Treasuring Waste: Predicting Covid Cases From Sewage

Jannick Akkermans, Sander van Gestel, Lauke Stoel, & Annemarie Timmers

Survey Data Analysis (201300001)

Under the supervision of Dr. P. Lugtig

Utrecht University

Treasuring Waste: Predicting Covid Cases From Sewage

In December 2019, a virus known as SARS-Cov-2 (COVID-19) was identified in Wuhan, China. This new variant of the SARS coronavirus caused a worldwide pandemic with far-reaching consequences. An important factor contributing to the silent spread of the virus is the fact that 20 to 40% of the patients show no symptoms, (Vallejo et al., 2020).

Although patients do not always show symptoms of the virus, they often do excrete RNA particles of the virus in their faeces as shown by for example (Pan, Zhang, Yang, Poon, & Wang, 2020). The virus can sustain itself for a long period within the faeces, in some cases even for one or more months after the respective patient has tested negative for COVID-19 in their respiratory samples (Vallejo et al., 2020). Therefore, the amount of RNA particles could be a valuable indicator of the true number of COVID-19 patients within a country or municipality. This information gives an indication of the true number of contagious people at any given moment. This is important, because it could give us a meaningful indication of the severity of the pandemic on a given moment. Since the number of positive tests on its own is not as informative, given contextual information, like testing capacity and willingness, is not considered alongside (Silver, 2020).

Ever since the start of the pandemic in the Netherlands, the National Institute for Public Health and the Environment (RIVM) has been collecting samples from sewage treatment plants (STPs) and testing them for RNA presence. Based on these data, we aim to evaluate whether RNA particles are an adequate predictor for the true number of covid cases. We work towards this goal through answering the following sub-questions:

1. How does the cumulative number of reported cases in the Netherlands as a whole develop over time?

2. What does this trend look like in the different security regions?

3. What is the mean level of RNA particles found in the water per 100,000 inhabitants?

4. How do RNA particles and the increase in the number of cases relate to each other?

The remainder of this report is structured as follows. Methods used by the RIVM to measure RNA and prior research in this field are discussed in Section 2. Section 3 describes the data and the challenges they provided. The methodology used to answer the subquestions, as well as the results are discussed in Section 4. Section 6 interprets, aggregates the results and subsequently provides a conclusion. Finally, Section 5 discusses certain things that could have gone better in this research.

2 Literature Review

As the aim of this project is to assess the predictive quality of RNA of the true number of positive COVID-19 cases, literature regarding RNA as a predictor of positive cases and PCR, by which RNA estimates are retrieved, should be reviewed. The following section therefore dives deeper into the usefulness of RNA particles as a predictor of COVID-19 cases, the sampling method used to retrieve sewage from the STPs, and the method by which the number of particles is determined.

2.1 Relevance of measuring RNA in sewage water

Early on in the pandemic, the RIVM began sampling wastewater. One of the reasons to do this, was the use of wastewater samples to detect diseases in the past. For example, wastewater has been used to detect and monitor the spread of polio since the 1980s, with the World Health Organization (WHO) issuing guidelines to do this (Center for Disease Control and Prevention, 2020; Mao et al., 2020).

Moreover, RNA was found to be present in the faeces of both symptomatic and asymptomatic carriers, hereby being able to capture both types of cases (Randazzo et al., 2020). Because of this, RNA can serve as an indicator for rises and falls in cases, independent of the number of positive tests. In addition, as RNA is measured locally, it can not only serve as a national, but also a regional indicator of the true number of cases. Altogether, wastewater sampling provides some valuable opportunities to improve the detection and monitoring of the spread of COVID-19.

2.2 Sampling of sewage data

The research into the sewage by the RIVM started small, with only 29 out of the 355 STPs in April 2020, but since the beginning of September of the same year, all STPs in the Netherlands are sampled once or multiple times a week. After the samples are taken, they are transported at a controlled temperature to the RIVM, where they are analysed by researchers for RNA particles. RNA is isolated and Polymerase Chain Reaction (PCR) is performed on the samples to determine the amount of RNA particles present in the wastewater (RIVM, 2020c). This technique is explained in the next section.

Multiple equations result in an estimate of the number of RNA particles per 100,000 inhabitants of the Netherlands, which was made possible by mapping the number of households connected to a STP. In these equations, the RIVM corrects for the amount of water that flows into the STPs (Rijksoverheid, 2020a). This is needed, because when it has rained, this results in more water in the sewage. More water would lower the concentration of RNA particles in the wastewater, hereby possibly distorting the estimate. The resulting numbers are presented on the Corona Dashboard (Rijksoverheid, 2020b).

2.3 The PCR method

As we have just established that measuring RNA is important, it is also important to understand the method by which the measuring is done. As mentioned in the introduction, sewage samples are transported at a controlled temperature to the RIVM. There, researchers isolate the RNA of the virus and perform a technique called PCR in order to determine the number of RNA in the sewage. PCR follows a three step process which consists of (1) denaturation of double-stranded DNA, (2) annealing of primers, and (3) primer extension (Schochetman, Ou, & Jones, 1988). Denaturation of double-stranded DNA involves separating the two strands that together form the complex DNA-sequence.

In the next step, primers are added to both strands. A primer is a single strand that is complementary to the DNA-sequence it is attached to. The new combinations of DNA strands are then synthesized together to form new DNA sequences. This process is iterated until there is not enough primer left to form new sequences. Afterwards, the amount of DNA is measured to provide the total amount of DNA in a given sample.

The amount of RNA is determined through an adaptation of this technique, called Reverse Transcription-Polymerase Chain Reaction (RT-PCR). The procedure is as follows according to Rio (2014): first, a primer is added to the RNA strand. This new synthesized RNA-DNA combination is then used as a template for Reverse Transcriptase, in which a single-stranded cDNA copy is created. However, this cDNA strand is only a proportion of the original RNA strand. The newly created strand is finally used in the PCR method to determine the amount of RNA particles in the population (i.e. the sewage of a security region in the Netherlands).

Although the PCR method can produce results relatively fast according to Garibyan and Avashia (2013), it does have a few limitations. First of all, PCR is a highly sensitive technique, so any contamination of the sample can lead to misleading results. Secondly, PCR depends on the addition of a primer to create new DNA sequences. The drawback in this case is that the creation of the primer requires prior knowledge of the target sequence you are attempting to create (Garibyan & Avashia, 2013). Therefore, PCR can only be applied to known pathogens or genes.

2.4 RNA as a predictor of COVID cases

Once the number of RNA has been measured, it could be used as a predictor of the true number of COVID-19 cases. However, there is something that needs to be taken into account when using this predictor. Peccia et al. (2020) attempted to track the spread of COVID-19 cases in Connecticut by measuring RNA flow in wastewater. They found that RNA concentrations in wastewater were six to eight days ahead of the corresponding reports of positive tests. Therefore, RNA cannot be compared to the number of positive tests on the same day, but rather, with a multiple day time-lag.

Another problem with the current data on RNA in sewage, is that it is unknown how much RNA particles need to be shed, for them to be detected during testing (Center for Disease Control and Prevention, 2020). This complicates prediction, as some RNA may go undetected, resulting in an underestimate of the true number of cases in the Netherlands as a whole (or a single region, for that matter).

Finally, when taking a sample of the sewage, the concentration of RNA that it contains may be lower than the original concentration in the sewage. This may be due to the physical and chemical properties of the environment the RNA is detected in (Lahrich et al., 2020). For example, some RNA particles may die due to the temperature of the wastewater or because of too much sunlight exposure. Since the RNA concentration can decrease, we need to be careful when using it as a predictor of the number of COVID-19 cases.

3 Data

3.1 Structure of the data

We are working with two datasets, both collected by the RIVM. The first dataset contains the total number of positive tests reported per day per municipality (RIVM, 2020a). It also contains information on key characteristics of the municipality, such as population density and which security region it is part of. The second dataset contains data recorded on the level of STPs (RIVM, 2020b). The key variable here is the average concentration of SARS-CoV-2 RNA measured in the daily amount of sewage per 100,000 inhabitants. This dataset also contains crucial metadata of the STPs, such as in which security region the area of responsibility of this STP falls. The two datasets were matched by the variable in which security region a municipality and a STP's area fall respectively.

