Outlier detection in digital PCR wastewater time series data

Master Thesis

Author(s):
McLeod, Rachel E.

Publication date:
2024-10-18

Permanent link:
https://doi.org/10.3929/ethz-b-000702298

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Outlier detection in digital PCR wastewater time series data

Master Thesis

Rachel McLeod

mcleodr@ethz.ch

Department of Biosystems Science and Engineering
Computational Evolution
ETH Zürich

Supervisors:
Adrian Lison
Prof. Dr. Tanja Stadler

October 18, 2024

Acknowledgements

Firstly, a huge thank you to Adrian Lison for his excellent supervision. I'm very grateful for all his input into the project and for his guidance, for interesting discussions and answering my many questions. Also, for providing motivation and encouragement when needed and generally helping me build skills and confidence.

I am grateful to Tanja Stadler for the opportunity to complete my masters thesis with the Computational Evolution group. It has been a really great experience and I have learnt a lot. Thank you to all the members of cEvo for being so welcoming and making my time in Basel so enjoyable.

I am also grateful to Tim Julian for his support, encouragement and providing me a with a desk at Eawag for those times when Basel has felt a long way away. Many thanks to the members of the Wastewater Monitoring Laboratory team at Eawag for their support and encouragement: their tireless effort to generate such high quality data is greatly appreciated. Thank you also to those of the team who took time to look through 70 graphs, labelling outliers for me.

Finally, thank you to my family for providing proofreading services, encouragement and occasional emotional support.

Abstract

Efforts to control the spread of disease in a population require data collection on its prevalence over time. Analysing this data can give insights into how disease spreads and provide feedback on the effectiveness of intervention methods. Wastewater-based epidemiology (WBE) involves sampling wastewater from a treatment plant to collect data on disease prevalence in the population served by a particular sewage catchment. This data can be used to estimate metrics such as effective reproduction numbers (R_t) which can inform public health decision making.

WBE contains several sources of uncertainty and the resulting data are noisy, requiring statistical analysis to gain meaningful measurements. This includes normalisation to account for population or system dynamics, and smoothing of signal to understand trends. When data containing outliers are used to inform models about the spread of disease, these outliers can have an impact on the model analysis. The identification and removal of such data points could improve modelling results and provide a more reliable basis for decision making for disease control.

The Wastewater Monitoring Laboratory, part of the Swiss Federal Institute of Aquatic Science and Technology, collects data on the concentration of four different respiratory pathogens in the wastewater from fourteen different treatment plants around Switzerland using digital PCR (dPCR). This data set was used to develop an outlier detection method able to deal with multiple challenges presented by the specific nature of wastewater-based dPCR data, thereby allowing for efficient outlier detection.

The outlier detection method was then used to study the impact of outliers on estimating R_t from wastewater viral concentration data. This highlighted the importance of robust model design and provided insight into the impact an outlier can have on estimates depending on when it occurs during an infection wave. Additionally, time series of wastewater data were simulated, providing a resource for further method evaluation.

Contents

Chapter 1

Introduction

Infectious disease epidemiology studies disease prevalence and transmission within a population[1]. Wastewater-based epidemiology (WBE) refers to the case where data collected from a sewage system are used to infer this prevalence and transmission information within the sewage system's catchment population[2]. Typically, bacterial or viral gene copies in the wastewater are measured using molecular techniques such as polymerase chain reaction (PCR). WBE data are inherently noisy and can contain outliers which pose challenges when it comes to data analysis[3]. There are three specific challenges for WBE data which would ideally be addressed by an outlier detection method.

1. Wastewater time series data are expected to contain fluctuations as waves of infections pass through a population. Therefore, a measurement which would be perfectly normal at the peak of infections could be considered an outlier if it occurs outwith a wave. In such a situation the measurement would be considered a local outlier, though not a global one.

2. Previous work on digital PCR (dPCR) data has shown that expected variation in measurements increases with increasing viral RNA concentration[4] and this should be taken into account when labelling data points as outliers. This has also been demonstrated for quantitative PCR (qPCR) data[5].

3. One of the benefits of wastewater data over case data is the speed with which data can be made available for public health decision making. In such a real-time monitoring programme, outlier detection should also be carried out in real time to ensure reliable data reporting for the public.

Though many data processing methods have been adapted and developed for WBE applications, including identifying outliers, currently none address these issues in their entirety.

A final aspect of outlier detection in WBE data is the difference in nature between high and low outliers. By this we mean outlier values which are, respectively, above or below the expected viral concentration. As negative concentration

measurements are a physical impossibility low outliers are lower-bounded, while the potential magnitude of high outliers is much greater. Low outliers seem to be identifiable by monitoring dilution events in the sewage system, best identified by analysing data used for signal normalisation rather than viral concentration data[6]. High outliers on the other hand seem to require outlier detection directly on the viral concentration data.

This thesis aims to develop a computationally inexpensive outlier detection method for identifying high outliers in real time. This will be done for dPCR wastewater time series data in the context of a Swiss monitoring programme. The method should be able to identify both local and global outliers and take into account changing measurement variation. The hope is that such a method will aid in the monitoring programme's data quality management and potentially improve modelling results. A long-term aspiration, beyond the scope of this thesis, is that the underlying processes contributing to outliers could be better understood. A first step towards this is to be able to efficiently identify outliers so they can be studied.

The following sections of this thesis present a short history of WBE together with some background information on current outlier detection methods used for WBE data. The data set used in this thesis is introduced, followed by an overview of the methodology developed. Results of method performance are shown and discussed and the effect of outlier removal on modelling are investigated. Finally, a simulation of a simple outlier generating process is carried out, with simulated data then used for further analysis of method performance.

Chapter 2

Background

2.1 Wastewater-based epidemiology

Wastewater-based surveillance was first proposed in 2001 as a method for determining population drug usage from analysis of drug metabolites in wastewater[7] and, in the decades since its conception, it has grown in use and applications. It works on the principle that metabolites, microorganisms, or other chemicals shed by individuals into the sewage system can be measured in the wastewater using techniques such as mass-spectrometry or polymerase chain reaction (PCR). Data collected from wastewater samples in this way can give information about drug consumption, antimicrobial resistance, or disease prevalence within the population contributing to the sewage system[2].

Epidemiology is the study of the incidence and distribution of a disease in a population. Traditionally it involves collecting data about the number of individuals who become sick over time[1]. These data can be used to estimate indicators such as the effective reproductive number (R_t), which is the average number of secondary infections an infected individual will cause, and which gives a measure of how quickly a disease is spreading through a population. Indicators like this are important in understanding the effectiveness of control measures which, along with understanding the cause of a disease and its method of transmission, is essential to combat disease[8].

The SARS-CoV-2 pandemic, in particular, shone a spotlight on how wastewater surveillance could be applied to epidemiology and the term wastewater-based epidemiology (WBE) is now in common usage. During the pandemic it was shown how data on the abundance of viral RNA shed from infected individuals in the wastewater could compliment case and hospitalisation data in public health decision making, with the correlation between cases and wastewater viral load being widely demonstrated[9–11]. Further applications of wastewater data include predicting hospitalisations and ICU admissions[12,13], and estimating R_t[10,14,15] with several R packages developed for this purpose[16–18]. Further to this, there have been efforts to sequence SARS-CoV-2 RNA extracted from wastewater, which

can enable the early detection of new viral variants[11].

Wastewater viral RNA concentration data are collected at the level of a sewage system's catchment population. As such this can be considered a non-intrusive data collection method due to the inability to trace incidence back to individuals. More importantly, from the perspective of data analysis and modelling, these data should contain less bias as they aggregate evenly over the sampled population. This is in contrast to testing efforts, which often sample from the sub-population of individuals presenting symptoms.

Challenges gaining epidemiological insights from wastewater data largely relate to the uncertainties within the data collection process and to relating measurements back to the sampled population. Sources of uncertainty include, but are not limited to: a fluctuating population contributing to the catchment; uneven or unknown shedding rates of infected individuals; uneven sampling from the sewage system; unknown degradation rates of RNA in the sewer and during sample transport; and uncertainties in lab processes, such as efficiency of the RNA extraction and PCR inhibition[3,19]. The culmination of these factors means that measurements obtained from wastewater are noisy.

Visualisations of wastewater data do indeed show this noise, along with the presence of apparent outliers. To relate this back to the uncertainties in the WBE process, it is conceivable that there are factors which only sporadically affect measurements. In such a situation, if the effect was large, a few isolated measurements could have extreme values which would align with the observation of such outlier data points in wastewater time series data.

For some sources of uncertainty, such as extraction efficiency or PCR inhibition, controls can be put in place to ensure that RNA has indeed been extracted and that the sample is not inhibited[20]. If these controls fail, the sample can be removed from the dataset or re-tested. However, there are many other sources of uncertainty which cannot be easily controlled for. Therefore, a method to identify outliers arising from such uncontrollable processes would be advantageous.

2.2 Outlier detection

As for any detection method, the first challenge of outlier detection is to know what it is you are looking for. A widely cited definition, which proves useful for this thesis, comes from Hawkins[21]:

an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism

In other words outliers are observations which do not follow expectations, and outlier detection is the process of identifying such data points.

Terminology surrounding outlier detection is varied, with data points being

referred to as *outliers*, *anomalies*, *noise*, or *peculiarities* to name a few. Attempts have been made to distinguish between different types of outliers, e.g. erroneous or missing data and deviations of interest. The former will generally be the target of data cleaning efforts while the latter often warrant further study as they may arise from some process we are unaware of or are not capturing in our models[22].

Aguinis *et al.* recommend several best practices for outlier detection. They highlight again the importance in distinguishing between error outliers and "interesting" outliers. Furthermore, they note that an intuitive observation of outliers in visualisations is often the step preceding the employment of statistical outlier detection methods[23]. When data sets are single dimensional, simple statistical methods such as a z-score or interquartile range (IQR) can be used to identify outliers[24].

Outliers in time series data present additional challenges due to the added dimension of time. Further classification of outliers into *global* or *local* outliers is made, accounting for the increased correlation of data points to their time-bound neighbours[25]. Statistical outlier detection methods can be modified to account for time information, which allows for outlier detection at all points along the time series[26].

Outlier detection in time series is widely applied to several fields including identifying credit card fraud[27,28] and environmental monitoring[29–31]. These fields come with the additional challenge of timeliness as, especially in the case of credit card fraud, the ability to detect outliers in real time is important. These are considerations that should also be applied to outlier detection in WBE, where near real-time data reporting is advantageous, especially in the context of epidemics[32].

2.3 Outlier detection in wastewater time series data

As with all developing fields, there are no standardised data analysis pipelines for WBE data. Normalisation of quantified viral RNA can be done in a couple of ways, such as using biomarker concentrations or flow data. Additionally, when it comes to smoothing noisy data to obtain meaningful signal from wastewater viral loads, a wide variety of approaches can be adopted[33].

To gain an understanding of currently employed outlier detection methods Google Scholar was queried with the key words *"wastewater-based epidemiology"* and *"outlier detection"*.

One approach to detect outliers is to apply outlier detection to the biomarker or flow data used for normalisation. Rauch *et al.* propose just this when they give recommendations of best practices when it comes to working with wastewater data in the context of modelling and data communication. They suggest simple metrics such as a box plot applied to the whole data set, but do also stress that first and foremost manual inspection of wastewater data is crucial because outliers

and errors can arise from several sources and not all can be easily tracked[6]. This approach of using only flow data to detect outliers was adopted by Markt *et al.*. In their study they had quantitative PCR (qPCR) data which were normalised with ammonia concentration as a biomarker. Outliers were identified as data points from days where the flow was greater than the 90th percentile over the whole time period[34]. A case study from Madrid adopted a similar approach using chemical oxygen demand as a marker to anticipate unusual dilution. Here both anomalously high and low values were used for data filtering[35]. Such methods, being independent from viral concentration data, are unable to take into account the changes in measurement variation with changing concentration and cannot identify outliers arising from a mechanism other than flow. By using simple metrics on the whole data set, the time series nature of the data is overlooked and local outliers are perhaps not identifiable. Additionally, they are difficult to apply in real-time during the initial stages of a monitoring programme where a sufficient period of data for calculating the 90th percentile or determining a "normal" range may be lacking.

Outlier detection methods applied directly to viral concentration data may overcome some of these limitations. Klaassen *et al.* used viral RNA concentrations measured in the wastewater to show that they can be used to improve the predictive performance of a nowcasting model for estimating infection and transmission rates. In their study they checked the robustness of their model by comparing results with and without outliers. Outliers were detected by fitting a spline to the non-normalised wastewater concentration data, taking the deviation of the measurement from the spline and filtering measurements with a deviation greater than three times the IQR of the deviations[36]. By applying outlier detection to the deviations rather than absolute values, local as well as global outliers can be identified. However, the specific challenge presented by dPCR measurements (increasing variation with increasing concentration) is not addressed. Additionally, this approach is not ideal for consistent real-time use because the IQR of the deviations changes as more data are collected. Despite the limitations, the authors showed that their models were robust to these outliers because their conclusions remained the same with and without the inclusion of outliers, though they note that variance in the model predictions was reduced upon removal of outliers.

Manuel *et al.* also take the deviation of measurements from their smoothed signal and use a resulting z-score to determine whether a particular time point is an outlier[37]. However, their application poses several limitations: they don't make direct use of raw measurements, instead they take a 7-day rolling average viral concentration as an input; they apply wastewater data to a model designed for case data; and, in the context of this thesis, their smoother requires to be fitted on historic data making it unsuitable in the early stages of a monitoring programme and generally more computationally expensive.

An example of outlier detection in the context of a monitoring programme comes from Scotland. Here they were interested in identifying high outlier "peaks" relative to case data. Viral loads were smoothed using smoothing splines which was done via a generalised additive model (GAM) which assumed the loads are Tweedie distributed. The probability of each point being a spike was calculated and a threshold for data removal was set by calibration on manually labelled data to achieve a desired level of false positives and negatives. This was done in a real-time manner where only data up to the day of observation were made use of[32]. The inclusion of case data in the model presents a limitation for contexts where such data is unavailable. In addition, it is not clear from the literature whether the GAM accounts for changing measurement variation with changing RNA concentration.

Most outlier detection identified in the literature made use of simple detection metrics applied either to data used for normalisation or to concentration measurement deviations from a smoothed trend line. In contrast to these approaches Courbariaux *et al.* showed how outliers could successfully be detected using autoregression. They modelled the time series as an autoregressive process where each measurement was either an outlier or not. The outlier status of a particular measurement was modelled as a Bernoulli distribution. In the case of being a non-outlier the measurement was modelled as the underlying real quantity plus normally distributed noise, whereas outliers were modelled to arise from a uniform distribution. A modified Kalman filter was used to determine the probability distributions of the underlying real values, along with the likelihood of each point being an outlier. A threshold was selected to turn outlier probabilities into binary classification according to a desired false positive and negative rate. Parameters of their model were fitted using expectation maximisation[38].

Taken together these studies highlight the absence of a computationally inexpensive method for outlier detection which fully accounts for the specific nature of wastewater time series data. Specifically, one able to find both global and local outliers, to account for dPCR measurement variation and applicable in a real-time monitoring situation. The next chapters present the development and evaluation of an outlier detection method able to deal with these three specifications.

Chapter 3

Methods

3.1 Wastewater time series data

The data used in this thesis were generated by the Wastewater Monitoring Laboratory at the Swiss Federal Institute of Aquatic Science and Technology (Eawag). This laboratory uses digital polymerase chain reaction (dPCR) to quantify viral RNA extracted from wastewater samples collected at various locations around Switzerland (Figure 3.1). Daily 24-hour composite wastewater samples are shipped from treatment plants once per week to the Monitoring Laboratory. On reception at the lab, samples are processed for viral RNA extraction, concentration and quantification.

Since February 2021 SARS-Cov-2 viral concentrations have been measured for six wastewater treatment plants using an assay targeting the N1 protein gene of the SARS-CoV-2 virus and Murine Hepititus Virus as internal control (*N1MHV* assay)[10]. In November 2022 an additional four targets were added to the monitoring program using an assay for Influenza A and B, Respiratory Syncytial Virus, and SARS-Cov-2 targeting the N2 protein gene (*RESPV4* assay)[15]. In July 2023 these two assays were combined into a single *RESPV6* assay [in preparation] and an additional eight treatment plants were added to the monitoring program. In total this gives data for five viral targets (SARS-N1, SARS-N2, IAV-M, IBV-M, RSV-N), from fourteen treatment plants (Figure 3.1), resulting in 70 time series of varying lengths which were available for method testing and evaluation (Figure 3.2).

To translate viral concentrations measured in the laboratory to a metric comparable between locations they must first be normalised. Normalisation is done by converting concentration measurements of viral gene copies per litre of wastewater (gc/Lww) into load measurements of viral gene copies per day per person (gc/day/person) by accounting for flow through the treatment plant system (data provided by the treatment plants) and catchment population size (taken from census data). Viral concentration and load measurements are used to model R_t estimates for the different treatment plants, shared with the Federal Office of Public Health and displayed publicly on a dashboard maintained by the project[39].

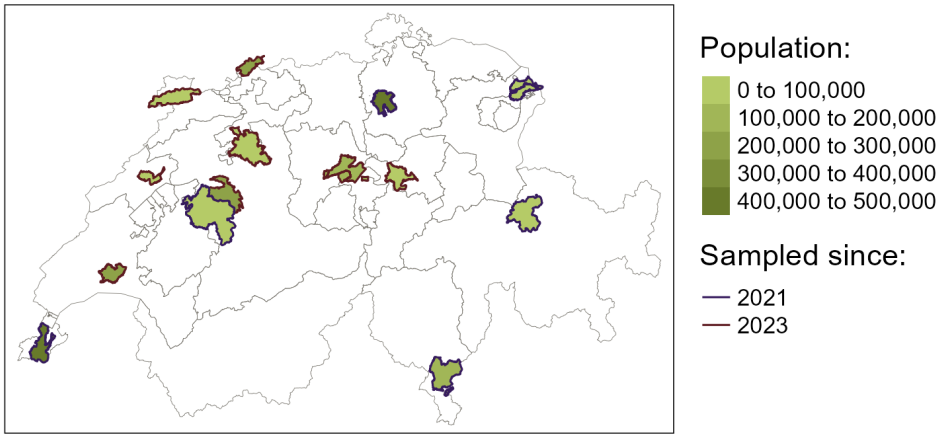

Figure 3.1: Catchments in Switzerland sampled by the Wastewater Monitoring Laboratory at Eawag. The border colour indicates when the site joined the monitoring programme. The fill colour indicates the catchment population size. About 27% of the Swiss population is covered by all of these catchments.

