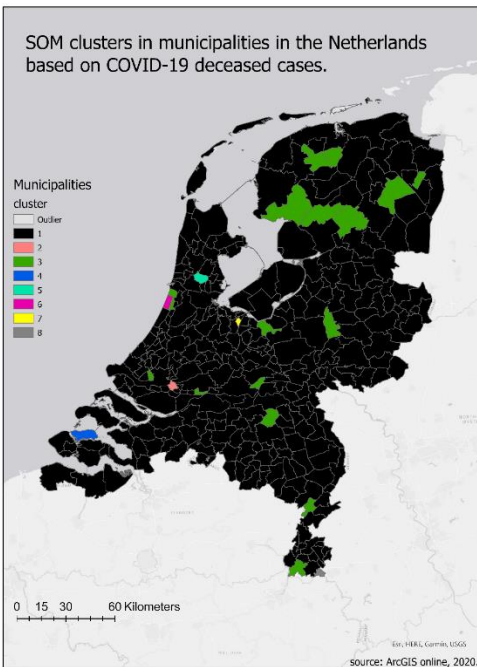
Figure 5.2.13: Codes plot with hierarchical cluster boundaries

Figure 5.2.14: Color visualization of the location of clusters



The spatial spread of the clusters can be seen in the map in Figure 5.2.15. The first cluster that stands out is a large number of municipalities in the Northern part of the Netherlands. This is cluster 3 and indicates that these municipalities experienced a peak late in time. These are all medium to small-sized cities. For the rest of the clusters, this map does not give a clear insight into what patterns there are. It is hard to determine which municipalities have developed similarly. This can be caused by several issues, namely: MAUP, lots of zero values, or overall unsuitability of the dataset.

Figure 5.2.15: SOM clusters in the Netherlands based on deceased cases



Sewage

For this dataset, a 6x6 size map of neuron is used. This is due to it containing neurons with zero values if the map of neurons is a larger size. To decide how many clusters should be assigned, the elbow method is used (Figure 5.2.17). This results in an unclear ideal number of clusters, as there are multiple options. After trying out multiple options, the best result was generated using seven clusters.

Figure 5.2.16: The number of municipalities within a neuron

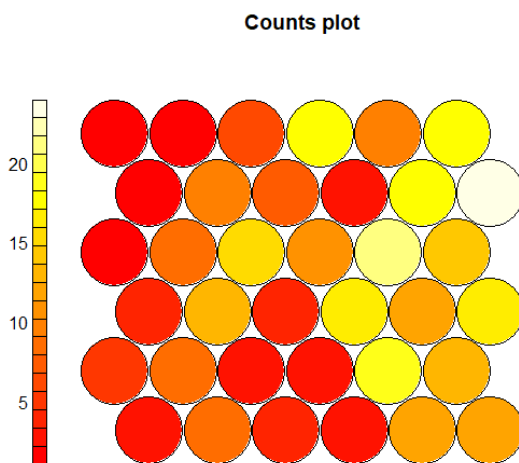
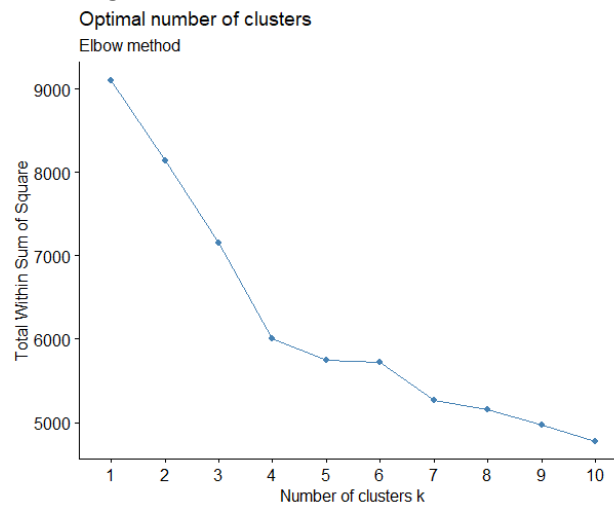


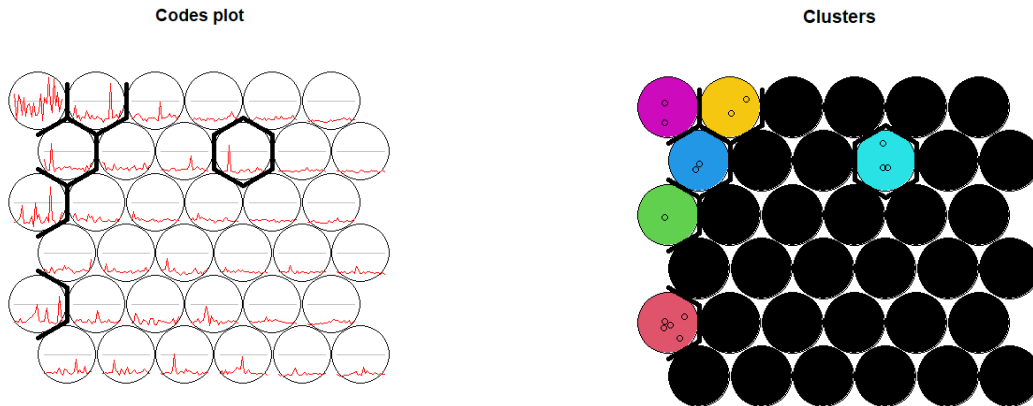
Figure 5.2.17: the ideal number of clusters according to the elbow method



These clusters are then visualized on the codes of the SOM (see Figure 5.2.18). Because it can be unclear which neurons belong to a certain cluster, in Figure 5.2.19 the clusters are visualized using colors. As can be seen in the codes plot (Figure 5.2.18), only the extreme cases have their own separate clusters. The other neurons have either no peaks (right) or a small peak in the beginning or middle (bottom).

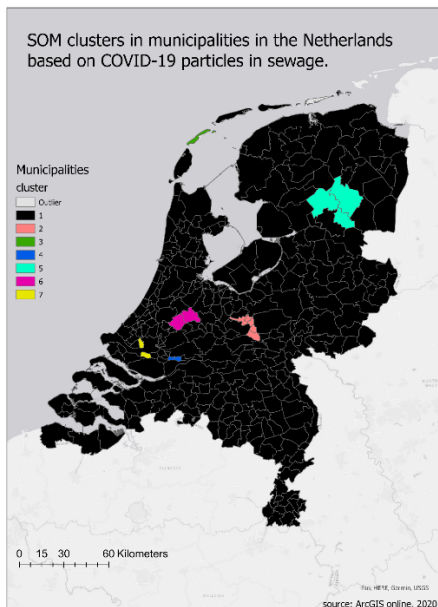
Figure 5.2.18: Codes plot with hierarchical cluster boundaries

Figure 5.2.19: Color visualization of the location of clusters



When looking at the map (Figure 5.2.20), it is clear that only a few municipalities are within other clusters than cluster 1. The municipalities that are in separate clusters are all very sparsely populated. Therefore it is reasonable to think that the number of inhabitants is very influential on the outcome of this map. For the rest, it is not possible to find a distinct pattern. It is hard to determine which municipalities have developed similarly. This can be caused by several issues, namely: MAUP, lots of zero values, or overall unsuitability of the dataset. It is important to realize that this dataset is originally displayed per treatment plant and was converted to municipalities. This conversion can be the cause of the lack of a distinct pattern as well.

Figure 5.2.20: SOM clusters in the Netherlands based on sewage data



5.3 Comparison of SOM results with hotspot analysis

In this section, the results of the SOM analyses for the datasets individually and the results of the hotspot analyses are compared. This is done to provide an overview of the differences between the two and to provide an insight into the added value of the SOM. In the hotspot analyses, some datasets were analyzed based on three peaks. The first peak of these three is however not relevant for the comparison, as the first peak is outside of the time period of the SOM.

Hospitalized cases

The hotspots for the hospitalized cases (Figures 5.3.1 and 5.3.2) show a clear hotspot in the southern part of the Randstad in both Figures. Moreover, a hotspot in the eastern part of the Netherlands is shown. The SOM map (Figure 5.3.3) shows less clear clusters, however, the majority of the southern part of the Randstad is colored green. This indicates that cluster 3 is mostly present here. This cluster corresponds to medium-high values throughout the time period. This confirms the findings in the hotspot analysis, as there are high values in the southern part of the Randstad in both Figures. The other found hotspot in the eastern part of the Netherlands can be seen in the SOM map as well. The municipalities that are within the hotspot found in Figures 5.3.1 and 5.3.2 show up in green in Figure 5.3.3. The cold spots do not show up clearly in Figure 5.3.3. Interestingly some of the cold spots are assigned to cluster 1, which is the black color. This cluster represents the municipalities that have a peak in infections at the end of the time period. This does not show up in the hotspot maps, as this is not within their time period.

Figure 5.3.1: Hotspots of hospitalized cases in the second peak

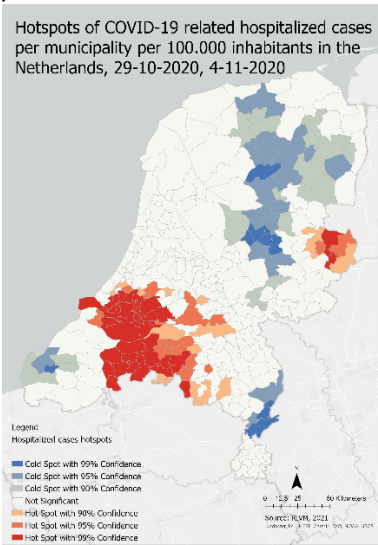


Figure 5.3.2: Hotspots of hospitalized cases in the third peak

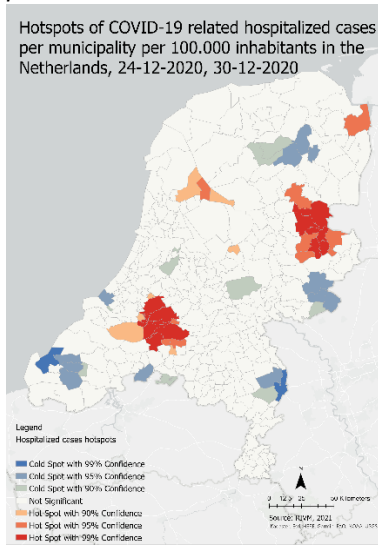
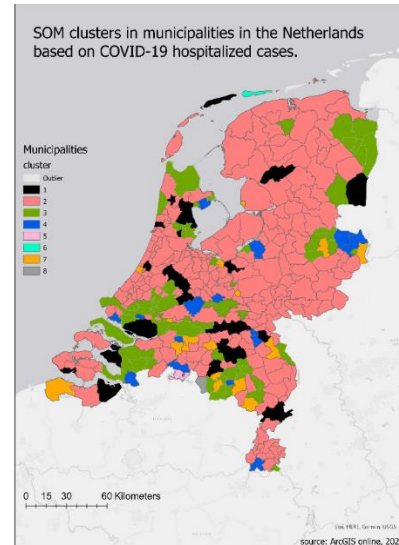


Figure 5.3.3: Clusters of the SOM of hospitalized cases



Positive percentage of tests

The hotspots for the positive percentage of tests (Figures 5.3.4 and 5.3.5) show very different hotspots. The first one shows hotspots in the southern Randstad and a part of Noord-Brabant. The second one shows a hotspot in the middle of the country. The similarity between the two maps is found in the eastern part of the Netherlands. The map of the SOM clusters (Figure 5.3.6) does not show any similarity with the first two maps. The SOM map is dominated by red and blue colors, which resemble a low amount of infections through time. Based on this, a vague resemblance can be seen with the cold spot in the north as seen in the hotspot maps. The north is more dominated by red and blue than any other part of the Netherlands. Therefore, this corresponds to the findings in the hotspot maps.

Figure 5.3.4: Hotspots of the positive percentage of tests in the first peak

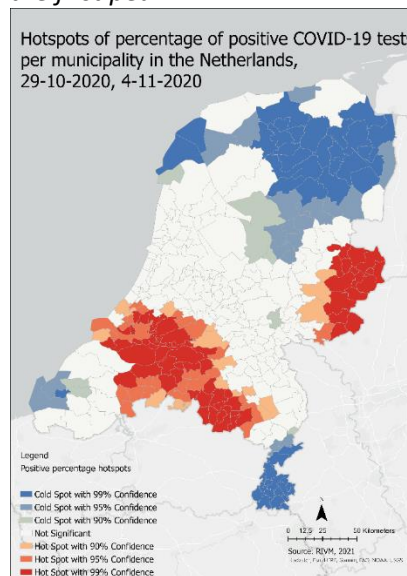


Figure 5.3.5: Hotspots of the positive percentage of tests in the second peak

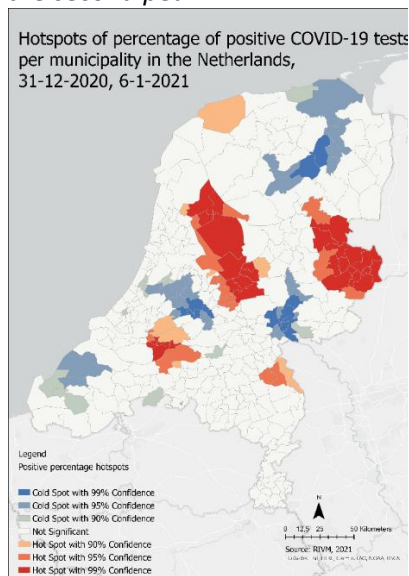
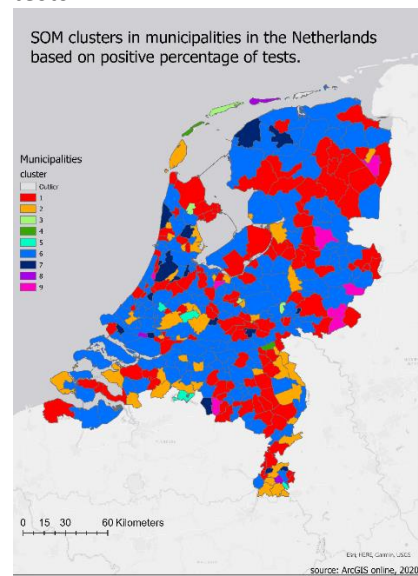


Figure 5.3.6: Clusters of the SOM of the positive percentage of tests



Deceased cases

The hotspots for the deceased cases (Figures 5.3.7 and 5.3.8) show clear hotspots in the southern part of the Netherlands in both Figures. Figure 5.3.7 shows a hotspot in the middle of the county, while Figure 5.3.8 shows hotspots to the east and west of the hotspot in Figure 5.3.7. In the map of the SOM results (Figure 5.3.9) the most visible pattern is that there is a grouping of the green cluster (cluster 3) in the north of the Netherlands. This indicates that there is a peak at the end of the time period. Therefore this finding is not directly related to the cold spot in Figure 5.3.6. However, it is possible that the people in the north did not build up immunity, which made them more vulnerable later on in the pandemic. This is however uncertain, as at the time of writing it is unsure how much immunity someone builds up after an infection.