3.2 Challenges in the data

Our goal is to assess the usefulness of RNA flow in sewage as a predictor of the true number of COVID-19 patients in the Netherlands at any given day. To do this, we first have to establish if there is a relationship between the RNA flow and the number of positive tests per day. Given the data structure, we are faced with a few challenges before we can take on this question.

3.2.1 Level of measurement. First of all, the data on the number of positive tests are recorded on the level of municipalities, whereas the data on RNA flow are recorded on the level of the STPs. In the most straightforward cases, we can aggregate the RNA flow data to municipality level data by virtue that the datasets were already matched by security region code. See SR1 of Figure 1.

However, some STPs also treat water from outside their primary security region, creating double entries in the dataset and making simple matching impossible. Fortunately, the dataset also provides information on what percentage of the water a STP processes comes from which security region, so we can weight the RNA flow by this variable. See SR2 of Figure 1.

Furthermore, some very large municipalities produce so much sewage in one day that multiple STPs are required to process it, causing a second kind of double entries. See SR3 of Figure 1. Unfortunately, there are no data available that specify how much of the water from these large municipalities goes to which treatment plant, making it impossible to establish if there is a relationship between the RNA flow and the number of positive tests on the municipality level.

We conclude that we have to aggregate our data to the security region level when establishing a relationship between RNA flow and number of positive tests, to get an accurate indicator.

3.2.2 Consistency of measurement. Another challenge of this dataset is that the way the RNA flow in the sewage is recorded changed during the period time frame that

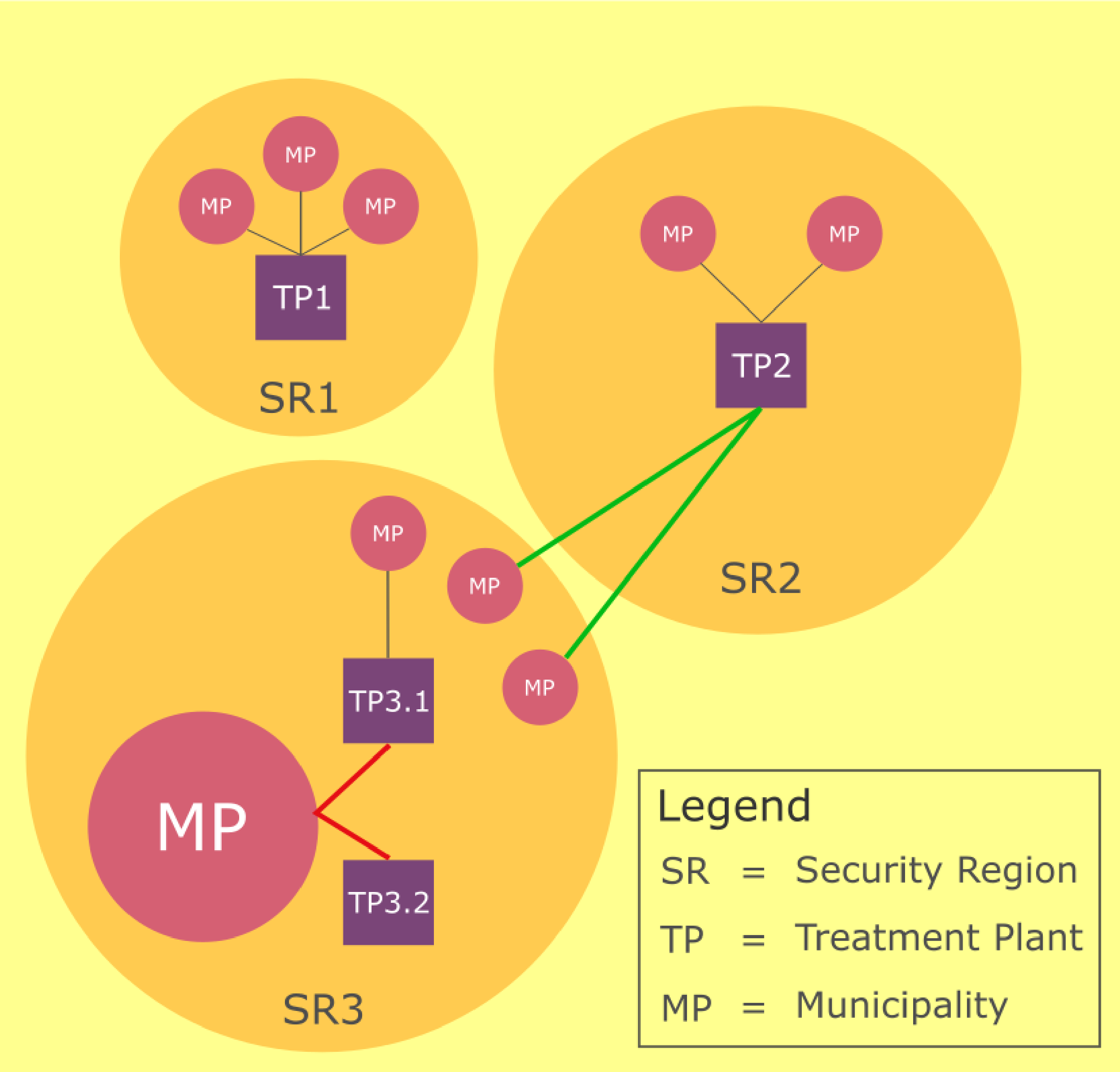

Figure 1. Illustration depicting the structure of the data.

data were collected. Up until the 7th of September, the RNA flow was measured as the average concentration of SARS-CoV-2 RNA per mL of unfiltered sewage. From the 7th of September onwards, it was recorded as the average concentration of SARS-CoV-2 RNA measured in the daily amount of sewage, per 100,000 inhabitants. This last measurement corrects for differences in the number of people an STP treats sewage for.

Moreover, it was not until the 7th of September that all STPs in the Netherlands started participating in reporting the average RNA flow. Furthermore, not every STP reports this measure daily or even consistently.

We decided to use a subset of the original dataset in determining the usefulness of RNA flow as a predictor of the true number of COVID patients, to foster internal consistency of measurement. The subset we used spanned from the 7th of September to the 11th of November, when we downloaded these data. We deal with the problem of missing data in the next section.

3.2.3 Missing data. This study suffers from a missing data problem, because not all STPs report their RNA flow on a daily basis or at a consistent frequency (i.e. once a week on set days). As a result, not every security region has data on RNA flow for each observation of reported cases seven days later. We need these combinations to assess the usefulness of the RNA flow as a predictor of the true number of cases. Since this problem lies at the core of our analysis, we decided to impute the missing RNA flow data.

We elected the Last Observation Carried Forward (LOCF) imputation method as best fitting to the scope of the project. LOCF is a simple imputation method that imputes missing data for a unit of observation by assigning the last observed value for that unit of observation to each missing value (Cook, Zeng, & Yi, 2004). We can regard the non-reporting by one STP for a period of time as attrition from a longitudinal study, until their next report. We deem it fitting to carry their last report forward, because we can assume that people who suffer from COVID-19 and excrete RNA particles into the sewage water will do so for at least as long as their symptoms last. Therefore, the change in the RNA flow that the STPs report can be expected to be gradual and the last observed value can be regarded as a reasonable placeholder for the actual value.

Although this solves the problem of missing data, the method does have some limitations. First of all, it is a conservative imputation method according to Streiner (2008). This means that - given that during the time span of our dataset, the number of COVID-19 cases has only increased - it likely underestimates the true RNA flow values, as this technique assumes that a unit of observation does not change at the times of missing values. Due to this underestimation, LOCF introduces some bias in the estimates. Nevertheless, it would still provide better estimates than if we used only the observed cases for the analyses. Second, LOCF has an ad hoc nature, which means that it has no measure of uncertainty concerning the imputed values (Kunzmann et al., 2020). This implies that no additional covariates can be included to reduce bias introduced by this imputation method. We further reflect on the implications of our choice of imputation method on our results in the discussion section.

4 Methodology and Results

In this section, the different subquestions are dealt with consecutively. First, the methodology that was used to answer the specific question is described. The results are displayed and discussed immediately afterwards. As to not clutter the flow of the text, we decided to display some of the less informative, though insightful figures in the appendix. It is mentioned in the text when this is indeed the case, after which the main takeaway of the figure is described shortly.

