COVID-19 Wastewater Monitoring in Michigan

SWEEP Data Analysis Guide

This guide walks through the data analysis methods used to create the Michigan COVID-19 Sentinel Wastewater Epidemiology Evaluation Project (SWEEP) Dashboard. The purpose of this guide is to explain how MDHHS performs wastewater data analysis and how other agencies can implement these methods, if desired. This guide was written to support local health departments, Native Nations, and other public health partners wishing to implement in-depth analyses for SARS-CoV-2 wastewater data in their jurisdictions.

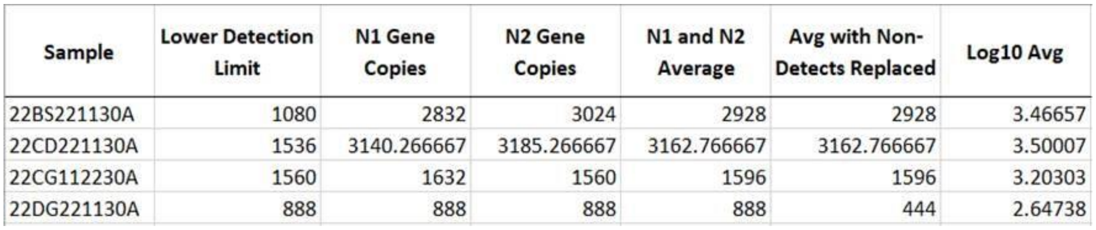
This guide covers:

- Data transformation
- Data normalization
- Percentiles
- 15-day trends
- Viral concentration graphs

MDHHS uses R to conduct this analysis and Tableau to visualize data and produce the dashboard. MDHHS would be happy to share code and other tips to help others implement this analysis. Please email any questions to MDHHS-SEWERNetwork@michigan.gov.

Data Transformation

Michigan's COVID-19 Wastewater Monitoring Project labs test wastewater samples for the N1 and N2 genes that are unique to the SARS-CoV-2 virus, so detection of either gene means the virus is present in the sample. The labs report the number of N1 and N2 gene copies per 100mL of sample. This data requires some cleaning and formatting before analysis can be done.



Sample	Lower Detection Limit	N1 Gene Copies	N2 Gene Copies	N1 and N2 Average	Avg with Non-Detects Replaced	Log10 Avg
22BS221130A	1080	2832	3024	2928	2928	3.46657
22CD221130A	1536	3140.266667	3185.266667	3162.766667	3162.766667	3.50007
22CG112230A	1560	1632	1560	1596	1596	3.20303
22DG221130A	888	888	888	888	444	2.64738

An example showing new variables created through the data transformation process and the impact of those variables on the data values.

1. Check for data duplication. Duplicates can be averaged to produce one N1 and one N2 concentration per sample. If the reason for duplication is known, consider removing the incorrect or outdated data.

2. Average the N1 and N2 values to create one measurement per sample.
   - This allows both measurements to be accounted for to improve accuracy.
   - This simplifies the analysis so it only has to be performed once instead of separately for each gene.

3. Replace non-detect values with half of the lower detection limit (LDL).
   - A non-detect value could mean that no viral RNA is present or that the viral RNA is present in the sample, but the concentration is too low to be detected. Therefore the true value could be between 0 and the lower detection limit, so half of the LDL is used as a proxy.
   - This ensures that non-detect concentrations will be lower than positive detections.

4. Log-transform the averages to put them on a linear scale. For the SWEEP dashboard, a log base 10 transformation is used.
   - In R, use the function *mutate('Log10 Avg' = log10('Avg with Non-Detects Replaced')).*
   - In SAS, use the function *'Log10 Avg' = LOG10('Avg with Non-Detects Replaced').*
   - In Excel, use the function *=LOG10('Avg with Non-Detects Replaced').*
   - Viral concentrations in wastewater are expected to exhibit exponential behavior because that is how the virus replicates in the human body. The data will behave linearly when it is log-transformed, which makes it easier to work with.

MDHHS Michigan Department of Health and Human Services

12/15/2022

COVID-19 Wastewater Monitoring in Michigan

SWEEP Data Analysis Guide

Normalization

Viral concentrations in wastewater should not be directly compared between sites because the range of values observed at each site is highly variable (for example, 1,000 genecopies/100mL may be a high concentration at Site 1, but a low concentration at Site 2). To make the results slightly more comparable across sites, MDHHS has opted to normalize the data by population size so that the viral concentrations are in units per person.

To normalize the data by population size, use the following equation:

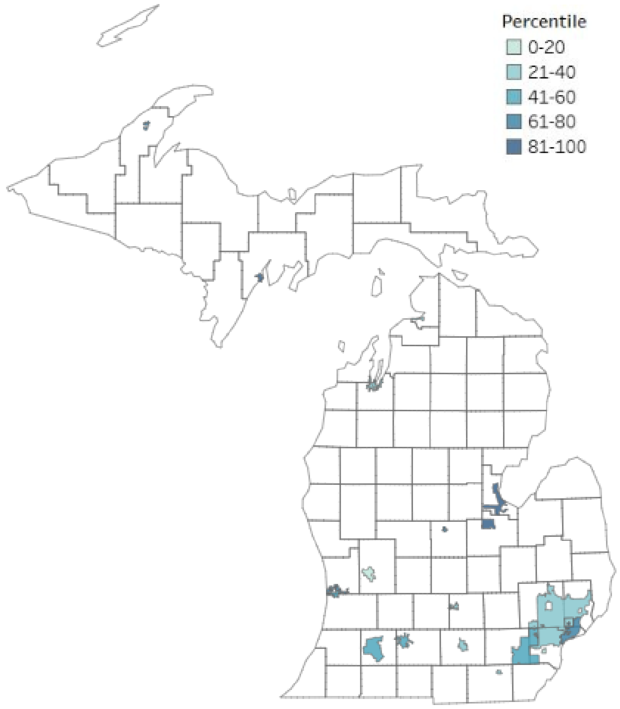
$$\frac{\text{N1/N2 averaged viral concentration}}{\text{Total number of people represented at this site}} \times 100{,}000 \longrightarrow$$

