Treasuring Waste: Predicting Covid Cases From Sewage

Jannick Akkermans[1], Sander van Gestel[1], Lauke Stoel[1], & Annemarie Timmers[1]

[1] Utrecht University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

Treasuring Waste: Predicting Covid Cases From Sewage

In December 2019, a virus known as SARS-Cov-2 (COVID-19) was initiated in Wuhan, China. This variant of the SARS coronavirus, which shocked the world in 2003, caused a worldwide pandemic with many consequences. However, this virus is more dangerous since 20 to 40% of the patients show no symptoms, contributing to the silent spread of the virus (Vallejo et al., 2020).

Although patients do not always show symptoms of the virus, they do leave RNA particles of the virus in their feces as shown by for example Pan, Zhang, Yang, Poon, and Wang (2020). The virus can sustain itself for a long period of time within the feces, in some cases even one or more months after the respective patient has tested negative for RNA particles in their feces (Vallejo et al., 2020). Therefore, the amount of RNA particles could be an indicator of the true number of COVID-19 patients within a country or security region. Given this information, this project attempted to build a model that estimates the true number of new positive cases in a security region on a given day based on the RNA flow seven days prior.

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. The research started small, with only 29 out of the 355 STPs in April, but since the beginning of September 2020, 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).

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).

The remainder of this report is structured as follows. Section 2 will discuss prior research in the field of measuring RNA. Section 3 describes the data and the analyses conducted in this research. The results of those analyses are described in Section 4. Section 5 interprets those results and subsequently provides a conclusion. Finally, Section 6 discusses certain things that could have gone better in this research.

2 Literature Review

As the aim of this project is to estimate the true number of positive corona tests in a security region from the RNA in the wastewater, 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 corona cases 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 feces of both symptomatic and asymptomatic carriers, hereby being able to capture both types of infections (Randazzo et al., 2020). Because of this, RNA can serve as an indicator for rises and falls in infections, 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 total number of infections. Altogether, wastewater sampling provides some valuable opportunities to improve the detection and monitoring of the spread of COVID-19.

2.2 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 water 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 water.

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 water 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. Nevertheless, this method has its advantages and is currently used by the RIVM.

2.3 RNA as a predictor for COVID infections

Once the number of RNA has been measured, it could be used as a predictor for the number of COVID infections. 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-infections 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 one-on-one, but rather, with a multiple day time-lag.

Another problem with the current data on RNA in sewage water, 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 total number of infections in the Netherlands as a whole (or a single region, for that matter).

Finally, when taking a sample of the sewage water, the concentration of RNA that it contains may be less than the original concentration in the sewage water. This may be due to the type of the environment or the physical and chemical properties of the environment (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 be lowered, we need to be careful when using it as a predictor to estimate the number of COVID infections.

3 Data and Methodology

3.1 Structure of the data

We are working with two datasets, both collected by the Dutch National Institute for Public Health and the Environment (RIVM). The first dataset contains the total number of positive tests reported per day per municipality. 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 sewage treatment plants. The key variable here is the average concentration of SARS-CoV-2 RNA measured in the daily amount of sewage water per 100,000 inhabitants. This dataset also contains crucial metadata of the sewage treatment plants, such as in which security region the area of responsibility of this treatment plant falls. The two datasets were matched to each other by the variable in which security region a municipality and a treatment plant's area fall respectively.

3.2 Challenges in the data

Our goal is to estimate the true number of COVID-19 patients at any given day, based on the data we have available describing the RNA flow in the sewage water. 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.

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 treatment plant level. 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.