4.1 How does the cumulative number of reported cases in the Netherlands as a whole develop over time?

In order to answer the first question, we needed a function that calculated the cumulative number of reported cases per day. Therefore, we wrote a function that aggregated the data by the number of reported cases in a municipality on a given day to the cumulative number of reported cases on that day in the Netherlands as a whole, discarding double entries per municipality. The resulting output is visualized in Figure 2, where you can see the cumulative number of reported cases in the Netherlands over time. The start of the second wave is also clearly visible, where the cumulative number of reported cases increases more rapidly in October compared to September.

We were also interested in the daily fluctuations of the number of reported cases in the Netherlands. To show this, we altered the function such that it would calculate the difference in cumulative number of reported cases between each day and the day before. The results are visualized in Figure 3. We see that the peak of new cases in the second wave is somewhere at the end of October, as we would expect from the previous figure.

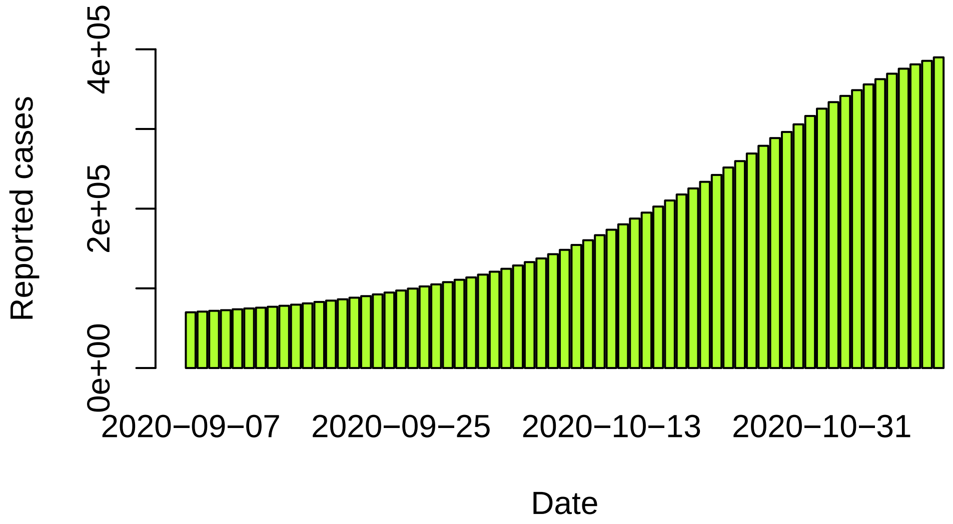

Figure 2. Cumulative number of reported cases per day in the Netherlands over time.

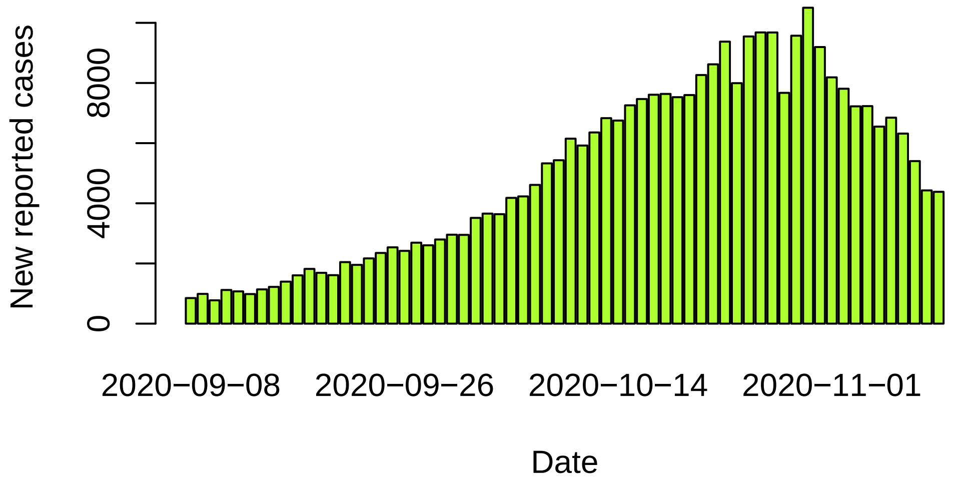

Figure 3. Number of new reported cases per day in the Netherlands over time

4.2 What does this trend look like in the different security regions?

Next, we examined if this trend differed between the various security regions in the Netherlands. We calculated both the cumulative number of reported cases per day for each of the 25 security regions. To be able to make a meaningful comparison between the different regions, we calculated this indicator per 100,000 inhabitants, correcting for differences in population density between security regions. See Figure A1 in Appendix A for a map of the Netherlands indicating where all municipalities are located, what security region they belong to, and how many cases are reported in each municipality. The higher concentration of municipalities and number of reported cases in for example the Randstad area supports our choice to calculate the cumulative number of cases in each security region per 100,000 inhabitants. We show the resulting graphs in Figure 4.

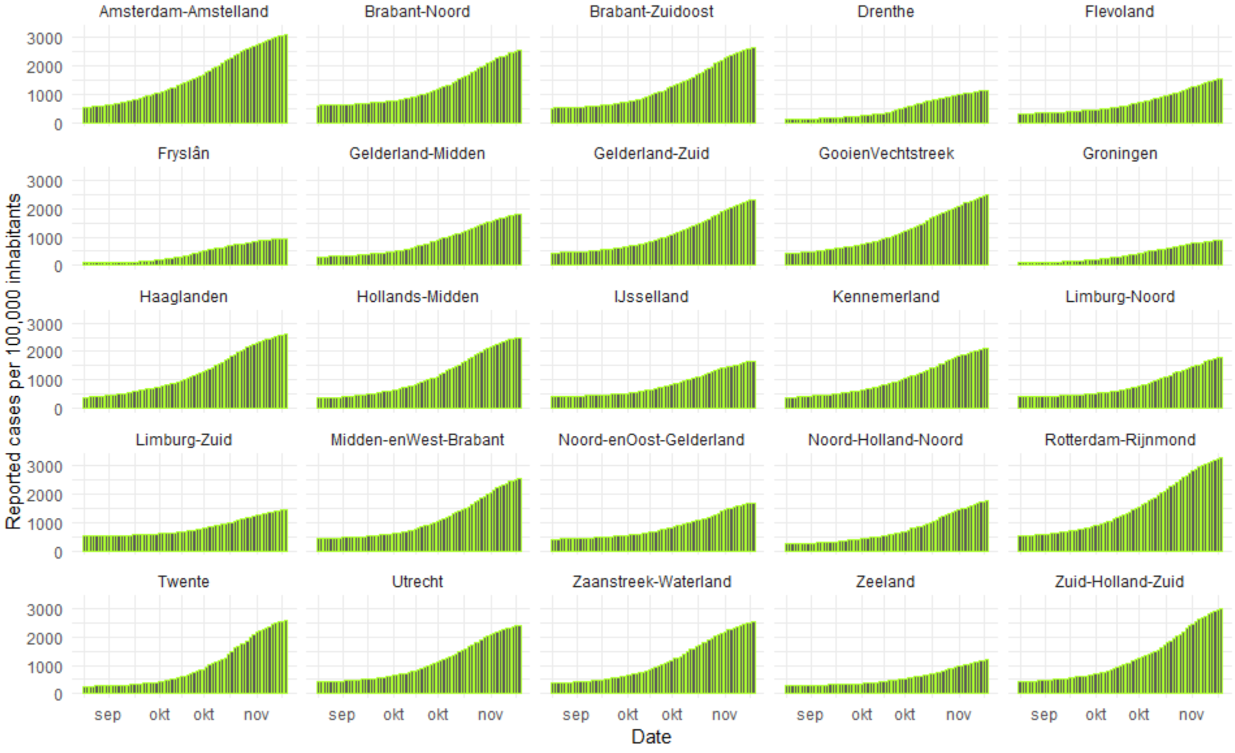

Figure 4. Cumulative number of cases per 100,000 inhabitants per security region

As can be seen in Figure 4, the number of cases per 100,000 inhabitants varies per security region both in magnitude and in development over time. For example, the number of cases in Amsterdam-Amstelland is much higher than in Zeeland. Moreover, the last recorded number of cases in Amsterdam-Amstelland is comparable to that of Rotterdam-Rijnmond, but the increase in cases had a much later onset and increased at a higher pace in the latter compared to the former.