The new unit is now **gene copies/100mL of sample per 100,000 people**.

Percentiles

A percentile is used to determine if viral concentrations in a sample are low or high compared to all previous viral concentrations measured at each site.

1. Sort the data by site so that each site is analyzed separately.
2. Use a percentile rank function (in R, SAS or Excel) to rank the concentrations from smallest to largest.
3. Multiply the percentile rank by 100 to get whole numbers, if necessary.
4. Categorize the percentiles into quintiles to make interpretation and comparison easier. For example, the SWEEP dashboard uses the following categories:
    - 0-20
    - 21-40
    - 41-60
    - 61-80
    - 81-100
5. Optional: To create a percentile for a given time period, average the percentiles for each sample collected during that time period.
    - For example, a time period that includes three samples with percentiles of 5, 7 and 18 would have an average percentile of 10 (5+7+18=30/3=10).



On the SWEEP dashboard map, percentiles are visualized using a light-to-dark color scheme. The color corresponds with the percentile of the most recent sample collected at each site. Darker colors represent higher percentiles. This provides a quick interpretation of which sites have a high percentile or a low percentile for the most recent sample.

COVID-19 Wastewater Monitoring in Michigan

SWEEP Data Analysis Guide

15-Day Trends

15-day trends are used to determine if viral concentrations are changing over time. Developed by the Centers for Disease Control and Prevention (CDC), this method uses linear regression to quantify the amount of change occurring over a 15-day period by transforming normalized gene copies and time into a total percent change metric.

1. Filter data to select only samples collected during the 15-day period of interest.

2. For each site individually, conduct a linear regression where y = normalized gene copies and x = date.
   - If the data was not log-transformed during the data transformation process, it must now undergo a log-transformation before being used in a linear regression equation to make the analysis work. The CDC uses log10 for transformation.

3. Calculate the modeled change using (slope * number of days elapsed).
   - “Slope” is the x coefficient for the date variable.
   - “Number of days elapsed” might be 15 or it could be smaller if samples are not collected on days 1 and 15 of the selected time frame. Subtract the first date from the second date to determine the number of days elapsed.

4. Calculate the total modeled percent change using ((10^modeled change) - 1)*100%.

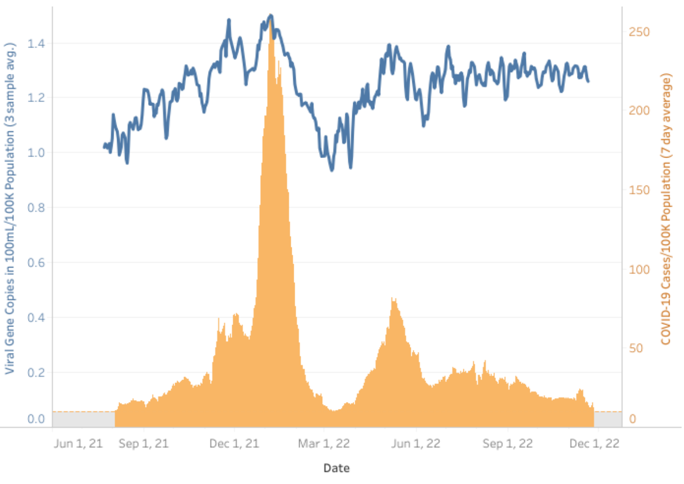
5. Categorize the total modeled percent change values to aid with interpretation. There will be a logarithmic nature to the total modeled percent change, so the categories should reflect that. The SWEEP dashboard uses the following 15-day trend categories:

⬆ 1000% or more
⬆ 100% to 999%
⬆ 10% to 99%
↗ 0% to 9%
↘ -1% to -9%
⬇ -10% to -99%
⬇ -100% to -999%
⬇ -1000% or more

- **Blue, up arrows** show an increasing trend.
- **Green, down arrows** show a decreasing trend.
- **Gray, right arrows** show a plateauing trend (-1% to -9% and 0% to 9% categories).

Viral Concentration Graphs

The SWEEP dashboard includes graphs of viral concentrations and corresponding clinical case counts over time.



Viral concentrations are shown as three-sample averages in gene copies per 100mL of sample per 100,000 people.
- Concentrations shown are **N1/N2 averages** and have been log-transformed.

Clinical COVID-19 cases are presented as seven-day averages in cases per 100,000 people.
- Cases are attributed to each sewershed by ZIP code.
- Cases have been calculated per 100,000 people to adjust for varying population size across sites.
- Case data is not shown on the graph when the average number of cases is fewer than 10 per 100,000 people to protect individual confidentiality.