Figure 5.3.7: Hotspots of deceased cases in the second peak

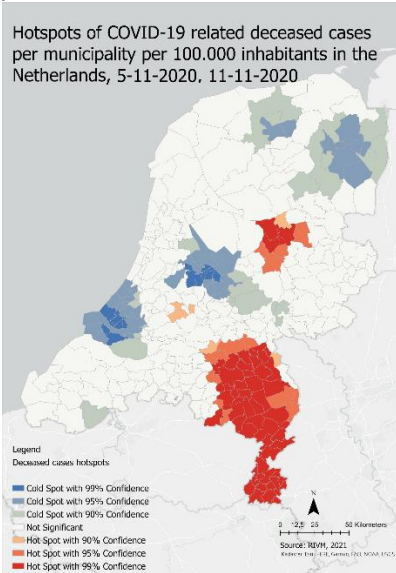


Figure 5.3.8: Hotspots of deceased cases in the third peak

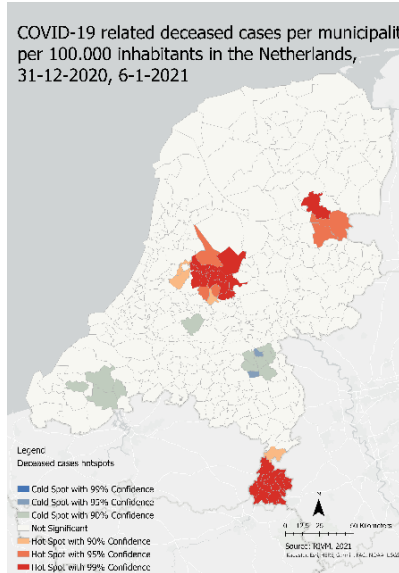
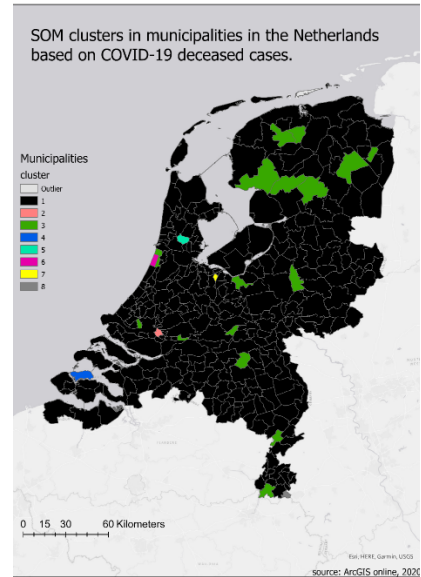


Figure 5.3.9: Clusters of the SOM of deceased cases



Sewage

The maps for the sewage data are harder to compare due to them being different formats. The hotspot maps (Figures 5.3.10 and 5.3.11) are displayed in their original form and the SOM results (5.3.12) are converted to the municipality level. However, it is already clear that the maps do not show many similarities. The hotspot maps show clear clusters in the middle of the country, while this is not the case for the SOM map. The clear cluster in the north on the cluster map also does not show in the hotspot map. Therefore can be concluded that the maps are very different, which can be caused by the conversion to the municipality level. Therefore, caution must be taken to base any conclusions on these maps.

Figure 5.3.10: Hotspots of virus particles in sewage in the second peak

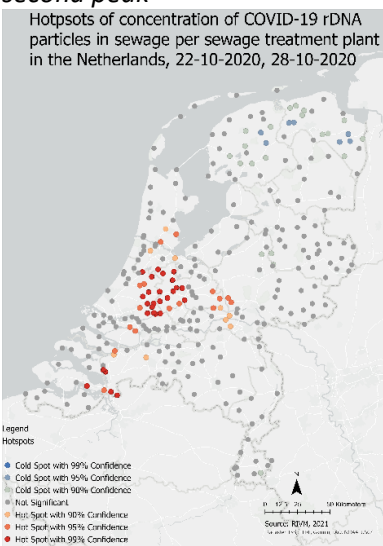


Figure 5.3.11: Hotspots of virus particles in sewage in the third peak

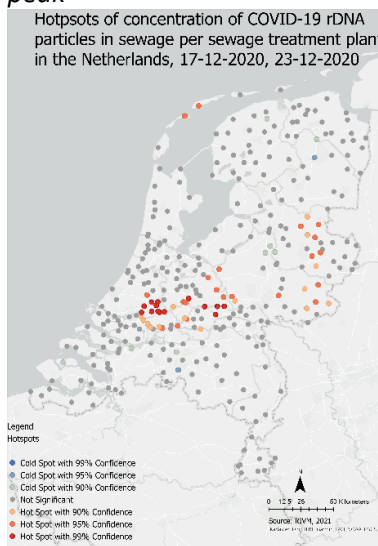
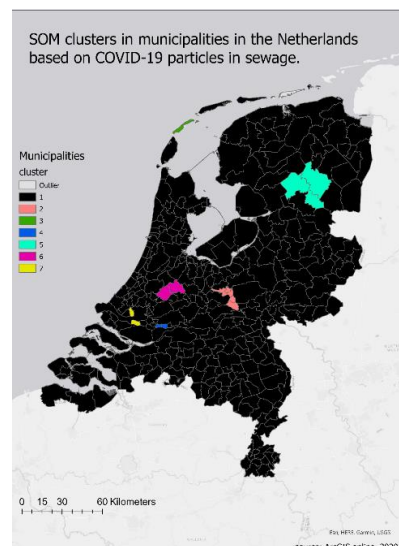


Figure 5.3.12: Clusters of the SOM of virus particles in sewage



Conclusion

Based on the comparison between the individual results and the hotspot maps, can be concluded that the SOM has added value. In the hospital comparison, it is clear that the SOM can provide an overview of the combination of multiple points in time in one map. This is a clear benefit of using SOM over some sort of analysis that can only capture one point in time. Moreover, the SOM could find local peaks better than the hotspot maps visualized based on the national hotspots. To conclude, the SOM has a clear added value and should be considered when evaluating the situation of COVID-19 in the Netherlands.

5.4 Results without delayed datasets

This section is an intermediate step in the research before the final results are generated. In this section, the results of a SOM based on the percentage of positive tests, infected percentage of the population, and the virus particles in the sewage systems are presented. Hospitalized cases and deceased cases are not included in this analysis, because literature has shown that both can be delayed in comparison with the three datasets named above. This delay also showed up in the analysis of the datasets in section 3.3. It is hard to correct the delay for every municipality without doing extensive research on how to do this effectively. An easy solution for this is to run the script for the datasets that do not have such a delay.

The three datasets have a different range of values that cannot be compared directly, therefore they are first scaled individually and thereafter merged. After this is done, it is possible to run the SOM with the new merged dataset. This dataset consists of 1053 rows (three times the municipalities) and 26 rows (the weeks that indicate time). The results of this SOM are shown below.

In Figure 5.4.1 the number of municipalities within a neuron is shown. The municipalities are evenly spread over the neurons except for neurons 16, 22, and 23. In these neurons, a lighter color is shown, which means that there are more values present in these neurons. The cause of this can be seen later in the codes plot. In Figure 2 the ideal number of clusters is shown. The ideal number according to the plot is three. However, this did not provide a comprehensible result, therefore the decision is made to deviate from the ideal number of clusters. As can be seen in Figure 5.4.2, there also is a slight dip at eight clusters. The results are therefore generated with eight clusters, this provided more meaningful results.

Figure 5.4.1: The number of municipalities within a neuron

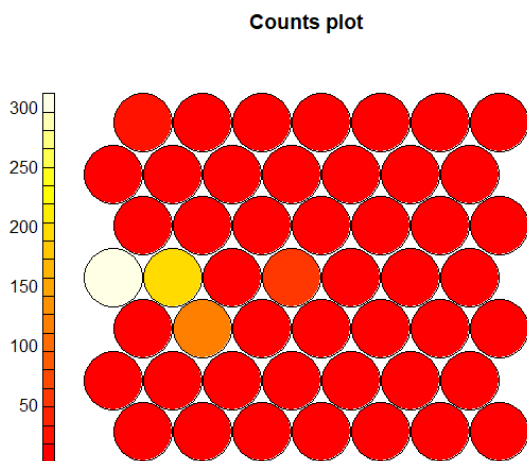
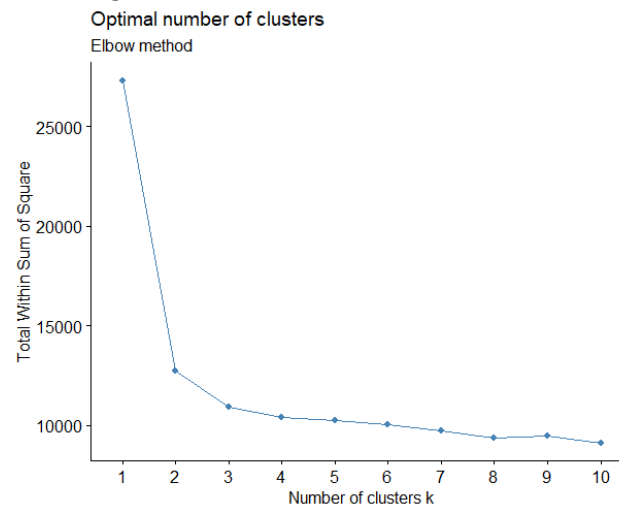


Figure 5.4.2: the ideal number of clusters according to the elbow method



In Figure 5.4.3 the codes plot is displayed. Neurons 16, 22 and 23, which were described in the previous section have a value of zero through time. In general, the neurons on the left have very low values through time, whereas the neurons on the right have higher values through time. In the top right, the values are the highest, with a constant high value. Remarkably, only neuron 10 (the dark blue neuron in Figure 5.4.4) has a distinct peak early in time.

Figure 5.4.3: Codes plot with hierarchical cluster boundaries

Codes plot

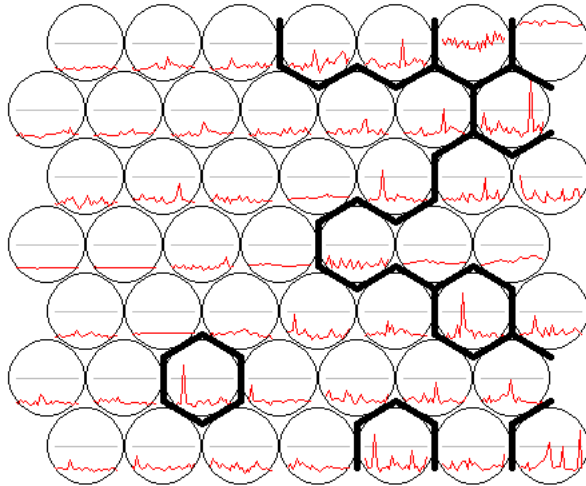
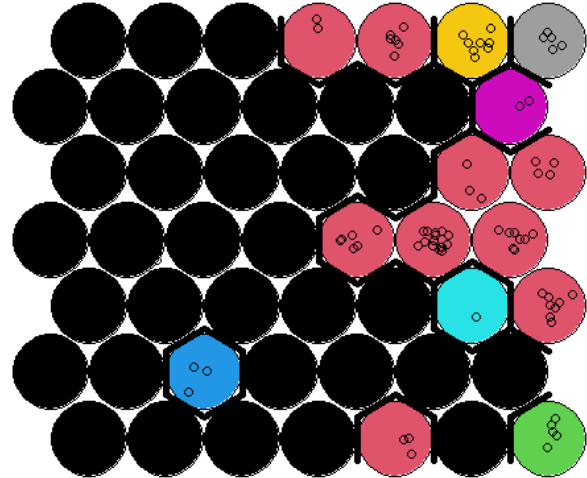


Figure 5.4.4: Color visualization of the location of clusters

Clusters



In Figures 5.4.5, 5.4.6, and 5.4.7 the location of the municipalities within the map of neurons is shown for the three datasets. Remarkably, the positive percentage of the population and the positive percentage of tests are extremely similar. Most of the similarity can be found in neurons 16, 22, and 23, which are as described earlier the neurons where the values are near zero all the time. The sewage dataset also has some values within these neurons, but this is far less than the other two datasets. It also stands out that the values of the sewage dataset are far more spread out than the other two datasets. This is most likely due to the sewage dataset measuring an infection with COVID-19 multiple times. The other datasets only measure an infection once. Moreover, the sewage datasets are constructed by assigning certain sewage treatment plants to municipalities, which can be influential on the results.

Figure 5.4.5: Location of the values of positive percentage of the population in the map of neuron.

Positive Percentage of Inhabitants

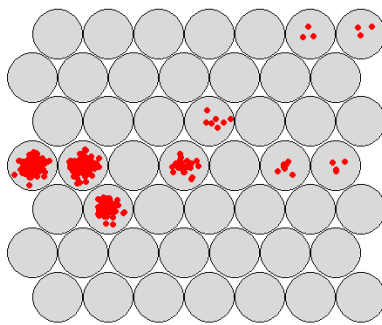


Figure 5.4.6: Location of the values of positive percentage of tests in the map of neuron.

Positive Percentage of Tests

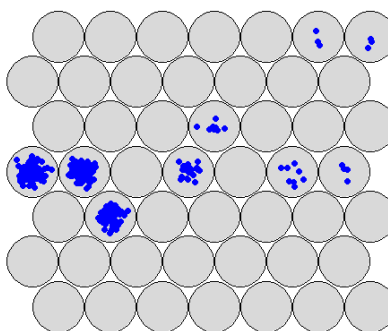
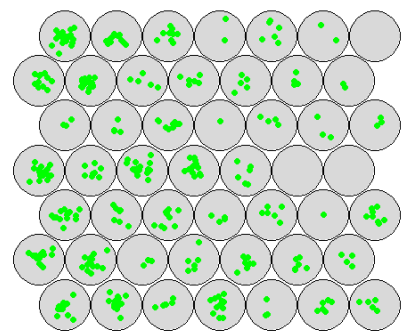


Figure 5.4.7: Location of the values of virus particles in sewage in the map of neuron.

Virus Particles in Sewage



In Figures 5.4.8, 5.4.9, and 5.4.10 the location of the clusters of the different datasets is displayed on a map. Again, the positive percentage of the population and the positive percentage of tests are very similar. This is to be expected, as Figures 5.4.5 and 5.4.6 showed that the location of the municipalities within the map of neurons is similar to each other. The municipalities that are not within cluster 1 are all small cities. This pattern is likely caused by the MAUP issue, where big spikes in small communities show up even larger. The location of the clusters of the sewage data is different from the other two in all clusters, except for cluster 1 (the black color). It is however remarkable that cluster 2 seems to be centered around Eindhoven and Amsterdam, both being the largest cities in the area. There is also a cluster in the north, which is centered around Assen, which is the capital of the province of Drenthe and therefore an important city in the area.