To be able to establish the predictive value of RNA for the total number of cases per region, we also needed to calculate the difference in the number of cases per region per day. This is because we expect the amount of RNA particles in the water not to have a strong relationship to the cumulative number of cases ever recorded, but to relate more closely to the change in the number of cases. We show the results of our calculations in Figure 5. Here we see confirmed that the number of cases peaked at different moments in the various regions.

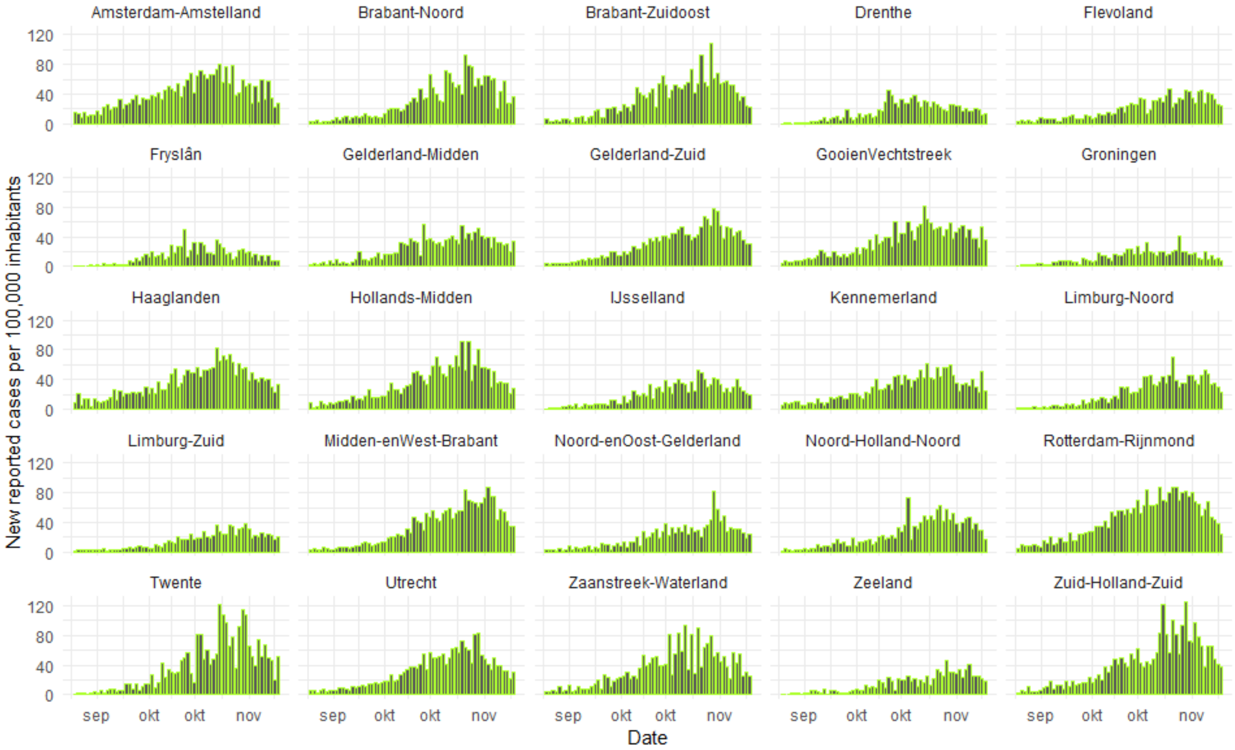

Figure 5. New number of reported cases per 100,000 inhabitants per security region

4.3 What is the mean level of RNA particles found in the water per 100,000 inhabitants?

As we mentioned in the introduction, the level of RNA particles in the sewage is measured by taking a sample from a STP. Figure 6 shows the locations of the STPs and shows the security region where the STP is located. The figure does not show from which security region the STP processes sewage. For example, in the security region Noord-Holland-Noord there are ten STPs.

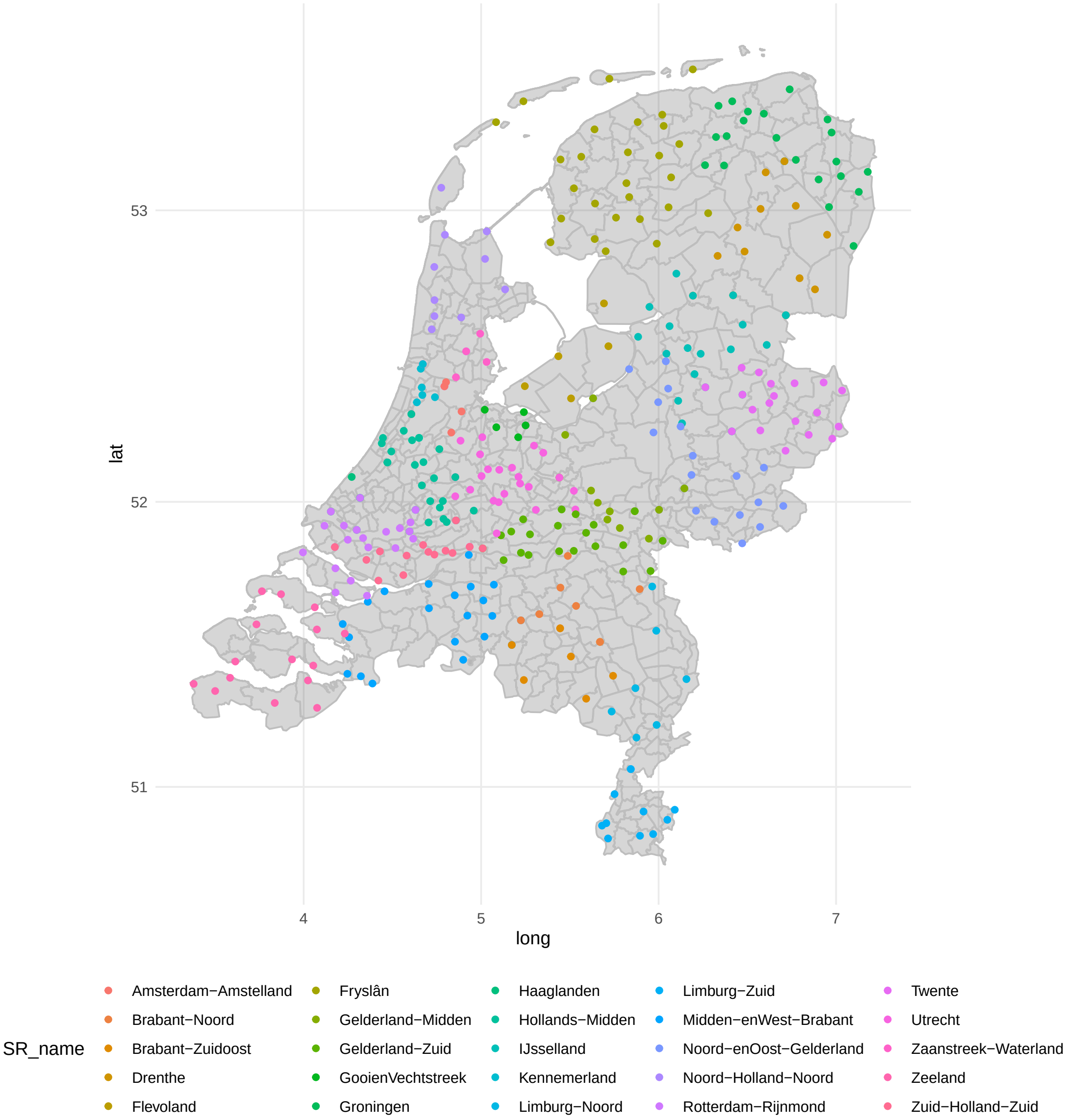

Figure 6. Treatment plants in the Netherlands

We saw in Figure 3 that the number of cases had been increasing since the beginning of September and peaked in October. As one only requests a test after experiencing symptoms, one would expect the amount of RNA particles in the sewage to precede the increase in the number of cases. To be able to compare the RNA flow with the daily increase, we first visualized the distribution of the RNA particles in the Netherlands over time, which can be seen in Appendix B. The shape of this plot loosely follows the same trend as Figure 3 displaying the increase in reported cases per day, but its shape is much less smooth. This could be due to the fact that the number of STPs that reported differs each day, which can be seen in Figure B2 in Appendix B. To correct for this we also plotted the mean level of RNA particles in the sewage per day in Figure 7.