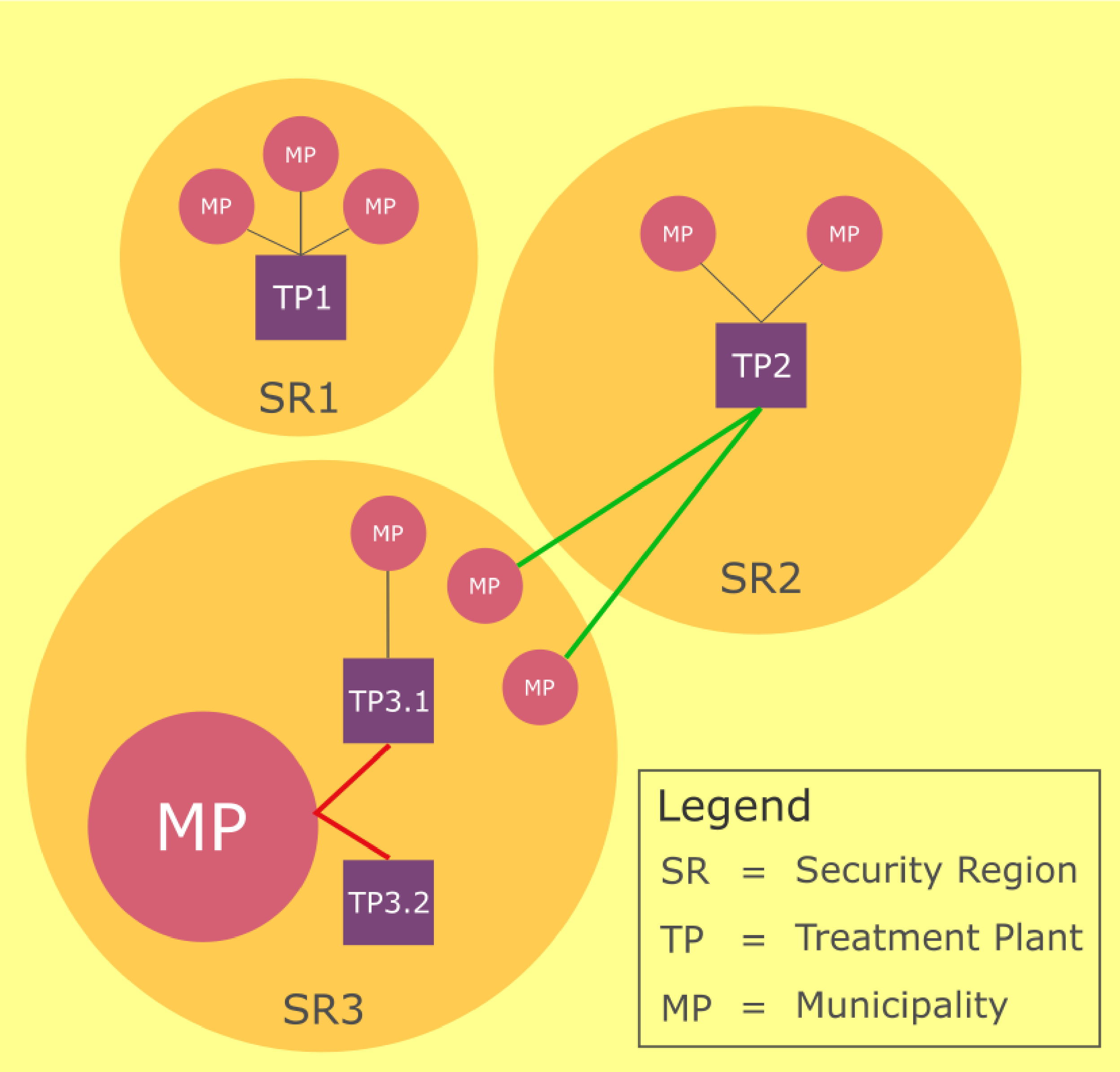

Figure 1. Illustration depicting the structure of the data.

However, some sewage treatment plants also treat water from outside their primary security region, creating double entries in the dataset and making simple matching impossible. Luckily, the dataset also provides information on what percentage of the water a treatment plant 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 water in one day, that multiple treatment plants 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 cases on the municipality level.

We take the above relationships into account in our analyses. Furthermore, 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 cases, to get an accurate indicator. Lastly, the data we use in our analyses are a subset of the original dataset, because the dataset only includes all sewage treatment plants from the 7th of September onwards.

