Text presentation

**Structure of the data**

We are working with two datasets, both collected by the 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. 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.

**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 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 is 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.

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.

Lastly, some very large municipalities produce so much sewage water in one day, that multiple treatment plants are required to process it. See SR3 of Figure 1. Unfortunately, there is no data available that specifies 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 answering our sub-questions. Furthermore, we concluded 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.