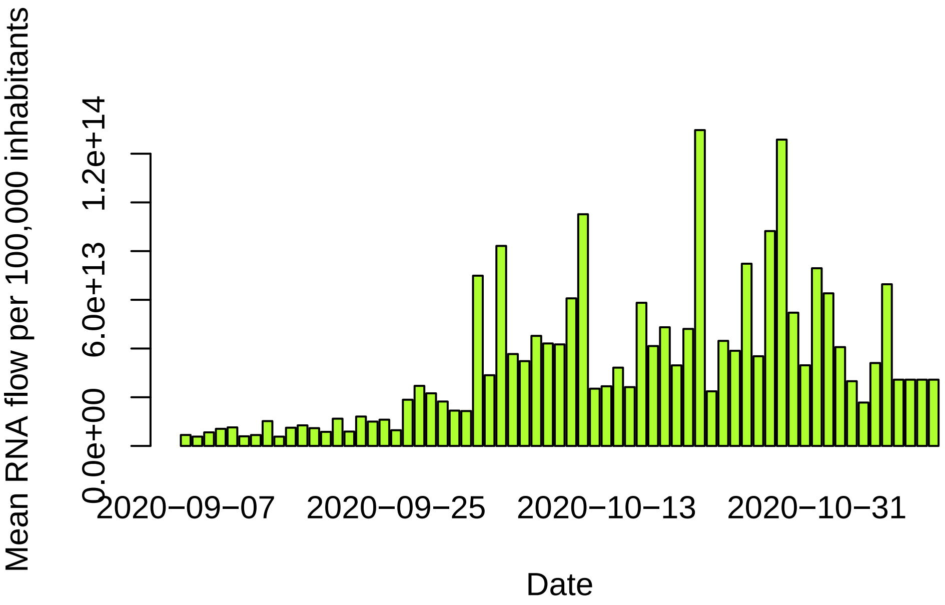

Figure 7. Mean RNA flow per 100,000 inhabitants in the Netherlands per day

The trend in this plot is smoother than the plot that showed the totals. However, there is still some fluctuation. If the same amount of RNA particles were found in all installations, you would not expect fluctuations. In that case the mean level would not be affected by a missing installation. This could be due to differing amounts of RNA particles found in different installations. Since different installations care for different regions, different regions may have different trends. In order to explore this, we plotted the mean amount of RNA particles found in the sewage per day for each region.

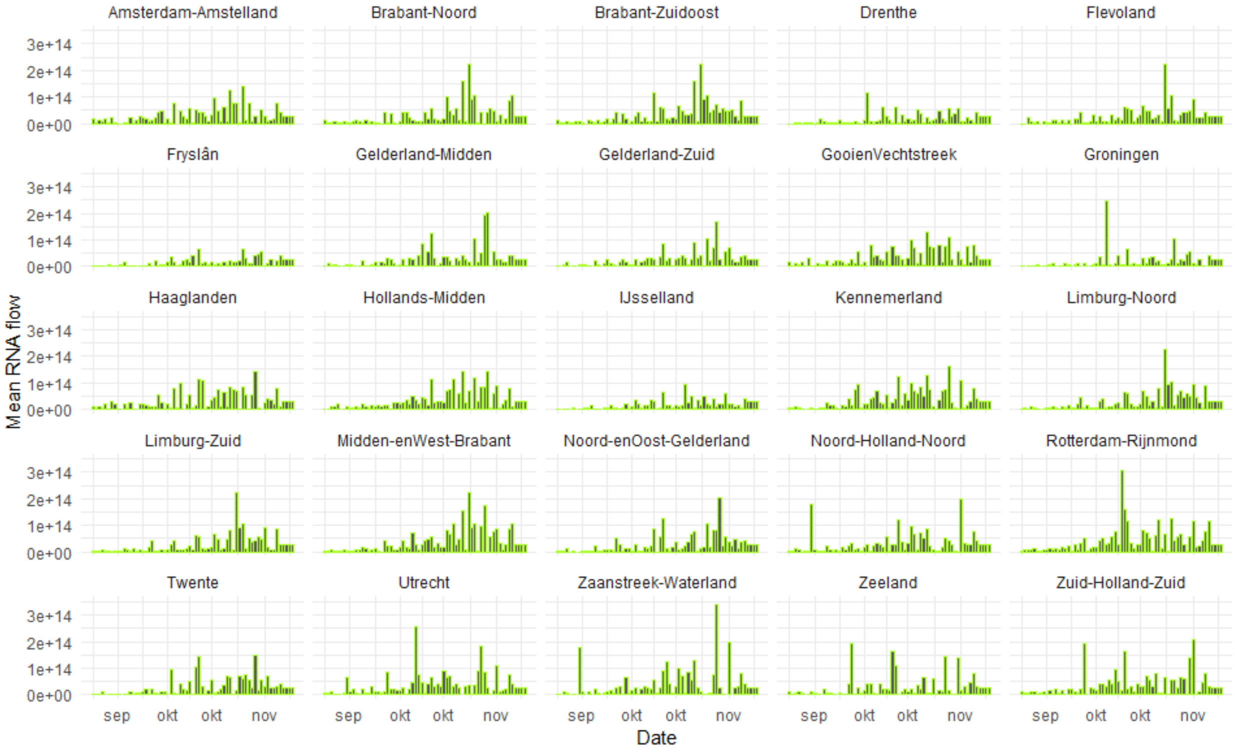

Figure 8. Mean RNA flow per 100,000 inhabitants per security region per day

As can be seen in Figure 8, the trend of the amount of RNA particles found differs per security region. For example, the amount of RNA particles found in Friesland is lower than in Midden- en West-Brabant. Also, in Friesland there is not really a peak, whereas there is a clear peak in Midden- en West-Brabant. The trends per region look similar to the overall trend. They loosely follow the trend of the reported cases, but there are still fluctuations.

4.4 How do RNA particles and the increase in the number of cases relate to each other?

As mentioned in the data description, some STPs process water from multiple security regions, rendering the current RNA flow variable unrepresentative. Another variable is available that contains the proportion of water from the security region processed by that respective STP. By multiplying these two, a representative RNA flow variable was created.

After weighting the RNA flow, we inspected how the RNA flow related to the increase in the number of cases on the level of the security regions. To achieve this, we altered the current functions for calculating the RNA flow and increased the number of cases per day so that these values are reported for each security region per day. Subsequently, we compared the RNA flow from one day to the increase in the number of reported cases from seven days later. The reason for this is that there is generally a seven-day time lag between a person contracting COVID – leaving RNA particles in the sewage water – and getting a positive test (Peccia et al., 2020).

Figures C1 and C2 in Appendix C show the correlation between the RNA flow on the 8th of September and the increase in the number of cases on the 15th of September and the rest of the days in the dataset, respectively. These figures displayed vastly differing scattered points, which was to be expected as we had established previously that the security regions displayed differences, both in number of new cases and mean RNA flow. We proceeded to plot the RNA flow and increase in cases separately for each security region. Then, we displayed the RNA flow on a given day and the increase in cases seven days later as a dot on a scatterplot. Doing this for all days in the dataset enabled us to calculate correlations between mean RNA flow and the increase in the number of cases for each security region. The resulting scatterplots and correlations are displayed in Figure 9.

As can be seen in the figure, the correlations vary, ranging from .19 in Groningen to .63 in Hollands-Midden. Overall, the results indicate a positive association between mean RNA flow and increase in number of cases. When the number of RNA particles in the sewage increases, so do the number of new cases. However, as the size of the correlation coefficient differs between the security regions, RNA flow cannot be aggregated to a single predictor and used to estimate the number of cases in the Netherlands as a whole in one go. We elaborate on the implications of this finding in the conclusion.

