Exploration of Wisconsin Based Wastewater Epidemiology

Marlin Lee, Kyllan Wunder, Abe Megahed
Data Science Institute, University of Wisconsin-Madison

Special thanks to Dagmara Antkiewicz, Adelaide Roguet
from the Wisconsin State Lab of Hygiene

Outline:

1. Problem Definition
    - What's WBE(Wastewater-Based Epidemiology)
    - Data sources - where is the data from (collection, analysers, reporters)
    - Data challenges
2. Investigations
    - Normalization
    - Cleaning
    - Analysis
3. Conclusions

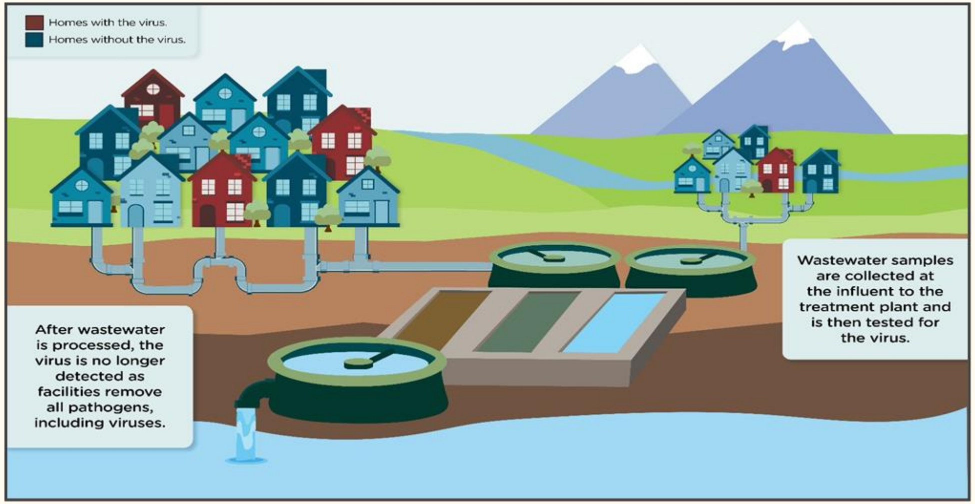
What is Wastewater-Based Epidemiology (WBE)?

Process:

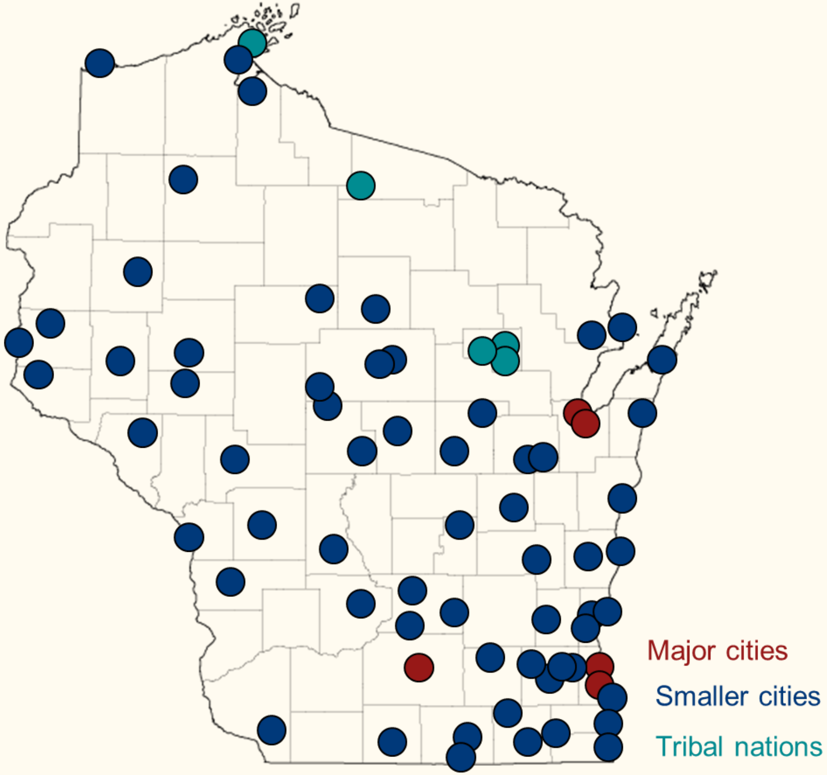
- Collect samples from sewersheds
- Analyze samples at State Lab of Hygiene
- Report results at Department of Health services

Goals:

- Help people make better policy
- Identify outbreaks before they manifest in the case data


🟥 Homes with the virus.
🟦 Homes without the virus.
Wastewater samples are collected at the influent to the treatment plant and is then tested for the virus.
After wastewater is processed, the virus is no longer detected as facilities remove all pathogens, including viruses.