3.1.1 Data processing

Before the data generated by the Wastewater Monitoring Laboratory are used for modelling purposes or made publicly available, they undergo several quality control checks. There are three automatic data checks and dPCR reaction results will be filtered out from the shared dataset when:

1. there are not enough partitions in the dPCR reaction ($<$15,000);
2. there is too much inhibition in the wastewater matrix ($>$40%); or
3. a negative control gives a positive signal ($\geq$3 positive partitions).

Further to these automatic checks, some metrics are tracked manually e.g. recovery (the proportion of viral RNA successfully extracted from the wastewater sample) and SARS-CoV-2 N1:N2 ratio. Samples are run in duplicate and only repeated in the case where both technical replicates fail quality checks. In the case where only one replicate passes quality checks this replicate will be reported. Data are also reported with a quality flag indicating if they fall below the limit of detection (lod). The data used in this thesis are data from dPCR reactions which have passed all internal lab quality controls, increasing our confidence in their veracity as an accurate measurement of the RNA concentration in the sample.

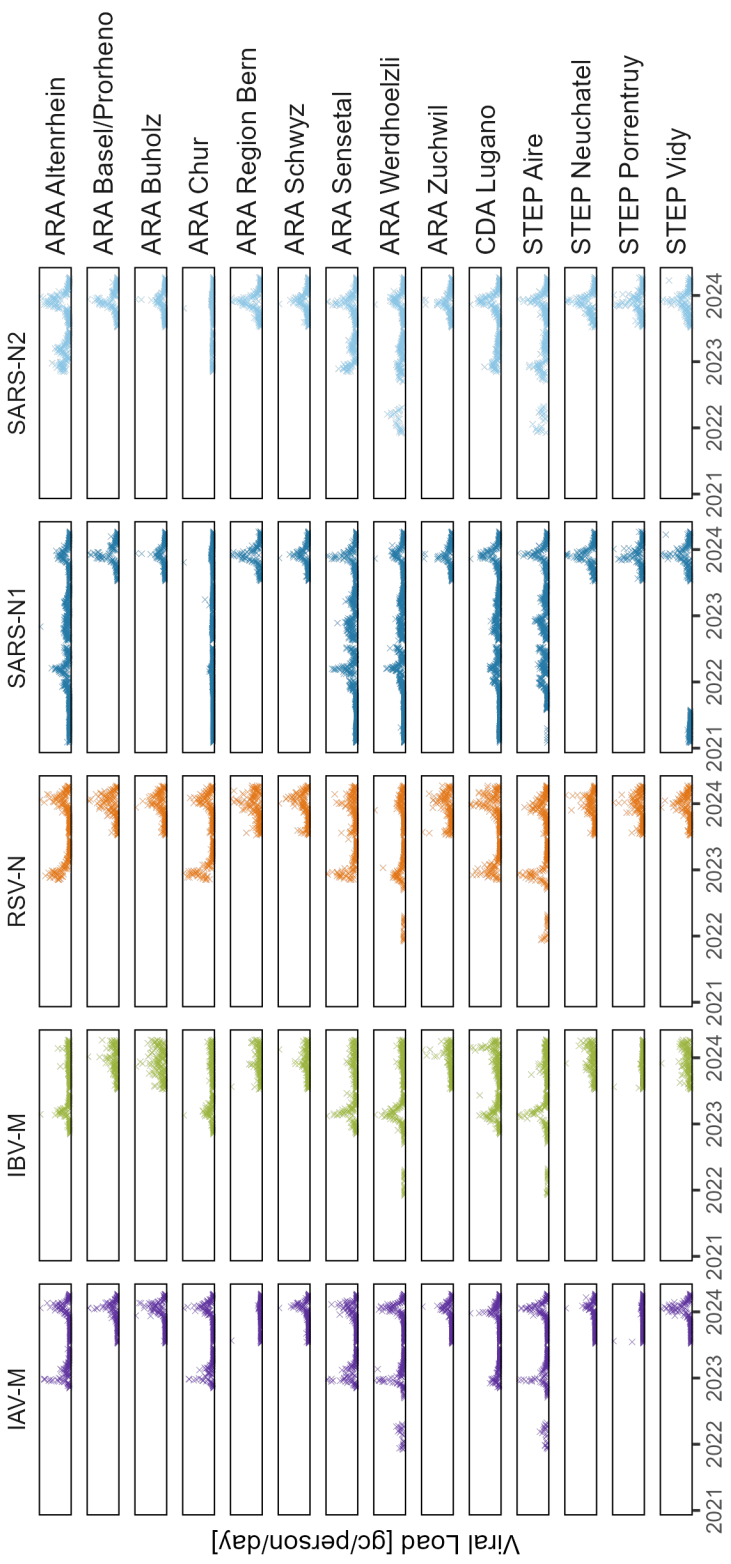

Figure 3.2: Concentration time series of the IAV-M, IBV-M, RSV-N, SARS-N1 and SARS-N2 genes at the 14 treatment plants in Switzerland.

3.1.2 Laboratory variables

In addition to viral concentration, load and the associated quality flag, several other variables are recorded in the data set. Mainly these are laboratory conditions from the dPCR experiment: a dilution factor, the number of replicates, total number of partitions, and the volume of a partition. This section describes these variables and their typical values.

Dilution factor

The dilution factor is the amount the "raw" RNA extract was diluted before being run on dPCR. Samples are diluted to reduce inhibition of the dPCR reactions. The dilution factor was known for each sample and was generally three or five.

Number of replicates

This is the number of technical replicates of a single sample run on the dPCR assay. Generally, it was two for our dataset, however, in cases where a technical replicate failed quality checks, or samples were tested again, this value could be higher or lower.

Total partitions

The dPCR set up in the laboratory uses Stilla sapphire chips for running dPCR. The manufacturer reports an upper limit of 30,000 partitions possible with these chips. The quality control removes samples with fewer than 15,000 partitions. On average our partitions were in the middle of this range, at around 22,000. The total number of partitions was known for each sample and for multiple technical replicates the average total partitions was taken.

Partition volume

The partition volume is the volume of a single dPCR partition and is dependent on the dPCR reaction reagents and system. For the set up used by the Wastewater Monitoring Laboratory this volume is 0.519 nL.

3.1.3 Data labelling

To aid method development a "ground truth" was required to allow different methods to be compared. Since a standard model to define expected measurements was lacking, an approach to generate these data was required. Taking inspiration from an outlier detection study in surface water temperature[31], domain experts who routinely work with the data were asked to manually label data points they would consider a high outlier. This also reflects the fact that the observation of outliers in data visualisations triggered the development of a statistical outlier detection method[23] and therefore any method developed should capture the intuitive models of researchers.

Six individuals were asked to go through the 70 time series graphs and identify points which they would consider a high outlier. Points which received a majority vote were then labelled as an outlier. To facilitate this process the average of technical replicates was taken so that there would be only one data point per day plotted and graphs were interactive, allowing zooming, in an attempt to minimise bias due to different scales.

3.2 Outlier detection method

In the following section the proposed method for outlier detection will be introduced, along with definitions of parameters and explanations of assumptions underlying the method.

The dPCR assay measures viral RNA concentration in the wastewater for a particular day, t. We consider the viral RNA concentration on a particular day, c_t, to be a random variable with normally distributed noise, which arises from variation in the sample collection and measurement process. Though this assumption is not entirely accurate due to negative concentrations being an impossibility, a normal distribution is nonetheless a good approximation in reality. Additionally, c_t, is calculated from dPCR reaction positive droplet counts which are binomially distributed[20] and a normal distribution is a good continuous approximation of this.

The viral load for a particular day, l_t, can be calculated as

$$l_t = \frac{c_t \cdot f_t}{P}, \tag{3.1}$$

where f_t is the flow through the wastewater treatment plant for that day in litres and P is the size of the population contributing to the catchment.

This normalisation removes noise in the time series from varying flow through the sewage system and allows comparison between locations with different catchment population sizes. We assume that l_t calculated for each day is an unbiased

estimate of the true load and therefore the expected value, $\mathrm{E}[l_t]$, is the true load in the system on a particular day.

With the benefit of being able to look back on at least one year of viral load data, plotted sequentially, the points could be described as following a smooth trend and for each day the value of this trend, $\mathrm{T}[l_t]$, can be calculated. We assume that $\mathrm{T}[l_t] \approx \mathrm{E}[l_t]$.

Generally, points manually labelled as outliers were single measurements which deviated from the trend of their neighbourhood. The first step of the method therefore attempts to capture this internal judgement by determining $\mathrm{T}[l_t]$ and calculating the deviation of the measured load from this trend, denoted

$$\delta_{l_t} = l_t - \mathrm{T}[l_t]. \tag{3.2}$$

If all the deviations came from a standard normal distribution, then a traditional z-score approach to outlier detection could be used[24,26]. Deviations should already have a mean of zero, as measurements are assumed to be randomly distributed around the trend line, however, the standard deviation is not one. Further normalisation is therefore required which takes into account the change in expected variation of measurements with changing concentration.

Here it is possible to benefit from the work demonstrating the change in measurement variation with concentration in dPCR reactions[4]. From this we have

$$CV_{c_t} = \begin{cases} \frac{\sqrt{\exp(c_t/d \cdot v) - 1}}{\sqrt{p \cdot n} \cdot c_t/d \cdot v} & \nu = 0 \\ \\ \sqrt{\nu^2 \cdot \frac{1}{c_t/d \cdot p \cdot n} + \frac{1+\nu^2}{2 \cdot p \cdot n}} & \text{otherwise,} \end{cases} \tag{3.3}$$

which captures the expected coefficient of variation at a particular concentration. Here n is the number of technical replicates, p is the (average) number of partitions in the dPCR reaction, v is the volume of a partition, d is the dilution factor of the sample and ν is the pre-PCR variation. All parameters, apart from ν, are laboratory variables explained in Section 3.1.2. The pre-PCR variation is unknown and therefore must be estimated. Section 3.2.1 describes how this was done.

Equation 3.3 can be used to estimate the expected variation around $\mathrm{T}[l_t]$ and normalise δ_{l_t}. Note, however, that this equation takes concentration rather than load as an input, and therefore a conversion is required. The inverse of equation 3.1 can be used to obtain $\mathrm{T}[c_t]$ from $\mathrm{T}[l_t]$, modified to avoid zero values by replacing any value below the lod with the lod:

$$\mathrm{T}[c_t] = \begin{cases} \frac{\mathrm{T}[l_t] \cdot P}{f_t} & \mathrm{T}[c_t] > \text{lod} \\ \text{lod} & \text{otherwise.} \end{cases} \tag{3.4}$$

Additionally, rather than calculating the load deviation, the concentration deviation can be calculated as

$$\delta_{c_t} = c_t - \mathrm{T}[c_t]. \tag{3.5}$$

$CV_{\mathrm{T}[c_t]}$ can be used to calculate the expected standard deviation at $\mathrm{T}[c_t]$ and this can be used to normalise the deviations:

$$\tilde{\delta}_{c_t} = \frac{\delta_{c_t}}{CV_{\mathrm{T}[c_t]} \cdot \mathrm{T}[c_t]}. \tag{3.6}$$

Ultimately this should result in deviations which are standard normal distributed ($\tilde{\delta}_{c_t} \sim \mathcal{N}(0, 1)$).

In this case the traditional z-score threshold of three can be applied and a measurement c_t is labelled as

$$\text{is outlier} = \begin{cases} 1 & \tilde{\delta}_{c_t} > 3 \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

where 1 represents an outlier observation and 0 a non-outlier observation (remember we are only interested in high outliers).

Two components remain to be further defined: the pre-PCR variation, ν, and the method used to determine $\mathrm{T}[l_t]$. These are discussed in Sections 3.2.1 and 3.2.2 respectively. The final method is summarised below and in Figure 3.3

1. Find the trend line through viral load measurements.
2. Convert the load trend line into concentration (equation 3.4).
3. Calculate all the deviations between measurement concentrations and the trend line (equation 3.5).
4. Normalise these deviations by dividing by the expected standard deviation at the trend line concentration (equations 3.3 and 3.6).
5. Label any data points with a normalised deviation greater than three as an outlier (equation 3.7).
6. Evaluate using manual outlier labels.

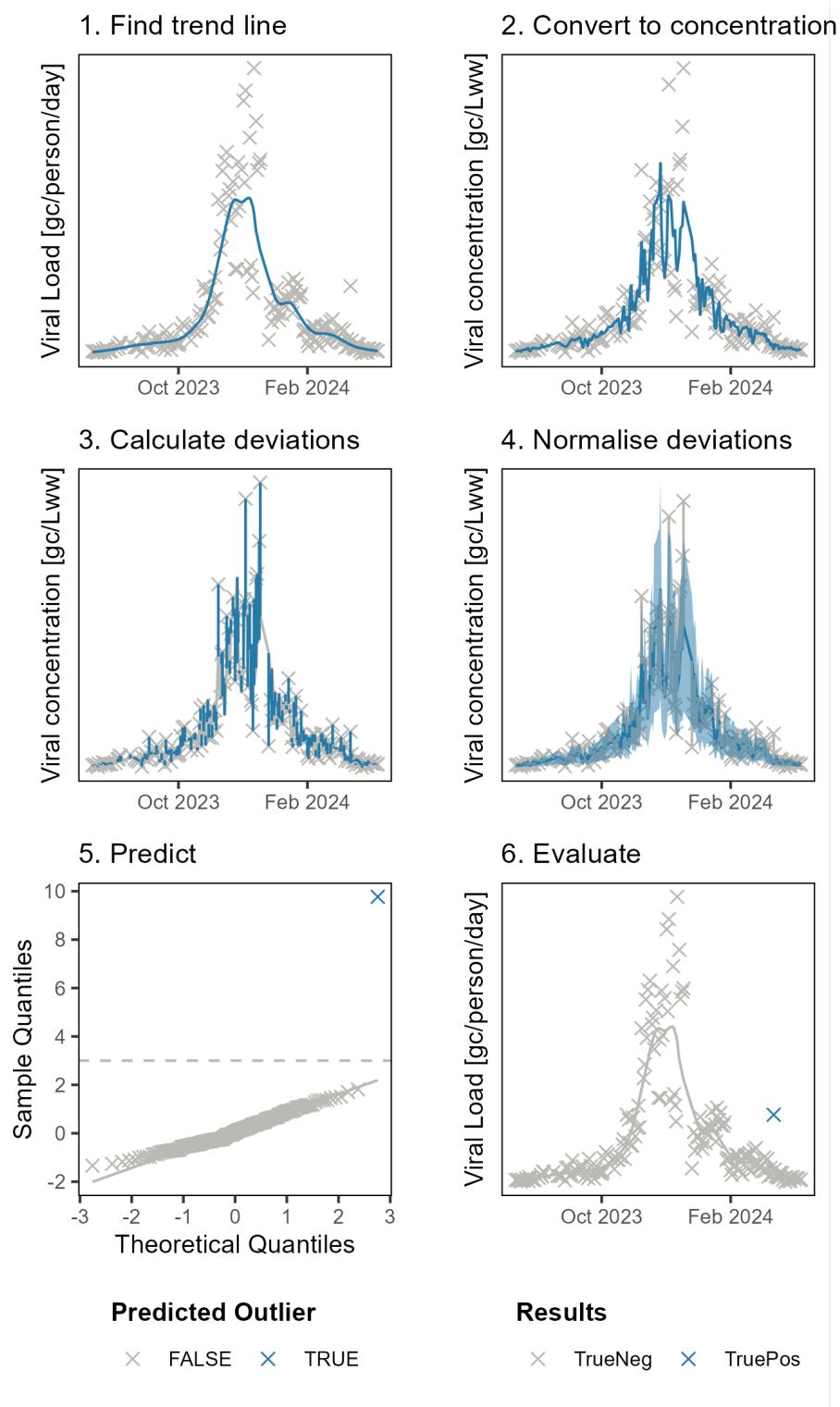

Figure 3.3: Visualisation of method steps using SARS-N1 data from ARA Basel/Prorheno with trend decomposition for calculating the trend line and a pre-PCR variation parameter of 0.6 for normalisation (see Sections 3.2.1 and 3.2.2 for details).

3.2.1 Estimating pre-PCR variation

The pre-PCR variation parameter represents the noise before PCR. In this case we assume this to include sources of variation from the sewage system and sample collection process. The goal of the normalisation - to standardise the deviations such that $\tilde{\delta}_{c_t} \sim \mathcal{N}(0, 1)$ - was used to define the value of ν. The standard deviations (sds) were used to select ν rather than the mean because the mean is expected to be slightly below zero. This is because the load to concentration conversion replaces zero values with the limit of detection (equation 3.4), skewing the distribution of measurements around the trend line to the right.

Each trend line method is run at ν values from zero to one at 0.1 intervals. The pre-PCR variation parameter value giving sds closest to one is used when further testing the methods. This was evaluated by taking the median standard deviation over all treatment plants for a particular trend line method and viral target. The median absolute deviation was used to determine the spread of values.

3.2.2 Trend line methods

Five methods were chosen for testing. The first three were selected because they are commonly used for plotting the trend of WBE data and are fast to run[33]: rolling mean, rolling median and loess, all of which rely on a centrally-aligned rolling window. Trend decomposition using loess was also tested. Compared to plain loess this involves the modelling of a seasonal component in addition to the trend line[40]. The final two methods tested make use of a right-aligned rolling window for determining the expected concentration. These are exponential smoothing and a weighted rolling median.

One criteria of the trend line method selected is that it must be able to tolerate missing values as we do not sample every day (generally five samples are taken each week). For the three rolling functions and loess, missing data were excluded (e.g. the median could be calculated from five values rather than seven) whereas for exponential smoothing missing data were interpolated using the method built into the R function. Trend decomposition was done using a function robust to missing data.

Additionally, for loess, trend decomposition and exponential smoothing, the data were transformed into log space before determining the trend line, to prevent negative trend line values. All five methods are detailed further in the following sections and summarised in Table 3.1.

Rolling mean

The rolling mean,

$$MA_k(t) = \frac{1}{k} \sum_{i=\frac{k-1}{2}}^{\frac{k-1}{2}} l_{t+i}, \tag{3.8}$$

was implemented using the `zoo` package `rollapply` function and `base R mean` function. A centrally-aligned rolling window of seven days ($k = 7$) was used as this is a widely adopted convention for reporting case data, and the mean was calculated excluding missing data values.

Rolling median

In the same way as for rolling mean, the rolling median,

$$RM_k(t) = \text{median}(l_{i:t-\frac{k-1}{2}<i\leq t+\frac{k-1}{2}}), \tag{3.9}$$

made use of the `rollapply` function from the `zoo` package and `median` function from `base R`. Again, a centrally-aligned rolling window of seven days was used ($k = 7$) and missing data values were excluded.