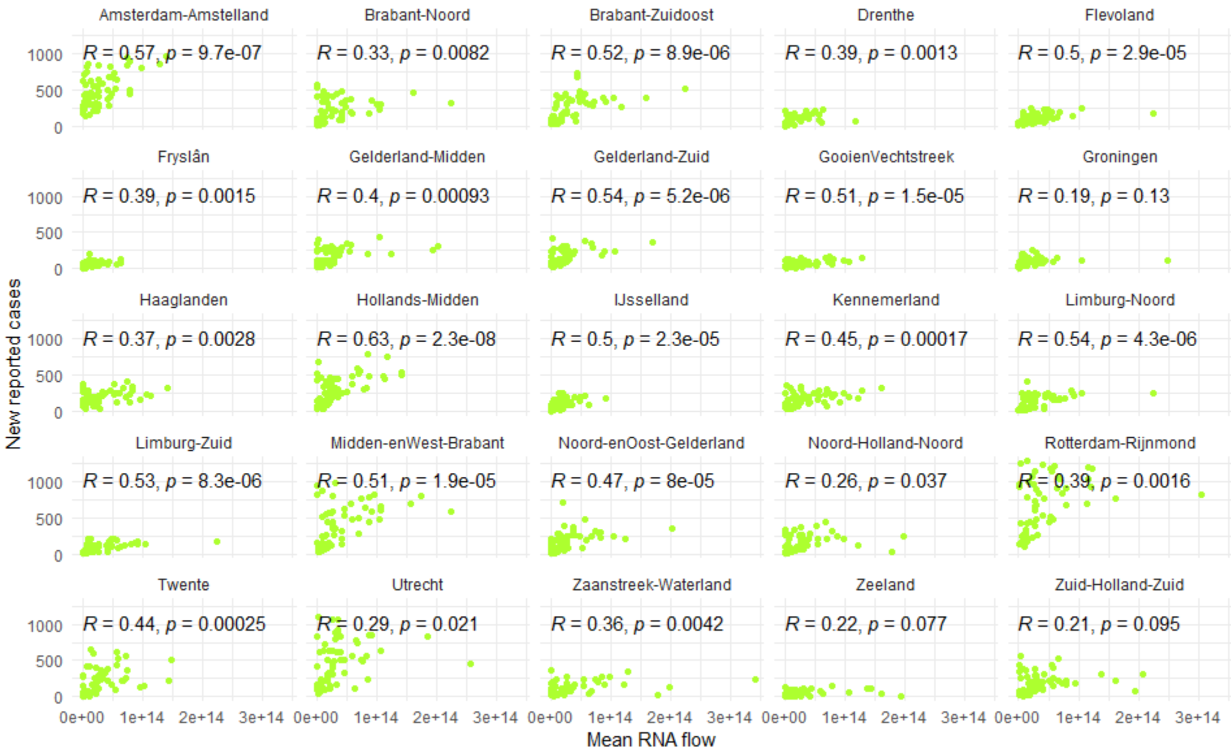

Figure 9. Relation between RNA flow and number of covid infections

5 Discussion

When we started this project, our aim was to build a model around RNA particles in the sewage in order to estimate the true number of COVID-19 cases in the Netherlands. After our initial exploratory analyses, we soon realized that RNA flow is a volatile indicator and we concluded that building a model within the scope of this project would be unrealistic. We therefore shifted our focus to understanding how the amount of RNA particles in the sewage and COVID-19 cases are related. We regard this report as a stepping stone to a model that can predict the number of COVID-19 cases.

As any research, our research has limitations. Our original dataset contained many missing values. We dealt with this by applying LOCF. This method has several drawbacks. As mentioned in the data-section, LOCF is a conservative method according to Streiner (2008). Applied to our results, this means that the estimates of the amount of RNA particles tend to maintain the status quo, and our results are less sensitive to changes between reports. A second problem with the method is that it does not provide a measure of uncertainty concerning the imputed values (Kunzmann et al., 2020). In our dataset, there are some STPs that report more frequently and consistently, and thus would suffer less bias as a result of the LOCF method. On the other hand, some STPs will suffer from this bias more significantly. The LOCF method does not enable us to quantify this difference in bias, meaning that we were unable to account for this in our analyses.

A second limitation of our study is that we discarded three outliers. In our analysis of the relationship between COVID-19 cases and RNA particles in the sewage we stumbled upon extreme values. As these values had such a different order of magnitude in comparison to all other data points, we suspect they were due to administrative errors and we decided to exclude them. Though, it is possible that these outliers were in fact not due to an administrative error and contained relevant information, but we are unable to form a definitive conclusion. Therefore, a correlation plot including these outliers is added in Figure C3 in Appendix C.

Moreover, we are still unable to explain why the trend of RNA particles in the sewage within each region fluctuates, where we would expect the curve to display a smooth trend. One possible explanation could be that, as indicated by the RIVM, people who contracted COVID-19 excrete RNA particles at different rates (RIVM, 2020c). Not every new case would therefore contribute the same increase in RNA particles to the sewage. Similarly, we also could not explain why different regions have different correlations between RNA particles in the sewage and COVID-19 cases. This could be a consequence of the fluctuations in the development of RNA particles over time. Another possible explanation could be that we chose one set time interval to calculate the correlations over, namely seven days. However, the time between contracting COVID-19 and excreting RNA into the sewage can vary greatly on an individual basis. Further research could explore this by calculating the correlation for multiple time intervals and taking their average.

Lastly, the correlations we found between RNA flow and the number of new

COVD-19 cases are relatively low. This could be explained by the fact that there is not a one-on-one relationship between RNA flow in the sewage and the change in number of cases. The RNA particles in the sewage represent the number of people that have COVID-19 at that point in time, not just the people who newly contracted the virus seven days later. To appropriately describe the relationship between the RNA particles and the true number of COVID-19 cases, we would need additional data on at which rates people that contracted the virus recover.

6 Conclusion

This project aimed to assess whether RNA flow is an adequate predictor of the true number of COVID-19 cases in the Netherlands. Based on the results in Section 4, we conclude that RNA flow could be a useful predictor, but that it is very volatile and that it does not have enough explanatory value on its own. The main reason for this conclusion is that the correlation between RNA flow and the increase in reported cases is unstable. It fluctuates over time and differs across regions as shown by figures C2 and 9, respectively . As a result, it would be a massive undertaking to build one model based on the RNA flow data to estimate the true number of COVID-19 cases in the Netherlands as a whole at any given time, because the strength of the correlation would be different for each region and for each time stamp. Nevertheless, the correlation in each of the security regions is positive and most of the correlations are moderate. This indicates for every security region that as the RNA flow increases, the increase in cases also rises. Therefore, there are grounds to believe that RNA flow can be a useful predictor for the true number of COVID-19 cases.

However, it is not a strong enough predictor by itself. As the RIVM stated, only approximately 40% of infected people leave traces of the virus RNA in the sewage water (RIVM, 2020c). Additionally, some infected people secrete higher concentrations of the RNA than others. Hence we conclude that RNA is too unstable a measure to be the only predictor of the true number of cases. It should be combined with other useful predictors to create better estimates of the true number of COVID-19 cases.

If we had infinite time and resources, we would like to build such a model and enrich it with other types of data. For example, we could use behavioural data on when people are more or less likely to get tested when displaying symptoms and include data on at which rates people recover from COVID-19, to improve the estimate of the true number of infected people. Another unexplored option is to - in addition to separating the analysis into different security regions - also control for population density on the municipal level. For now, we conclude that the RNA flow is an imperfect indicator of the true number of infected people in the Netherlands, but that it could potentially be very valuable in combination with other data sources.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Center for Disease Control and Prevention. (2020). National wastewater surveillance system (nwss). Retrieved December 15, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/wastewater-surveillance.html

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on locf imputation. *Biometrics*, *60*(3), 820–828.

Garibyan, L., & Avashia, N. (2013). Research techniques made simple: Polymerase chain reaction (pcr). *The Journal of Investigative Dermatology*, *133*(3), e6.

Kunzmann, K., Wernisch, L., Richardson, S., Steyerberg, E. W., Lingsma, H., Ercole, A., … Wilson, L. (2020). Imputation of ordinal outcomes: A comparison of approaches in traumatic brain injury. *Journal of Neurotrauma*.

Lahrich, S., Laghrib, F., Farahi, A., Bakasse, M., Saqrane, S., & El Mhammedi, M. (2020). Review on the contamination of wastewater by covid-19 virus: Impact and treatment. *Science of the Total Environment*, *751*, 142325.

