Technical Guide Part 4 on SARS-CoV-2 Wastewater Surveillance - Data Processing

This Technical Guideline is part of four working documents. The working documents are closely related from a technical point of view and should be considered together:

- Technical Guideline Part 1 for SARS-CoV-2 Wastewater Surveillance
  - Sampling Wastewater
- Technical Guideline Part 2 for SARS-CoV-2 Wastewater Surveillance
  - Molecular Biological Analysis
- Technical Guideline Part 3 for SARS-CoV-2 Wastewater Surveillance
  - Sample Logistics and Data Submission
- Technical Guideline Part 4 for SARS-CoV-2 Wastewater Surveillance
  - Data Processing

This guideline provides the basics of data quality testing, data calculation and statistical analysis as part of SARS-CoV-2 wastewater surveillance. It includes the following chapters:

1. **Calculation of PCR Data**
2. **Normalization**
3. **Statistical analysis**
4. **Software**

The raw wastewater inflow from wastewater treatment plants is subject to daily, weekday and seasonal fluctuations. In addition, there are changes in the volumetric inflow rate due to continuous or discontinuous indirect discharges, percolating and precipitation water. These changes in the volumetric flow rate result in a dilution of the respective test parameters - including the levels of SARS-CoV-2 gene fragments.

It is therefore necessary to normalize the raw data before they can be used for statistical analysis. The normalization currently includes an average PCR measurement value of the raw wastewater per location and sampling time (date of start of sampling) as well as information on volumetric raw wastewater flow rates (flow data) of the wastewater treatment plant. Alternative normalization parameters (including faecal reference viruses, chemical human markers) are also possible, but are not currently used in the presentation of results.

Measured viral loads in wastewater typically show a high degree of variation over time, so that many forms of representation and statistical analyzes are not based on the measured (possibly normalized) values, but on smoothed values. A variety of statistical methods exist for smoothing the measured values.

1. Calculation of PCR data

Only samples that were analyzed for at least two different of the possible six gene sequences (N1, N2, N3, E, ORF and RdRp) are included in the data processing in order to exclude false positive

values. If there are not two measured values, i.e. values for at least two gene sequences, the sample is generally excluded from further evaluation and marked as “not determined”.

The quantification limits of the PCR workflow/PCR measurement used are queried for each gene sequence when data is transmitted.

If the measured values are above the limit of quantification (LOQ), further evaluation is performed as described below.

If one or more of the measured values are below the quantification limit, the value / values are replaced by 0.5 * LOQ.

If the value for the LOQ is not specified, the value of 4,000 gene copies/l is assumed for normalization, regardless of the preparation or PCR method used. However, in the medium and long term, the laboratories' own values should be used.

A geometric mean is currently calculated from (at least) two available values for the above-mentioned gene sequences:

$$Gene\ copies\ mean = \sqrt[n]{x_1 * x_2 * \dots * x_n}$$

with $x_1, x_2, \dots, x_n$*: Gene copies/l per gene*

2. Normalization of PCR Data

For normalization, the dry weather influent to the wastewater treatment plant must be determined as a reference. The dry weather influent is the volume of water that is not influenced by precipitation events or thaw.

With the currently implemented and applicable normalization method, to normalize the PCR data, the ratio of the volumetric flow rate in the sampling period ($Q_{WWTP,current}$) and the dry weather inflow is formed and then multiplied by the geometric mean of the available PCR measurement results (gene copy mean). To determine the dry weather inflow, the median of all previously transmitted values of the volumetric flow rate of the wastewater treatment plant ($Q_{WWTP,median}$) is used.

Based on the assumptions described above, the normalized value is calculated as follows:

$$Gene\ copies\ normalized = \frac{Q_{WWTP,current}}{Q_{WWTP,median}} * Gene\ copies\ mean$$

with $Q_{WWTP,current}$*: Volume flow of the wastewater treatment plant during sampling period;*

$Q_{WWTP,median}$*: Median of Volumen flow of the wastewater treatment plant*

The volumetric flow rate can also be used as the average flow rate of the wastewater treatment plant during the sampling period.

The PCR data normalized in the manner described are included in the statistical analyses.

3. Statistical Analyses

For SARS-CoV-2 wastewater surveillance, smoothed values are calculated for each site using a locally weighted regression (LOESS method, from Locally Estimated Scatterplot Smoothing). This calculation is based on the logarithmized (and normalized, see above) viral loads in the wastewater.

The LOESS method uses a locally weighted regression function so that a window with a certain number of observed viral loads around the measured value (e.g. the 15 closest measured values) is included in the prediction of a viral load on the LOESS curve, with the influence decreasing with the distance to the time of the viral load to be predicted. The proportion of viral loads to be included is determined for each sampling location using the generalized cross-validation method, which optimizes the predictive quality of the curve. The idea of the cross-validation method is to remove an observed viral load from the data, predict it based on a number of other observed viral loads around the measured value using the LOESS method and calculate the resulting deviation between observed and predicted viral load. This procedure must be performed for all measurement times and the calculated deviations are added up. The optimal number of viral loads to include according to the cross-validation criterion is the one that minimizes this sum.

This results in a smooth curve for each location and a predicted viral load for each time point (including between the measurement times), which is then transformed back to the original scale. The percentage change between the predicted viral load on Wednesday of a week and the predicted viral load on Wednesday of the previous week is calculated which allows to indicate at how many locations or sites, for example, have an increasing viral load (trend category "increasing", defined as a change > 15%), a decreasing viral load (trend category "decreasing", change < -15%) or no relevant change (trend category "unchanged", change between -15% and 15%) over these seven days. It is also possible to define the trend categories differently.

In addition to the individual sites, the time series of the sites are aggregated to obtain a Germany-wide trend of SARS-CoV-2 viral load in wastewater. For this purpose, weekly mean values are first formed for the logarithmized measured values of each site. The resulting weekly values are then averaged over all sites for each week, weighted according to the number of inhabitants connected to the respective site. The resulting time series are also smoothed using a LOESS regression as described above.

4. Software

The PCR data are processed and normalized in a browser-based web application: Pathogens in wastewater (PiA; https://app.pia-monitor.de). The application also includes an automated quality control of the data to be imported. Login data for the application and the description of the application must be requested separately.

The LOESS method used for the trend calculation is implemented in statistical software packages such as R. If a fixed window width is selected for the estimation (without a cross-validation criterion), the implementation in Excel or similar is also comparatively easy to implement manually.

Contact	Umweltbundesamt, sarscov2@uba.de Robert Koch-Institut, abwassersurveillance@rki.de	
Funding	The BMG is funding wastewater monitoring until the end of 2024 as part of AMELAG, the federal German wastewater monitoring project for epidemiological situation assessment (German: "Abwassermonitoring für die epidemiologische Lagebewertung")	AMELAG
	The technical guide cointains information/findings that were developed as part of the ESI-CorA project. ESI-CorA was funded by the European Commission under the Emergency Support Instrument (ESI; No 060701/2021/864650/SUB/ENV.C2).	