Wisconsin Covid-19 Data Sources



Wastewater Data

- 82 Sites
- ~3 years
- 1 - 6 days / week
- 64,000 samples

Case Data

- 82 Sites
- ~3 years
- 7 days / week

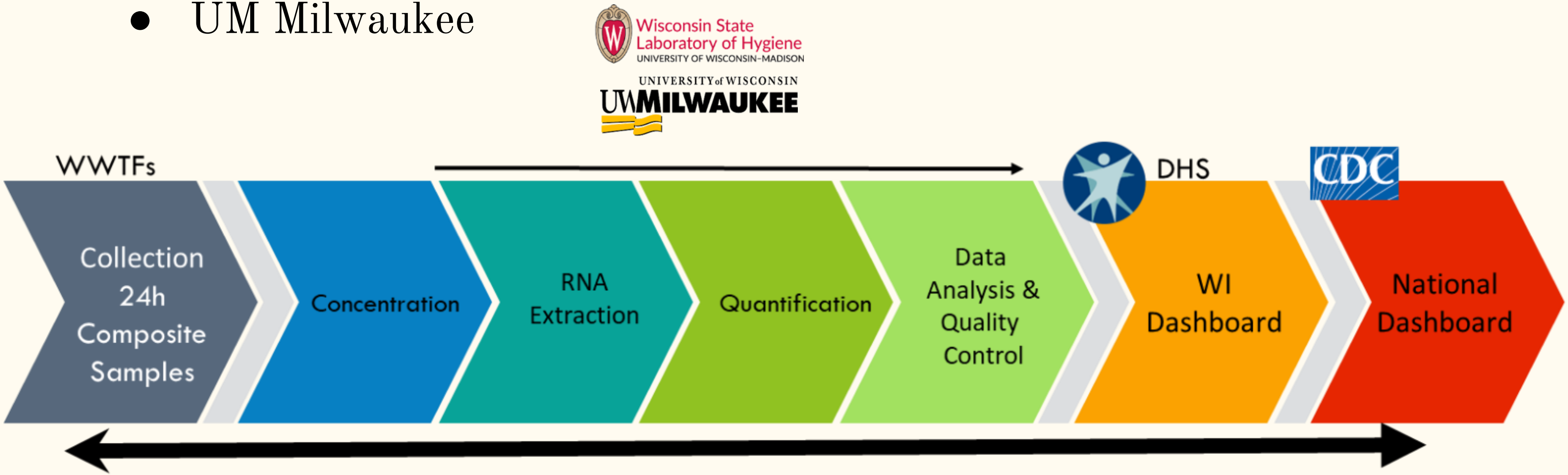
WBE is a Multi-Organization Collaboration

Regional

- DHS
- WSLH
- UM Milwaukee

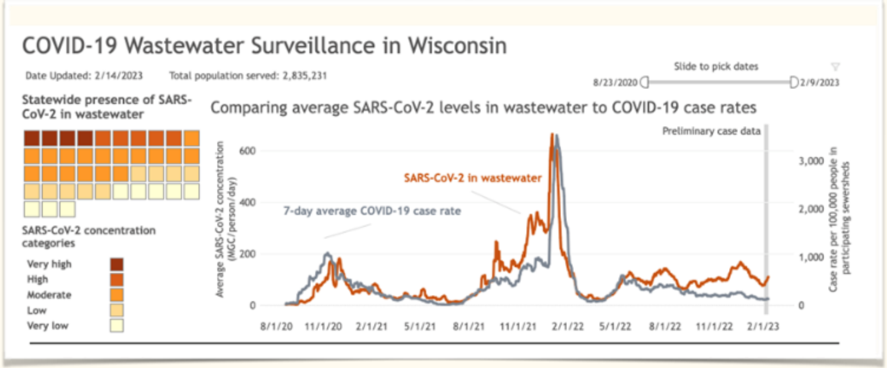
National

- CDC



Wisconsin State Laboratory of Hygiene
UNIVERSITY OF WISCONSIN–MADISON
UNIVERSITY of WISCONSIN
UWMILWAUKEE
WWTFs
DHS
CDC
Collection 24h Composite Samples
Concentration
RNA Extraction
Quantification
Data Analysis & Quality Control
WI Dashboard
National Dashboard