Mao, K., Zhang, K., Du, W., Ali, W., Feng, X., & Zhang, H. (2020). The potential of wastewater-based epidemiology as surveillance and early warning of infectious disease outbreaks. *Current Opinion in Environmental Science & Health*.

Pan, Y., Zhang, D., Yang, P., Poon, L. L., & Wang, Q. (2020). Viral load of sars-cov-2 in clinical samples. *The Lancet Infectious Diseases*, *20*(4), 411–412.

Peccia, J., Zulli, A., Brackney, D. E., Grubaugh, N. D., Kaplan, E. H., Casanovas-Massana, A., … others. (2020). Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. *Nature Biotechnology*, *38*(10), 1164–1167.

Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simón, P., Allende, A., & Sánchez, G. (2020). SARS-cov-2 rna in wastewater anticipated covid-19 occurrence in a low prevalence area. *Water Research*, 115942.

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rijksoverheid. (2020a). Cijferverantwoording. Retrieved December 15, 2020, from https://coronadashboard.rijksoverheid.nl/verantwoording#rioolwater

Rijksoverheid. (2020b). Dashboard coronavirus. Retrieved December 15, 2020, from https://coronadashboard.rijksoverheid.nl/

Rio, D. C. (2014). Reverse transcription–polymerase chain reaction. *Cold Spring Harbor Protocols*, *2014*(11), pdb–prot080887.

RIVM. (2020a). Covid-19 aantallen per gemeente per publicatiedatum [data file]. Retrieved November 10, 2020, from https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/5f6bc429-1596-490e-8618-1ed8fd768427

RIVM. (2020b). Covid-19 nationale sars-cov-2 afvalwatersurveillance [date file]. Retrieved November 10, 2020, from https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/a2960b68-9d3f4dc3-9485-600570cd52b9?tab=general

RIVM. (2020c). Rioolwateronderzoek. Retrieved December 10, 2020, from https://www.rivm.nl/coronavirus-covid-19/onderzoek/rioolwater

Schochetman, G., Ou, C.-Y., & Jones, W. K. (1988). Polymerase chain reaction. *The Journal of Infectious Diseases*, *158*(6), 1154–1157.

Silver, N. (2020). Coronavirus case counts are meaningless. *FiveThirtyEight. April, 4.*

Streiner, D. L. (2008). Missing data and the trouble with locf. *Evidence-Based Mental Health*, *11*(1), 3–5.

Vallejo, J. A., Rumbo-Feal, S., Conde-Pérez, K., López-Oriona, Á., Tarrío, J., Reif, R., ...

others. (2020). Highly predictive regression model of active cases of covid-19 in a population by screening wastewater viral load. *MedRxiv*.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Appendix A

Map of the Netherlands displaying municipalities, securite regions and number of cases