Figure 5.4.8: Location of the clusters of the positive percentage of population displayed on a map

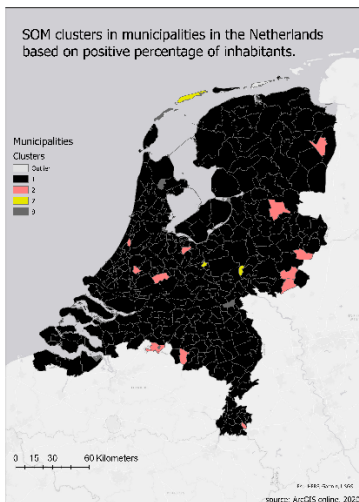


Figure 5.4.9: Location of the clusters of the positive percentage of tests displayed on a map

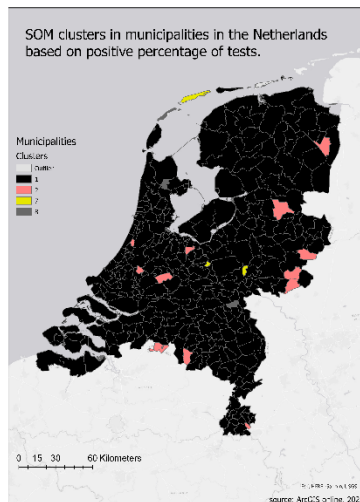
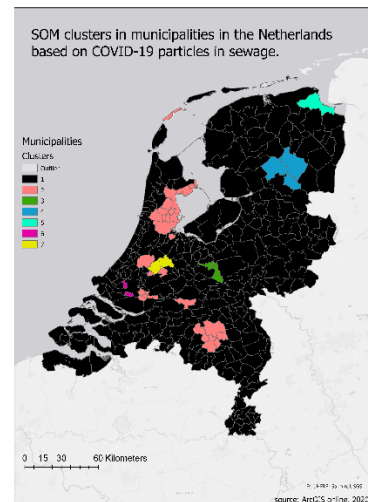


Figure 5.4.10: Location of the clusters of virus particles in sewage displayed on a map



In table 5.4.1 up to table 5.4.3, the figure of merit for all the datasets is shown. This is calculated both for the neurons and the clusters. The higher the number, the more similar the datasets are to each other. The maximum score is 1. In table 5.4.1 the average of all the figures of merits per dataset is shown for a quick overview. Again the percentage of tests and percentage of the population has the highest score, which means they are the most similar. The clusters show a high score throughout all the datasets. This is the result of most municipalities begin assigned to the first cluster. The higher scores for the percentage of tests and percentage of the population are caused by the fact that these datasets are almost identical. This is confirmed by table 5.4.2 and table 5.4.3. In this table, the datasets have a figure of merit of 1 and nearly 1. The other datasets all show fewer similarities and are roughly assigned the same number. This corresponds to the finding in the plots in Figures 5.4.5, 5.4.6, and 5.4.7.

Table 5.4.1: Figure of merit for the datasets without a delay

	Percentage tests	Percentage population	Sewage
Neuron	0,480114	0,480114	0,036932
Cluster	0,913352	0,913352	0,826705

Table 5.4.2: Figure of merit for the datasets without a delay calculated with clusters

	Percentage tests	Percentage population	Sewage
Percentage tests	1	1	0,826705
Percentage population	1	1	0,826705
Sewage	0,826705	0,826705	1

Table 5.4.3: Figure of merit for the datasets without a delay calculated with neurons

	Percentage tests	Percentage population	Sewage
Percentage tests	1	0,923295	0,036932
Percentage population	0,923295	1	0,036932
Sewage	0,036932	0,036932	1

In general, the conclusion can be drawn that the positive percentage of tests and the positive percentage of the population are more similar to each other than they are to the virus particles in sewage.

Sammon's projection

The next step is to analyze the datasets in a Time over Space structure (TxS) as proposed by Andrienko et al. (2010). This provides insight into the temporal variations of the datasets. The structure is created by transposing the previous structure of the datasets. This results in columns as the weeks and the rows as the municipalities. After training the SOM, this results in the codebook vector as shown in Figure 5.4.11. This codebook vector is displayed as a Sammon's projection in Figure 5.4.12. This shows the distance between the vectors, however, the topology is not necessarily maintained. The numbers in Sammon's projection refer to the numbers of the neurons. The neuron in the left bottom corner is neuron 1, the neuron in the right bottom corner is neuron 5. The neuron in the top right corner is number 25. The neurons around number 15 indicate a low amount of COVID-19 presence and the neurons around number 12 indicate a high amount of COVID-19 cases.

Figure 5.4.11: Codes plot with hierarchical cluster boundaries

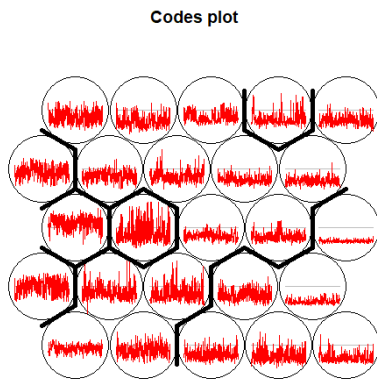
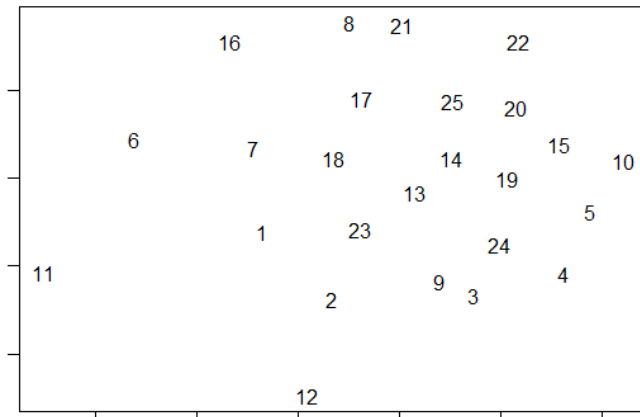


Figure 5.4.12: Sammon's projection of the codebook vector



In Figure 5.4.13 Sammon's trajectory of the positive percentage of tests can be seen. The trajectory starts at neuron 22, which indicates that there is a medium amount of COVID-19 presence. The general trend is that it slowly fades until it reaches neuron 1, after which it climbs again in neuron 2 up till 5. The trajectory ends with low values in neuron number 10.

In Figure 5.4.14 Sammon's trajectory of the positive percentage of inhabitants can be seen. Unlike the positive percentage of tests, this trajectory starts at a low number of COVID-19 presence in neuron 15. After which it follows roughly the same trajectory as the positive percentage of tests. One big difference is the step to neuron 11, which indicates high values, after which it returns to 1 to follow a similar trajectory again.

In Figure 5.4.15 the Sammon's trajectory of the COVID-19 particles in sewage can be seen. The first thing that stands out is that the pattern is much more chaotic than the two previously described trajectories. This chaos and the fact that the lines cross often indicate that the pattern of sewage experiences fewer ups and downs. As concluded before, the sewage measures a COVID-19 infection for a longer amount of time instead of one moment. This causes the dataset to be generally stable. The peaks that are caused are therefore less extreme than the previous two datasets. These smaller peaks are located in neurons closer to the prior neurons, which can cause the trajectory to visit the same neuron more than once.

In general, the conclusion can be drawn that the first two trajectories are similar to each other and the trajectory of sewage is very different. This corresponds to the previous findings that the positive percentages are more similar to each other. As stated before, this is caused by the positive percentages tracking an infection once and the sewage data tracking an infection for a longer period of time.

Figure 5.4.13: Sammon's projection and trajectory of the positive percentage of tests

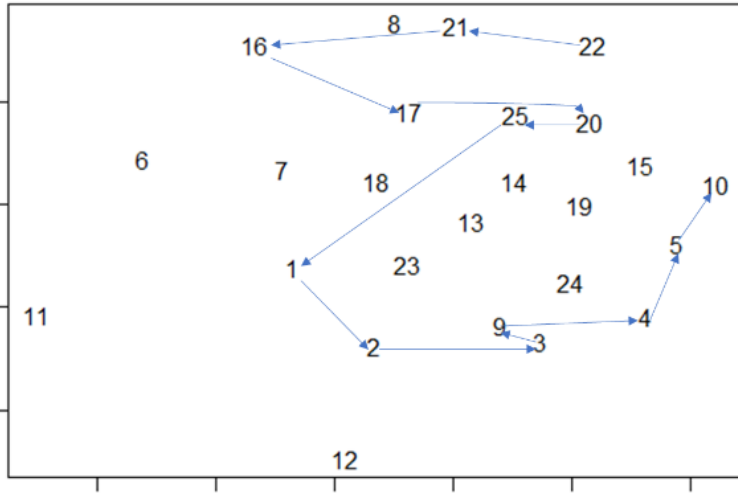


Figure 5.4.14: Sammon's projection and trajectory of the positive percentage of inhabitants

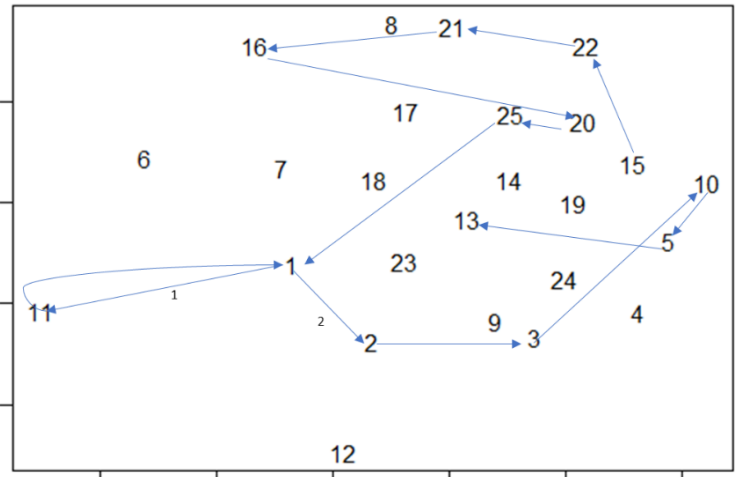
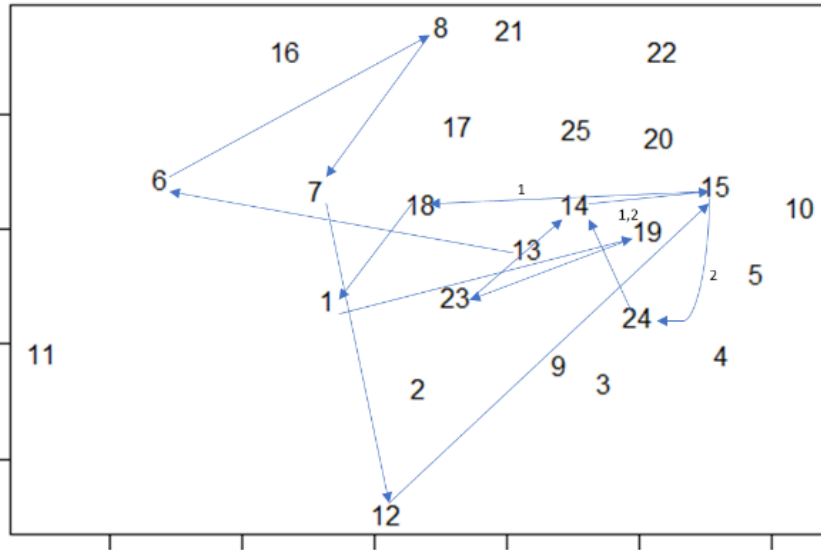


Figure 5.4.13: Sammon's projection and trajectory of the COVID-19 particles in sewage



5.5 Results with delayed datasets

This section describes the results of the comparison of all the datasets. This includes the datasets without a delay as described in section 5.2 and the datasets which are supposedly delayed (hospitalized cases and deceased cases). Therefore the difference between datasets is expected to be larger than in section 5.2. The map of neuron size of the SOM is in this case 9x9, which is much smaller than the formula in 5.1.1. This is due to a map of neurons any larger than 9x9 returning too many neurons without any values.

In this analysis, all of the datasets are first scaled and thereafter merged. This is due to them begin incomparable in their original state. This dataset consists of 1755 rows (the municipalities for the five datasets) and 26 columns (the weeks that indicate time).

In Figure 5.5.1 the counts plot of the SOM is shown. The municipalities are in general equally divided over the neurons, except for the white neuron (neuron 39) in the middle which is overrepresented. This makes sense when looking at the codebook vector, which is displayed in Figure 5.5.3. Neuron 39 indicates that the value is near zero throughout the entire timespan. The elbow method, which shows the ideal number of clusters is shown in Figure 5.5.2. The method indicates that three clusters should be ideal, however, after testing this out it did not provide meaningful results. These results were meaningful at nine clusters, which is also a valid option due to the small bump that can be seen at the end of the graph.

Figure 5.5.1: The number of municipalities within a neuron

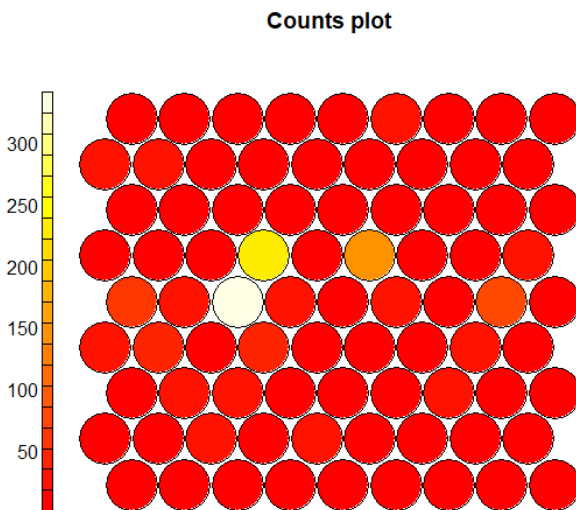
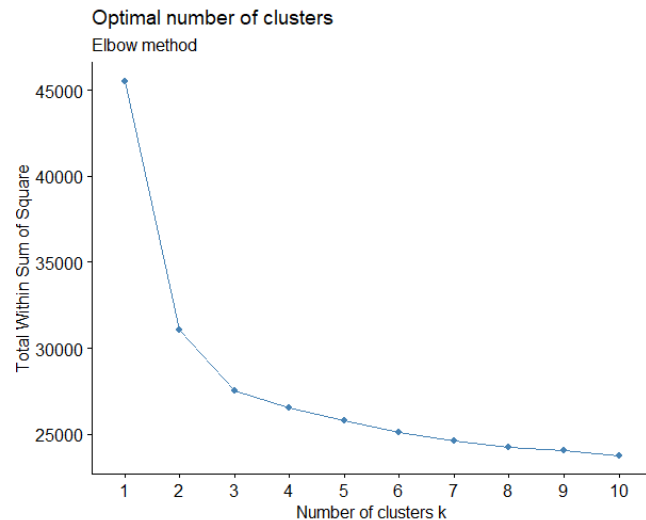


Figure 5.5.2: the ideal number of clusters according to the elbow method



In Figure 5.5.3 the codebook vector is shown in which the values per neuron are displayed through time. In general, the extreme values are displayed at the edges of the codebook vector. In general, on the bottom left the neurons experience a peak in the beginning, and on the top right, the neurons experience a peak late in time. The top left and the bottom right both show a peak in the middle. The clusters which are drawn over the neurons generally are located around the extremes at the edges of the codebook vector. This is the result of hierarchical clustering which is sensitive to extremes. This is another reason why it is better to use more clusters than indicated in the elbow method as shown in Figure 5.5.2. In Figure 5.5.3 the clusters are plotted, which shows that the clusters that are not red have few values in them. This can be caused by the input data, which consists of multiple datasets. These datasets all have their own extremes, which can be grouped in a small cluster.