Loess

The `loess` function from the `stats` package was used to calculate the smoothed trend line[41]. In order to prevent negative results, the data were first log transformed. Zero measurements were replaced with 1000 gc/person/day which is a load corresponding to around half of the lod. Missing data values were excluded. The span parameter was calculated such that a window of 28 days was used for smoothing. This is a centrally-aligned window with decreasing weights given to the observations at the edge of the window.

Trend decomposition

The `stlplus` function from the package of the same name is used to decompose the time series into seasonal and trend components. This function has been adapted from the `stats` package `stl` function to handle missing data points[40]. As with loess, data were first log transformed and zero measurements were replaced with 1000 gc/person/day. The "seasonal" component was set to seven days, considering sample delay effects that we know vary throughout the week. The trend component was determined using loess with a window size of 28 days.

Exponential smoothing

Exponential smoothing was carried out using the `ets` function from the `forecast` package[42]. Again, data were first log transformed and zero measurements were replaced with 1000 gc/person/day. Missing values were then linearly interpolated using the `na.interp` function, also from the `forecast` package. A model with additive errors and trend, no seasonal component, and no damping was used. The first smoothing parameter for the level, $\alpha = 0.3$, was selected such that 94% of the weight is given to the most recent seven days. The second smoothing parameter for the trend, $\beta = 0.1$, has 90% of the weight over 28 days. This is because we consider the level to be more closely associated with measurements closer in time, while smoothing effects of the trend are greater.

Weighted rolling median

Using the `zoo` package `rollapply` function, a custom function was written to calculate the weighted median:

$$WRM_k(t) = \text{weighted\_median}(l_{i:t-k<i\leq t}, w_{j:1\leq j\leq k}), \tag{3.10}$$

where the weights $\sum_{j=1}^{k} w_j = 1$ and the weighted median is defined as the load l_i satisfying

$$\sum_{j=1}^{i-1} w_j \leq \frac{1}{2} \quad \text{and} \quad \sum_{j=i+1}^{k} w_j \leq \frac{1}{2}. \tag{3.11}$$

Exponentially decreasing weights were given to observations, with most recent observations receiving the highest weight. Missing data points were then removed and the observation at half of the remaining weight was selected as the median value. Again, a seven-day window was used ($k = 7$), this time right-aligned.

Table 3.1: Overview of trend line methods and their parametrisation.

Method	Conditions
Rolling mean	`zoo::rollapply()` centred window of seven days, missing values excluded
Rolling median	`zoo::rollapply()` centred window of seven days, missing values excluded
Loess	`stats::loess()` log transformation, span = 28/number of observations, na.action = exclude
Trend decomposition	`stlplus::stlplus()` log transformation, seasonal window of seven days, trend window of 28 days
Exponential smoothing	`forecast::ets()` log transformation, AAN model, $\alpha = 0.3$, $\beta = 0.1$, dampened = FALSE, na.action = interpolate
Weighted rolling median	`zoo::rollapply()` right-aligned window of seven days with exponentially decreasing weights for historic measurements, missing values excluded

3.3 Evaluation of method performance

The main evaluation was to compare the performance of the outlier detection when the different trend line methods were used. Initially this was done retrospectively, and then in a simulated real-time fashion. Results were evaluated for individual time series (the 70 unique combinations of treatment plant and viral target) by comparing predicted and manual outlier labels.

Using the manual outlier labels as the ground truth, the predicted labels were classified as either true positive (TP), true negative (TN), false positive (FP) or false negative (FN). These values can then be used to calculate various performance metrics such as

$$\text{sensitivity} = \frac{\#TP}{\#TP + \#FN} \tag{3.12}$$

and

$$\text{specificity} = \frac{\#TN}{\#TN + \#FP}. \tag{3.13}$$

The sensitivity or true positive rate gives a measure of the proportion of outliers correctly identified by the method. Specificity or true negative rate gives a measure of the proportion of non-outliers correctly labelled[43]. Both of these values can be used to calculate

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}, \tag{3.14}$$

which gives a single metric to compare methods. Balanced accuracy is preferred over accuracy, due to the large class imbalance between outliers and non-outliers[43].

3.3.1 Retrospective outlier detection

Retrospective outlier detection describes the fact that a single trend line for each time series is calculated using the whole data set available and each data point is labelled based on this single trend line. A single pre-PCR variation parameter as defined in Section 3.2.1 is used and the performance of all six trend line methods was evaluated by comparison with the manual outlier labels and calculation of the described metrics.

3.3.2 Real-time outlier detection

To gain an understanding of how the methods would perform if used in the context of a monitoring program, a real-time analysis was simulated. The outlier status of each data point was predicted as if it was the most recent measurement, making use of only data up to that particular day to calculate the trend line. The real-time simulation iterates through the whole data set predicting the status of each measurement and these results are compared with the retrospective outlier prediction when the whole data set was made available. Following the approach for the retrospective outlier detection, a single pre-PCR variation parameter was used, and the performance of all six trend line methods was evaluated with the same metrics described above.

3.4 Effect of outlier removal on modelling

R_t was estimated for five selected time series, with and without outliers, using the R packages estimateR[17] and EpiSewer[18], which are both used to generate R_t estimates for the wastewater monitoring dashboard[39]. This was done to gain an understanding of the influence of outliers on the modelling output.

EstimateR takes noisy and delayed observations of infection events to estimate R_t. It does this by smoothing the data using loess, reconstructing a time series of infection events from this smoothed data before finally using the EpiEstim package to estimate R_t from the inferred series of infection events. Uncertainty intervals are generated by creating bootstrap replicates from the data which then undergo the same estimation steps. In addition to wastewater measurements, the package is compatible with observations including case confirmations, hospital admissions or deaths.

The EpiSewer package has been specifically designed for WBE. It adopts a Bayesian generative model for estimating R_t. Concentration measurements are taken as observations and are normalised using flow data. Measurement uncertainty is incorporated via Markov Chain Monte Carlo sampling, and estimates are reported with credible intervals. One aspect of the model is the *measurement distribution*, which is the distribution from which the measurement observations can be expected to arise from. It accounts for the dependence of the measurement noise on the concentration but does not model outliers. Two set ups were tested: one where this measurement distribution was modelled as a truncated normal distribution, and one where it was modelled as a gamma distribution.

For selected time series, R_t estimates were calculated with and without outlier data points labelled retrospectively using the trend decomposition method. Initially this was done for the time series containing the largest magnitude outlier. An additional four time series were selected which contained outliers in the beginning of a wave, the peak of a wave, the end of a wave, or in a stationary

period of time to see if the location of an outlier impacts its effect on model results. Additionally, to see the "real-time" impact of outliers on the estimation, the time series data were either used in their entirety or truncated one week after the last observed outlier. Overall, this resulted in 54 modelled R_t time series. Results were analysed by visualisation of the R_t estimates.

3.5 Simulating outliers

A simulated data set was created for further method validation and to test assumptions made about the outlier generation process. As input, the simulation took the expected viral load trend line, and used this to generate measurement data points for each day. It was assumed that there was a constant daily probability of a measurement being an outlier and that the "outlier status" of any day was independent from any other day.

An example viral load trend line to use as an input was generated from the real-world data (using trend decomposition). This simulation input was taken as the expected viral load, $\mathrm{E}[l_t]$. It was converted into a concentration

$$\mathrm{E}[c_t] = \frac{\mathrm{E}[l_t] \cdot P}{f_t,} \tag{3.15}$$

where P is the catchment population and f_t is the flow through the treatment plant on a particular day. These parameters were taken from the real-world data.

The status of each time point as either an outlier or not, o_t, was determined by randomly sampling a Bernoulli distribution,

$$\begin{gathered} o_t \sim B(1, p) \\ \text{where} \quad p = \text{outlier proportion.} \end{gathered} \tag{3.16}$$

An average outlier proportion (i.e. average proportion of outlier data points from all measurements) from the manually labelled data was taken, giving a probability of 0.005.

Measurements were simulated as a combination of pre-PCR noise, an outlier addition component, and measurement noise. First the pre-PCR noise around $\mathrm{E}[c_t]$ was simulated to define a pre-PCR level, s_t. This was done by drawing values from a log-normal distribution with a mean of the expected concentration and a coefficient of variation equal to the pre-PCR variation parameter ν defined in Section 3.2.1:

$$\begin{gathered} s_t \sim \text{Lognormal}(\mu, \sigma^2) \\ \text{where} \quad \mu = \log(\mathrm{E}[c_t]) - \frac{\sigma^2}{2}, \quad \sigma^2 = \log(1 + \nu^2). \end{gathered} \tag{3.17}$$

Though the log-normal distribution deviates from the assumption of a normal distribution made in Section 3.2, it was selected to ensure only positive values were simulated.

An "outlier addition" concentration, a_t was drawn from an exponential distribution for each outlier time point:

$$a_t = \begin{cases} 0 & o_t = 0 \\ \sim \exp(\lambda) & o_t = 1 \end{cases} \\ \text{where} \quad \lambda = \frac{-\log(1-0.99)}{4 \cdot \max(\mathrm{E}[c_t])}. \tag{3.18}$$

An exponential distribution was chosen as it can take only positive values, has a single rate parameter to define, and has an easily usable quantile function. The maximum outlier magnitude above the maximum trend value was determined from the real-world data and used along with the quantile function to define λ.

A measurement level,

$$m_t = s_t + a_t, \tag{3.19}$$

was calculated for each time point as a combination of the pre-PCR noise and outlier addition. The final simulated concentration c_t was taken from a log-normal distribution again, with mean equal to m_t and a coefficient of variation calculated using the measurement error (case when $\nu = 0$ from equation 3.3):

$$c_t \sim \text{Lognormal}(\mu, \sigma^2) \\ \text{where} \quad \mu = \log(m_t) - \frac{\sigma^2}{2}, \quad \sigma^2 = \log\left(1 + \left(\frac{\sqrt{\exp(c_t/d \cdot v) - 1}}{\sqrt{p \cdot n} \cdot c_t/d \cdot v}\right)^2\right). \tag{3.20}$$

A simulated data set was generated for each of the 70 time series. To compare how well the simulation recreated real-world data, the distribution of the deviations of the real-world data and simulated data from the input trend line were plotted. The complete retrospective and real-time analysis were repeated and evaluated on the simulated dataset.

Finally, to gain insight into the uncertainty of the method, several simulations were carried out on a single time series. One set of 50 simulations was done with the same parameters as before and an additional 50 where the outlier frequency was set to zero.

Chapter 4

Results

4.1 Manually labelled outliers

Out of the 70 time series given to experts for manual outlier labelling 47 contained at least one outlier data point. The average outlier frequency was 1.8 per year for these time series, which decreases to 1.2 per year when the average of all 70 time series is taken. An initial analysis of these manually labelled data points indicated co-occurrence of N1 and N2 outliers and showed no observable patterns regarding sample processing days in the laboratory (Figure 4.1).

It was also of interest to see if there were any initial potential explanatory factors for variations in outlier frequency. Catchment population was hypothesised to play a role in the frequency of observing outliers. Figure 4.2 shows outlier frequency plotted against catchment population size. While there is no correlation between outlier frequency and the catchment population size, there is a decrease in variation of outlier frequency as catchment population increases.

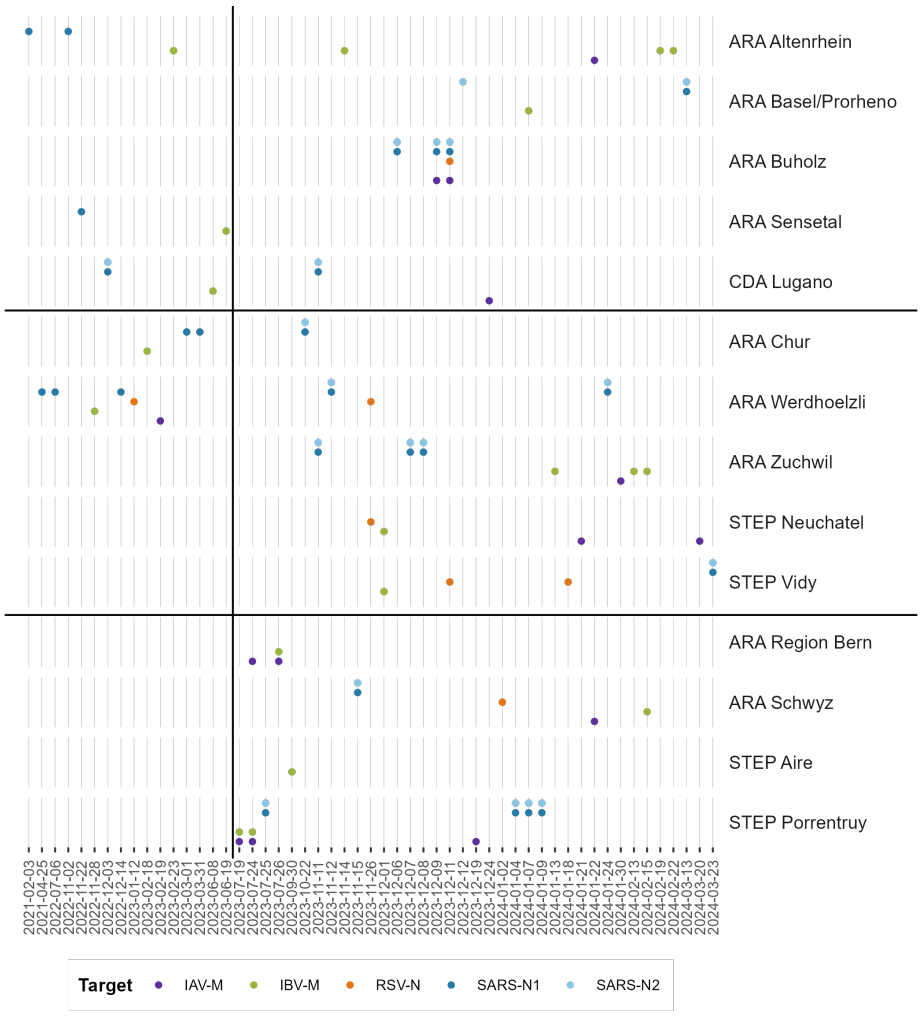

Figure 4.1: Overview of data points marked as high outliers by domain experts. The vertical line marks when monitoring increased from six treatment plants to fourteen and the horizontal lines group treatment plants based on when they are processed in the lab. The colours indicate the different viral targets.

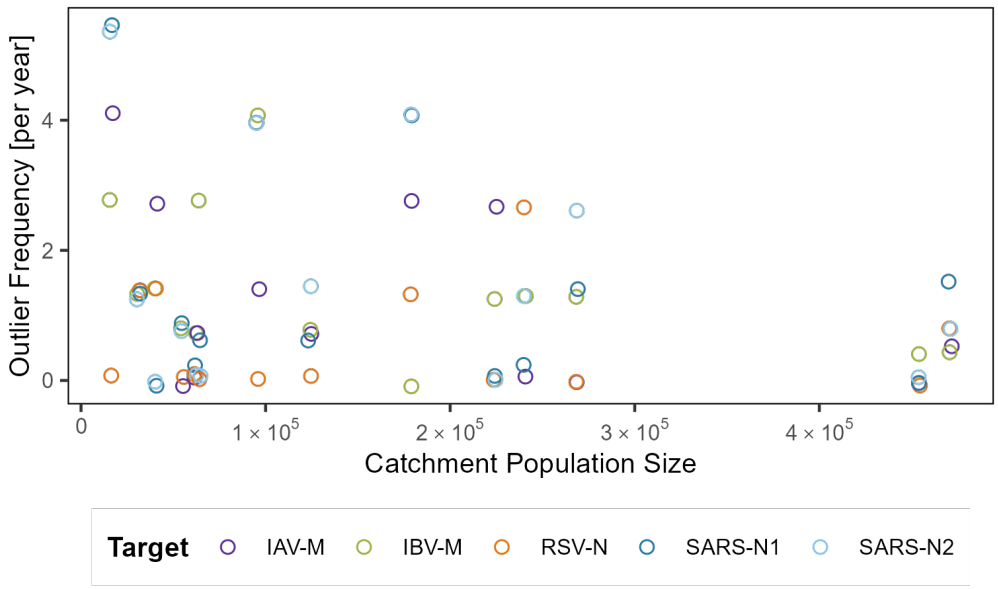

Figure 4.2: Outlier frequency plotted against catchment population size. The colours indicate the different viral targets. A small amount of random noise has been added to the points to reduce overplotting.

4.2 Parameter selection

As described in Section 3.2, the goal of the deviation normalisation is to have a collection of measurement deviations from the trend line which are standard normal distributed. The pre-PCR variance parameter was calibrated to yield a standard deviation of one. Figure 4.3 shows how standard deviation decreases with increasing pre-PCR variance for each of the trend line methods tested. All targets follow this trend, however the SARS-CoV-2 curves are generally lower compared to the other viruses, with the exception of exponential smoothing where they are higher. Differences between the viral targets could be due to their different average concentrations in the wastewater. The optimal parameter value varies between 0.4 - 1 for the different trend line methods and viral targets. A value of 0.6 was selected for testing because it gave values closest to one for the most methods and targets.