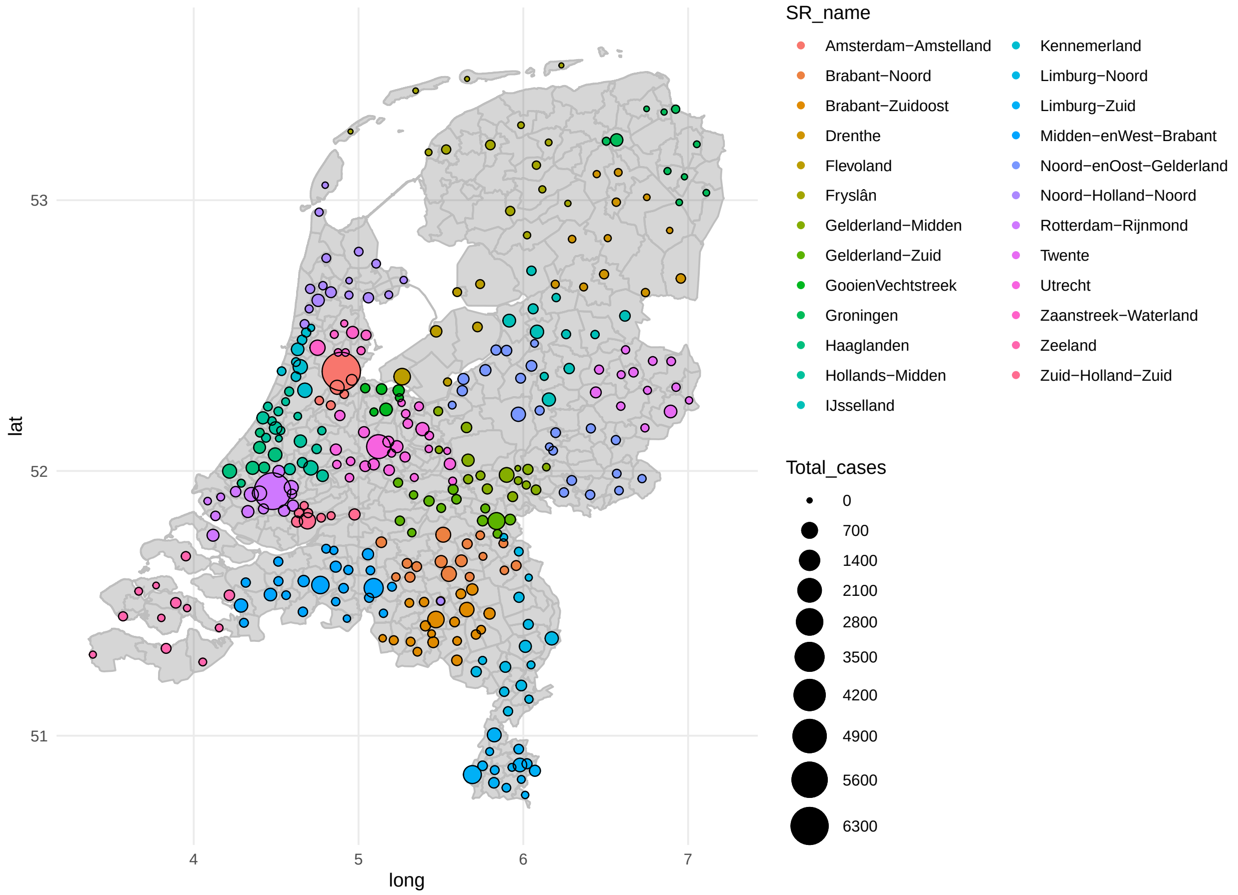

Figure A1. Municipalities and security regions in the Netherlands

Appendix B

Extra plots for subquestion three

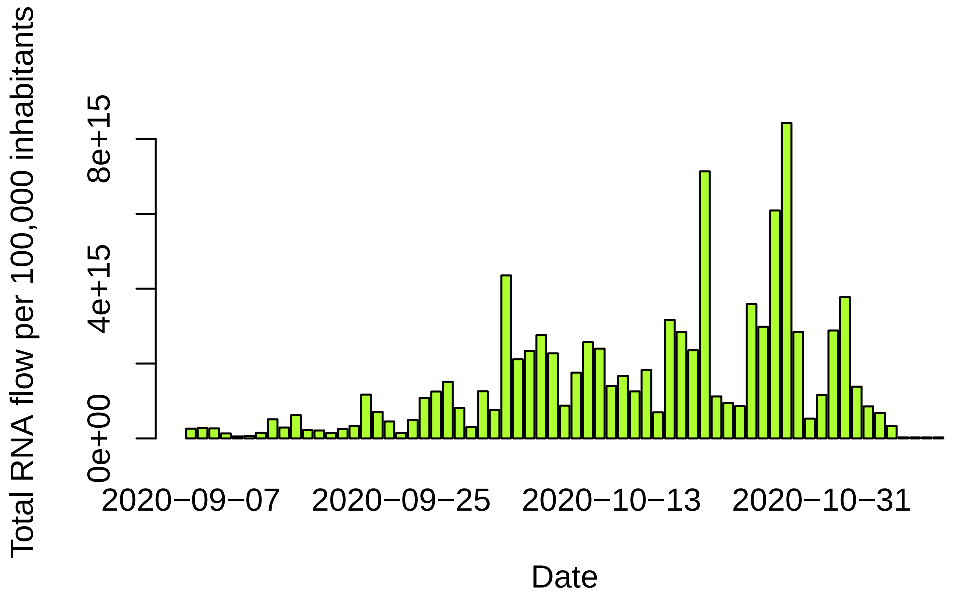

Figure B1. Total RNA flow per 100,000 inhabitants in the Netherlands per day

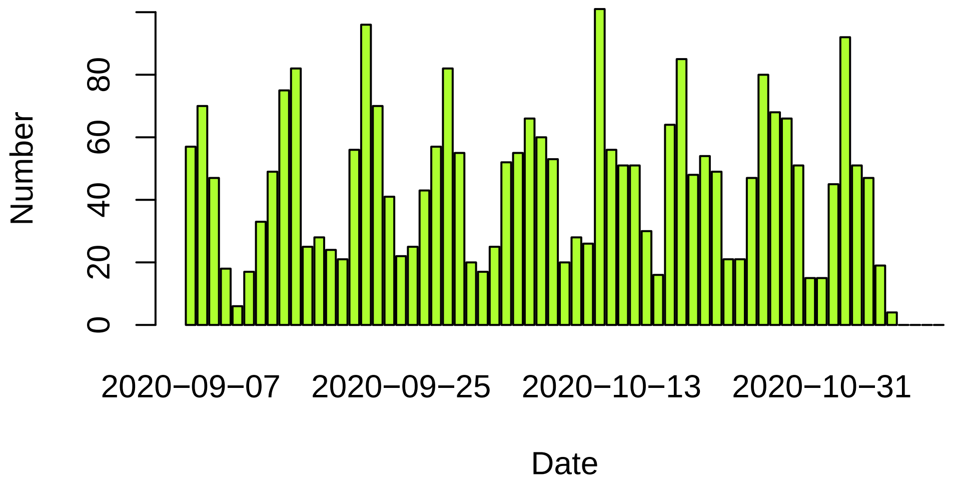

Figure B2. Number of installations that have provided data per day

Appendix C

Extra plots for subquestion four

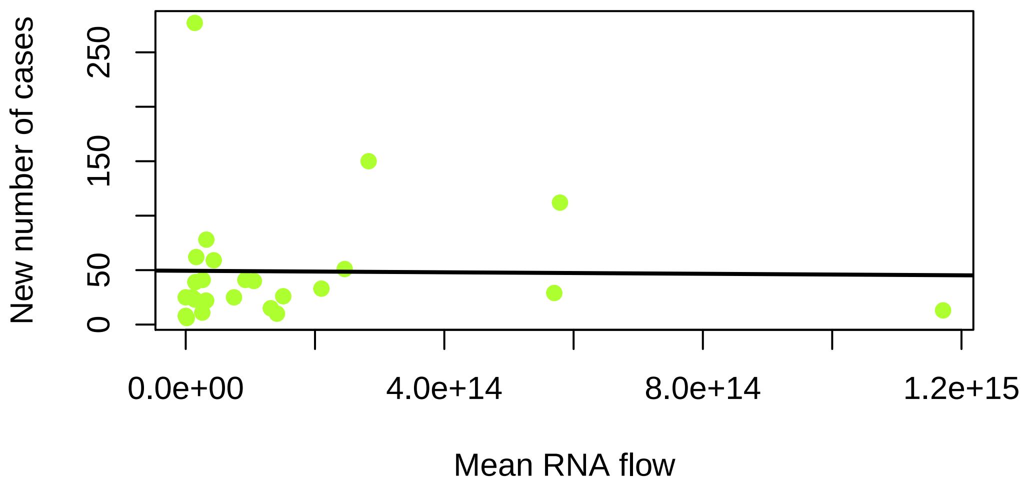

Figure C1. Relation between RNA flow and new number of reported cases for one timestamp

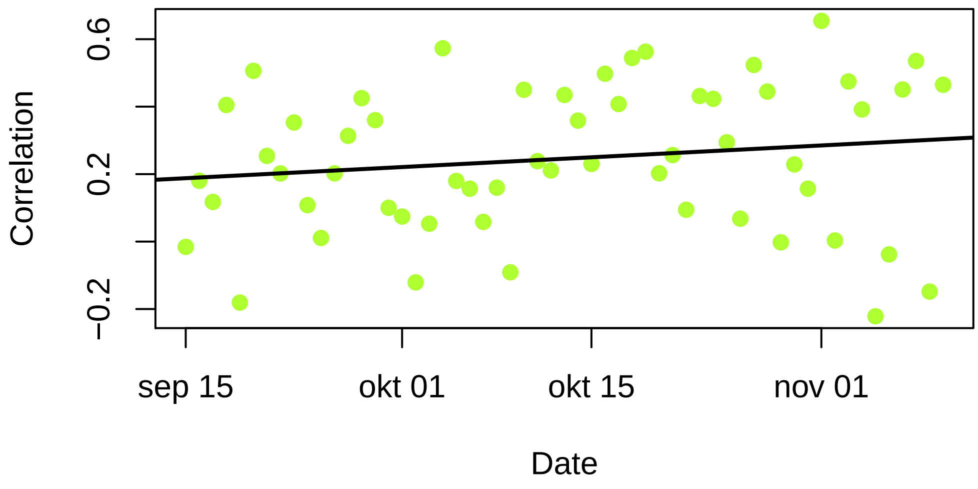

Figure C2. Correlations between RNA flow and new number of cases over time

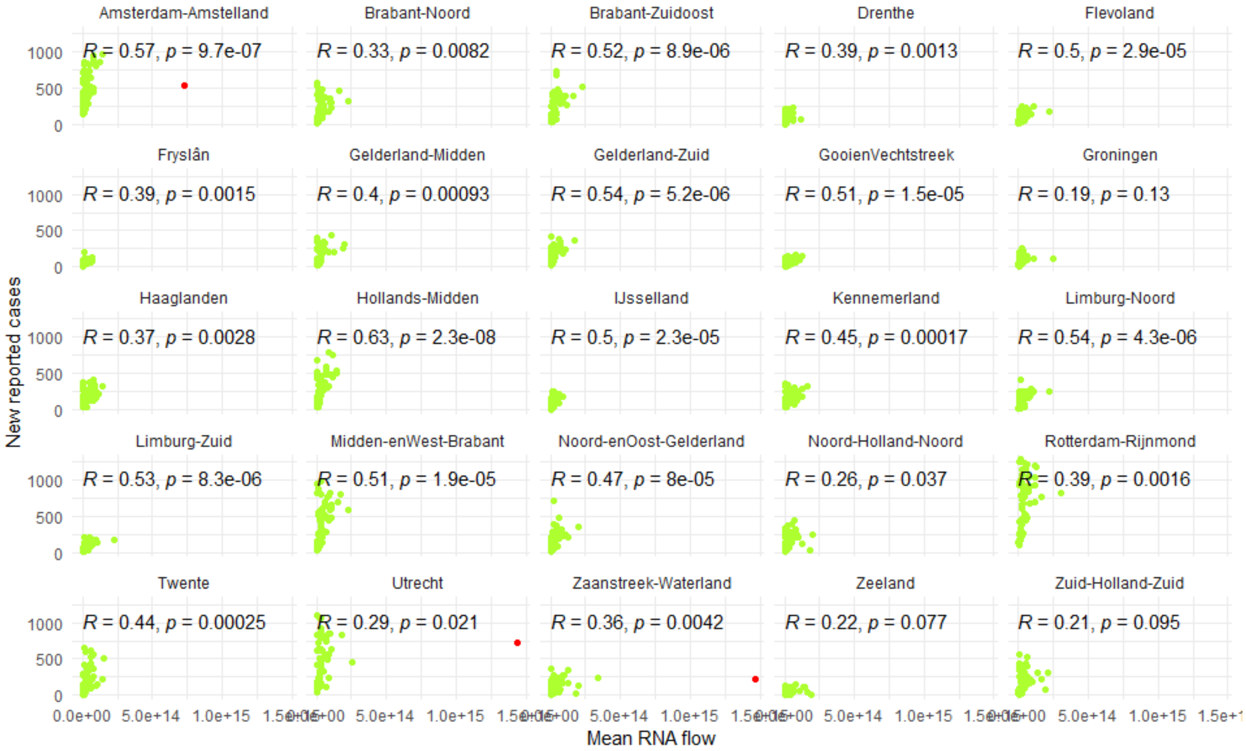

Figure C3. Correlation between RNA flow and new number of cases per region with outliers