We work towards our goal through asking and answering several sub-questions. The following section contains the analysis for answering the following questions:

1. How does the total number of reported infections in the Netherlands as a whole develop over time? (Section 3.3)

2. What does this trend look like displayed per 100,000 inhabitants of the Netherlands? (Section 3.4)

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

4. What is the relationship between RNA particles and the total number of infections on one given moment?

5. How does this relationship look over time?

3.3 How does the total number of reported infections in the Netherlands as a whole develop over time?

In order to answer this question we need a function that calculates the number of total reported infections per day. We wrote a function to aggregate the data by the number of reported infections in a municipality on a given day to the total number of reported infections 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 total number of reported infections we have had over time in the Netherlands. The second wave is also clearly visible, where you see the total number of reported infections increasing more rapidly in October compared to September.

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

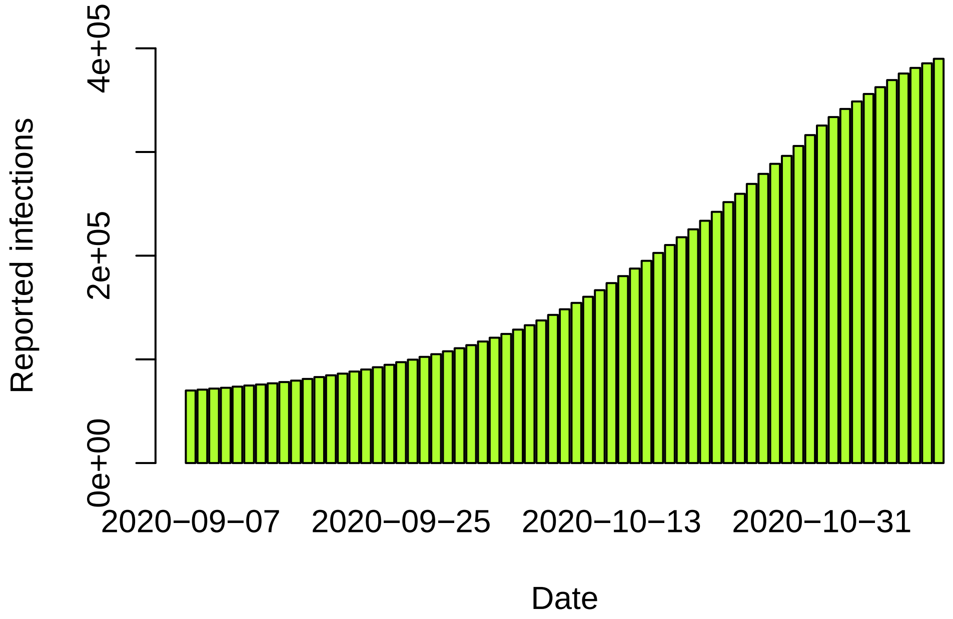

Figure 2. Cumulative number of reported infections per day in the Netherlands over time.