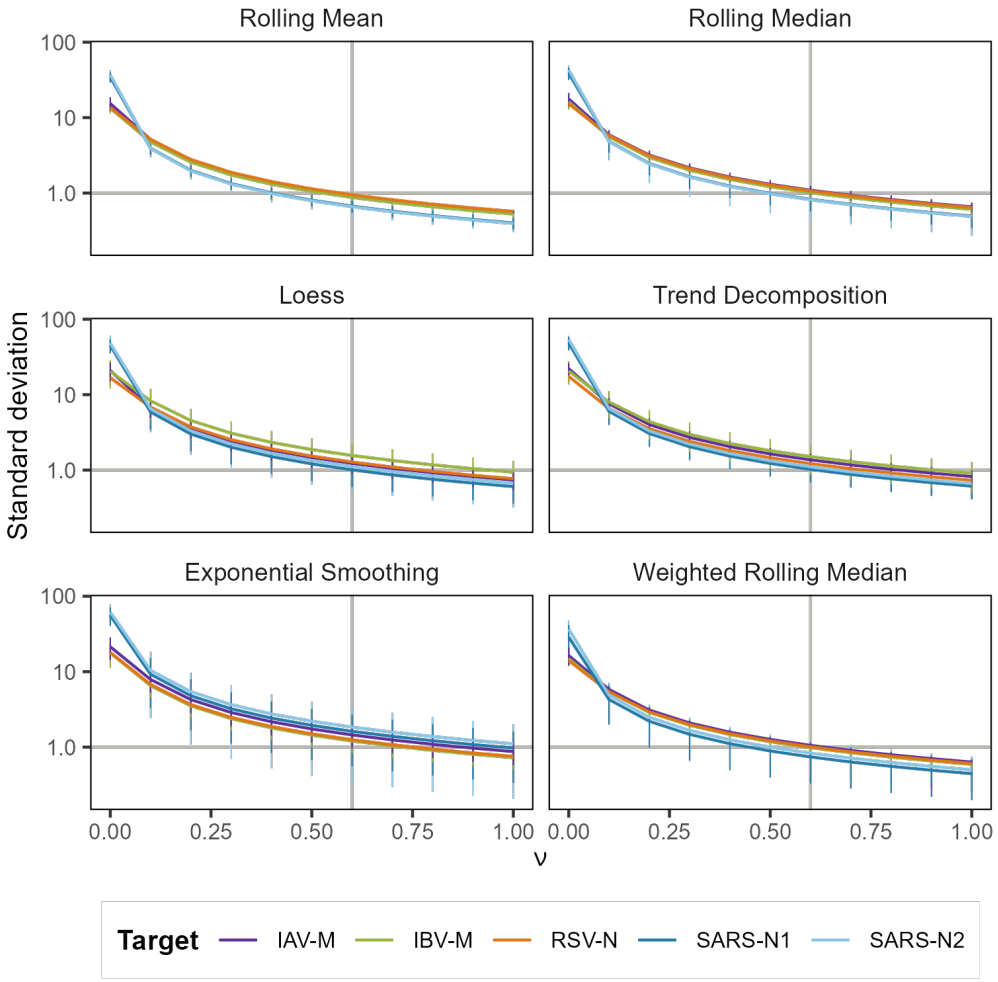

Figure 4.3: Standard deviations of measurement deviations normalised with different pre-PCR variance parameters. Lines show the median standard deviation over the 70 time series with error bars showing the median average deviation of this range. The colours indicate the different viral targets.

4.3 Outlier detection method performance

As discussed in Section 3.3, the performance of the different trend line methods was evaluated by calculating a confusion matrix using the manually labelled outliers as a reference. This was done for the retrospective and real-time results. Performance metrics for each method can be seen in Figure 4.4. A perfect trend line method would have high sensitivity and specificity equal to one. As can be seen from the methods tested, there is a trade-off between these two values. High specificity means fewer false positives while high sensitivity means fewer false negatives.

From the retrospective results, it can be seen that, apart from the rolling mean, the trend line methods making use of a centrally-aligned window (trend decomposition, loess and rolling median) have higher sensitivity compared to the right-aligned methods (exponential smoothing, weighted rolling median). Since the right-aligned methods have less data available to them for calculating the trend line (only data before the observation), it is intuitive that they would perform less well than methods which have data before and after the observation available.

When the centrally-aligned window methods are applied in real time their performance decreases. This is due to the reduction in data available to them for calculating the trend line "at the edge". For the final time point, only half of the window was available for calculations and as such in the real-time simulation every outlier label was predicted with half the amount of data compared to the retrospective approach. Since the right-aligned methods do not make use of this data in the first place, their performance is unchanged when used in real time.

Due to the large class imbalance (very few outliers compared to many normal measurements) changes in specificity are very slight, (note the scale of the y-axis in Figure 4.4). For this reason the balanced accuracy results shown in Figure 4.5 are largely determined by the sensitivity of the method. Again, the decrease in performance for the centrally-aligned methods in real time compared to retrospectively can be observed.

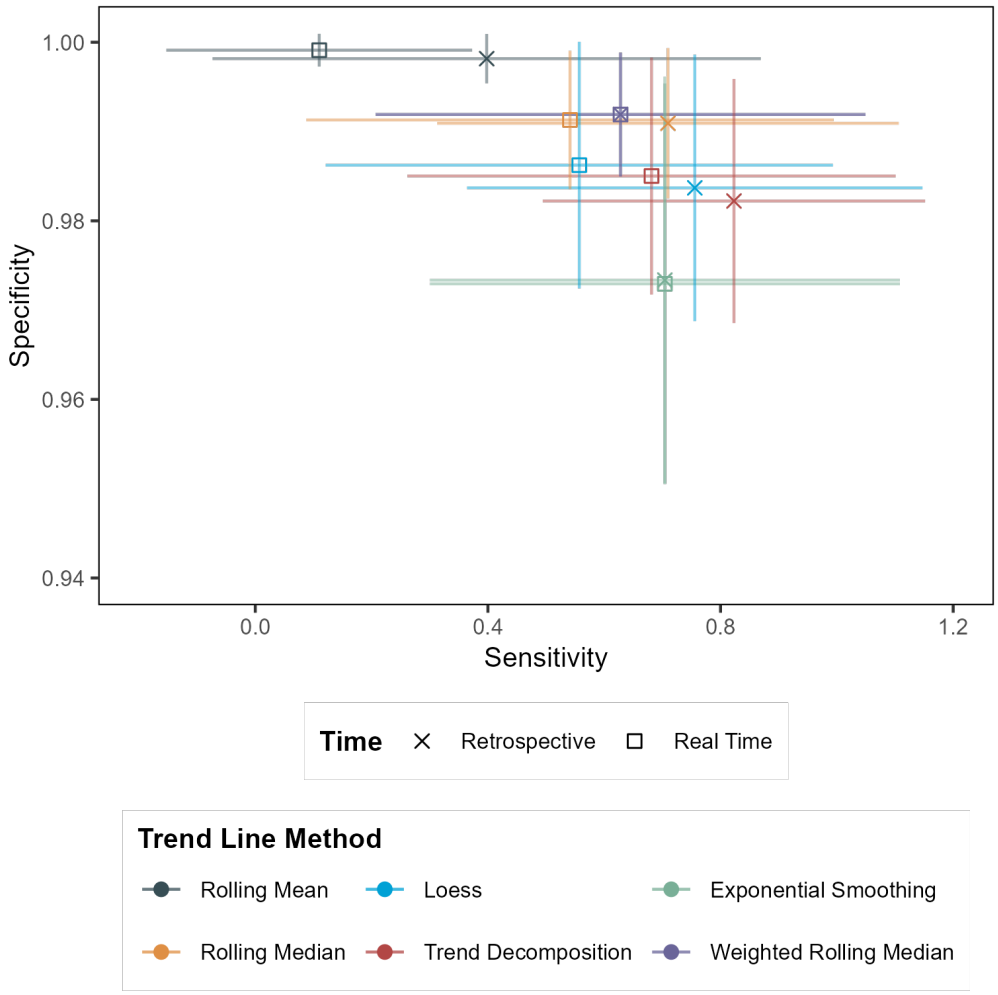

Figure 4.4: Mean sensitivity and specificity for each method (different colours) when applied retrospectively (crosses) and in simulated real time (squares). Error bars show standard deviation of the 70 time series.

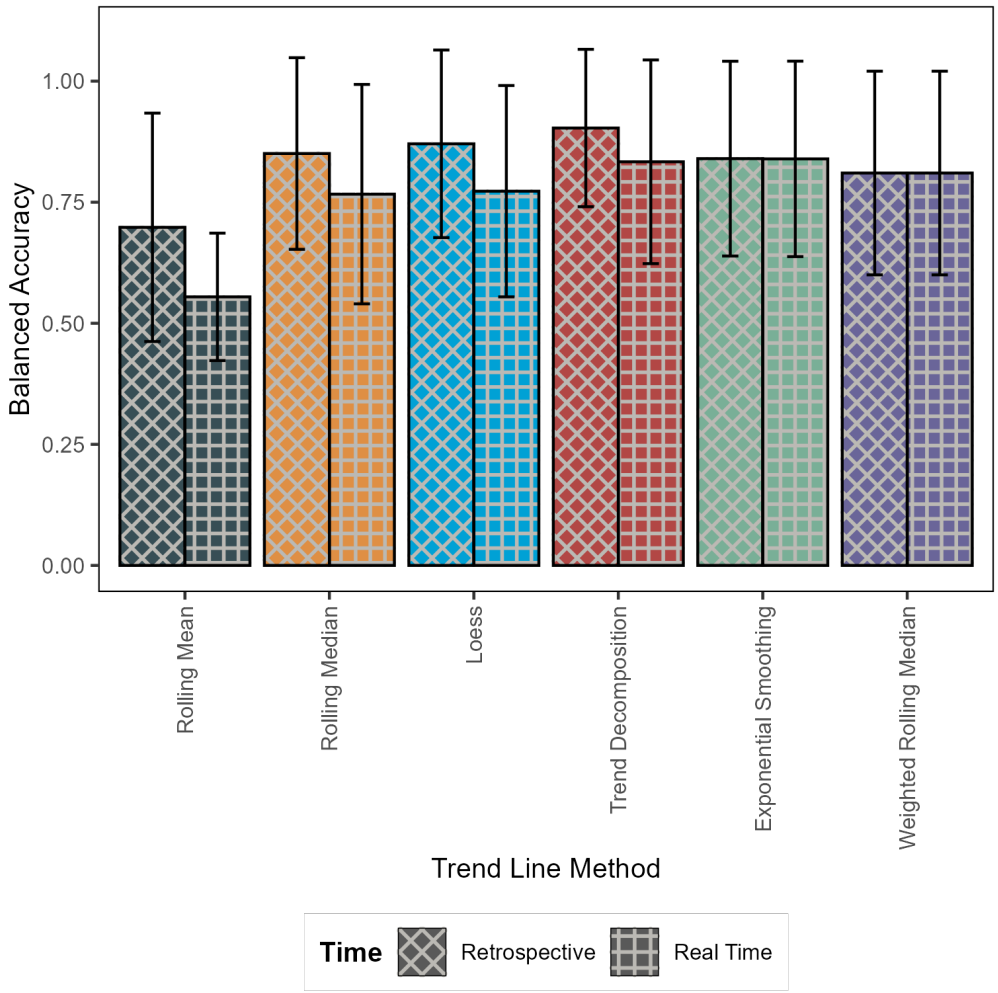

Figure 4.5: Mean balanced accuracy scores for each of the trend line methods. Cross-hatched bars show results from the retrospective application while square-hatched bars show results from the real-time simulation. Error bars show standard deviation of the 70 time series.

Outlier frequency was also calculated for each time series and again plotted against catchment population size (Figure 4.6). Generally, the same trend of decreasing variation with increasing catchment size can be observed. For the exponential smoothing and loess methods it appears that SARS-CoV-2 time series have a higher outlier frequency compared to the other targets. The rolling mean results in an outlier frequency generally below the manual average.

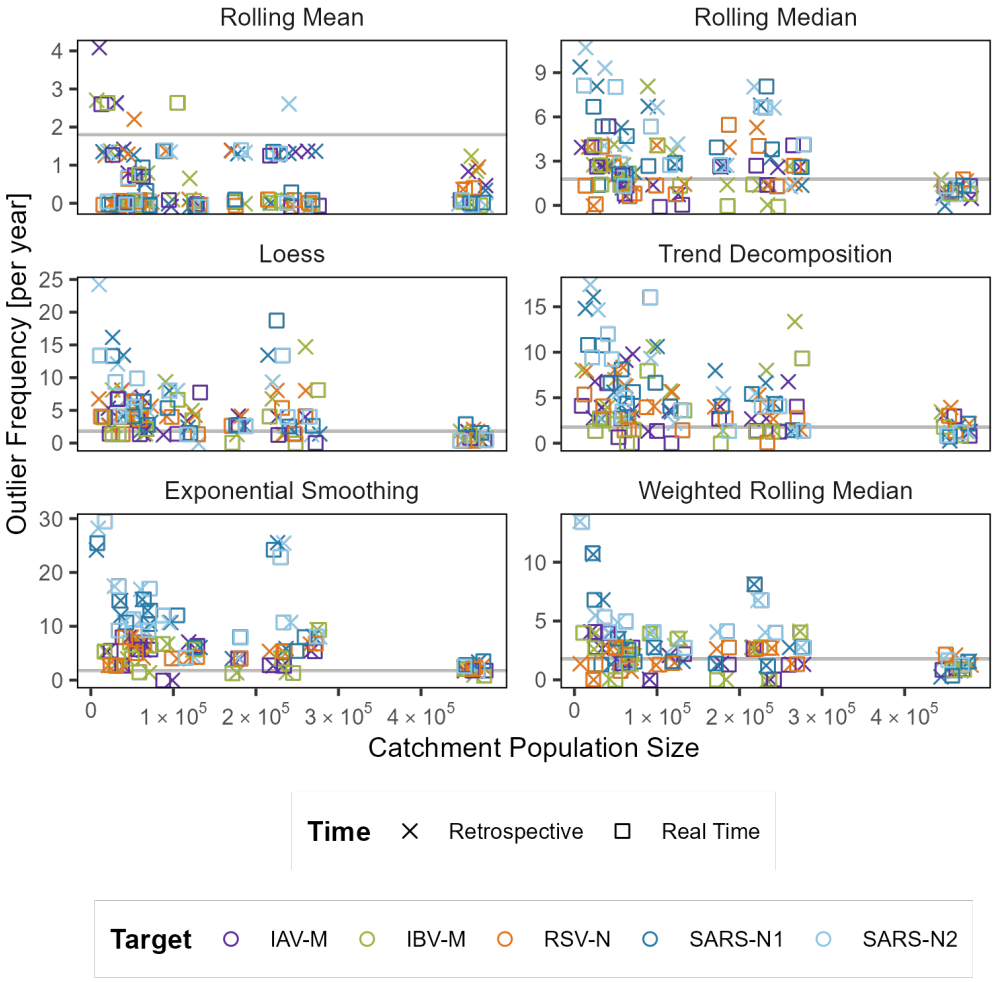

Figure 4.6: Outlier frequency plotted against catchment population size for each of the methods tested retrospectively (crosses) or in real time (squares). The colours indicate the different viral targets. The horizontal grey line marks the average outlier frequency from the manually labelled data set (1.8). A small amount of random noise has been added to the points to reduce overplotting.

Specific time series have been examined more closely to gain a deeper understanding of the results, and Figure 4.7 shows the trend lines plotted for a time series where no data points were manually identified as an outlier. In this case the trends determined by all six trend line methods are comparable.

In the presence of outliers, this agreement changes, and we can see that the robustness of the method is an important factor in performance. For example, the rolling median identifies twice as many outliers compared to rolling mean (60 vs. 30). Figure 4.8 shows that in the presence of outliers, the rolling mean can introduce large fluctuations to the trend line. In this case the value is so extreme that it is still correctly labelled by all methods as an outlier, however, in several less extreme cases the outlier goes undetected. Exponential smoothing also appears to be more susceptible to the extreme outlier, likely due to interpolation of missing data points around the outlier.

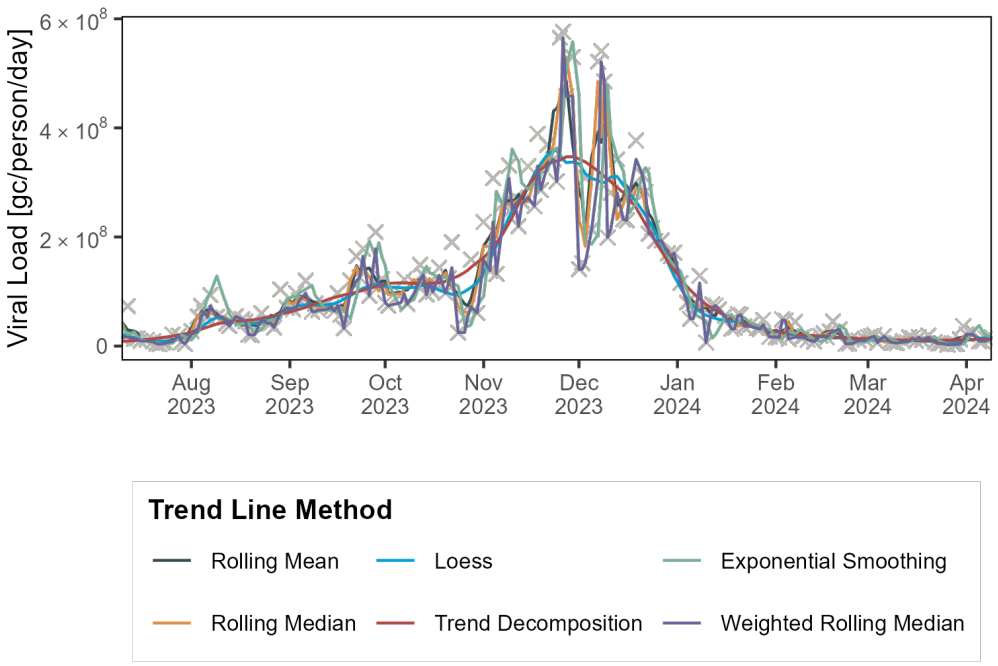

Figure 4.7: Trend lines generated from the different methods (different colours) when given SARS-N2 data from STEP Neuchatel. This time series contained no data points manually labelled as an outlier.

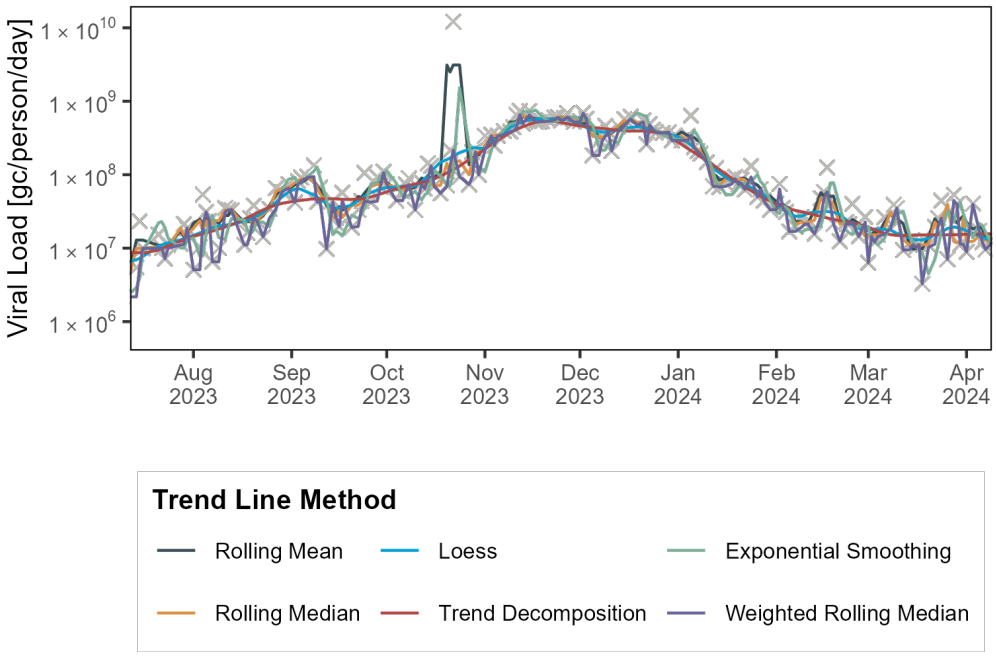

Figure 4.8: Trend lines calculated using SARS-N2 data for ARA Chur with different trend line methods (different colours).