Figure 5.5.3: Codes plot with hierarchical cluster boundaries

Codes plot

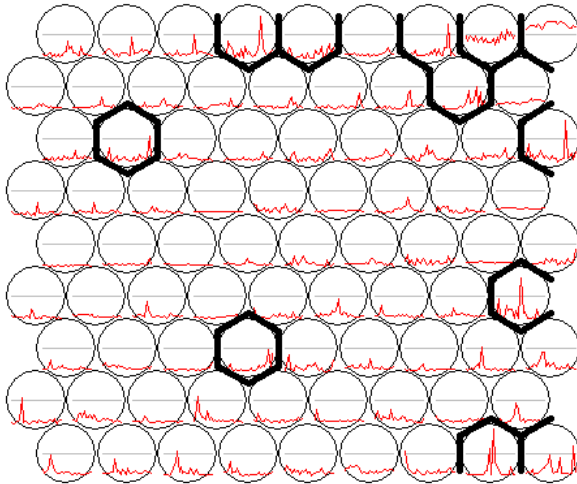
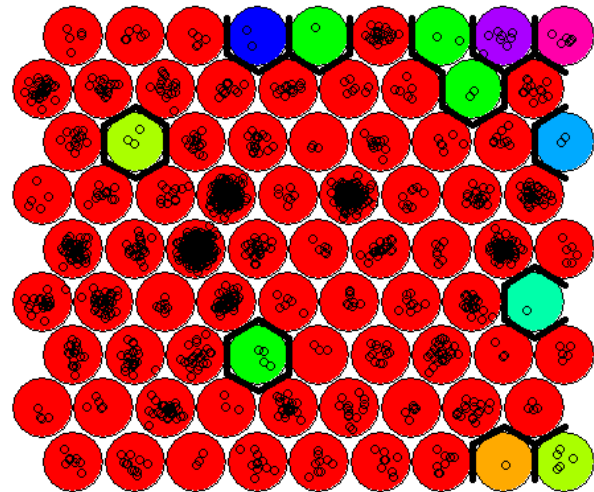


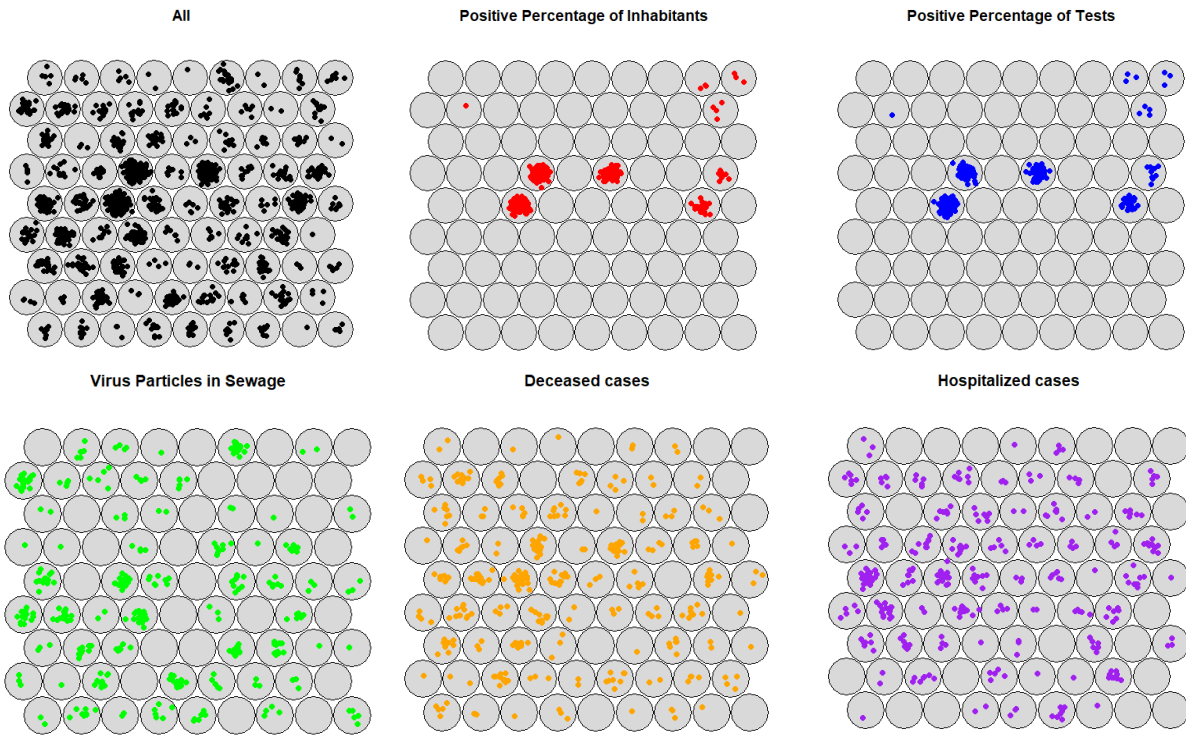
Figure 5.5.4: Color visualization of the location of clusters

Clusters



In Figure 5.5.5 the mappings of the datasets are plotted. The first two datasets, which are the positive percentage of Inhabitants and positive percentage of tests are again very similar with a distribution that is comparable to that in the previous SOM described in section 5.3. This is due to the same reason given in that section. The virus particles in sewage, deceased cases, and hospitalized cases give similar results. This might seem strange, due to deceased cases and hospitalized cases being recorded once in time and the virus particles in sewage recorded longer throughout time. However, this can be caused by the fact that deceased cases and hospitalized cases do not suffer from the same issue as the percentage datasets. The positive percentage datasets show much higher peaks, which results in a lot of municipalities being reduced to values close to zero after scaling. Due to the lower peaks in deceased cases and hospitalized cases, this is less of an issue in these datasets.

Figure 5.5.5: Visualization of the spread of the datasets within the SOM map of neurons.



In Figures 5.5.6 up to 5.5.10, the location of the clusters of the five datasets is shown. The first thing that stands out is the overwhelming majority of the map is colored red. This is as expected as the majority of Figure 5.5.3 is also red and it has the neurons with the most values in it. The maps of the positive percentages are again nearly identical, which is expected based on the results in section 5.3. The other maps do not show any similarity between them in the smaller clusters. The only similarity here is the municipality of Urk in the deceased cases and the hospitalized cases. This corresponds to a large peak in the middle according to the codebook vector in Figure 5.5.3. Interestingly, this does not show up in any other dataset. This can be due to the inhabitants of Urk not testing for COVID-19 at testing facilities as Urk is a very religious and traditional village. However, it still should show up in the sewage dataset if this was the case. However, the only nearby treatment plant is located in the Noord Oost Polder to the west of Urk. This means that the sewage virus particles from Urk are merged with the sewage particles of the Noord Oost Polder. This is an issue that can play a role in multiple municipalities that do not have their own sewage treatment plant.

Figure 5.5.6: Location of the clusters of the positive percentage of tests displayed on a map

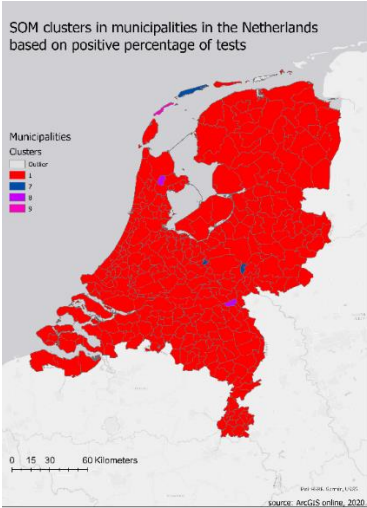


Figure 5.5.7: Location of the clusters of the positive percentage of inhabitants displayed on a map

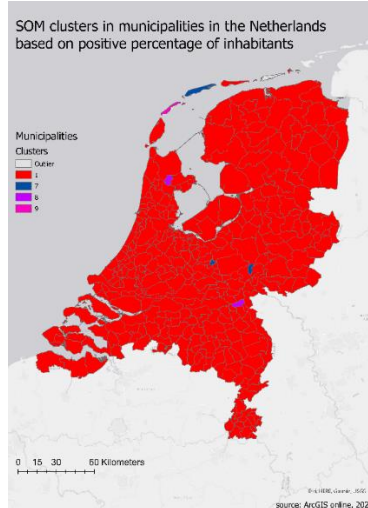


Figure 5.5.8: Location of the clusters of virus particles in sewage displayed on a map

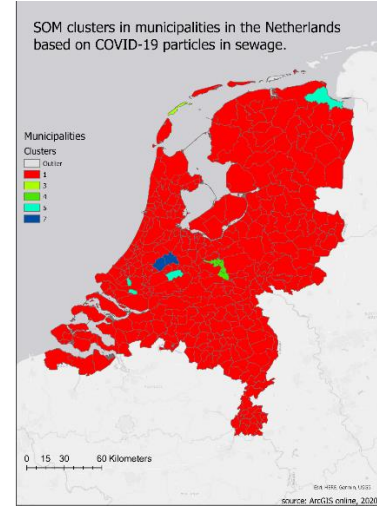


Figure 5.5.9: Location of the clusters of deceased cases displayed on a map

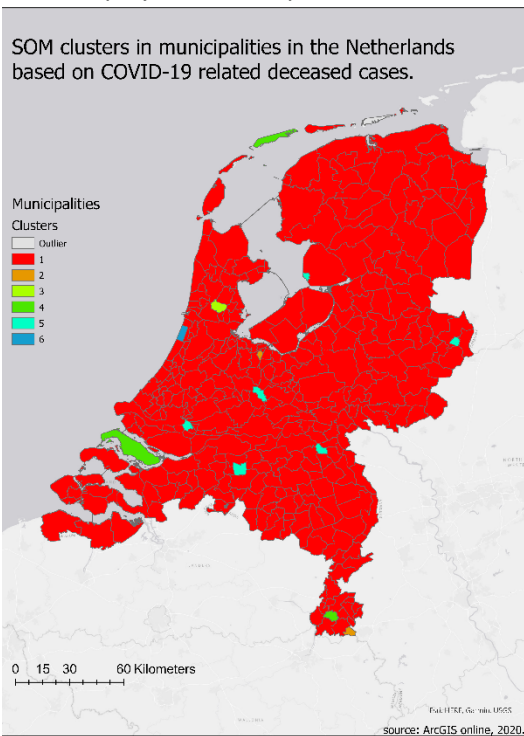


Figure 5.5.10: Location of the clusters of hospitalized cases displayed on a map



In table 5.5.1 up to table 5.5.3, the figure of merit for all the datasets is shown. This is calculated both for the neurons and the clusters. The higher the number, the more similar the datasets are to each other. The maximum score is 1. In table 5.5.1 the average of all the figures of merits per dataset is shown for a quick overview. Again the percentage of tests and percentage of the population has the highest score, which means they are the most similar. The clusters show a high score throughout all the datasets. This is the result of most municipalities begin assigned to the first cluster. The higher scores for the percentage of tests and percentage of the population are caused by the fact that these datasets are almost identical. This is confirmed by table 5.5.2 and table 5.5.3. In this table, the datasets have a figure of merit of 1 and nearly 1. The other datasets all show fewer similarities and are roughly assigned the same number.

Table 5.5.1: Figure of merit for all datasets

	Percentage tests	Percentage population	Sewage	Deceased	Hospital
Neuron	0,256392045	0,257102273	0,03125	0,040482955	0,035511364
Cluster	0,97017	0,97017	0,949574	0,947443	0,96946

Table 5.5.2: Figure of merit for all datasets calculated with clusters

	Percentage tests	Percentage population	Sewage	Deceased	Hospital
Percentage tests	1	1	0,954545	0,948864	0,977273
Percentage population	1	1	0,954545	0,948864	0,977273
Sewage	0,954545	0,954545	1	0,928977	0,960227
Deceased	0,948864	0,948864	0,928977	1	0,963068
Hospital	0,977273	0,977273	0,960227	0,963068	1

Table 5.5.3: Figure of merit for all datasets calculated with neurons

	Percentage tests	Percentage population	Sewage	Deceased	Hospital
Percentage tests	1	0,934659	0,022727	0,042614	0,025568182
Percentage population	0,934659	1	0,022727	0,042614	0,028409091
Sewage	0,022727	0,022727	1	0,034091	0,045454545
Deceased	0,042614	0,042614	0,034091	1	0,042613636
Hospital	0,025568	0,028409	0,045455	0,042614	1

Sammon's projection

To compare the disease diffusion of the five datasets the datasets are structured in a Time over Space (TxS) model as proposed by Andrienko et al. (2010). This provides insight into the temporal variations of the datasets. The structure is created by transposing the previous structure of the datasets. This results in columns as the weeks and the rows as the municipalities. After training the SOM, this results in the codebook vector as shown in Figure 5.5.11. This codebook vector is displayed as a Sammon's projection in Figure 5.5.12. This shows the distance between the vectors, however, the topology is not necessarily maintained. The numbers in Sammon's projection refer to the numbers of the neurons. The neuron in the left bottom corner is neuron 1, the neuron in the right bottom corner is neuron 5. The neuron in the top right corner is number 25. In general, the neurons on the left (again neuron 15) are low values and the neurons on the right are high values. Neurons 1 and 6 in the bottom left corner are very different from their neighboring neurons, this can be seen in the distance in Sammon's projection.