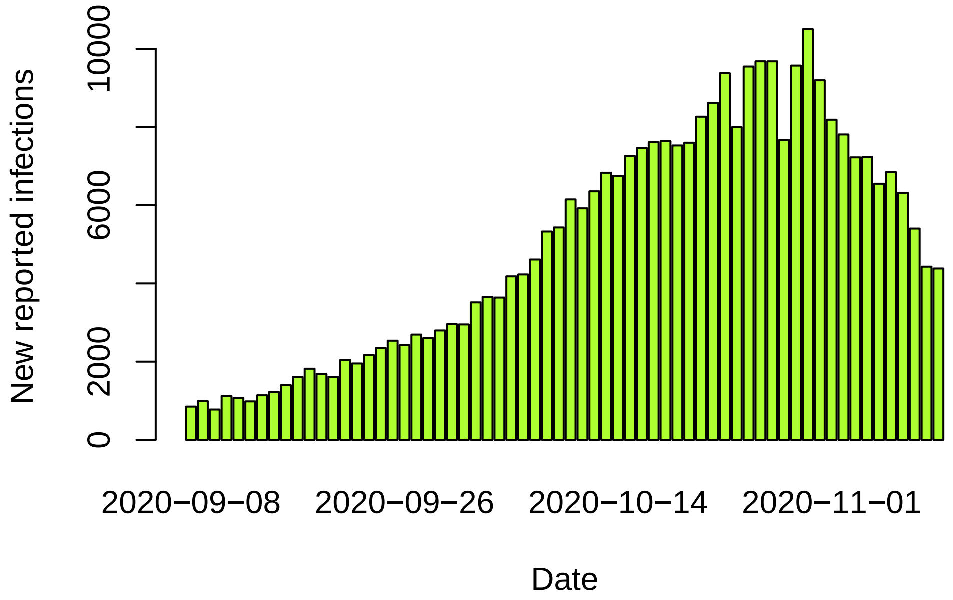

Figure 3. Number of new reported infections per day in the Netherlands over time

3.4 What does this trend look like displayed per 100,000 inhabitants of the Netherlands?

Next, we visualized the number of reported cases per 100,000 inhabitants of the Netherlands. Again, we decided to do this for both the total number of infections we have had over time, as well as the daily differences. As you can see in the figures below, the shape of the distribution is the same, but the scale on the y-axis is different. This is expected, as we took the functions of the previous step and multiplied the total number of reported infections by the number of inhabitants of the municipality divided by 100,000.

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

We saw that the number of infections 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 water to precede the increase in the number of infections. Therefore, we have visualized the distribution of the RNA particles in the Netherlands over time. In the figure below, the mean number of RNA particles per 100,000 inhabitants summed over all sewage water installations for every day is plotted. The shape of the plot loosely follows the same trend as the plot of the reported infections per day, but its shape is much less smooth.

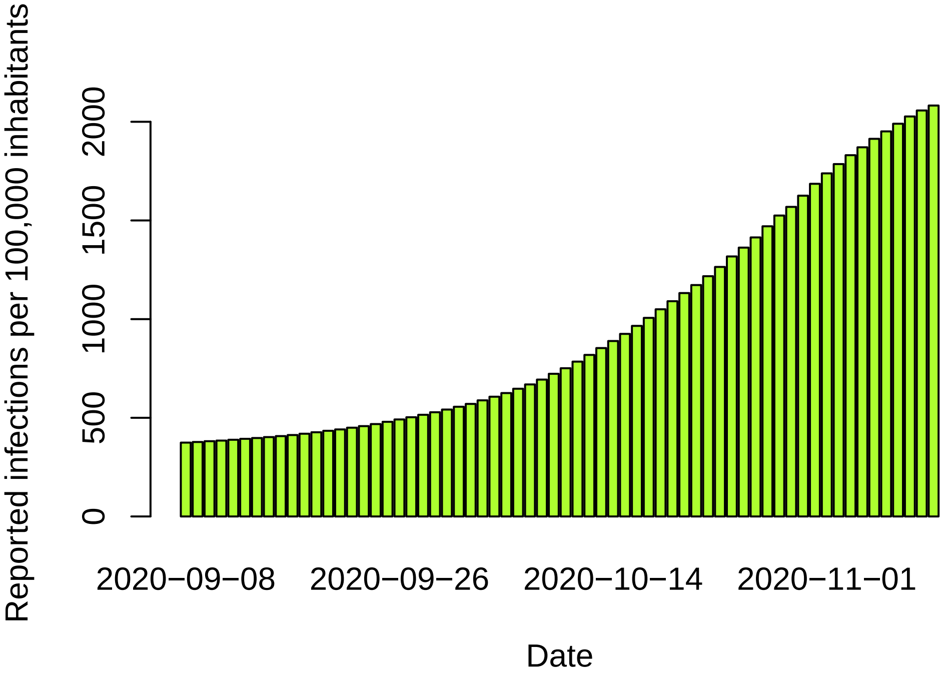

Figure 4. Cumulative number of reported infections per day per 100,000 inhabitants in the Netherlands over time