How WBE Information is Communicated to the Public



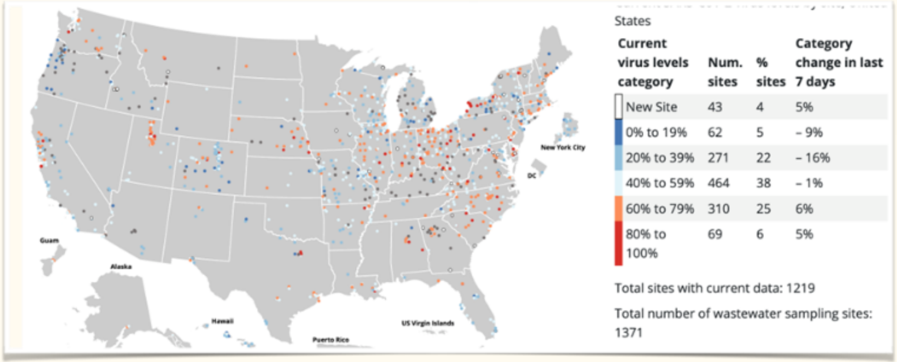
COVID-19 Wastewater Surveillance in Wisconsin

Date Updated: 2/14/2023 Total population served: 2,835,231

Slide to pick dates
8/23/2020
2/9/2023
Statewide presence of SARS-CoV-2 in wastewater
SARS-CoV-2 concentration categories
Very high
High
Moderate
Low
Very low
Comparing average SARS-CoV-2 levels in wastewater to COVID-19 case rates
Preliminary case data
SARS-CoV-2 in wastewater
7-day average COVID-19 case rate
Average SARS-CoV-2 concentration (MGC/person/day)
600
400
200
0
Case rate per 100,000 people in participating sewersheds
3,000
2,000
1,000
0
8/1/20
11/1/20
2/1/21
5/1/21
8/1/21
11/1/21
2/1/22
5/1/22
8/1/22
11/1/22
2/1/23

WI Department of Health Services (**DHS**) dashboard

https://dhs.wisconsin.gov/covid-19/wastewater.htm



States

Current virus levels category	Num. sites	% sites	Category change in last 7 days
New Site	43	4	5%
0% to 19%	62	5	- 9%
20% to 39%	271	22	- 16%
40% to 59%	464	38	- 1%
60% to 79%	310	25	6%
80% to 100%	69	6	5%

Total sites with current data: 1219

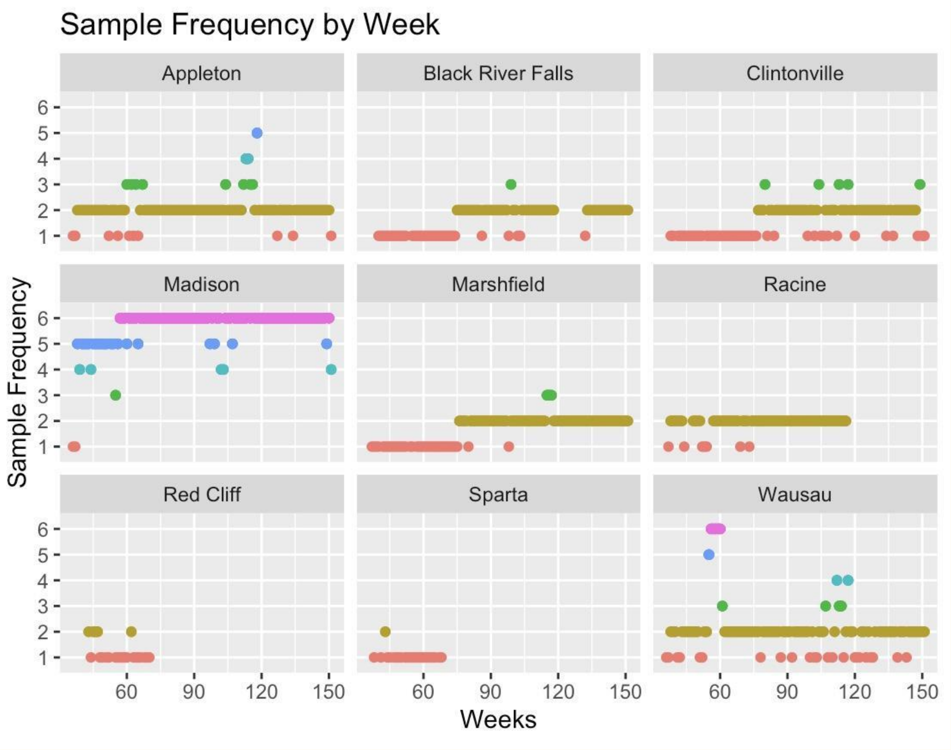
Total number of wastewater sampling sites: 1371