Figure 5.5.11: Codes plot with hierarchical cluster boundaries

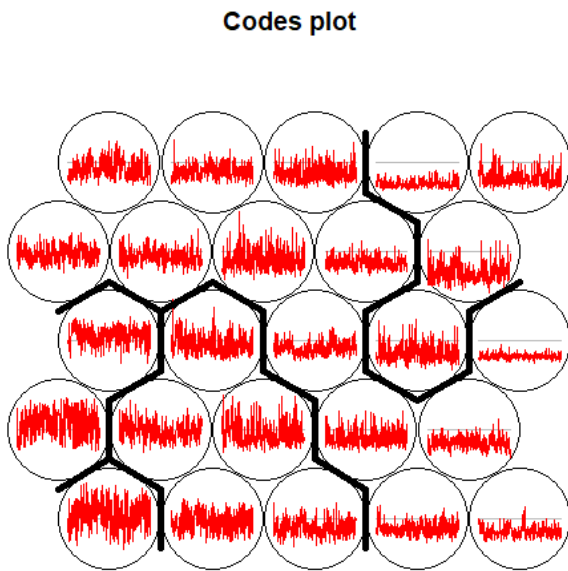
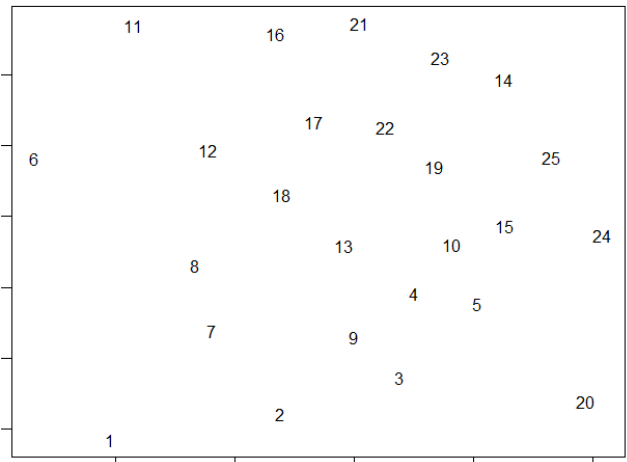


Figure 5.5.12: Sammon's projection of the codebook vector



The trajectories of Figures 5.5.13 And 5.5.14 Are again very similar to each other, starting at neuron 20 with low values throughout most municipalities. Continuing to neurons 3,2 and 1 gradually increasing through time. Then shortly decreasing to number 10, after which it increases again to neuron 11. After that, the positive percentage of tests slowly decreased from neurons 23,25 and thereafter 24. The positive percentage of inhabitants increases again and ends at neuron 17.

The trajectory of COVID-19 particles in sewage is again rather chaotic due to its concentration in neurons that are close to each other. The trajectory starts low at neuron 19 and increases rapidly through neurons 21 and 6. It returns close to zero at neuron 15 again. Thereafter it slowly increases again through neurons 13 and 8. The trajectory eventually passes neuron 15 again and then shows stable medium values in neurons 9, 4, and 5.

Figure 5.5.13 Sammon's projection and trajectory of the positive percentage of tests

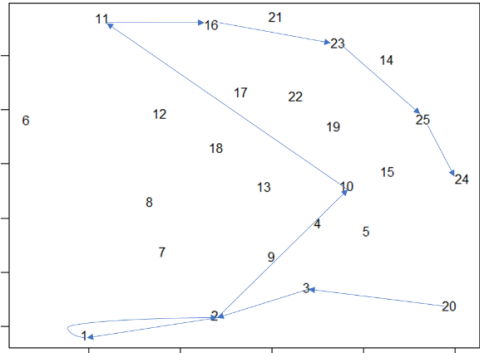


Figure 5.5.14: Sammon's projection and trajectory of the positive percentage of inhabitants

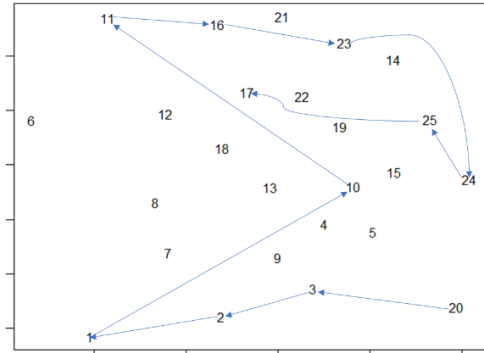


Figure 5.3.15: Sammon's projection and trajectory of the COVID-19 particles in sewage

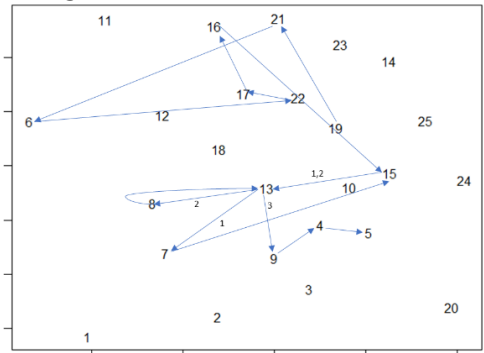


Figure 5.5.16 Sammon's projection and trajectory of the deceased cases

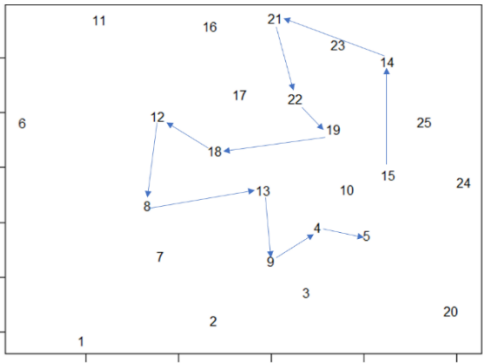
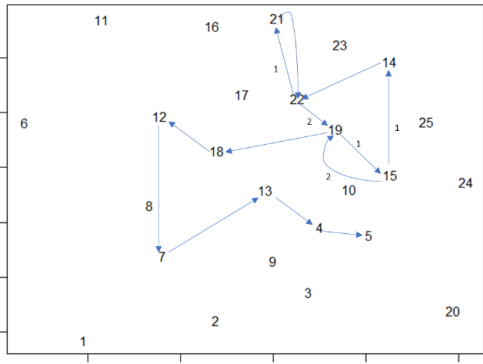


Figure 5.5.17 Sammon's projection and trajectory of the hospitalized cases



In Figure 5.5.16, the trajectory for the deceased cases is shown. This trajectory makes small steps in between neurons, which indicates that there should be very few large jumps from small presence to high presence. The first step is from neurons 15 to 14 which shows a small increase, after which it stabilizes. This stabilization continues throughout the end, at which it reaches low values again. This stability can be seen in the trajectory from the fact that it stays in the middle of the plot and therefore is very compact.

In Figure 5.5.17, the trajectory for the hospitalized cases can be seen. Again the trajectory seems to be present primarily in the middle of the plot. It follows a similar trajectory to the deceased cases. The trajectory appears different since some neurons appear multiple times. This causes the trajectory to look different from the deceased cases. The major difference is that neuron 15 is visited more often. This is likely the result of the scaling of the data. The low values in the hospitalized cases get scaled closer to zero than the low values in deceased cases. This is due to the hospitalized cases having a higher maximum value than the deceased cases.

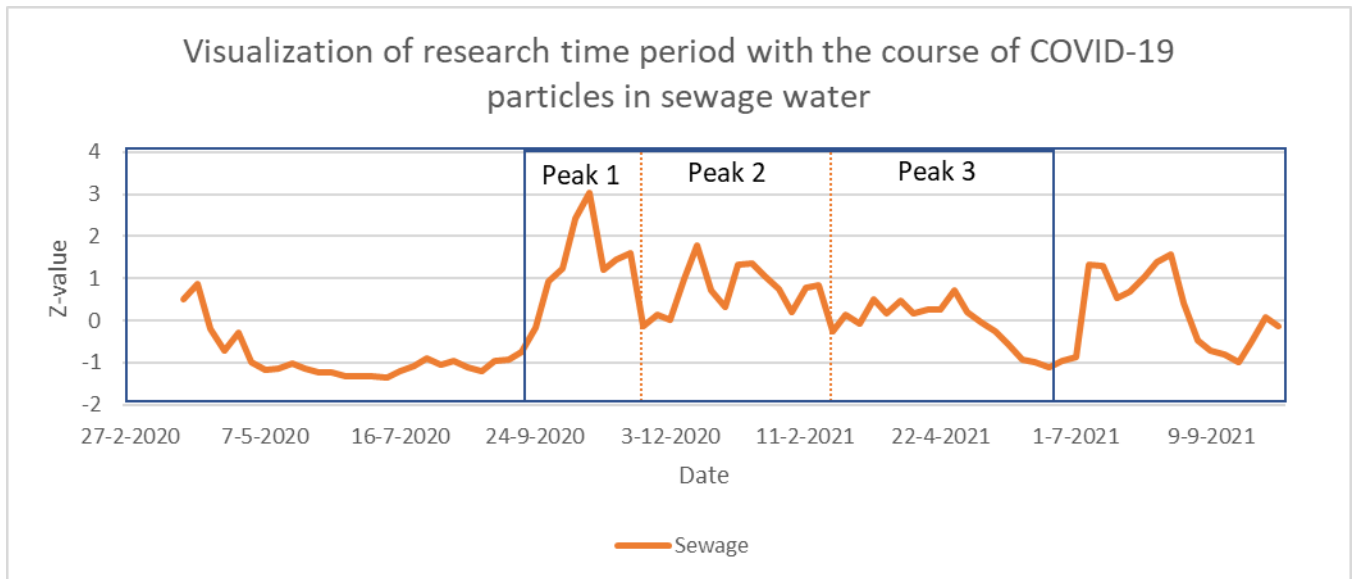
To conclude, the same findings can be seen in the TxS construction as in the SxT structure. The positive percentages are very similar to each other with more extreme values than the rest of the datasets. The sewage dataset is different from all datasets, with both extreme values and low values. The deceased cases and hospitalized cases are also similar to each other, with hospitalized cases showing more extreme values.

5.6 Comparison of waves in COVID-19 based on sewage data

In this section, multiple waves of the pandemic are compared to each other. This is done with the use of the virus particles in the sewage dataset. This dataset is chosen, as it is influenced the least by testing capacity and vaccinations. Moreover, this dataset could not be used in the previous sections due to it being collected per sewage treatment plant instead of per municipality.

The input data of this SOM is constructed according to the peaks shown in Figure 5.6.1. Every sewage treatment plant is present three times in the dataset, once for every peak. This is the same construction for the comparison of the datasets as seen in sections 5.2 and 5.3, the only difference is that waves are compared in this section.

Figure 5.6.1: The peaks in the sewage datasets on which the waves are based



In Figure 5.6.2 the counts plot is shown. The lighter color represents the most frequently occurring neurons. In this case, the neurons in the middle of the top rows are presented most often. This corresponds to the neurons in the codebook vector (Figure 5.6.4) with a low number throughout the entire timespan. In Figure 5.6.3 the ideal number of clusters according to the elbow method is shown. The optimal number of clusters would be three in this case. However, the choice is made to work with seven clusters. This is due to three clusters not providing results that could be used to find patterns in the data. Moreover in the graph of the elbow method can be seen that there is an increase in added value from six clusters to seven clusters. Therefore, seven clusters are still a strong choice.

Figure 5.6.2: The number of municipalities within a neuron

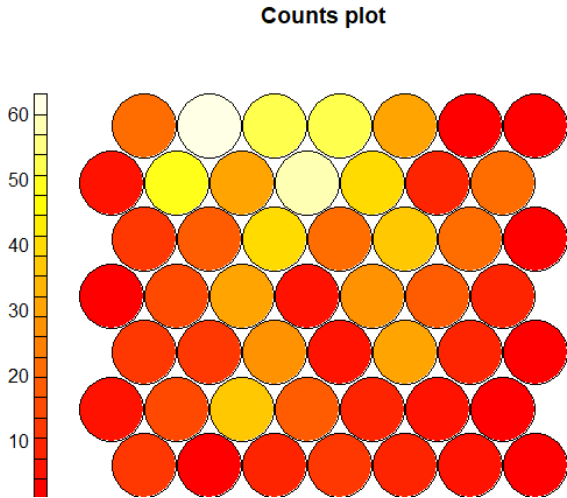
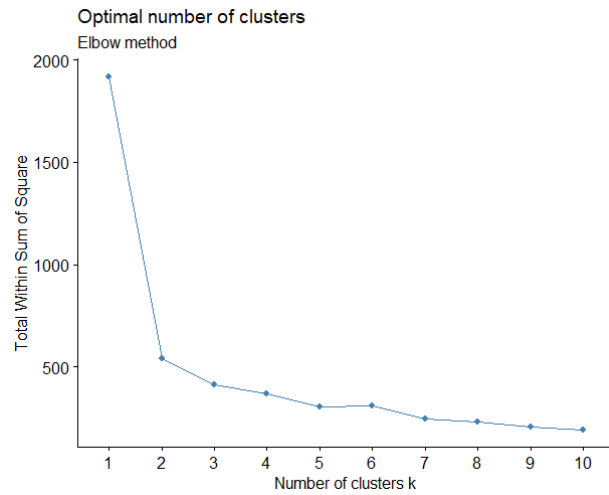


Figure 5.6.3: the ideal number of clusters according to the elbow method



In Figure 5.6.4 the codebook vector is shown. The first thing that stands out is that most of the neurons have low values over time. The neurons that have high peaks, therefore, are assigned to a separate cluster. This is the result of the hierarchical clustering method. At the top of the codebook vector, the neurons have rather low values throughout the entire timespan. On the left, the neurons have a peak late in time, and on the right, the peak occurs earlier in time. In Figure 5.6.5 the number of sewage treatment plants within a neuron can be seen. Most of the small clusters have a very low number of treatment plants within them. More detailed information on the spread of the treatment plants over the neurons can be seen in figure 5.6.6.