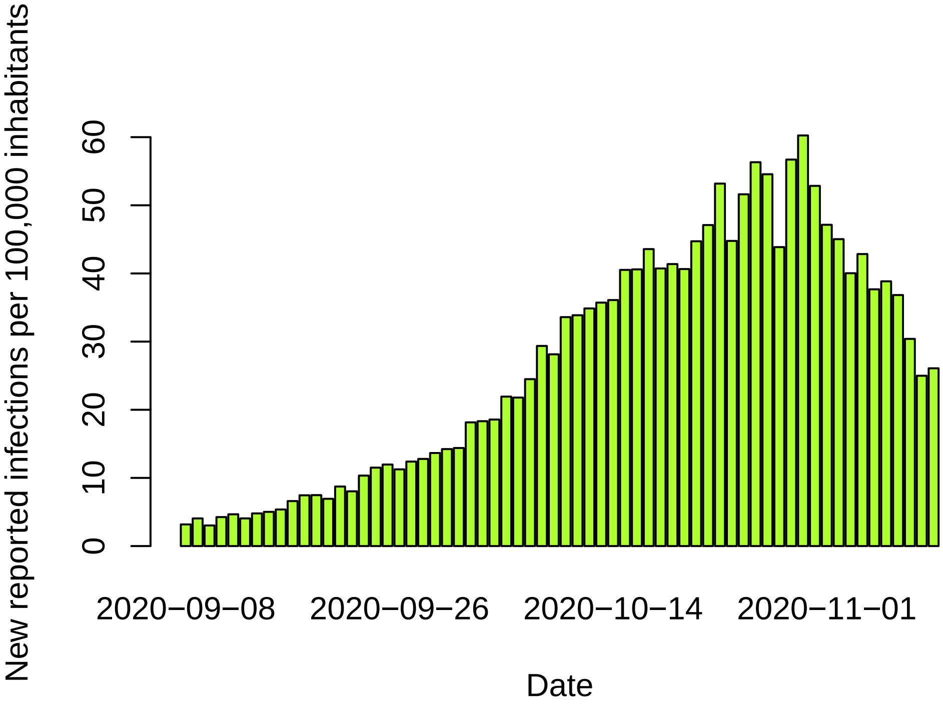

Figure 5. Number of new reported infections per day per 100,000 inhabitants in the Netherlands over time