Of the four methods making use of a centrally-aligned window, trend decomposition performs the best when using the manual labelling for evaluation. With the manual labels used as a reference, 16 data points are considered false negatives, the majority of which (12) are found at the peaks of trend line waves. This highlights a potential downside of the manual labelling, i.e. that experts perhaps struggle to intuit the increased variation at higher concentrations compared to lower concentration measurements, especially when viewed on the same graph, and this could be a contributing factor to the false negative values. Figure 4.9 shows how one manually labelled outlier at the peak of the wave is in fact within the expected measurement variation whereas there are points at lower concentrations which are outwith this range that were not manually identified as outliers.

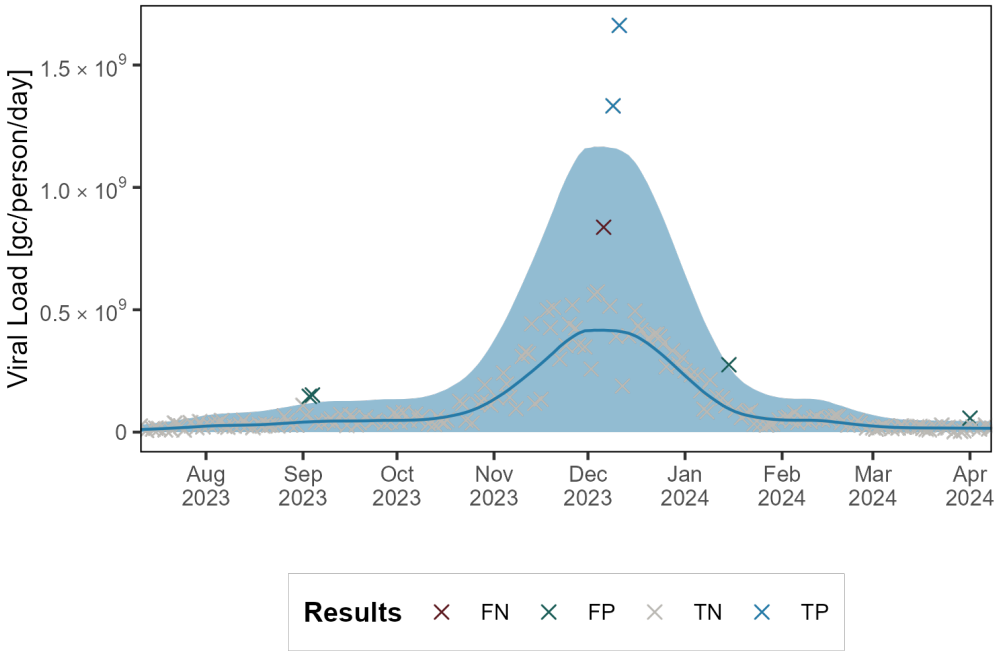

Figure 4.9: Trend line calculated using SARS-N1 data for ARA Buholz using trend decomposition. Points are coloured by the result of the outlier detection metric. The ribbon shows three times the expected standard deviation at the trend line concentration - points which fall out of this range are marked as outliers by the algorithm.

Loess, compared to trend decomposition, identifies 5 fewer outliers and rolling median a further 5. All 10 of these data points are found at the peaks of waves highlighting how the "strength" of the trend line method's smoothing effect impacts outlier identification. Crucially, the trend decomposition and loess methods make use of a larger window for smoothing compared to the rolling median.

This effect of increased smoothing strength can be seen in results where the rolling window length of the rolling median is increased. Figure 4.10 shows the increasingly smooth trend lines and Figure 4.11 shows this impact on the sensitivity and specificity. By increasing the rolling median window from 7 to 14 days, sensitivity is already increased to match that of the retrospective trend decomposition, with sensitivity only slightly improved by further window size increases. Specificity remains higher for rolling median at all window sizes compared to the trend decomposition.

Interestingly the reverse effect is observed when increasing the window size of the right-aligned weighted rolling median method (Figure 4.11). Here specificity increases with increasing rolling window while sensitivity decreases. Figure 4.12 shows how the trend line becomes more affected by outlier points as a larger window is considered for calculation.

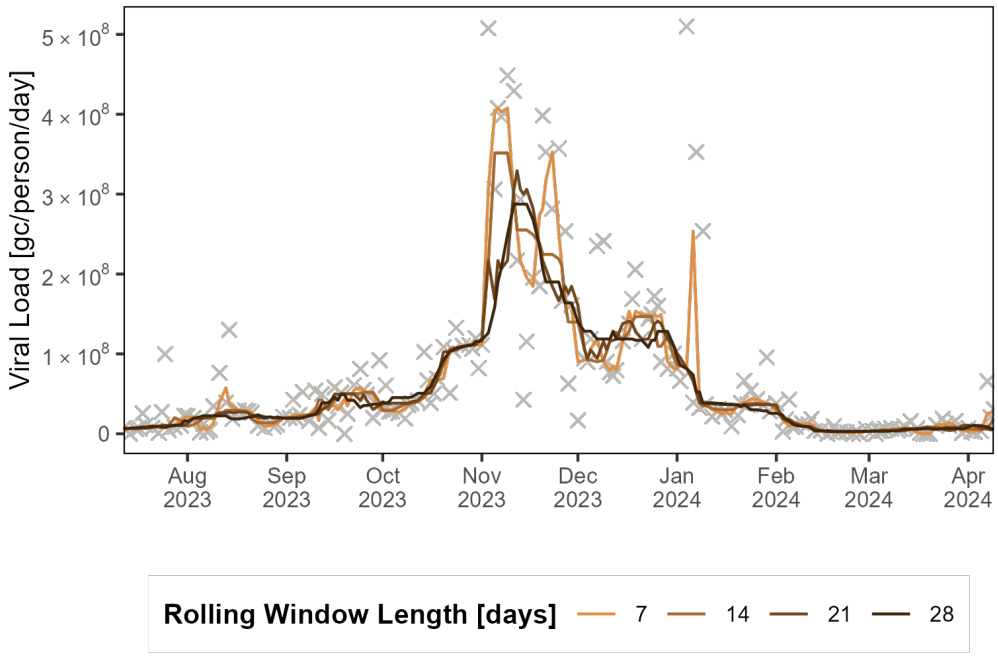

Figure 4.10: Trend line calculated using SARS-N2 data for STEP Porrentruy using rolling median with increasing window size (darker colours).

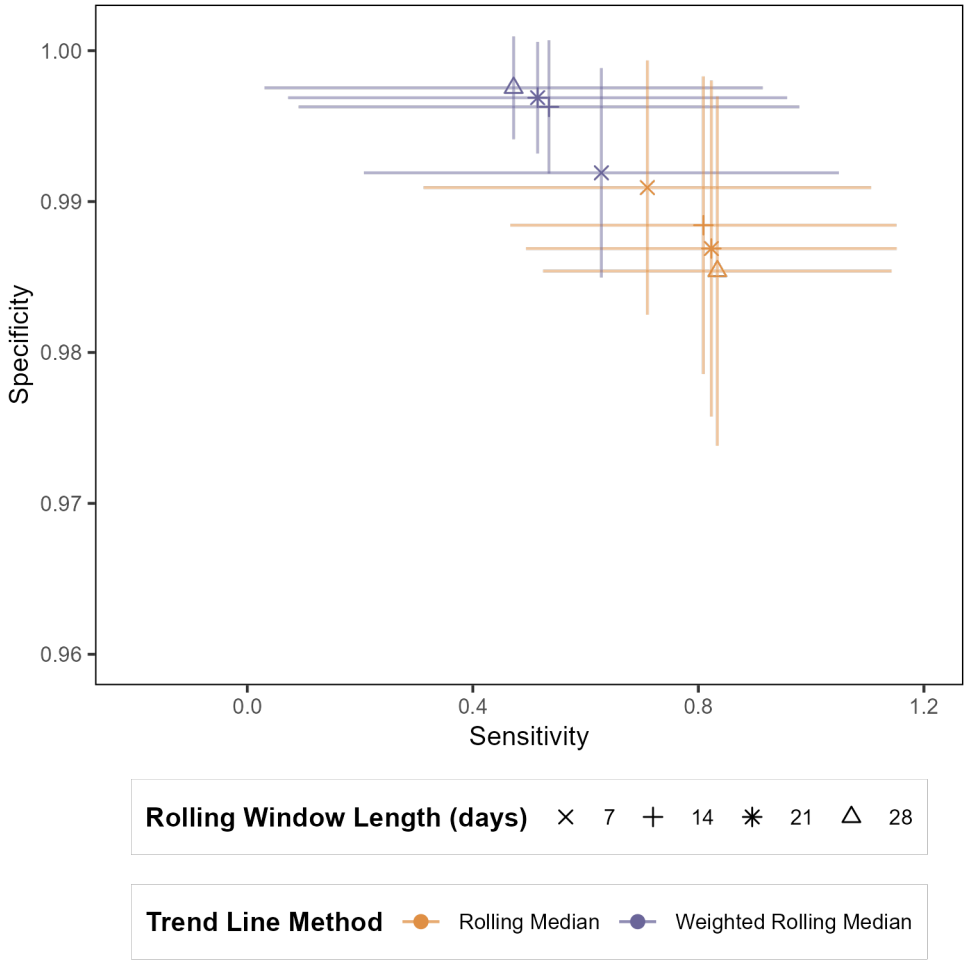

Figure 4.11: Mean sensitivity and specificity for rolling median (yellow) and weighted rolling median (lilac) when different rolling window sizes are used (different shapes). Error bars show standard deviation of the 70 time series.

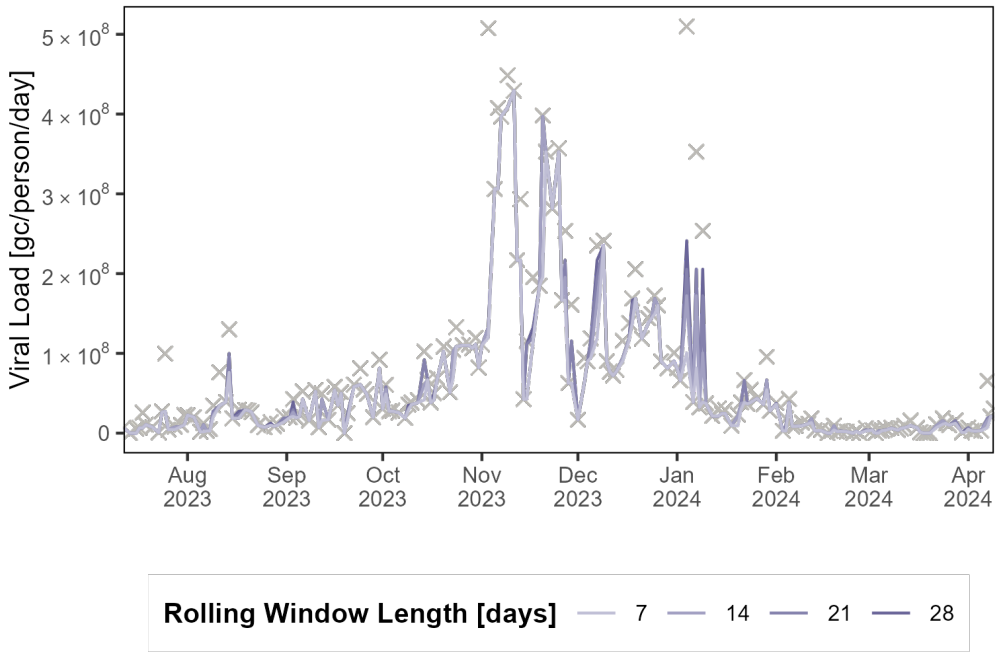

Figure 4.12: Trend line calculated using SARS-N2 data for STEP Porrentruy using weighted rolling median with increasing window size (darker colours).

There are some false negatives which do not occur at the peak of a wave (4 for trend decomposition) and therefore not explainable by the intuition of measurement variation at the peak. These appear as outliers in terms of viral load likely due to the flow normalisation, where a high flow value inflates the viral load. Figure 4.13 shows an example of this where the false negative value occurs on a day when the flow measurement is greater than the 99.7% quantile. The panel showing the concentration measurements shows that the false negative point fell within the expected measurement variation and is therefore not marked as an outlier.

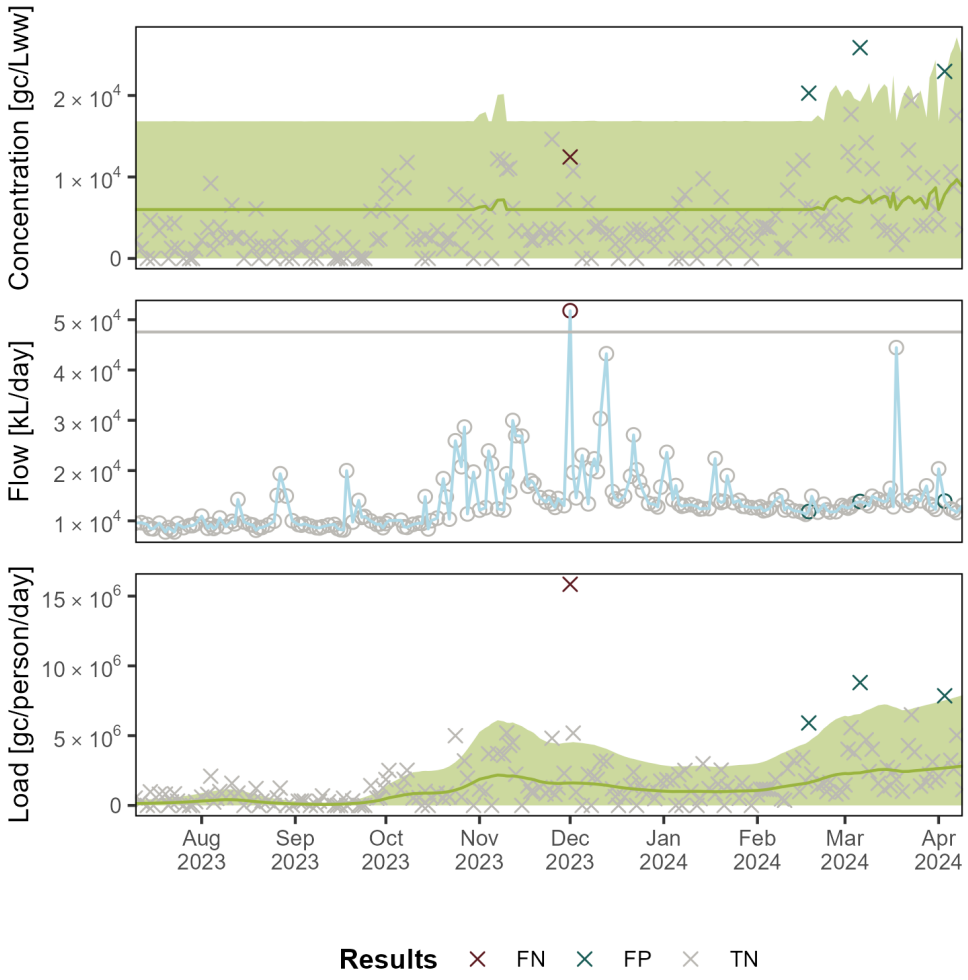

Figure 4.13: Viral loads for IBV-M from STEP Neuchatel (bottom panel) are a combination of viral concentration measurements (top panel) and flow readings (middle panel). Colours of markers indicate classification result, the horizontal line on the middle panel marks the 99.7% quantile of the flow data.

4.4 Effect of outlier removal on modelling

To investigate the effect of outlier removal on modelling, first the R_t estimates from the time series with the greatest magnitude outlier ($\tilde{\delta}_{c_t} = 170$) were compared. Figure 4.14 shows how the presence of outliers impacted the results. For estimateR and EpiSewer with the truncated normal distribution, the presence of outliers impacts results throughout the whole time series, whereas for EpiSewer with a gamma distribution, the impact of outlier measurements is limited to their location in time. The EpiSewer estimate without outliers from the truncated normal distribution more closely resembles the estimation with outliers when the gamma distribution is used.

For all model set ups the removal of outliers can be seen to have an effect. Generally, outlier removal "pulls" R_t estimates closer to one compared to when outliers are included. It can also be seen that the R_t "peak" around November 2023 is shifted in time on outlier removal, occurring later and causing a shallower rise and fall in estimated values.

As mentioned in Section 3.4, comparisons of R_t estimates were also made for time series with outliers occurring in different points of a wave ($7 < \tilde{\delta}_{c_t} < 25$) and real-time effects were investigated by truncating the time series one week after the last observed outlier. The EpiSewer estimates using a truncated normal measurement distribution are shown for illustrative purposes. The effects are also apparent for the other set ups, to a greater (estimateR) or lesser (EpiSewer: gamma distribution) extent.

The removal of outliers at the beginning of the wave decreases the rate of increase of R_t and smooths the peak (Figure 4.15). In the middle, meaning around the maximum of viral loads in the wave, the rate of decrease of R_t is increased and the lowest R_t value is found sooner in time (Figure 4.16). At the end of the wave the effects are limited, but generally the estimates are closer to one (Figure 4.17). Finally, when concentrations are reasonably stationary, there appears to be no effect of outlier inclusion or exclusion on R_t estimates (Figure 4.18), though this does not necessarily preclude an effect in the case of a more extreme outlier.

When considering the real time effects, the first thing to note is that model uncertainty is greater closer to the present, regardless of the inclusion or exclusion of outliers. The changes still follow the same patterns as described for the whole time series at the start, middle and end of a wave, however, the effects are noticeably larger (less overlap between the 50% credible intervals). For the stationary phase (Figure 4.18) in real time, the presence of the outlier has a noticeable effect to slightly raise the R_t estimate.