Centers for Disease Control and Prevention (**CDC**) dashboard

https://covid.cdc.gov/covid-data-tracker/#wastewater-surveillance

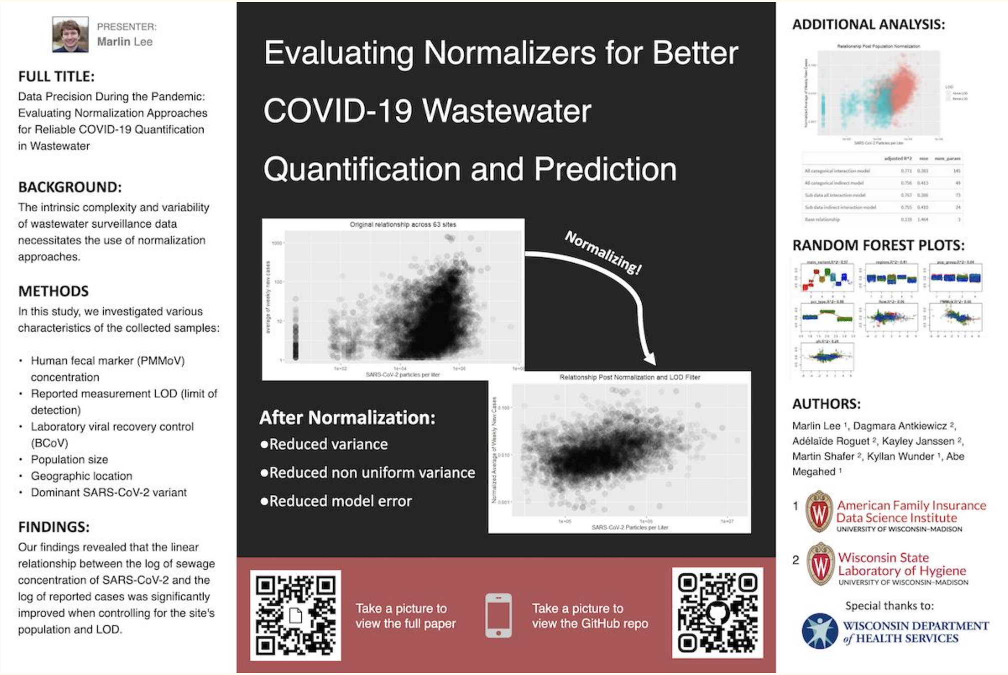
Challenges of Working with the Data

- Data is inherently noisy
- Sampling differences (once / week vs. many times / week)
- Analysis method differences (qpcr vs. dpcr)
- Many cofactors (population, time etc.)



Investigations

- Cleaning
  - Outliers
  - Smoothing
- Normalization
- Analysis
  - Variance analysis
  - Offset analysis



PRESENTER:
Marlin Lee

FULL TITLE:
Data Precision During the Pandemic: Evaluating Normalization Approaches for Reliable COVID-19 Quantification in Wastewater

BACKGROUND:
The intrinsic complexity and variability of wastewater surveillance data necessitates the use of normalization approaches.

METHODS
In this study, we investigated various characteristics of the collected samples:

- Human fecal marker (PMMoV) concentration
- Reported measurement LOD (limit of detection)
- Laboratory viral recovery control (BCoV)
- Population size
- Geographic location
- Dominant SARS-CoV-2 variant

FINDINGS:
Our findings revealed that the linear relationship between the log of sewage concentration of SARS-CoV-2 and the log of reported cases was significantly improved when controlling for the site's population and LOD.