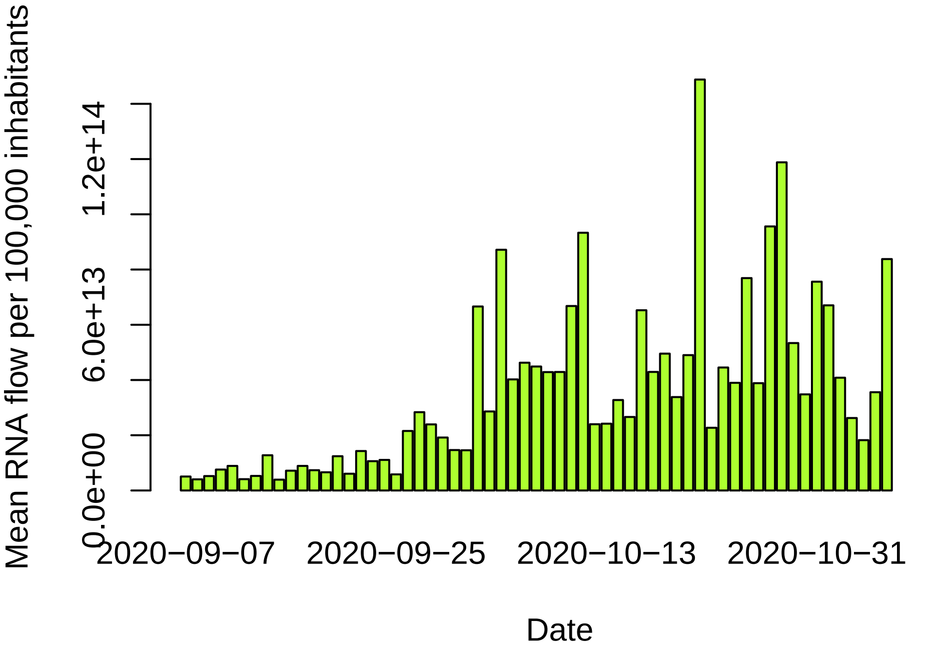

Figure 6. Mean RNA flow per 100,000 inhabitants in the Netherlands per day

The reason for the different shapes of the plot showing the mean RNA flow and the plot showing the increase in infections could be the fact that not all treatment plants report the RNA flow every day, and that the number of reporting treatment plants differs per day and per region. In order to explore this idea we plotted the number of installations from which a sample was taken over time. The figure below shows that the number of installations from which a sample is taken differs per day. The figure also shows that the maximum number of sewage water installations from which samples are taken lies somewhere around one hundred. Since there are more than three hundred installations from which a sample can be taken (RIVM, 2020c), there is not a single day where a sample is taken from all installations. So every day, samples are taken from different installations that process waste water from different regions. If every region produced the same amount of RNA particles it would not matter that we do not observe each installation each day. Since we take the mean of the amount of RNA particles, the estimate would not be harmed by a missing installation. However, the figure that shows the mean number of RNA particles observed each day suggests that the estimate of the mean number of RNA particles is not correct. If it were correct, then the values would not fluctuate as much as they do now and would follow the trend of new infections more clearly. Thus, the plot supports our claim that the different shapes of the plot showing the mean RNA flow and the plot showing the increase in infections could be explained by the fact that not all treatment plants report the RNA flow every day, and that the number of reporting treatment plants differs per day and per region.

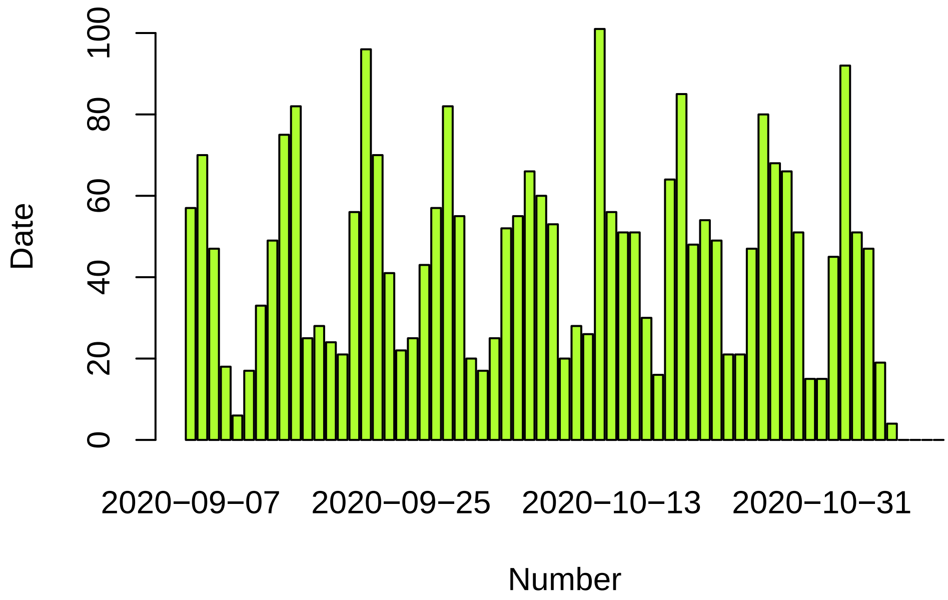

Figure 7. Number of installations per day

3.6 What is the relationship between RNA particles and the total number of infections on one given moment?

As mentioned in the data description, some installations process water from multiple security regions, rendering the current RNA flow variable unrepresentative. We do have a variable available that contains the proportion of water from the security region processed by that respective installation. We multiplied these two columns to create a representative RNA flow variable.

Now that the RNA flow has been weighted, we want to inspect how the RNA flow is related to the increase in the number of infections on the level of the security regions. To achieve this, we altered the current functions for calculating the RNA flow and increase in number of infections 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 infections 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). A scatterplot with the aforementioned data should provide a first impression of how they are related to each other. Since not every security region has data about the RNA flow on any given day and data about the increase in the number of infections seven days later, only security regions that have data for both days are selected. Subsequently, the correlation for those data is calculated.