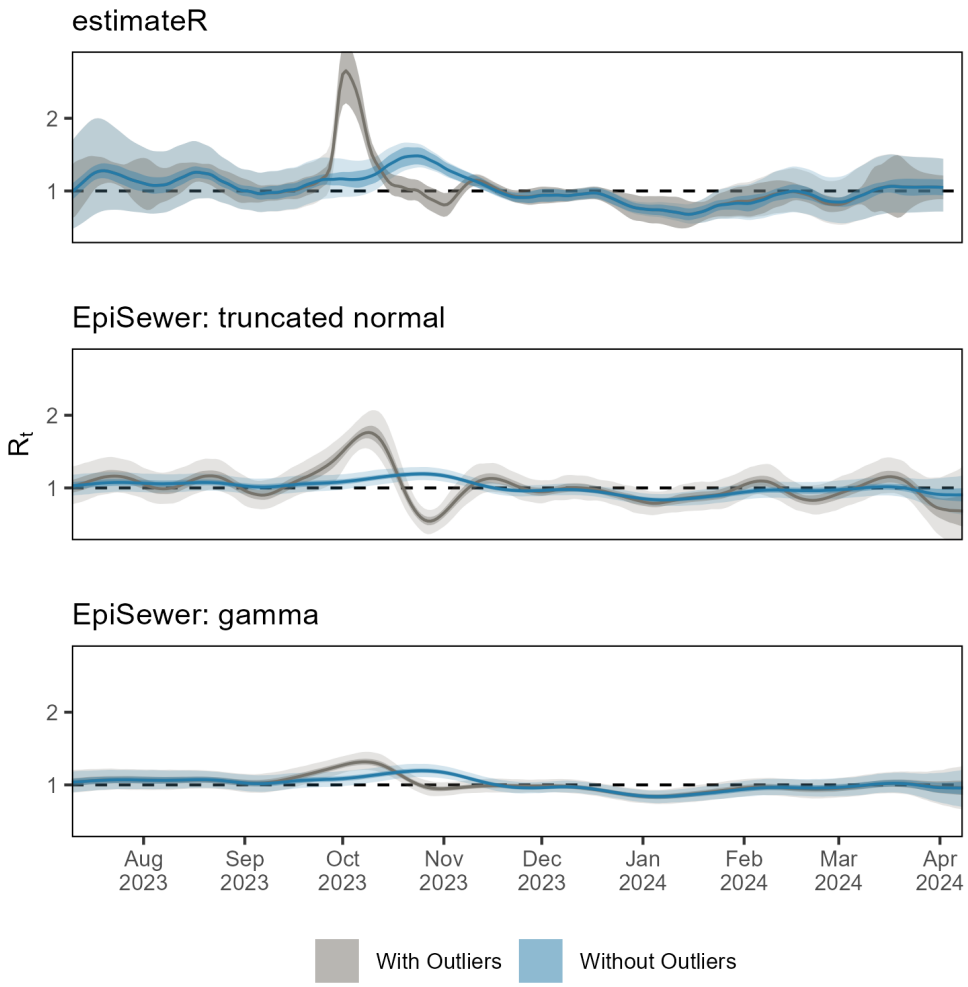

Figure 4.14: R_t estimates from SARS-N1 data for ARA Chur using either a truncated normal distribution or a gamma distribution as the measurement distribution in the model. The data set used either contained outliers (grey) or had outliers removed before estimation (blue). For the estimateR panel the solid line represents the mean and the ribbons represent inner and outer uncertainty bounds. For the EpiSewer panels the solid line represents the median, the darker ribbon the 50% credible interval and the lighter ribbon the 95% credible interval.

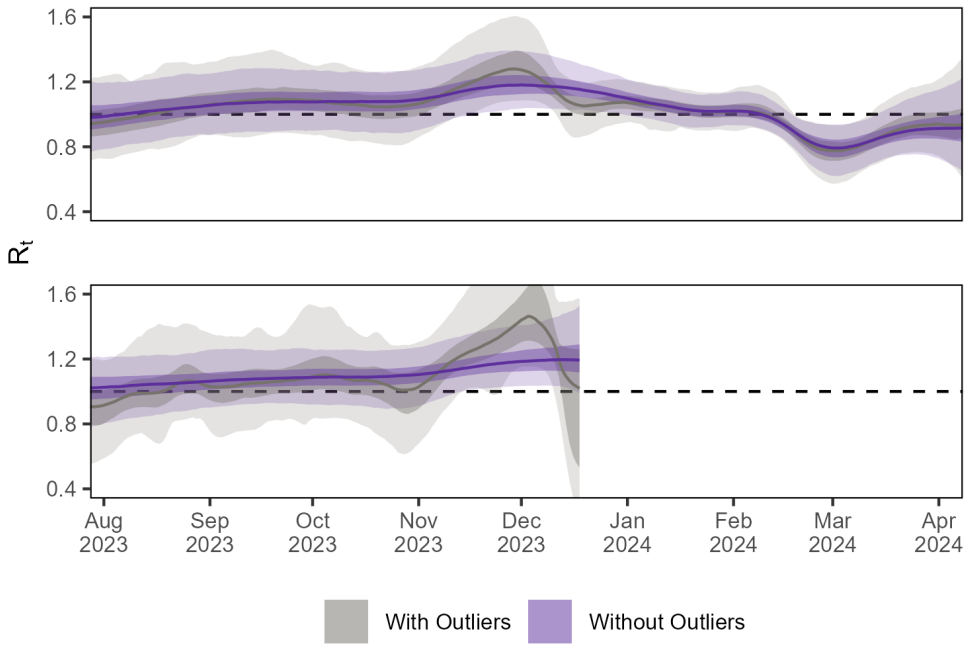

Figure 4.15: R_t estimates where outliers occur at the start of an infection wave. Data are IAV-M measurements for ARA Buholz and a truncated normal distribution was used as the measurement distribution in the model. The top panel shows estimates for the whole time series, while the bottom panel shows estimates up until one week after the last outlier was identified. The data set used either contained outliers (grey) or had outliers removed before estimation (purple). The solid line represents the median, the darker ribbon the 50% credible interval and the lighter ribbon the 95% credible interval.

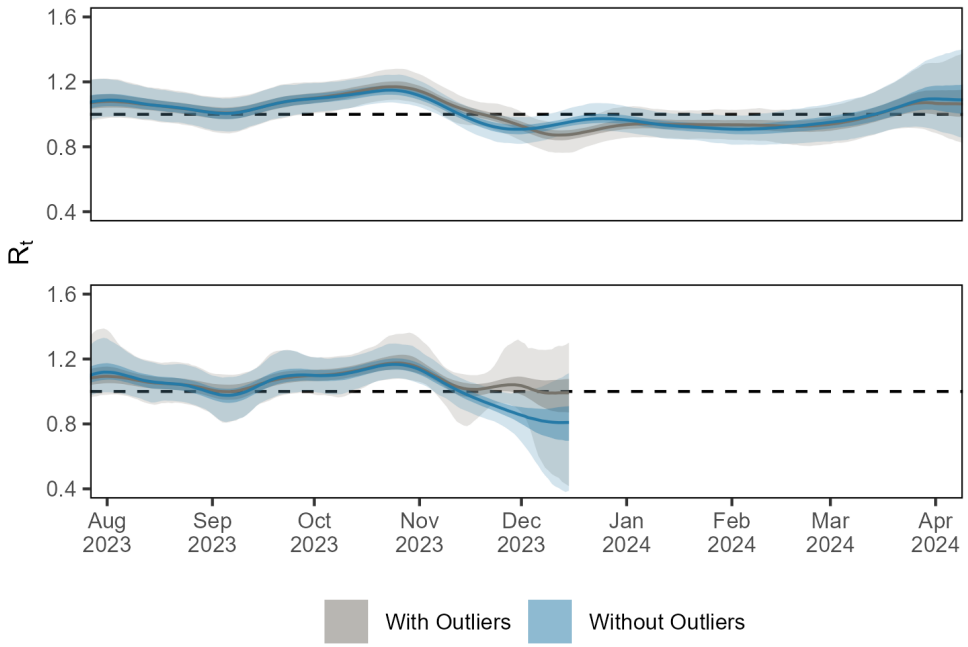

Figure 4.16: R_t estimates where outliers occur in the middle of an infection wave. Data are SARS-N1 measurements for ARA Zuchwil and a truncated normal distribution was used as the measurement distribution in the model. The top panel shows estimates for the whole time series, while the bottom panel shows estimates up until one week after the last outlier was identified. The data set used either contained outliers (grey) or had outliers removed before estimation (blue). The solid line represents the median, the darker ribbon the 50% credible interval and the lighter ribbon the 95% credible interval.

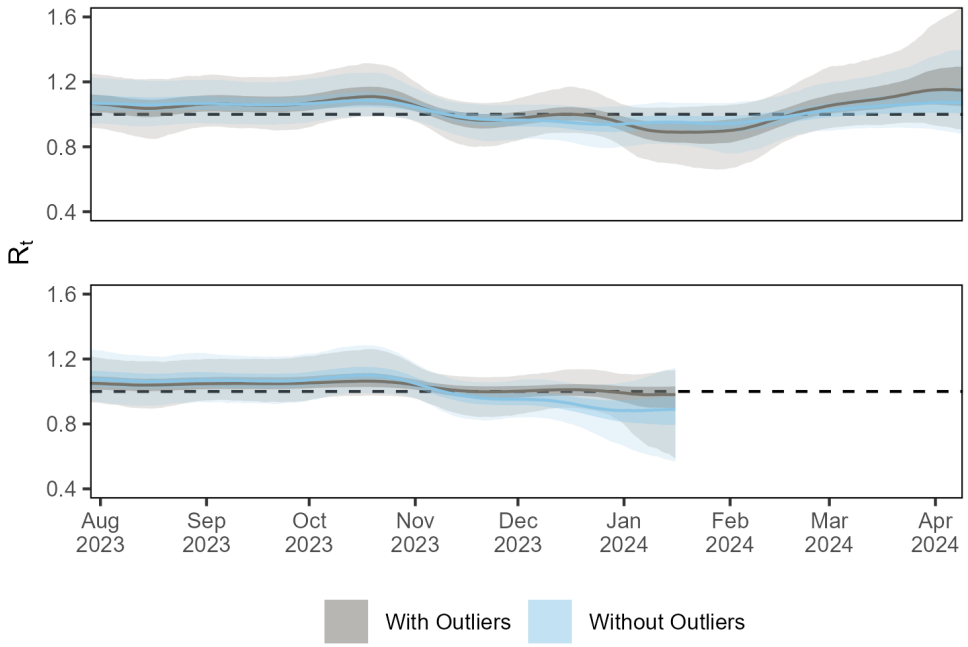

Figure 4.17: R_t estimates where outliers occur at the end of an infection wave. Data are SARS-N2 measurements for STEP Porrentruy and a truncated normal distribution was used as the measurement distribution in the model. The top panel shows estimates for the whole time series, while the bottom panel shows estimates up until one week after the last outlier was identified. The data set used either contained outliers (grey) or had outliers removed before estimation (blue). The solid line represents the median, the darker ribbon the 50% credible interval and the lighter ribbon the 95% credible interval.

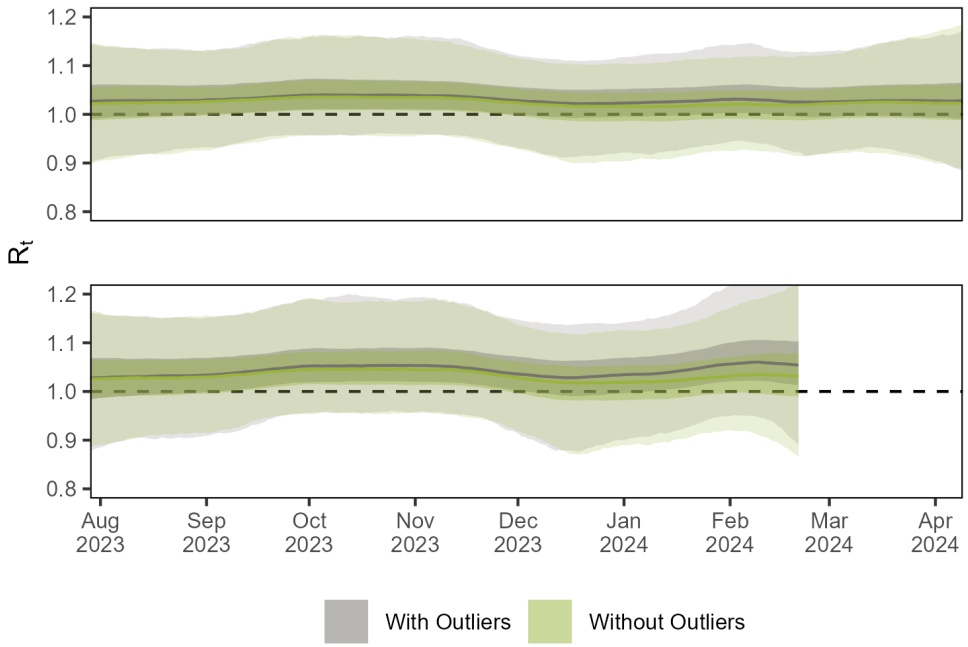

Figure 4.18: R_t estimates where outliers occur in a stationary phase. Data are IBV-M measurements for ARA Schwyz and a truncated normal distribution was used as the measurement distribution in the model. The top panel shows estimates for the whole time series, while the bottom panel shows estimates up until one week after the last outlier was identified. The data set used either contained outliers (grey) or had outliers removed before estimation (green). The solid line represents the median, the darker ribbon the 50% credible interval and the lighter ribbon the 95% credible interval.

4.5 Outlier simulation

An initial comparison of simulated data with real data was done by comparing the deviations of concentration measurements from the input trend line. Figure 4.19 shows the density of these deviations for measurements above the limit of detection. For most time series these are in good agreement, however for time series where measurements are at lower concentration and with less defined trends, agreement is poorer. The peak of deviations for most time series is around zero, i.e. close to the trend line, as expected.

The retrospective and real-time analyses were repeated for the simulated data. Figure 4.20 shows the sensitivity and specificity of the different trend line methods on the simulated data set. Compared to the real-world data, specificity is broadly similar. Sensitivity on the other hand is lower for all methods except rolling mean. Additionally, differences in the retrospective performance of the different methods are minimal, all having a balanced accuracy of around 0.75 (Figure 4.21). Here it can also be seen that the difference between retrospective and real-time application of the method is reduced for the centrally-aligned methods, with trend decomposition showing no difference between the two.

The performance of the different trend line methods was again compared for different time series. In this case we know the input - our expected viral load - so it is possible to see how well the trend line methods can capture this. Figure 4.22 shows the different trend lines compared to the input. The greater smoothing strength of the trend decomposition compared to loess and rolling median can be seen, but all methods can be said to capture the input.

Since the overall data distribution and trend line performance were comparable with the real data, the defining of measurements as outliers was investigated. The outlier addition concentration in several cases is likely too low to generate measurements which fall outside the expected range (Figure 4.23). This would explain the decrease in sensitivity for the simulation compared to the manually labelled real data.

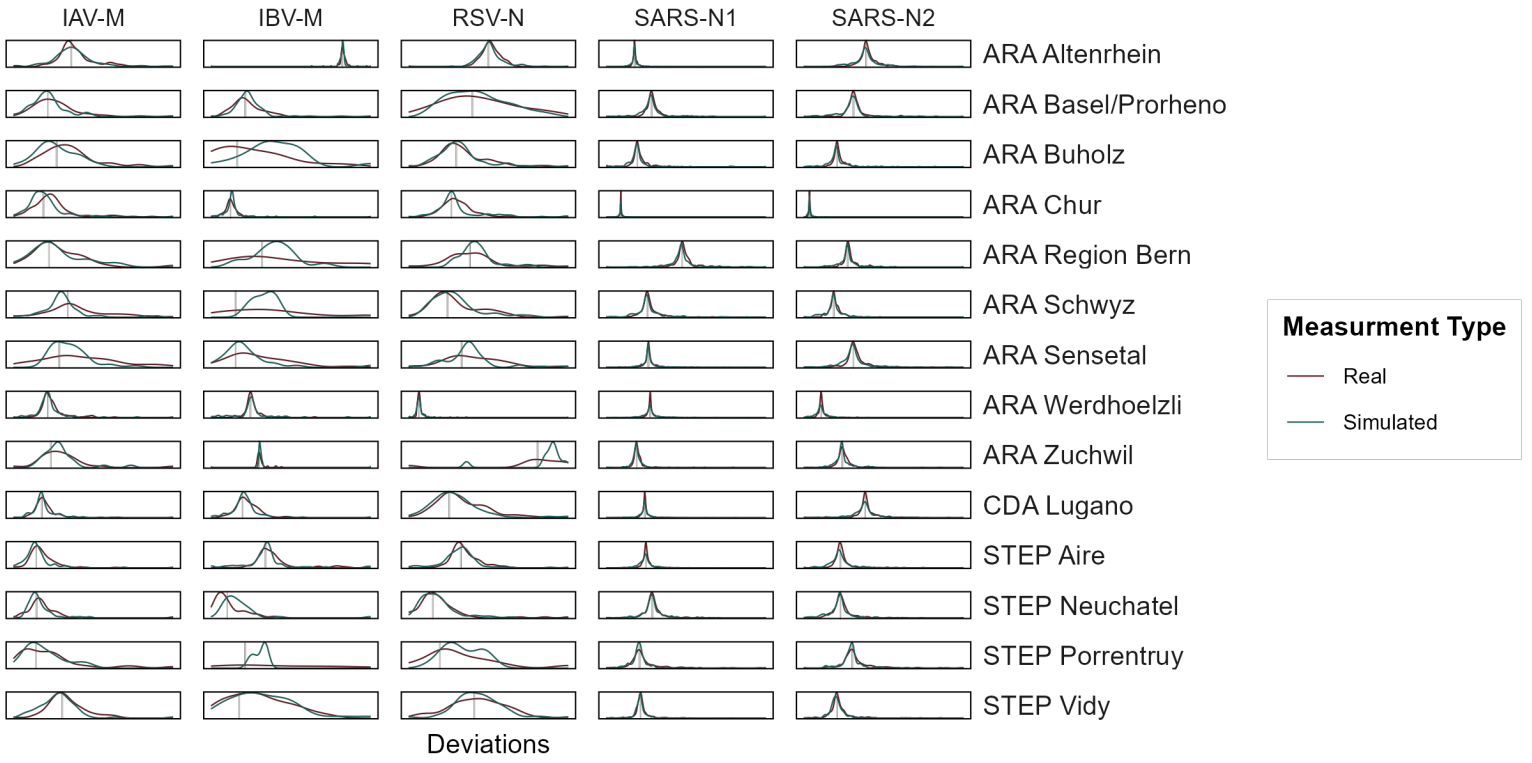

Figure 4.19: Measurement deviations from the input trend line for the real data (red) and the simulated data (green). The vertical grey line marks a deviation of zero, i.e. the measurement is on the trend line. Data are filtered for measurements above the lod. Empty plots indicate treatment plants where the specific target was measured below lod for the entire time series. Columns correspond to the viral targets and rows to the different treatment plants.