Figure 5.6.4: Codes plot with hierarchical cluster boundaries

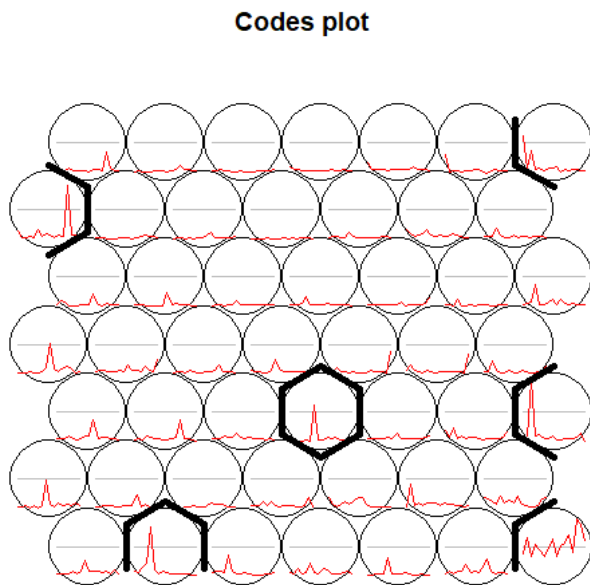
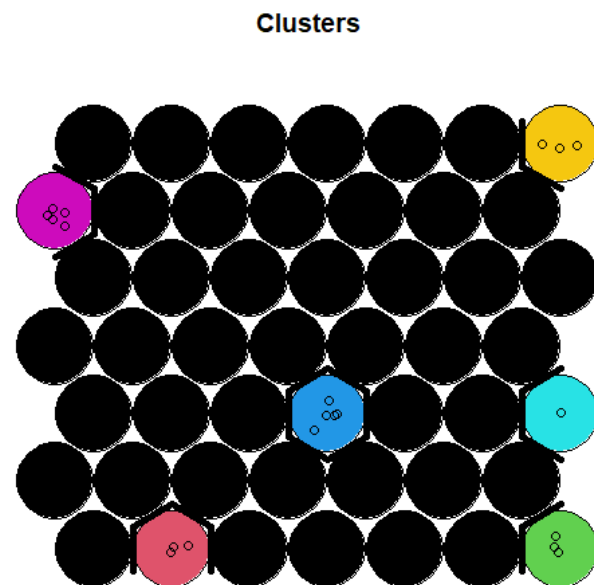
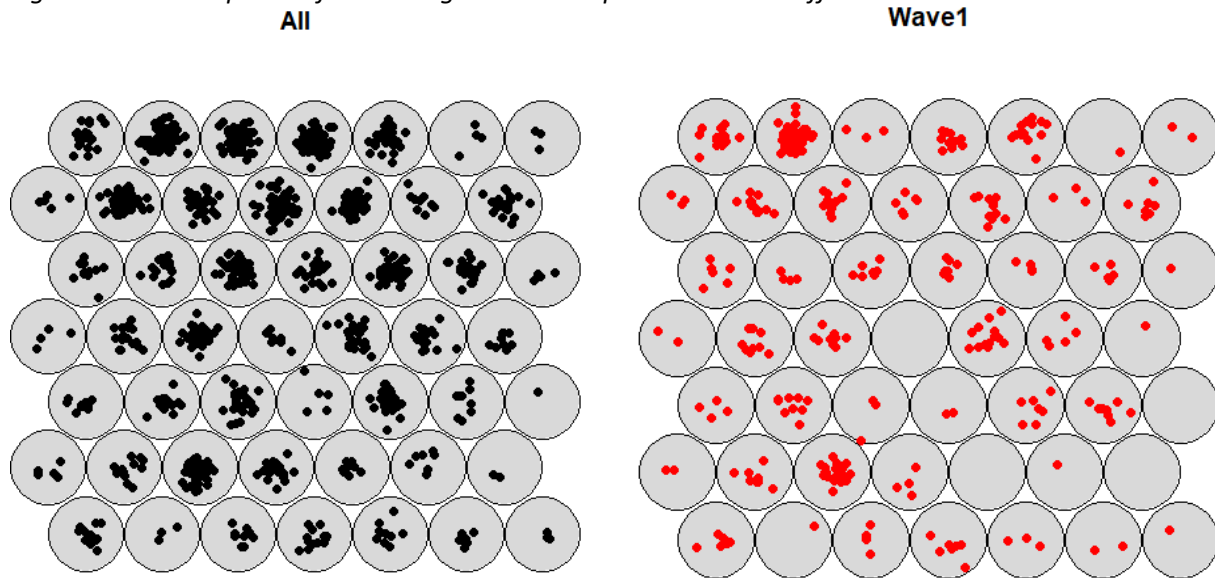


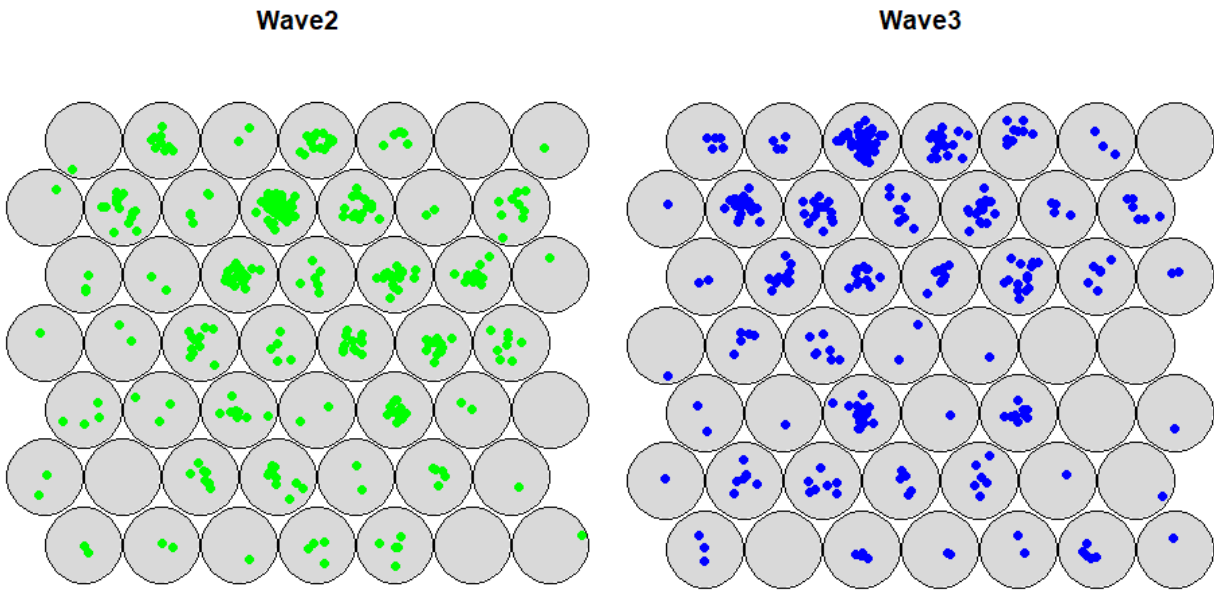
Figure 5.6.5: Color visualization of the location of clusters



The spread of the sewage treatment plants over the different waves is shown below (Figure 5.6.6). The Top left picture shows the mapping of all of the waves combined. The majority of treatment plants are located at the top of the middle columns, this corresponds to the counts plot. When looking at the mappings of the different waves, differences can be seen. For all three waves, the majority of the treatment plants are located at the top of the middle columns. However, the most occurring neuron is different each time. For Wave 1 the most occurring neuron is neuron 44, for wave 2 this is neuron 39 and for wave 3 this is neuron 45. This shows that every wave is relatively similar to each other, however, they are not exact matches. When comparing the mappings with the codebook vector (Figure 5.6.4) the most occurring neuron for wave 1 (neuron 44) has a small peak at the end of the wave. The most occurring neuron of wave 2 (neuron 39) has a peak in the middle of the wave. The most occurring neuron of wave 3 (neuron 45) does not show a peak and remains close to zero for the entire wave. The absence of a peak can be explained by looking at Figure 5.5.1, which shows that the peaks in wave 3 are lower than the peaks in the other two waves. Although they are quite similar based on the mappings this might not be the case in space, in Figures 5.6.7, 5.6.8, and 5.6.9 the waves are compared in space.

Figure 5.6.6: The spread of the sewage treatment plants over the different waves.





Although the mappings of the SOM show similar results between the waves, this similarity is harder to find in the spatial patterns of the waves. In the maps below cluster 1 is depicted as a smaller black circle to emphasize the other clusters. Moreover, this adds to the readability of the map which would otherwise be dominated by black circles. The similarity between the maps is that most of the sewage treatment plants are all within cluster 1, which is indicated by the black color. The only other similarity shared by the three waves is that the sewage treatment plant ‘Woerden’ is in cluster 3 for all of them. Cluster 3 (neuron 7 in Figure 5.4.4) shows remarkably high values for the entire timespan. This could be caused by the issue that it is a small city and the virus particles are calculated per 100.000 inhabitants. The other clusters are not in the same location during the three waves and therefore do not show many similarities.

Figure 5.6.7: Location of the clusters wave 1 displayed on a map

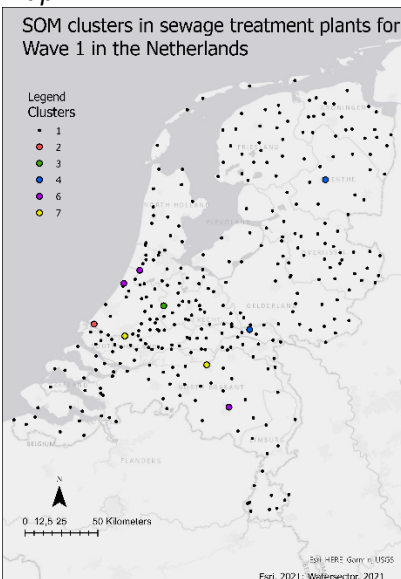


Figure 5.6.8: Location of the clusters wave 2 displayed on a map

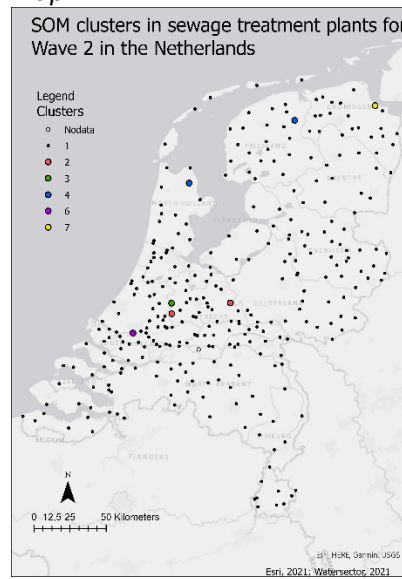
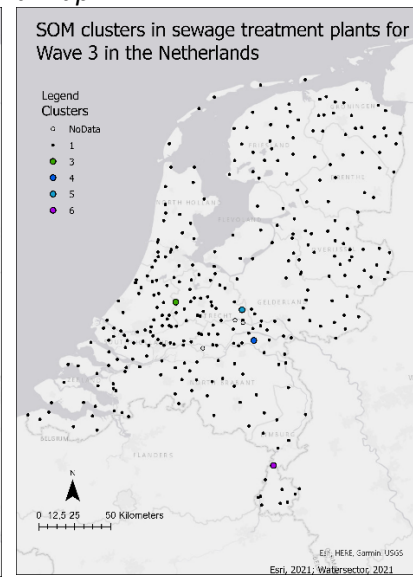


Figure 5.6.9: Location of the clusters of wave 3 displayed on a map



The findings of the maps in Figures 5.6.7, 5.6.8, and 5.6.9 are confirmed in the figure of merit between the waves. The higher the figure of merit, the more the waves have in common. In table 5.6.1 the average figure of merit for both the neurons and the clusters is shown. Again the clusters are assigned a very high score and the neurons have a really low score. Moreover, it looks like the waves all have roughly the same similarity.

Table 5.6.1: Average figure of merit between waves based on sewage data

	Wave1	Wave2	Wave3
Neuron	0,036277603	0,031545741	0,033123028
Cluster	0,955835962	0,960567823	0,960567823

Table 5.6.2 shows the figure of merit based on the neurons of the SOM. In the table can be seen that waves 2 and 3 are the least similar to each other. This is in line with what can be seen in Figure 5.6.6, as it shows a lot of values at the right side of the map of neurons for wave 2 and not for wave 3.

Table 5.5.2: Figure of merit between waves based on the SOM neurons of the sewage data

	Wave1	Wave2	Wave3
Wave1	1	0,034700315	0,03785489
Wave2	0,034700315	1	0,028391167
Wave3	0,03785489	0,028391167	1

Table 5.6.3 shows the figure of merit based on the clusters of the SOM. In this table, the difference between the waves is not as clear as in table 5.6.2. This is likely due to the hierarchical clustering used. This clustering is sensitive to high values and creates clusters around them, which causes all medium values to be assigned to the same cluster. Therefore, it is clear the neurons are a better measure in this case.

Table 5.6.3: Figure of merit between waves based on the SOM clusters of the sewage data

	Wave1	Wave2	Wave3
Wave1	1	0,955835962	0,955835962
Wave2	0,955835962	1	0,965299685
Wave3	0,955835962	0,965299685	1

Sammon's projection

To compare the different waves over time, a Time over Space (TxS) is created. This provides insight into the temporal variations over time. In other words, the TxS structure can be used to see the variations in the diffusion patterns. The structure is created by transposing the previous dataset. In Figure 5.6.10 the codebook vector of the SOM is displayed. There is a general pattern that can be seen. The left side of the codebook vector generally has low values, with neuron 16 being the lowest. The codebook vector is displayed as a Sammon's projection in Figure 5.6.10. This projection shows the distance between the vectors, however, the topology is not necessarily maintained. The numbers in Sammon's projection refer

to the numbers of the neurons. The neuron in the left bottom corner is neuron 1, the neuron in the right bottom corner is neuron 5. The neuron in the top right corner is number 25. Sammon's projection is much more compact than the projections seen in sections 5.3 and 5.4. This is due to the dataset being created from only one source, contrary to the three or five sources from the previous sections. This results in a projection that is more concentrated but can cause the trajectories to look more chaotic due to the smaller distances between neurons.

Figure 5.6.10: Codebook vector of the sewage TxS dataset with hierarchical cluster boundaries

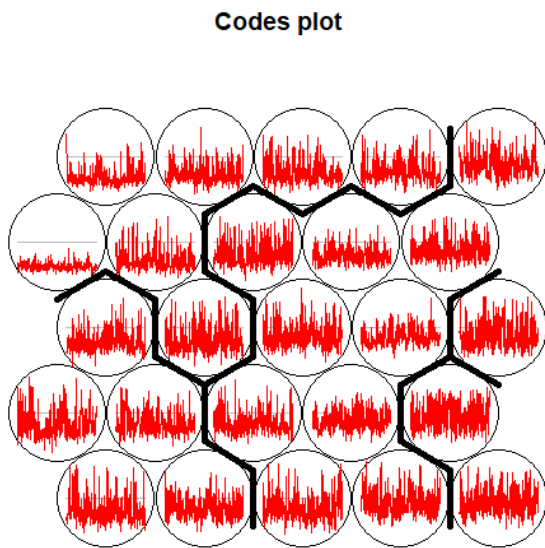
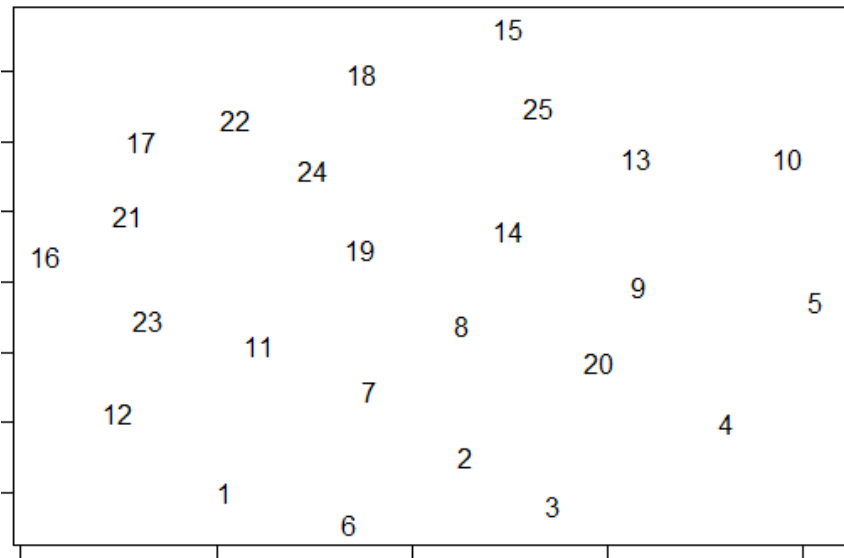


Figure 5.6.10: Sammon's projection of the codebook vector of the sewage TxS dataset



In Figure 5.6.11 Sammon's projection and trajectory of wave 1 are shown. The trajectory starts at neuron 16, which contains low values through time. It then slowly increases until it decreases again in neuron 9 in the fourth week. After that, it increases up until the very end, at which it is relatively low in neuron 21.