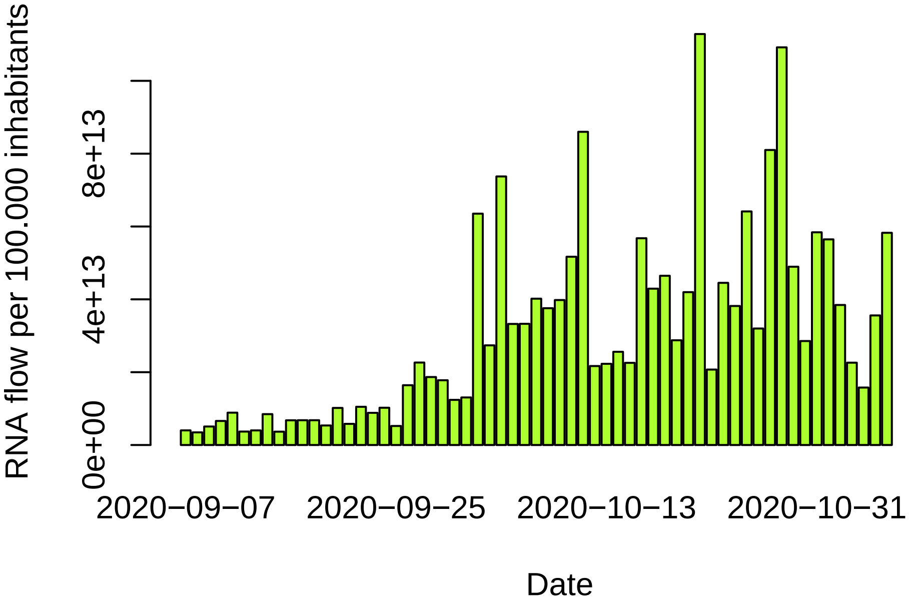

Figure 8. RNA flow per 100.000 inhabitants in the Netherlands over time

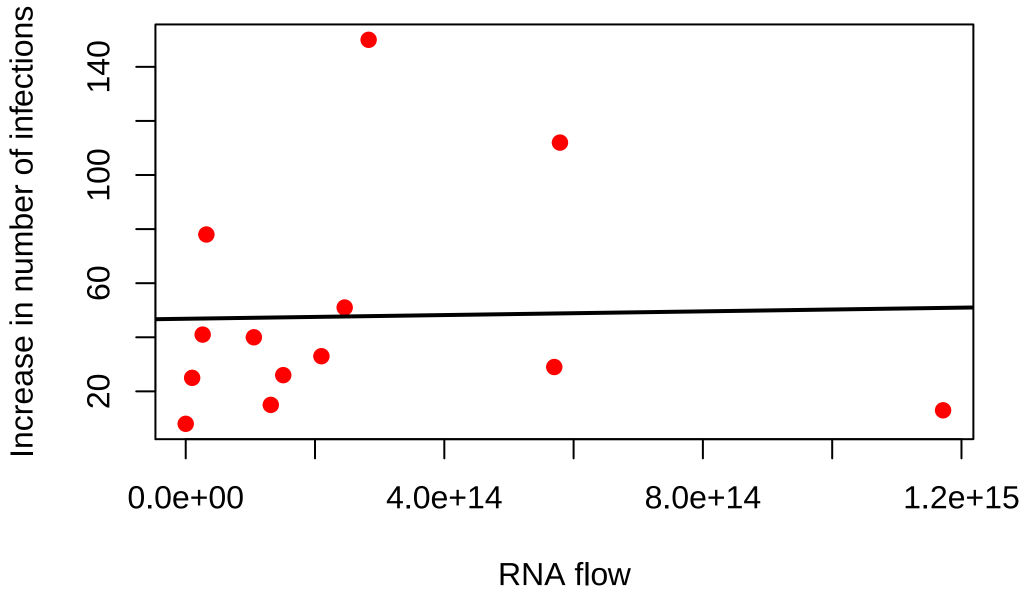

Figure 9. Relation between RNA flow and increase in infections for one timestamp

Figure 9 shows the correlation between the RNA flow on the 8th of September and the increase in the number of infections on the 15th of September. As can been seen from the scatterplot, these data do not seem be strongly related to each other. Additionally, the correlation associated with this scatterplot is 0.027, which indicates that these data almost show no relationship to each other. This strikes us as very surprising, since one would expect a relationship between the RNA flow and the increase in the number of infections. In the next step, we extend the analysis conducted over the period of September until November to see if this trend persists.