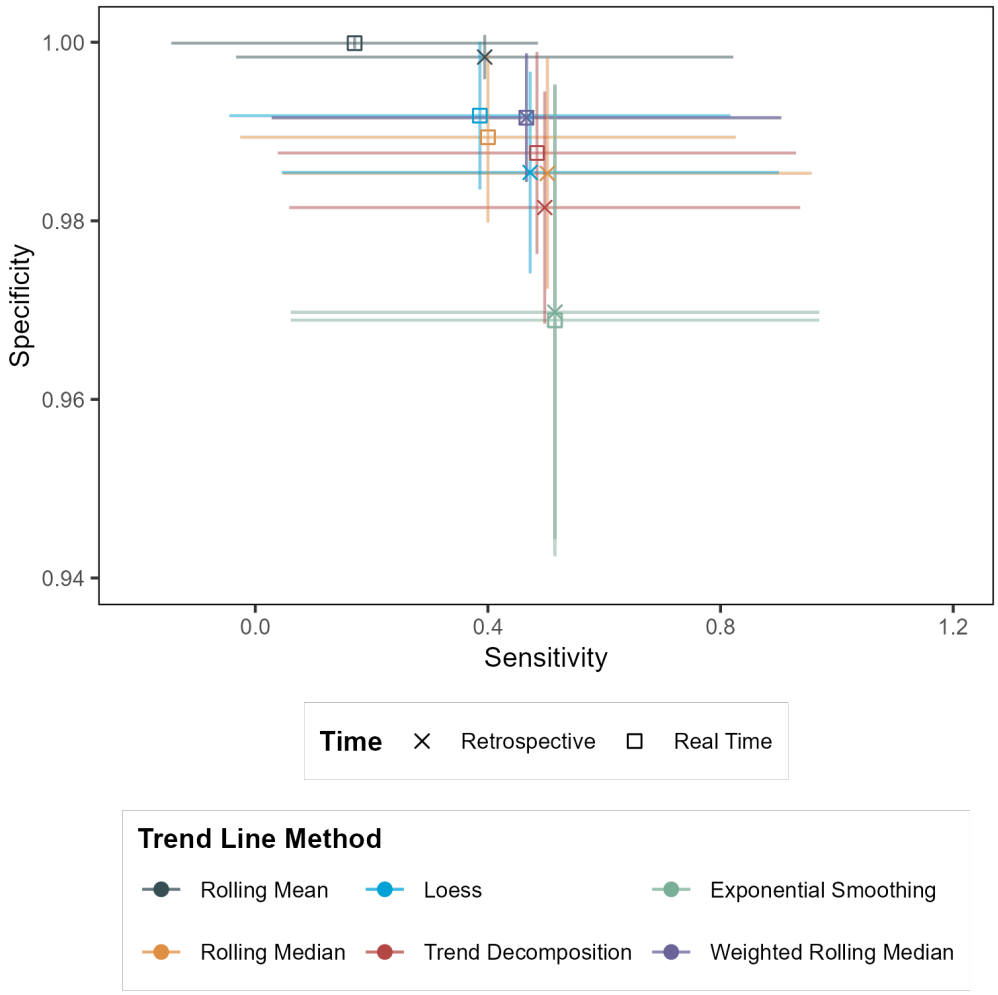

Figure 4.20: Mean sensitivity and specificity for each method (different colours) when applied retrospectively (crosses) and in simulated real time (squares) on simulated data. Error bars show standard deviation of the 70 time series.

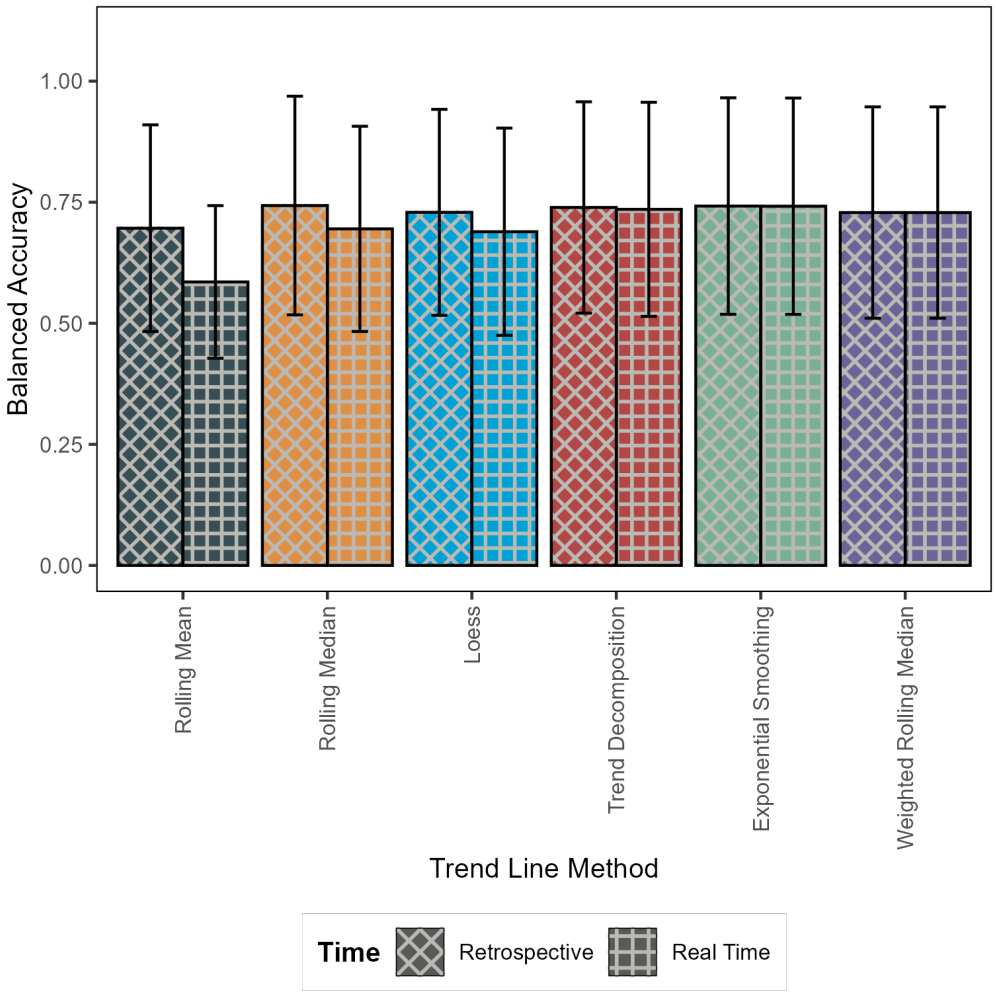

Figure 4.21: Mean balanced accuracy scores for each of the trend line methods for simulated data. Cross-hatched bars show results from the retrospective application while square-hatched bars show results from the real-time simulation. Error bars show standard deviation of the 70 time series.

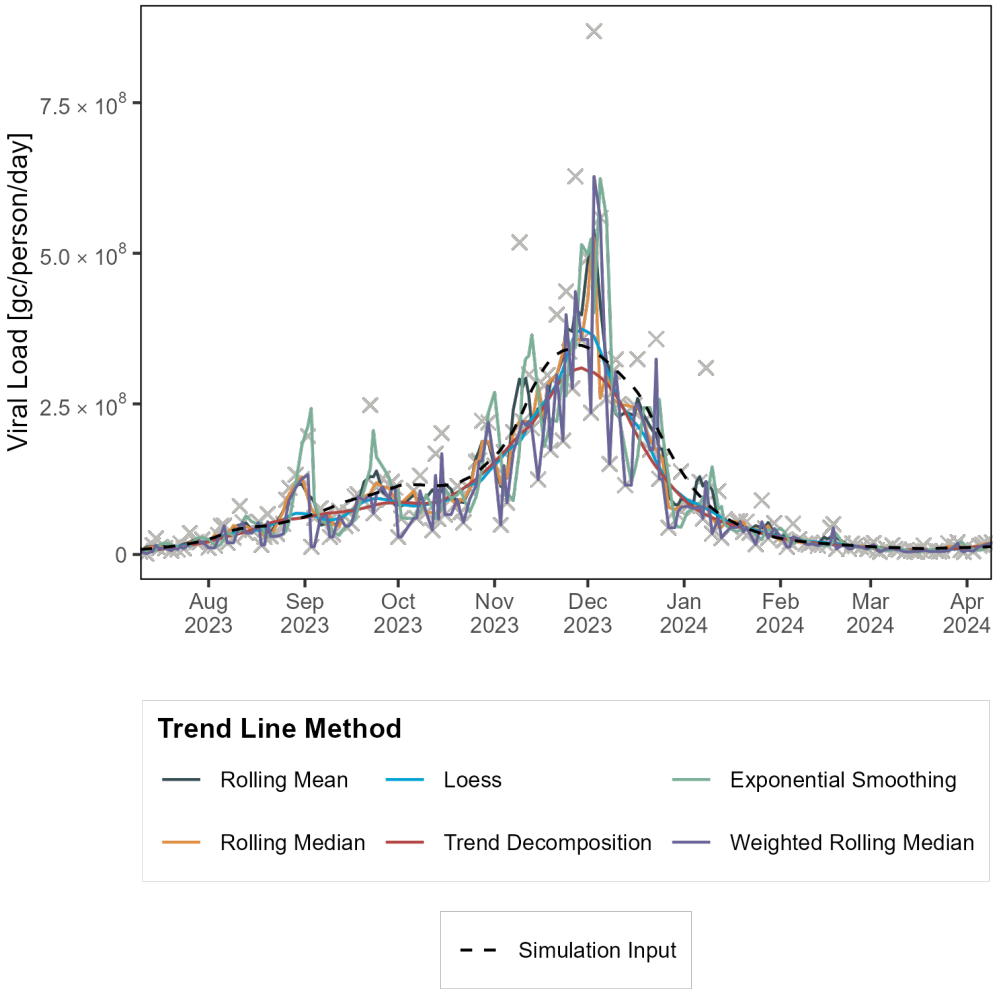

Figure 4.22: Trend lines generated from the different methods (different colours) given simulated data from an input trend line (black).

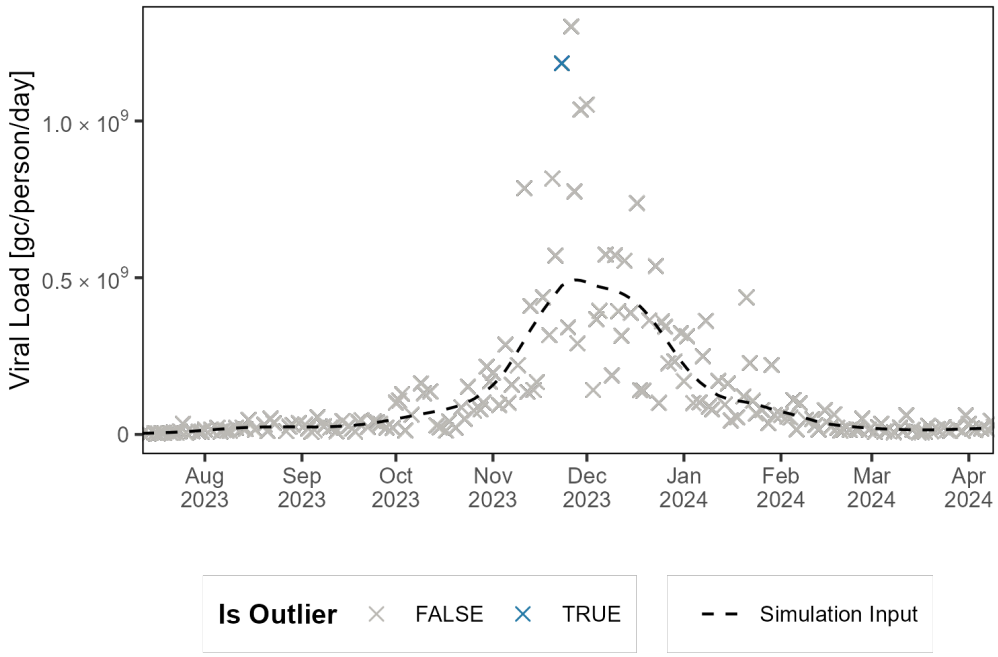

Figure 4.23: Simulated measurements when given an input trend line (black), either considered normal (grey) or an outlier (blue). The plot shows that the outlier value falls clearly within the range of normally simulated measurements.

The final analysis was to carry out multiple simulations from the same initial input to gain a measure of the uncertainty in the process. Since simulating outlier measurements was not consistently generating measurements which would be considered an outlier, the analysis focussed on specificity and made use of the set of simulations where outlier frequency was set to zero. Figure 4.24 shows the variability of the false positive rate (inverse of specificity) for the different trend line methods. Assuming the method assumption of normality around the trend line holds, we would expect a false positive rate of 0.3% since we take normalised deviations greater than three standard deviations (corresponding to the 99.7% quantile). Rolling mean falls below this expected value, while trend decomposition is closer to 2.5% with other methods falling in between the two.

However, the assumption of normality is not strictly true. Since it is not possible to have negative concentration measurements, data are simulated from a log-normal distribution rather than a normal distribution. Figure 4.25 shows that a greater proportion of values from a log-normal distribution of equal mean and variance fall above the normal distribution 99.7% quantile. For the range of viral concentrations in the data set, this value corresponds to about 2.1%. Therefore, most of the trend line methods fall within the expected range of false positive values.

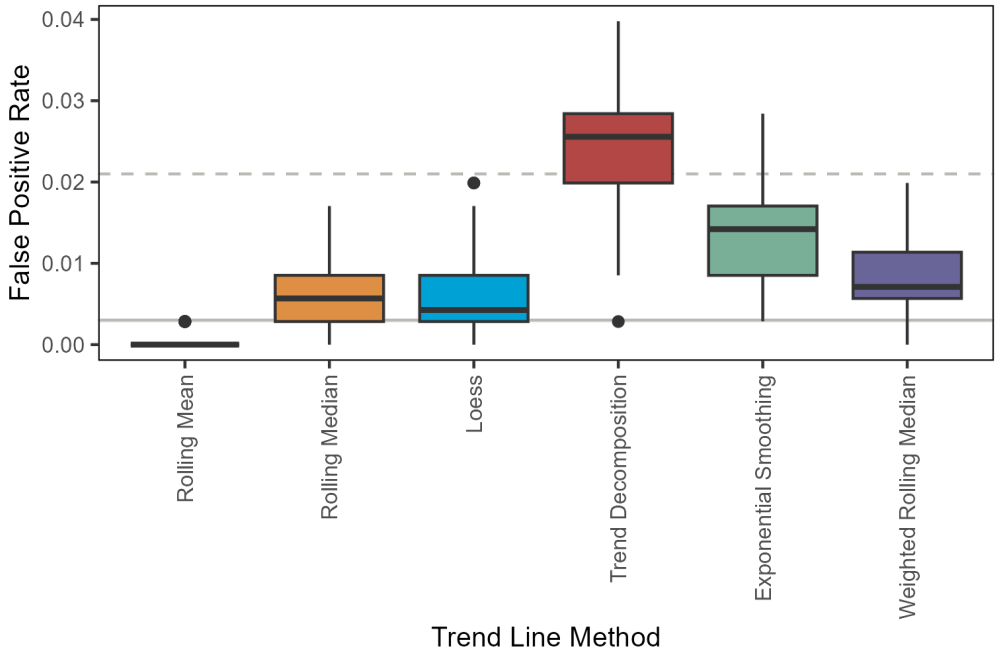

Figure 4.24: False positive rate for simulated data with outlier frequency fixed to zero for the different trend line methods (different colours). Box plots show results for 50 simulations from a single input trend line. The solid horizontal grey line is the expected false positive rate assuming a normal distribution (0.003) while the dashed grey line is the corresponding expected rate for a log-normal distribution (0.021).

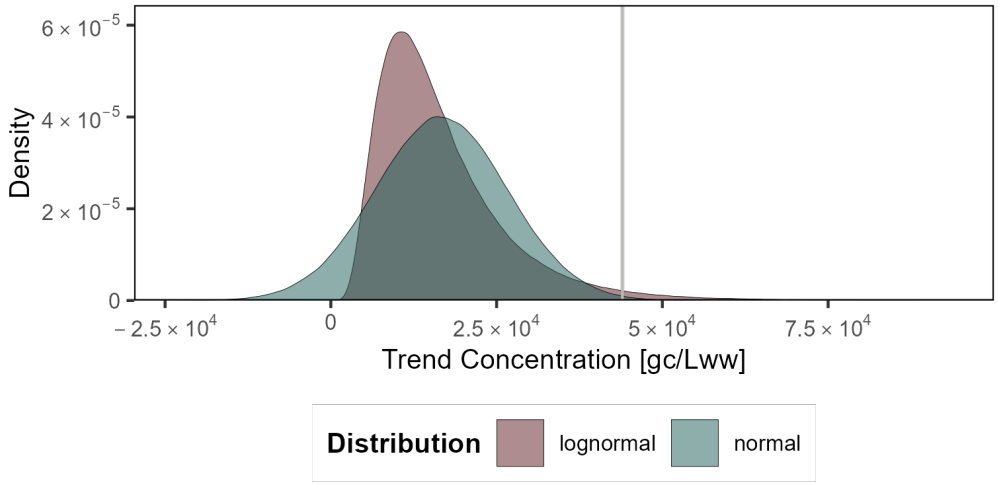

Figure 4.25: Log normal (red) and normal distribution (green) of same mean and coefficient of variation. The grey vertical line indicates the 99.7% quantile of the normal distribution, corresponding to around the 97.9% quantile for the log normal distribution.

Chapter 5

Discussion

This thesis has presented the development of an outlier detection method which can be applied in real time to digital polymerase chain reaction (dPCR) wastewater measurement data collected as part of a monitoring programme. While conceptually similar to previous outlier detection methods discussed in Chapter 2, the method presented in this study is uniquely able to deal with multiple challenges presented by the specific nature of wastewater-based dPCR data to allow for efficient outlier detection: local, as well as global outliers are identified, expected dPCR measurement variation is accounted for, and the method is applicable in real time. The impact of outliers on modelling results has been demonstrated using two methods for estimating R_t values. It has been shown that not only does the magnitude of an outlier measurement impact results in different ways, but outputs can also be influenced by the specific location of an outlier in an infection wave. Finally, an approach for simulating wastewater concentration measurements has been established which is able to generate time series of concentration data that shows agreement with real-world data, and which can aid further method development and evaluation. In the following sections results and limitations of the analysis are discussed along with possible directions of further study.

5.1 Labelling outliers

In the absence of a ground truth of outliers, this thesis adopted the approach of asking experts to label the data set to aid method evaluation. The results suggest that limitations of this approach are likely due to experts struggling to intuit the expected variation at different concentrations, especially when plots cover a wide scale range. Despite this, the most extreme outliers were identified, and the labelled data set provided a consistent means to compare trend line method performance.