In Figure 5.6.12 Sammon's projection and trajectory of wave 2 are shown. It starts at week 10, so it follows up wave 1 directly. This wave is also much longer than wave 1, which can be seen as it shows much more lines than wave 1. However, a longer duration is not necessarily causally related to the number of lines, which is shown later in wave 3. Wave 2 seems to have multiple circular patterns in Sammon's trajectory. The first one is at the top, after which it follows from neuron 9 to neuron 16, which indicates low values. After that, a short circle goes from 16 to 1 to 6 and back to 16 again. This corresponds to the three peaks as shown in Figure 5.6.1.

In Figure 5.6.13 Sammon's projection and trajectory of wave 3 are shown. Although the duration of this wave is longer, it is assigned to fewer neurons than wave 2. This is due to small peaks in wave 3 being smaller than the peaks in wave 2. This shows up again in the trajectory, as it is concentrated much more in the middle than waves 1 and 2. It shows a relatively stable trajectory through time and only really decreases at the end.

Figure 5.5.11: Sammon's projection and trajectory of Wave 1 of the COVID-19 particles in sewage dataset

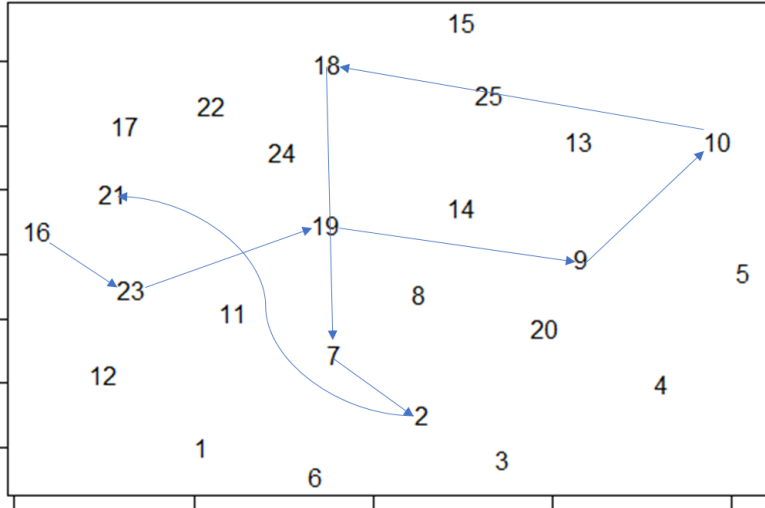


Figure 5.5.12: Sammon's projection and trajectory of Wave 2 of the COVID-19 particles in sewage dataset

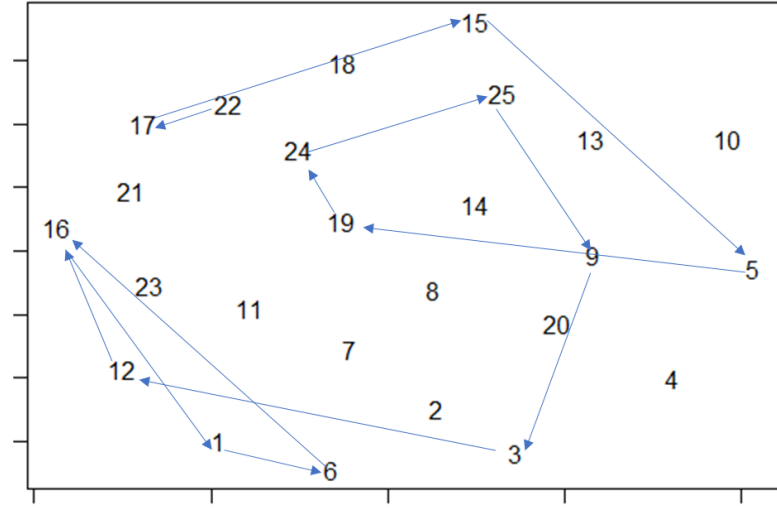
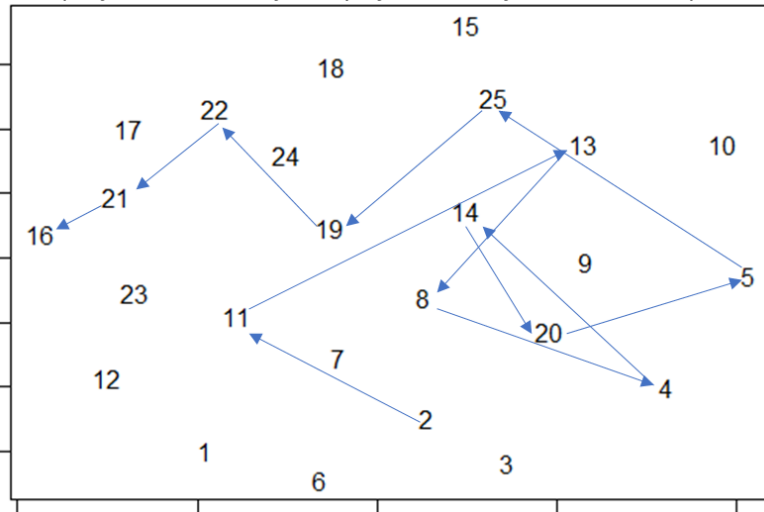


Figure 5.5.13: Sammon's projection and trajectory of Wave 3 of the COVID-19 particles in sewage dataset



To conclude, the waves do not share many similarities. The one pattern that they have in common is that a circular pattern can be distinguished. This pattern is related to the number of peaks that every wave experiences. The first and the third wave do not clearly show multiple peaks, which can be seen in Figure 5.6.1. Wave 2 clearly shows multiple circles and in its pattern.

The findings in the temporal patterns (TxS structure) are the same as in the spatial patterns (SxT structure). The waves do not show a lot of similarities in either the spatial or the temporal patterns. In the SxT this can be caused by the clustering method, but this is not the case in the TxS structure. Therefore the conclusion can be drawn that the waves are very different from each other.

6. Discussion & Conclusion

In this chapter, the research questions are answered, and the research process is reflected upon. In the end, suggestions for further research are provided.

The main research question that is answered is the following:

Which spatiotemporal patterns can be found in the spread of COVID-19 in the Netherlands?

The methodology used to answer this question is a Self Organizing Map (SOM). With the use of dually structured data, two different ways to analyze spatiotemporal data can be used (Andrienko et al., 2010). The data structures are as follows: A temporally ordered sequence of spatial situations (Space over time). The second data structure is a set of spatially arranged places where each place is characterized by its particular temporal variation of attribute values (Time over space). This allows for analyzing the change of the spatial situation over time and analyzing the distribution of the local temporal variations over space.

To be able to answer the main research question, it is divided into three sub-questions. The first one is answered to be able to understand which datasets are suitable for the research. The question is phrased as follows:

Which datasets are suitable to explore spatiotemporal COVID -19 patterns in the Netherlands and what is the implication of using these datasets?

Five different datasets were identified that describe Covid-19 incidence in the Netherlands. Two of these datasets are traditional notification data: The percentage of infected individuals and the percentage of positive tests and two other datasets describe the severity of the pandemic in the Netherlands: Number of hospitalized cases and the number of deceased. The fifth dataset is the Sewage dataset. The most important difference between the infection rate and the number of hospitalized or deceased individuals is the delay effect. Hospitalization and deaths will occur sometime after infection. There is a difference between the spatial units of the traditional datasets and the sewage dataset.

What patterns can be found in different datasets?

This question was answered in two different ways. Hotspot analysis and SOM analysis were conducted on the various datasets to detect clusters. These analyses were not performed to compare between datasets, but only to identify if clusters are found within single datasets.

The implication of using the different datasets is tested by training a SOM on all combined data and mapping back the individual datasets to detect differences.

Two different experiments were conducted. The first on comparing the patterns of the sewage data with the percentage tested and the percentage positive and the second experiment compares all the datasets. This is done to be able to experiment without the influence of the delay mentioned above. The Space over Time analysis resulted in clear differences and similarities between datasets. The positive percentage of tests and the positive percentage of inhabitants were almost perfectly similar. However, they were very different from the other datasets. The deceased cases, hospitalized cases, and COVID-19 particles in sewage also showed similarities between them. However, these were not as similar as the percentages datasets.

The Time over Space analysis resulted in similar findings as to the Space over Time analysis. The positive percentage of tests and the positive percentage of inhabitants were again almost perfectly similar. They stand out because their trajectories include more extreme values compared to the other datasets. The hospitalized cases and the deceased cases are also very similar with less extreme values than the percentage datasets. The sewage dataset does not show a lot of resemblance with the other datasets. The diffusion pattern of the sewage dataset is more chaotic than the other datasets due to it containing more peaks.

After the differences between datasets have been evaluated, the difference between waves was analyzed. To be able to explore the differences between different waves, the following question is asked.

What differences can be found in the patterns of different waves of COVID-19 in the Netherlands with the use of samples of sewage water?

This sub-question was again analyzed with a Time over Space and a Space over Time structure. The Space over Time resulted in little similarity between the three waves. This was to be expected, as the graph of the virus particles in sewage water through time was very different for every wave. This resulted in the maps showing different patterns. The same is true for the Time over Space analysis. The diffusion trajectories showed different patterns. This was again in line with the expectations. The higher peaks showed higher values in the diffusion patterns. To conclude, the researched waves in the sewage water dataset showed little similarity.

Conclusions and Discussion

Based on these results and the findings in the literature, the main question can be answered. A variety of patterns can be found in COVID-19 data in the Netherlands. The patterns that are found are heavily reliant on the datasets used. The positive percentage of tests and the positive percentage of the population both show patterns with high peaks and low minimum values. The COVID-19 particles in sewage, deceased cases, and hospitalized cases show a pattern with moderate peaks. Moreover, the percentage of population and tests, deceased cases, and hospitalized show an infection with COVID-19 at one point in time. The percentage datasets notice an infection earlier in time than both the hospitalized cases and deceased cases. The virus particles in sewage water track infection for a longer period of time, which results in more moderate peaks. Furthermore, the sewage dataset picks up an infection at the same time or even earlier than the percentage of tests or percentage of the population. The selected waves of COVID-19 show few similarities. This conclusion is based on a small sample from the COVID-19 particles in sewage water.

The results are partly in line with the expectations based on the literature used. The differences between the hospitalized cases and the deceased cases relative to the positive percentage of population and tests were expected and showed up clearly in the results. However, it was expected that the sewage dataset was similar to the percentage of population and tests. This was not the case. The sewage dataset showed more similarities to the deceased cases and hospitalized cases.

However, multiple points should be addressed that influenced the research process and therefore the results. Firstly, the comparison between the sewage dataset and the other datasets is influenced by several factors. The dataset is constructed by converting the location of the sewage datasets to municipalities. This is done based on the Euclidean distance between the location of the sewage plants

and the center of municipalities. This is not how these treatment plants function in the real world. Treatment plants have a catchment area, which is a different size for every treatment plant. However, this catchment area is not known, and therefore it is not possible to use it. This construction has influenced the results, but it is unclear to what extent.

The results are also influenced by the clustering methodology. Although the hierarchical clustering showed success in other research (Augustijn & Zurita-Milla, 2013) it was not as successful in this research. The previous research focuses on Iceland, which is an island isolated from the rest of the world with just one major city. The situation in the Netherlands is not comparable to that of Iceland and therefore the results are quite different. The Netherlands is much more densely populated than Iceland and the hierarchy in cities is not as distinct in the Netherlands.

Moreover, the comparison of datasets has influenced the use of hierarchical clustering. Because five datasets are compared, there are five times as many outliers as there would be with the comparison of waves. This is due to every dataset having its own outliers. The hierarchical clustering method is sensitive to outliers and therefore this might not have generated the best possible results.

The elbow method, which was used to determine the number of clusters has proven to be not as successful as anticipated. Most of the time the number of clusters suggested by the elbow method provided results that were not suitable. Better results were generated after differing from the number of clusters as suggested by the elbow method. Therefore, the conclusion can be drawn that the elbow method was not suitable for determining the number of clusters in this specific case.

Although there are areas in which this research could improve upon, the research provides a good basis in the analysis of spatiotemporal patterns of COVID-19 in the Netherlands.

Recommendations for further work

There are multiple aspects of the research that could be improved. In this research, a good start has been made in the evaluation of the different datasets and their implications. In the next research, this could be expanded upon. More advanced data preparation could benefit the results. This preparation could exist for the removal of more outliers. This could be done by only including larger cities, which diminishes the issue that small municipalities are assigned large values due to the data being recorded per 100.000 inhabitants. This provides better results in hierarchical clustering. Moreover, other clustering methodologies could be explored so the small municipalities can still be included. The research showed that hierarchical clustering might not be the most suitable method.

Exploration of other datasets is also a possibility. In this research, the choice was made to only include datasets from official sources provided by the government. However, there are also unofficial sources available, which can be investigated upon. Furthermore, other countries might have different datasets available. It would be interesting to explore those as well.

To conclude, this research provides an insight into the spatiotemporal patterns of COVID-19 data in the Netherlands. The results have proved that Self Organizing Maps provide useful information about COVID-19 data in the Netherlands. With the suggestions provided the research can be expanded upon and provide even more information. Ultimately the Self Organizing Map could be used as an alternative to traditional ways to track COVID-19.