3.7 How does this relationship look over time?

Figure 10 shows that the correlations fluctuate over time, which seems very interesting. If the data points would have formed a straight horizontal line, the estimation of the true number of infections in the Netherlands would have been more straightforward, since it would mean that the relationship between the RNA flow and the increase in infections would have been the same for every day. The mean correlation between the datapoints in this scatterplot is 0.33, which is a moderate positive correlation. This means that as the pandemic progresses, the relationship between the RNA flow and the increase in the number of infections becomes stronger. We have two hypotheses as to why this would be the case: it could be a result of increased consistency and precision in the measurements as we get more rigorous in reporting these data, or – as the number of cases has only increased over the period we analysed – it could mean that RNA flow becomes a better predictor of the total number of reported cases, as both measures have higher values.

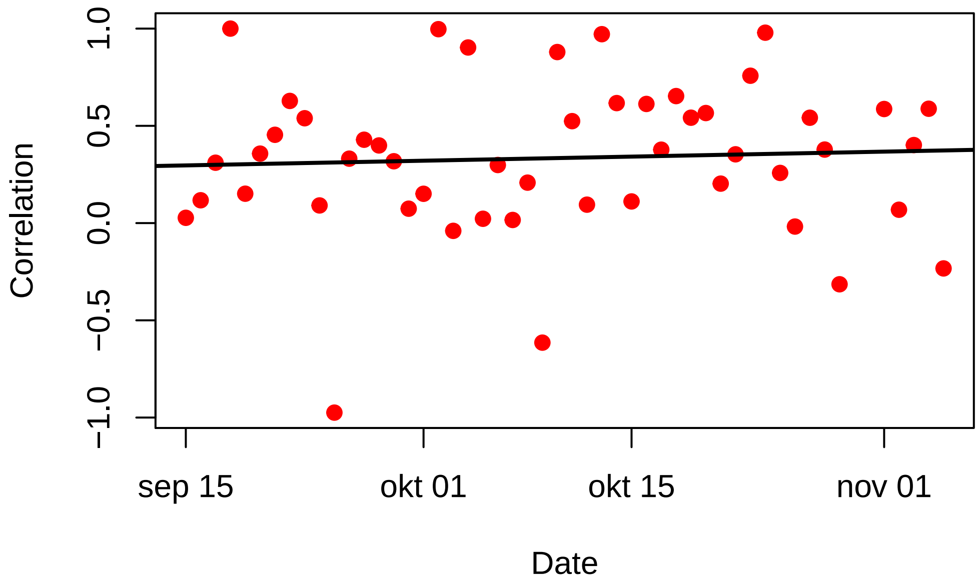

Figure 10. Correlations between RNA flow and increase in infections over time

4 Results

5 Conclusion

As of yet, we are unable to make a reliable estimate of the true number of infected people in the Netherlands. First and foremost, this is due to the unstable nature of the correlation between the RNA flow and the total number of recorded infections. However, the moderate, positive correlation between these two variables indicates that the RNA flow becomes an increasingly better predictor as we get further along in the pandemic. If we do essay to construct a prediction model, we recommend to use only the most recent RNA flow data to build it, as they are likely the most reliable.

Our next line of investigation will be to try to combine the RNA flow as a predictor with other available variables in our dataset, such as population density. We would also like to separate this analysis by security region, to isolate the effect of geographic location on the predictor.

If we had infinite time and resources, we would like to enrich our analysis by incorporating other types of data in our analysis. For example, we could use behavioural data on when people are more or less likely to get tested when displaying symptoms, to improve the estimate of the true number of infected people. Another option would be to include environmental data on how much rain fell in each region on each day, to correct for the influence this might have on the concentration of RNA particles in the sewage water. 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.

6 Discussion

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

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

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-covid19/onderzoek/rioolwater

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

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