Taken together, the lack of a pattern in outlier occurrence with lab processing days and the high level of quality control the data are subject to, lead us to the working assumption that outliers in the data set are *extreme but true* values. It

should be noted here that by *true* we mean the value is a true dPCR measurement of what was in the sample. This does not directly translate to an assumption that this is an accurate measurement of the virus present in the catchment, because, as discussed in Chapter 2, there are many stages of the pipeline where unknown processes could contribute to outlier generation and result in a measurement that is not representative of the current viral load of the catchment.

Our understanding of these sources of uncertainty could be further developed by laboratory experiments to understand variability in specific parts of the wastewater-based epidemiology (WBE) pipeline. For example, measuring viral concentrations at different points along a sewage pipeline to understand how concentration is affected by in-sewer transport; measuring concentration degradation under different storage conditions to understand how sample-processing delays can impact measurements; or comparing sequencing data from outlier and non-outlier days to see if there are differences in variant proportions. Having more information about these aspects of the process can inform model design and theories about sources of outliers.

Improved model design could ultimately mean that outlier data points could be incorporated into analyses, a goal supported by our working assumption that the outlier data points in this data set are extreme but true values. Despite this assumption of truth, however, currently the processes generating outliers are not specifically captured by models, and therefore the preferred approach is still to remove these data points beforehand.

5.2 Applying the method

Although methods making use of a centrally-aligned rolling window were able to perform better compared to right-aligned methods, they all exhibited a drop in performance when used in a real-time manner. Therefore, a right aligned method should be taken forward for consistent use in a monitoring programme. Weighted rolling median has slightly better specificity, whereas exponential smoothing has better sensitivity. This sensitivity-specificity trade-off means exponential smoothing is arguably the better method to be used in real time. Since the presence of outliers can have such an impact on model results (Figure 4.14) it is desirable to tolerate a few more missing values than to have outliers in the data set.

Another consideration for a real-time monitoring programme is the use of quality checks on data sources used for normalisation. As discussed in Chapter 2, extreme flow values can lead to non-representative measurements, corroborated by the data set used in this thesis, where some viral load outliers were attributable to high flow values (Figure 4.13). Current practices which could be adopted along with the outlier detection method for concentration data, include filtering all measurements where flow is above the 90th percentile[34] or values are greater

than three times the IQR range from the median[6]. However, it should be noted that this is not a common occurrence, and possible only around 5% of manually labelled outliers are explainable by high flow values.

Applying data filtering on the normalisation flow data is a method that may detect low outliers induced by dilution events. As can be seen from Figure 4.9, the expected variation includes zero measurements, even at the peak of the infection wave, and therefore the method is not applicable for the detection of low concentration outliers.

In situations where real-time outlier detection is not a requirement, trend decomposition or rolling median with a 14-28 day window could be used to give better results. These methods, which have a stronger smoothing effect, correctly identify more manually labelled outliers at the peaks of infection waves. These methods could also be used in combination with a right-aligned method in real-time, to update labels retrospectively. This has the potential to capture more outliers, though, depending on the reporting structures of a monitoring programme, has potential challenges when communicating data changes to ensure updated information is shared to all stakeholders.

As shown in the final analysis (Figures 4.24 and 4.25), by using a z-score of 3 as the threshold for labelling outliers, we can expect some degree of false positives, (no false positives would mean perfect specificity). If we want to reduce the number of points marked as an outlier by the detection method, the z-score threshold could be increased. Doing so still ensures that the most extreme outliers are identified.

Another means of only capturing extreme outliers could be to incorporate a measure of uncertainty into the method, so that rather than assigning a binary prediction of outlier status, a proportion score is given to each point. One way to do this would be to make use of a trend line method which has some measure of uncertainty and from which multiple trend lines can be drawn. These trend lines could each be used to determine the outlier status of data points and then an average, over all trend line labels, taken to give a proportional measure.

In this thesis the average of technical replicates was taken to generate time series with a single measurement per day. It would be interesting to make use of all the data available and apply the method to individual technical replicates. In its current form, the method can be applied to individual replicates, though the trend line is still calculated by first taking the average of the individual replicates. Knowing the proportion of time points that have technical measurements differing in their outlier classification, could help inform assumptions about the outlier generation process and sources of variation. It is likely such a method could benefit from an alternative way to calculate the expected measurement concentration from technical replicates, not just taking the average.

On the other hand, consideration should be made for situations where less

data are available. The method requires several parameters for the normalisation calculation (equation 3.3 and Section 3.1.2) which may not be available or known for other data sets. Further sensitivity analysis should be carried out to assess the impact estimating or randomly selecting such lab parameters has on results. If results are sensitive to changes in these parameters, it would be useful to investigate ways to estimate these parameters for a particular data set. Here, however, the presence of outliers is likely to pose challenges.

The pre-PCR parameter was estimated as a single value for all time series. However, it is likely that it is different for different locations and viral targets, and could be optimised for individual time series. This too could form part of a sensitivity analysis of the method to see how much results are impacted by changing pre-PCR variation. Also of interest is that the estimated value ($\nu = 0.6$) was higher than that estimated by the EpiSewer model ($\nu = 0.17$)[4]. Since estimated pre-PCR variation depends on the smoothing of measurements, stronger smoothing will be compensated with higher pre-PCR variation. This can be seen in Figure 4.3 where the range of optimal ν parameter values is higher for trend decomposition compared to those for rolling mean. Furthermore, the trend line methods used are deterministic, whereas the EpiSewer trend contains uncertainty which could contribute to a weaker smoothing effect. A final consideration for why these values differ, is the underlying likelihood distribution of measurements, which was shown to impact estimated results. The pre-PCR median value of 0.17 was obtained by fitting data to the specific dPCR likelihood, whereas the method presented in this thesis implicitly adopts a normal distribution.

One final consideration for wider applicability of the method presented in this thesis would be to adapt it for quantitative PCR (qPCR) data as this is widely used in WBE for data collection. The current normalisation calculation is specific for dPCR data and therefore another equation for estimating the expected coefficient of variation for a particular concentration would have to be used to make the method applicable to qPCR data sets.

5.3 Effect on modelling

As shown in Section 4.4, when data are used to estimate R_t, for a subset of the 70 time series, changes in model outputs are observed. The magnitude of these effects is partially due to a combination of the robustness of the model used and the magnitude of the outlier data point(s), but may also be affected by where, within an infection wave, they occur. Especially at the start of a wave, the presence of outliers can show R_t estimates as rising more sharply than they perhaps are. Since this thesis focused on only one time series with outliers for each location, it is not possible to conclude how extreme an outlier must be before it has an effect on results. Further understanding of how the magnitude of an outlier affects results at different locations could aid in decisions about the

sensitivity-specificity trade-off for the different trend line methods.

EstimateR uncertainty ranges seem to be greater when concentration measurements are lower and there are no clear trends. Differences in the two EpiSewer set ups are due to the different distributions that the measurements are assumed to arise from, either a truncated normal distribution or a gamma distribution. There is a greater change in results on exclusion of outlier data points when a truncated normal distribution is used rather than a gamma distribution. This is likely because, under the truncated normal distribution, the modelled variance must increase dramatically for the same mean to include these high data points. Due to its longer tail, the gamma distribution can accept more of these extreme data points without increasing the modelled variance as much. This highlights the importance of assessing the sensitivity of models to outliers.

5.4 Outlier simulation

The simulation worked well to generate a time series of concentration measurements analogous to real-world measurements and the outlier detection method performed well on this data. Sensitivity scores were lower compared to those for the real data due to the "outlier addition" component of the simulation. Here a value was drawn from an exponential distribution where the rate was defined with the quantile function, such that 99% of the values were below four times the maximum expected concentration (see Section 3.5). While allowing for the high, positive values associated with an outlier, this distribution also allows for smaller values which, as can be seen in Figure 4.23, can result in outlier measurements which are not in fact outliers.

This aspect of the simulation could be improved by drawing outlier addition values from a Weibull distribution instead. Like the exponential distribution, this takes only positive continuous values and has an easily useable quantile function for defining the rate. The required modification would be definition of the shape parameter which would reduce the likelihood of low outlier addition values. It would be interesting to see if, when outlier points are more appropriately simulated, the different trend line methods continue to have a similar performance, or whether they perform in the same pattern as with the real world data.

Outlier points aside, simulation of regular measurements seems to work well. All trend line methods were able to capture the initial input: the trend decomposition trend line calculated from the real-world data. It would be interesting to see how simulation results change when given a different input trend line, such as the weighted rolling median. Additionally, some parameter sensitivity could be investigated using the simulation. Since the pre-PCR variation parameter is defined for the simulated data set, the impact of selecting a wrong parameter value for outlier detection can be observed.

5.5 Outlook

Continued development and improvement of outlier detection methods can help ensure quality data reporting from monitoring programmes, which are becoming a crucial part of many countries' public health surveillance schemes[44]. Additionally, this can improve modelling results which often guide public health responses to disease outbreak. As methods in wastewater-based epidemiology advance, there is hope that underlying processes, especially those contributing to outliers, could be better understood. Being able to identify outliers enables further study of them, which could help inform theories about their sources and perhaps shed light on unknown processes. The outlier detection method presented in this thesis is a small contribution towards that goal.

References

1. Bouter, L., Zeegers, M. & Li, T. *Textbook of Epidemiology* 260 pp. ISBN: 978-1-119-77602-4 (John Wiley & Sons, Jan. 4, 2023).
2. Bowes, D. A. Towards a Precision Model for Environmental Public Health: Wastewater-based Epidemiology to Assess Population-level Exposures and Related Diseases. *Current Epidemiology Reports.* ISSN: 2196-2995 (June 24, 2024).
3. Li, X., Zhang, S., Shi, J., Luby, S. P. & Jiang, G. Uncertainties in estimating SARS-CoV-2 prevalence by wastewater-based epidemiology. *Chemical Engineering Journal* **415,** 129039. ISSN: 1385-8947 (July 1, 2021).
4. Lison, A., Julian, T. R. & Stadler, T. *Improving inference in wastewater-based epidemiology by modelling the statistical features of digital PCR* 2024.
5. Forootan, A. *et al.* Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). *Biomolecular Detection and Quantification* **12,** 1–6. ISSN: 2214-7535 (June 1, 2017).
6. Rauch, W., Schenk, H., Insam, H., Markt, R. & Kreuzinger, N. Data modelling recipes for SARS-CoV-2 wastewater-based epidemiology. *Environmental Research* **214,** 113809. ISSN: 0013-9351 (Nov. 1, 2022).
7. Daughton, C. G. Emerging pollutants, and communicating the science of environmental chemistry and mass spectrometry: pharmaceuticals in the environment. *Journal of the American Society for Mass Spectrometry* **12,** 1067–1076. ISSN: 1044-0305 (Oct. 1, 2001).
8. Nishiura, H. & Chowell, G. in *Mathematical and Statistical Estimation Approaches in Epidemiology* (eds Chowell, G., Hyman, J. M., Bettencourt, L. M. A. & Castillo-Chavez, C.) 103–121 (Springer Netherlands, Dordrecht, 2009). ISBN: 978-90-481-2313-1.
9. Rabe, A. *et al.* Correlation between wastewater and COVID-19 case incidence rates in major California sewersheds across three variant periods. *Journal of Water and Health* **21,** 1303–1317. ISSN: 1477-8920 (Sept. 11, 2023).
10. Huisman, J. S. *et al.* Wastewater-Based Estimation of the Effective Reproductive Number of SARS-CoV-2. *Environmental Health Perspectives* **130,** 057011 (May 2022).

11. Bagutti, C. *et al.* Wastewater monitoring of SARS-CoV-2 shows high correlation with COVID-19 case numbers and allowed early detection of the first confirmed B.1.1.529 infection in Switzerland: results of an observational surveillance study. *Swiss Medical Weekly* **152,** w30202–w30202. ISSN: 1424-3997 (June 27, 2022).

12. Schenk, H. *et al.* Prediction of hospitalisations based on wastewater-based SARS-CoV-2 epidemiology. *Science of The Total Environment* **873,** 162149. ISSN: 0048-9697 (May 15, 2023).

13. Galani, A. *et al.* SARS-CoV-2 wastewater surveillance data can predict hospitalizations and ICU admissions. *Science of The Total Environment* **804,** 150151. ISSN: 0048-9697 (Jan. 15, 2022).

14. Okada, Y. & Nishiura, H. Estimating the effective reproduction number of COVID-19 from population-wide wastewater data: An application in Kagawa, Japan. *Infectious Disease Modelling* **9,** 645–656. ISSN: 2468-0427 (Sept. 1, 2024).

15. Nadeau, S. *et al.* Influenza transmission dynamics quantified from RNA in wastewater in Switzerland. *Swiss Medical Weekly* **154,** 3503–3503. ISSN: 1424-3997 (Jan. 3, 2024).

16. Champredon, D., Papst, I. & Yusuf, W. ern: An R package to estimate the effective reproduction number using clinical and wastewater surveillance data. *PLOS ONE* **19,** e0305550. ISSN: 1932-6203 (June 21, 2024).

17. Scire, J. *et al.* estimateR: an R package to estimate and monitor the effective reproductive number. *BMC Bioinformatics* **24,** 310. ISSN: 1471-2105 (Aug. 11, 2023).

18. Lison, A. *adrian-lison/EpiSewer: EpiSewer 0.0.1* version v0.0.1. Jan. 25, 2024.

19. Wade, M. J. *et al.* Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the United Kingdom national COVID-19 surveillance programmes. *Journal of Hazardous Materials* **424,** 127456. ISSN: 0304-3894 (Feb. 15, 2022).

20. The dMIQE Group & Huggett, J. F. The Digital MIQE Guidelines Update: Minimum Information for Publication of Quantitative Digital PCR Experiments for 2020. *Clinical Chemistry* **66,** 1012–1029. ISSN: 0009-9147 (Aug. 1, 2020).

21. Hawkins, D. M. *Identification of Outliers* ISBN: 978-94-015-3996-8 978-94-015-3994-4 (Springer Netherlands, Dordrecht, 1980).

22. Smiti, A. A critical overview of outlier detection methods. *Computer Science Review* **38,** 100306. ISSN: 1574-0137 (Nov. 1, 2020).

23. Aguinis, H., Gottfredson, R. K. & Joo, H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods* **16,** 270–301. ISSN: 1094-4281 (Apr. 1, 2013).

24. Hodge, V. & Austin, J. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* **22,** 85–126. ISSN: 1573-7462 (Oct. 1, 2004).

25. Blázquez-García, A., Conde, A., Mori, U. & Lozano, J. A. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Comput. Surv.* **54,** 56:1–56:33. ISSN: 0360-0300 (Apr. 17, 2021).

26. Yaro, A. S., Maly, F., Prazak, P. & Malý, K. Outlier Detection Performance of a Modified Z-Score Method in Time-Series RSS Observation With Hybrid Scale Estimators. *IEEE Access* **12,** 12785–12796. ISSN: 2169-3536 (2024).

27. Seyedhossein, L. & Hashemi, M. R. *Mining information from credit card time series for timelier fraud detection* in *2010 5th International Symposium on Telecommunications* 2010 5th International Symposium on Telecommunications (Dec. 2010), 619–624.

28. Moschini, G., Houssou, R., Bovay, J. & Robert-Nicoud, S. Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model. *Engineering Proceedings* **5,** 56. ISSN: 2673-4591 (2021).

29. Basu, S. & Meckesheimer, M. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems* **11,** 137–154. ISSN: 0219-3116 (Feb. 1, 2007).

30. Yu, Y., Zhu, Y., Li, S. & Wan, D. Time Series Outlier Detection Based on Sliding Window Prediction. *Mathematical Problems in Engineering* **2014,** 879736. ISSN: 1563-5147 (2014).

31. Jamshidi, E. J., Yusup, Y., Kayode, J. S. & Kamaruddin, M. A. Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature. *Ecological Informatics* **69,** 101672. ISSN: 1574-9541 (July 1, 2022).

32. Fang, Z. *et al.* Wastewater monitoring of COVID-19: a perspective from Scotland. *Journal of Water and Health* **20,** 1688–1700. ISSN: 1477-8920 (Dec. 14, 2022).

33. Arabzadeh, R. *et al.* Data filtering methods for SARS-CoV-2 wastewater surveillance. *Water Science and Technology* **84,** 1324–1339. ISSN: 0273-1223 (Aug. 30, 2021).

34. Markt, R. *et al.* Detection and abundance of SARS-CoV-2 in wastewater in Liechtenstein, and the estimation of prevalence and impact of the B.1.1.7 variant. *Journal of Water and Health* **20,** 114–125. ISSN: 1477-8920 (Nov. 30, 2021).

35. Lastra, A. *et al.* SARS-CoV-2 detection in wastewater as an early warning indicator for COVID-19 pandemic. Madrid region case study. *Environmental Research* **203,** 111852. ISSN: 0013-9351 (Jan. 1, 2022).

36. Klaassen, F. *et al.* Predictive power of wastewater for nowcasting infectious disease transmission: A retrospective case study of five sewershed areas in Louisville, Kentucky. *Environmental Research* **240,** 117395. ISSN: 0013-9351 (Jan. 1, 2024).

37. Manuel, D. G. *et al.* Wastewater-based surveillance of SARS-CoV-2: Short-term projection (forecasting), smoothing and outlier identification using Bayesian smoothing. *Science of The Total Environment* **949,** 174937. ISSN: 0048-9697 (Nov. 1, 2024).

38. Courbariaux, M. *et al.* A Flexible Smoother Adapted to Censored Data With Outliers and Its Application to SARS-CoV-2 Monitoring in Wastewater. *Frontiers in Applied Mathematics and Statistics* **8.** ISSN: 2297-4687 (Feb. 9, 2022).

39. *WISE* https://wise.ethz.ch/ (2024).

40. Hafen, R. P. *Local regression models: Advancements, applications, and new methods* ISBN: 9781124151717. PhD thesis (Purdue University, United States – Indiana, 2010). 306 pp.

41. Cleveland, W. S., Grosse, E. & Shyu, W. M. in *Statistical Models in S* (Routledge, 1992). ISBN: 978-0-203-73853-5.

42. Hyndman, R. J., Koehler, A. B., Snyder, R. D. & Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18,** 439–454. ISSN: 0169-2070 (July 1, 2002).

43. Tharwat, A. Classification assessment methods. *Applied Computing and Informatics* **17,** 168–192. ISSN: 2634-1964, 2210-8327 (Jan. 1, 2021).

44. Naughton, C. C. *et al.* Show us the data: global COVID-19 wastewater monitoring efforts, equity, and gaps. *FEMS Microbes* **4,** xtad003. ISSN: 2633-6685 (Jan. 1, 2023).