7 Literature

- Algobeans. (2017). *Self-Organizing maps tutorial*. <https://algobeans.com/2017/11/02/self-organizing-map/>
- Andrienko, G., Andrienko, N., Bremm, S., Schreck, T., von Landesberger, T., Bak, P., & Keim, D. (2010). Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum*, 29(3), 913–922. <https://doi.org/10.1111/j.1467-8659.2009.01664.x>
- Assunção, R. M., Reis, I. A., & Oliveira, C. D. L. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine*, 20(15). <https://doi.org/10.1002/sim.844>
- Augustijn, E.-W., & Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International Journal of Health Geographics*, 12(1), 1–14.
- Augustijn-Beckers, E.-W. (2018). *Revealing patterns : spatio-temporal pattern detection and reproduction*. <https://doi.org/10.3990/1.9789036545785>
- Bergwerk, M., Gonen, T., Lustig, Y., Amit, S., Lipsitch, M., Cohen, C., Mandelboim, M., Levin, E. G., Rubin, C., Indenbaum, V., Tal, I., Zavitan, M., Zuckerman, N., Bar-Chaim, A., Kreiss, Y., & Regev-Yochay, G. (2021). Covid-19 Breakthrough Infections in Vaccinated Health Care Workers. *New England Journal of Medicine*, 385(16), 1474–1484. <https://doi.org/10.1056/NEJMoa2109072>
- Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. In *International Journal of Computer Applications* (Vol. 105, Issue 9).
- Birhane, M., Bressler, S., Chang, G., Clark, T., Dorrough, L., Fischer, M., Watkins, L. F., Goldstein, J. M., Kugeler, K., Langley, G., Lecy, K., Martin, S., Medalla, F., Mitruka, K., Nolen, L., Sadigh, K., Spratling, R., Thompson, G., & Trujillo, A. (2021). COVID-19 Vaccine Breakthrough Infections Reported to CDC — United States, January 1–April 30, 2021. *MMWR. Morbidity and Mortality Weekly Report*, 70(21), 792–793. <https://doi.org/10.15585/mmwr.mm7021e3>
- Brownlee, J. (2016, April 18). *Learning Vector Quantization for Machine Learning*. Machine Learning Mastery. <https://machinelearningmastery.com/learning-vector-quantization-for-machine-learning/>
- Burki, T. (2020). Outbreak of coronavirus disease 2019. *The Lancet Infectious Diseases*, 20(3), 292–293.
- Chen, Q., Toorop, M. M. A., de Boer, M. G. J., Rosendaal, F. R., Lijfering, W. M., & Group, L.-C.-19 R. (2020). Why crowding matters in the time of COVID-19 pandemic? - a lesson from the carnival effect on the 2017/2018 influenza epidemic in the Netherlands. *BMC Public Health*, 20(1), 1516. <https://doi.org/10.1186/s12889-020-09612-6>
- Chitungo, I., Dzobo, M., Hlongwa, M., & Dzinamarira, T. (2020). COVID-19: Unpacking the low number of cases in Africa. *Public Health in Practice*, 1, 100038. <https://doi.org/10.1016/j.puhip.2020.100038>
- Chou, Y.-H. (1995). Spatial pattern and spatial autocorrelation. In A. U. Frank & W. Kuhn (Eds.), *Spatial Information Theory A Theoretical Basis for GIS* (pp. 365–376). Springer Berlin Heidelberg.
- Cliff, A. D., Ord, J. K., Haggett, P., & Versey, G. R. (1981). *Spatial diffusion: an historical geography of epidemics in an island community* (Vol. 14). CUP Archive.

- Consuegra, D., Seidner-Isaacs, Y., Larios-Sanjuan, D., Ibarra, J., Benavides-Rodríguez, P., Viloria, S., Buendía, E., & Viasus, D. (2021). Unexpected high frequency of early mortality in <scp>COVID</scp> -19: a single-centre experience during the first wave of the pandemic. *Internal Medicine Journal*, 51(1), 102–105. <https://doi.org/10.1111/imj.15134>
- Covid, C. D. C., Team, R., Covid, C. D. C., Team, R., COVID, C. D. C., Team, R., Bialek, S., Gierke, R., Hughes, M., & McNamara, L. A. (2020). Coronavirus disease 2019 in children—United States, february 12–april 2, 2020. *Morbidity and Mortality Weekly Report*, 69(14), 422.
- Data farmers. (2019). *Sammon mapping A non-linear mapping for data visualization*.
- de Ridder, D., & Duin, R. P. W. (1997). Sammon's mapping using neural networks: A comparison. *Pattern Recognition Letters*, 18(11–13). [https://doi.org/10.1016/S0167-8655\(97\)00093-7](https://doi.org/10.1016/S0167-8655(97)00093-7)
- de Smith, M. J., Goodchild, M. F., & Longley, P. (2007). *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador publishing ltd.
- Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2008). Introduction to Artificial Neural Network. In *Certified International Journal of Engineering and Innovative Technology (IJEIT)* (Vol. 9001, Issue 1).
- el Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In *Machine Learning in Radiation Oncology* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1
- Esri Nederland. (2021). *Gemeenten (Bestuurlijke Grenzen 2021)*. Gemeenten (Bestuurlijke Grenzen 2021)
- Faes, C., Abrams, S., van Beckhoven, D., Meyfroidt, G., Vlieghe, E., Hens, N., Aouachria, A. S., Bafort, K., Belkhir, L., Bossuyt, N., Colombie, V., Dauby, N., de Munter, P., Deblonde, J., Delmarcelle, D., Delvallee, M., Demeester, R., Dugernier, T., Holemans, X., ... Wyndham-Thomas, C. (2020). Time between symptom onset, hospitalisation and recovery or death: A statistical analysis of different time-delay distributions in Belgian COVID-19 patients. *International Journal of Environmental Research and Public Health*. <https://doi.org/10.1101/2020.07.18.20156307>
- Franch-Pardo, I., Napoletano, B. M., Rosete-Verges, F., & Billa, L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. *Science of The Total Environment*, 739, 140033. <https://doi.org/https://doi-org.proxy.library.uu.nl/10.1016/j.scitotenv.2020.140033>
- Galbadage, T., Peterson, B. M., & Gunasekera, R. S. (2020). Does COVID-19 Spread Through Droplets Alone? *Frontiers in Public Health*, 8. <https://doi.org/10.3389/fpubh.2020.00163>
- Gould, P., Kabel, J., Gorr, W., & Golub, A. (1991). AIDS: Predicting the Next Map. *Interfaces*, 21(3), 80–92. <https://doi.org/10.1287/inte.21.3.80>
- Gupta, S., Cantor, J., Simon, K. I., Bento, A. I., Wing, C., & Whaley, C. M. (2021). Vaccinations Against COVID-19 May Have Averted Up To 140,000 Deaths In The United States. *Health Affairs*, 40(9), 1465–1472. <https://doi.org/10.1377/hlthaff.2021.00619>
- Huisman, C. (2021, January 6). *Eerste vaccinatie in Nederland is een historisch moment: de spuiten gaan naar het museum*. <https://www.volkskrant.nl/nieuws-achtergrond/eerste-vaccinatie-in-nederland-is-een-historisch-moment-de-sputen-gaan-naar-het-museum~be685b91/>

- Institute for Health Metrics and Evaluation. (2021, September 29). *COVID-19 Projections*.
<https://Covid19.Healthdata.Org/Global>.
- Ives, A. R., & Bozzuto, C. (2021). Estimating and explaining the spread of COVID-19 at the county level in the USA. *Communications Biology*, 4(1). <https://doi.org/10.1038/s42003-020-01609-6>
- Joanna Roberts. (2021, June 24). *How we prepare for an 'age of pandemics.'* European Commission.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Kashyap, S., Gombar, S., Yadlowsky, S., Callahan, A., Fries, J., Pinsky, B. A., & Shah, N. H. (2020). Measure what matters: counts of hospitalized patients are a better metric for health system capacity planning for a reopening. *Journal of the American Medical Informatics Association*, 27(7), 1026–1131.
- Keijsers, N. L. W. (2010). Neural Networks. In *Encyclopedia of Movement Disorders* (pp. 257–259). Elsevier. <https://doi.org/10.1016/B978-0-12-374105-9.00493-7>
- Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21(1–3). [https://doi.org/10.1016/S0925-2312\(98\)00030-7](https://doi.org/10.1016/S0925-2312(98)00030-7)
- Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, 37, 52–65. <https://doi.org/10.1016/j.neunet.2012.09.018>
- Kuo, F.-Y., Wen, T.-H., & Sabel, C. E. (2018). Characterizing Diffusion Dynamics of Disease Clustering: A Modified Space–Time DBSCAN (MST-DBSCAN) Algorithm. *Annals of the American Association of Geographers*, 108(4), 1168–1186. <https://doi.org/10.1080/24694452.2017.1407630>
- LaFree, G., Xie, M., & Matanock, A. M. (2018). The Contagious Diffusion of Worldwide Terrorism: Is It Less Common Than We Might Think? *Studies in Conflict & Terrorism*, 41(4), 261–280. <https://doi.org/10.1080/1057610X.2017.1290428>
- Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11), 2573–2593.
- Madabhavi, I., Sarkar, M., & Kadakol, N. (2020). COVID-19. A review. *Monaldi Archives for Chest Disease*, 90(2). <https://doi.org/10.4081/monaldi.2020.1298>
- Mallapaty, S. (2020). How sewage could reveal true scale of coronavirus outbreak. *Nature*, 580(7802), 176–177.
- Marini, F. (2009). 3.14 - Neural Networks. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive Chemometrics* (pp. 477–505). Elsevier. <https://doi.org/10.1016/B978-044452701-1.00128-9>
- Mayo Clinic. (n.d.). *Pneumonia*. Retrieved November 28, 2021, from <https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204>

- Melin, P., Monica, J. C., Sanchez, D., & Castillo, O. (2020). Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps. *Chaos, Solitons & Fractals*, *138*, 109917.
- Ministerie van Algemene Zaken. (2020). *February 2020: Eerste coronabesmetting in Nederland. Coronavirus tijdslijn | Rijksoverheid.nl*. . <https://www.rijksoverheid.nl/onderwerpen/coronavirus-tijdslijn/februari-2020-eerste-coronabesmetting-in-nederland>
- Ngoc Thach, N., Bao-Toan Ngo, D., Xuan-Canh, P., Hong-Thi, N., Hang Thi, B., Nhat-Duc, H., & Dieu, T. B. (2018). Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecological Informatics*, *46*, 74–85. <https://doi.org/10.1016/j.ecoinf.2018.05.009>
- Over Morgen. (2018). *AWZI_CO2_emissie_warmteatlas_20160203*. <https://www.arcgis.com/home/item.html?id=ae133e5ef0054a179213d18ecf6691b1>
- Parisi, D. R., Mariani, M. C., & Laborde, M. A. (2003). Solving differential equations with unsupervised neural networks. *Chemical Engineering and Processing: Process Intensification*, *42*(8–9), 715–721. [https://doi.org/10.1016/S0255-2701\(02\)00207-6](https://doi.org/10.1016/S0255-2701(02)00207-6)
- Peccia, J., Zulli, A., Brackney, D. E., Grubaugh, N. D., Kaplan, E. H., Casanovas-Massana, A., Ko, A. I., Malik, A. A., Wang, D., & Wang, M. (2020). SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. *MedRxiv*.
- Pitzer, V. E., Chitwood, M., Havumaki, J., Menzies, N. A., Perniciaro, S., Warren, J. L., Weinberger, D. M., & Cohen, T. (2021). The Impact of Changes in Diagnostic Testing Practices on Estimates of COVID-19 Transmission in the United States. *American Journal of Epidemiology*, *190*(9), 1908–1917. <https://doi.org/10.1093/aje/kwab089>
- Pollakowski, H. O., & Ray, T. S. (1997). Housing Price Diffusion Patterns at Different Aggregation Levels: An Examination of Housing Market Efficiency. *Journal of Housing Research*, *8*(1), 107–124. <http://www.jstor.org.proxy.library.uu.nl/stable/24833634>
- R project. (n.d.). *The R project for statistical computing*.
- Ritchie, H. (2020). *Coronavirus Pandemic (COVID-19) - Statistics and Research*. <https://ourworldindata.org/coronavirus#note-1>
- Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hassel, J., Macdonald, B., Beltekian, D., & Roser, M. (2020). Coronavirus (COVID-19) Deaths. *OurWorldInData*.
- RIVM. (n.d.-a). *Covid-19 Nationale SARS-CoV-2 Afvalwatersurveillance*.
- RIVM. (n.d.-b). *Ontwikkeling COVID-19 in grafieken*.
- RIVM. (2021, October 5). *Ontwikkeling COVID-19 in grafieken*. <https://Www.Rivm.Nl/Coronavirus-Covid-19/Grafieken>.
- Rosenberg, E. S., Holtgrave, D. R., Dorabawila, V., Conroy, M., Greene, D., Lutterloh, E., Backenson, B., Hoefler, D., Morne, J., Bauer, U., & Zucker, H. A. (2021). New COVID-19 Cases and Hospitalizations

- Among Adults, by Vaccination Status — New York, May 3–July 25, 2021. *MMWR. Morbidity and Mortality Weekly Report*, 70(37), 1306–1311. <https://doi.org/10.15585/mmwr.mm7037a7>
- Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus disease (COVID-19)—Statistics and research. *Our World in Data*, 4.
- RTL. (2020). *Corona in cijfers*. <https://www.rtlnieuws.nl/nieuws/nederland/artikel/5069661/corona-cijfers-besmettingen-doden-intensive-care-rivm-hoe-veel>
- Sammon, J. W. (1969). A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18(5), 401–409. <https://doi.org/10.1109/T-C.1969.222678>
- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- Schærström, A. (2009). Disease Diffusion. In *International Encyclopedia of Human Geography*. Elsevier. <https://doi.org/10.1016/B978-008044910-4.00330-8>
- Schechtman K. (2021, May 19). *How Lagging Death Counts Muddled Our View of the COVID-19 Pandemic*. The Covid Tracking Project. <https://covidtracking.com/analysis-updates/how-lagging-death-counts-muddled-our-view-of-the-pandemic>
- Schutte, S., & Weidmann, N. B. (2011). Diffusion patterns of violence in civil wars. *Political Geography*, 30(3). <https://doi.org/10.1016/j.polgeo.2011.03.005>
- Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315.
- Singh, J. (2005). Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Management Science*, 51(5). <https://doi.org/10.1287/mnsc.1040.0349>
- Stoecklin, S. B., Rolland, P., Silue, Y., Mailles, A., Campese, C., Simondon, A., Mechain, M., Meurice, L., Nguyen, M., & Bassi, C. (2020). First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. *Eurosurveillance*, 25(6), 2000094.
- Tian, J., Azarian, M. H., & Pecht, M. (2014). *Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm*. <https://doi.org/https://doi.org/10.36001/phme.2014.v2i1.1554>
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1), 234–240.
- Utsch, A. (1990). Kohonen's self organizing feature maps for exploratory data analysis. *Proc. INNC90*, 305–308.
- Vekaria, B., Overton, C., Wiśniowski, A., Ahmad, S., Aparicio-Castro, A., Curran-Sebastian, J., Eddleston, J., Hanley, N. A., House, T., Kim, J., Olsen, W., Pampaka, M., Pellis, L., Ruiz, D. P., Schofield, J., Shryane, N., & Elliot, M. J. (2021). Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. *BMC Infectious Diseases*, 21(1), 700. <https://doi.org/10.1186/s12879-021-06371-6>

Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical Medicine & International Health*, 25(3), 278–280. <https://doi.org/10.1111/tmi.13383>

World Health Organization. (2020a). *Cancer trends in the African Region*. WHO.Com. . <https://whotogo-whoafroccmaster.newsweaver.com/JournalEnglishNewsletter/dy0ovpaxeqty48iiujdam4?a=2&p=56444925&t=31103707>

World Health Organization. (2020b). *COVID-19 - virtual press conference*. https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-press-conference-full-30mar2020.pdf?sfvrsn=6b68bc4a_2

World Health Organization. (2020c). *WHO COVID-19: Case Definitions*.

Yang, X., Yu, Y., Xu, J., Shu, H., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., & Yu, T. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*, 8(5), 475–481.