Evaluating Normalizers for Better COVID-19 Wastewater Quantification and Prediction

After Normalization:

- Reduced variance
- Reduced non uniform variance
- Reduced model error

Take a picture to view the full paper

Take a picture to view the GitHub repo

ADDITIONAL ANALYSIS:

	adjusted R^2	mse	num_param
All categorical interaction model	0.771	0.383	145
All categorical indirect model	0.756	0.413	49
Sub data all interaction model	0.767	0.386	73
Sub data indirect interaction model	0.755	0.410	24
Base relationship	0.139	1.464	3

RANDOM FOREST PLOTS:

AUTHORS:
Marlin Lee [1], Dagmara Antkiewicz [2], Adélaïde Roguet [2], Kayley Janssen [2], Martin Shafer [2], Kyllan Wunder [1], Abe Megahed [1]

1 American Family Insurance Data Science Institute UNIVERSITY OF WISCONSIN–MADISON

2 Wisconsin State Laboratory of Hygiene UNIVERSITY OF WISCONSIN–MADISON

Special thanks to:
WISCONSIN DEPARTMENT of HEALTH SERVICES

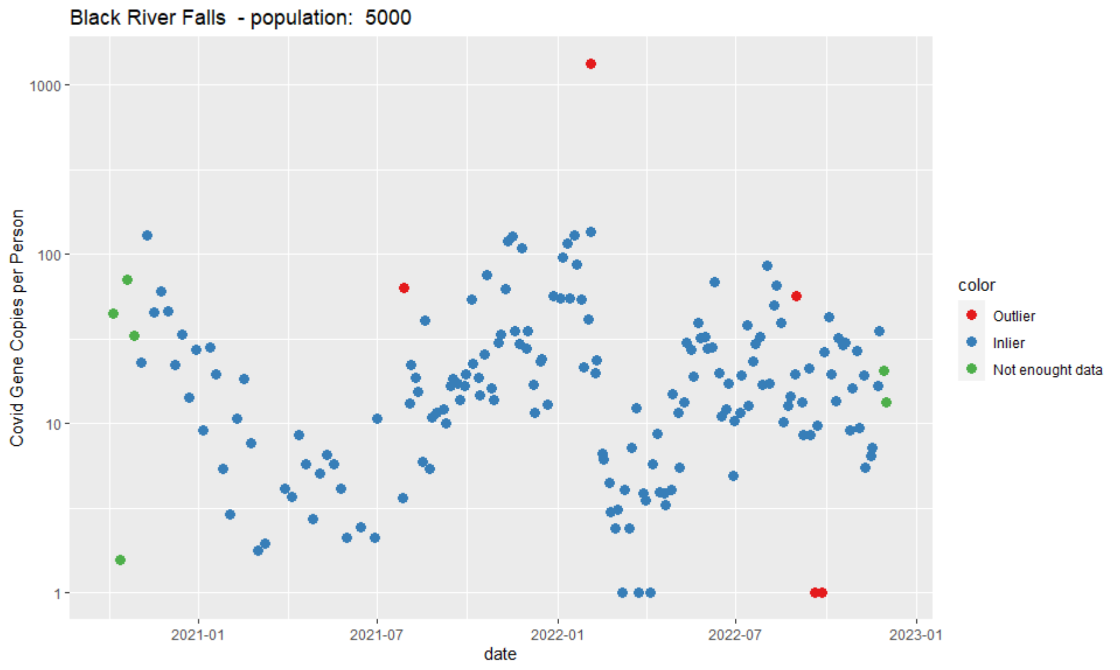
Original relationship across 63 sites
average of weekly new cases
1000
100
10
1
1e+02
1e+04
1e+06
SARS-CoV-2 particles per liter
Normalizing!
Relationship Post Normalization and LOD Filter
Normalized Average of Weekly New Cases
0.100
0.010
0.001
1e+05
1e+06
1e+07
SARS-CoV-2 Particles per Liter
Relationship Post Population Normalization
SARS-CoV-2 Particles per Liter
LOD

Presented on April 17, 2023 Midwest Chapter SETAC

Outliers

- Wastewater measurement data tends to be noisy

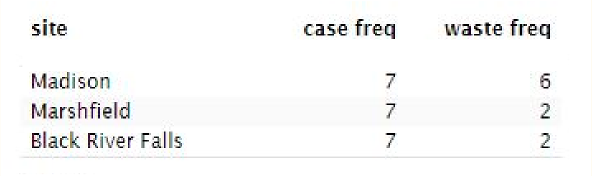
- Removing outliers allows improved detection of trends in the data



Outlier Detection

Solution:

- Look at local spikes
    - Only requires 7 data points / at most 2 week future knowledge
- Pick threshold based on historical spike data
    - Controls for difference frequencies



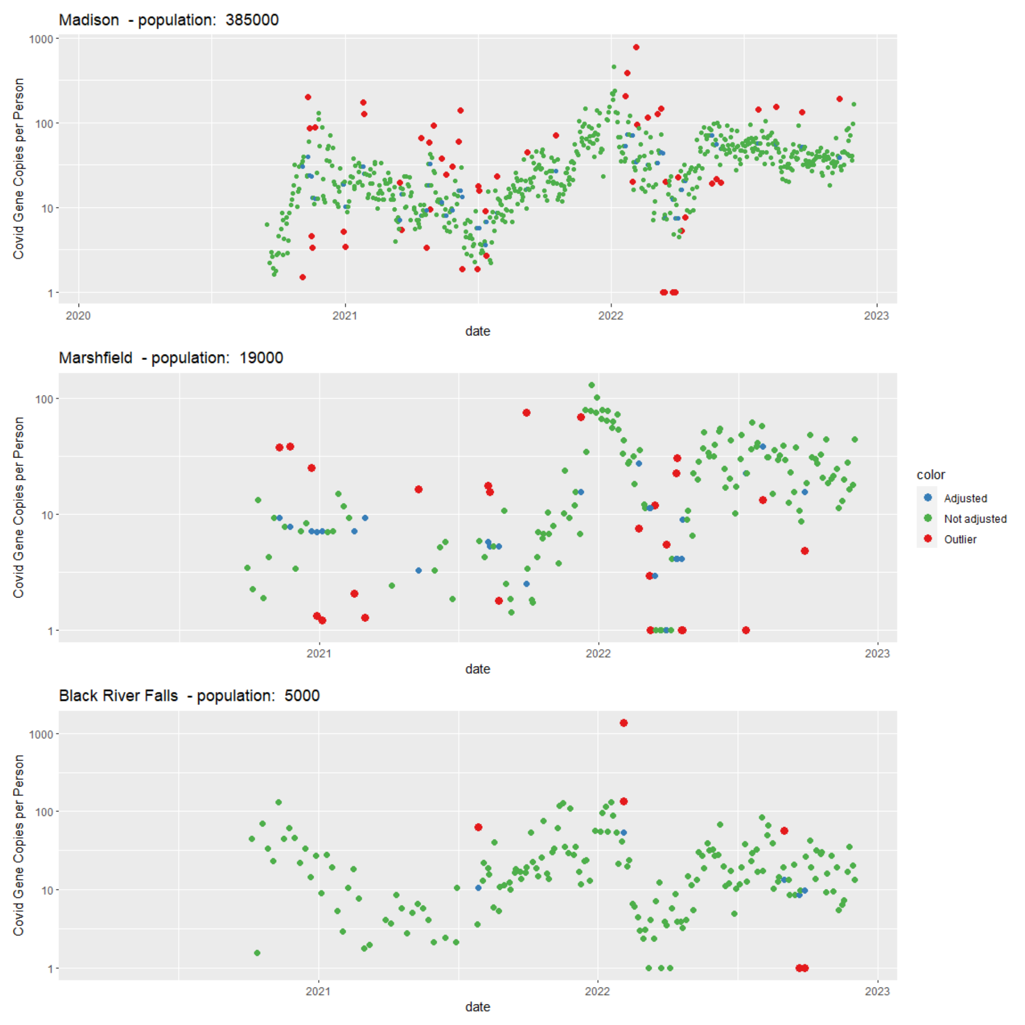
site	case freq	waste freq
Madison	7	6
Marshfield	7	2
Black River Falls	7	2

Problems:

- Inconsistent frequency of wastewater collection
    a. Solution must be able to handle this
- Low minimum frequency
    a. One measurement / week means 2 future data points is a soft cap on influence
    b. Removing data means no measurement for the week

Outlier Removal

- Removing outliers fixes issues on a small time scale
- Does not meaningfully improve overall trend



Smoothing

- Smoothing removes high frequency noise in the data

Approach:

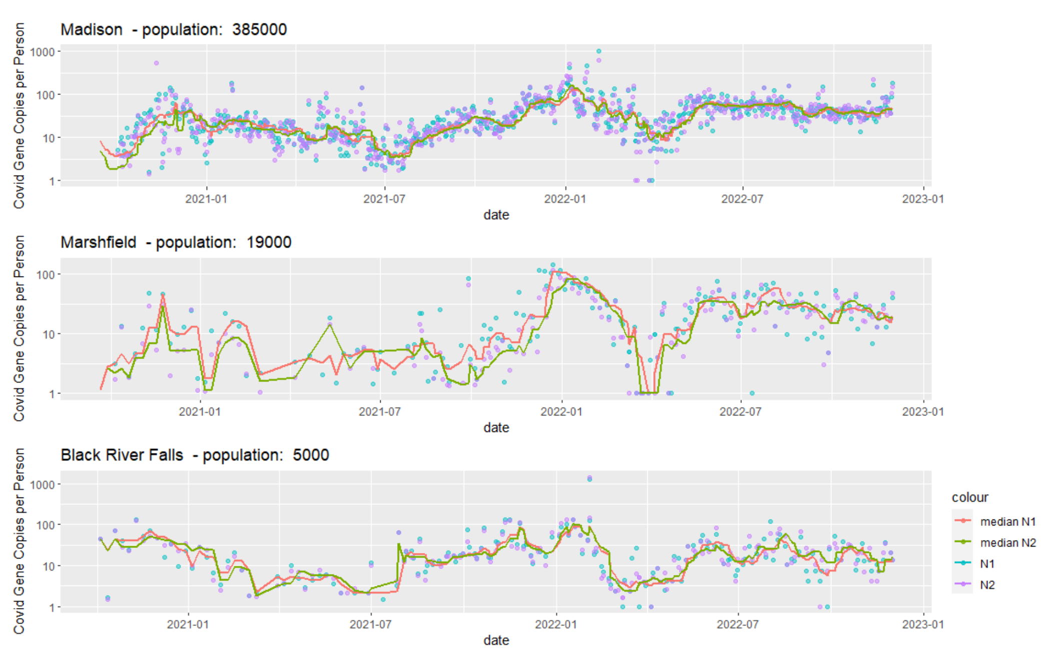
- Use right aligned Median smoothing

Issues:

- Sitewise Inconsistent frequency of wastewater collection

Results:

- Smoothed data (lines) provide better trending indicators than non-smoothed data (points)



Normalization

Goals:

- We want measurement to represent the true presence of Covid in the community

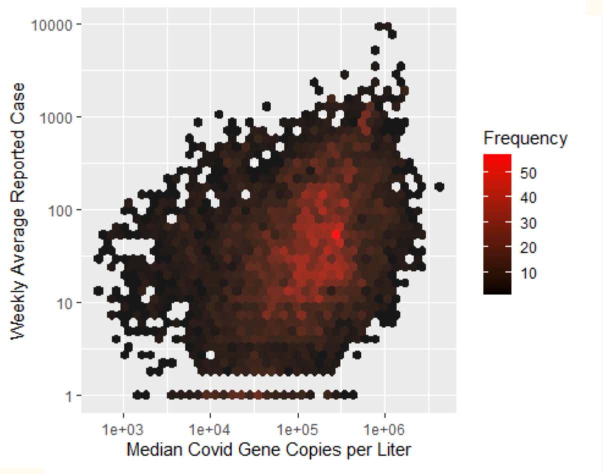
Concerns:

- Does the signal scale with population?
- Does the signal scale with the collection method (tests, flow)?

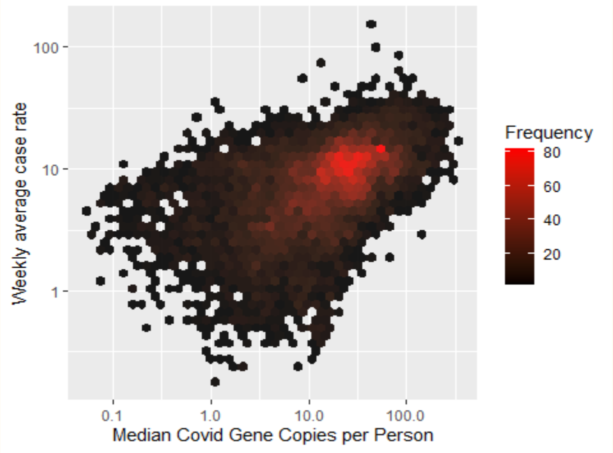
Results:

- Reduced variance
- Better captures underlying signal
- Reduced model error

Before:



After:



Normalization of Gene Copies

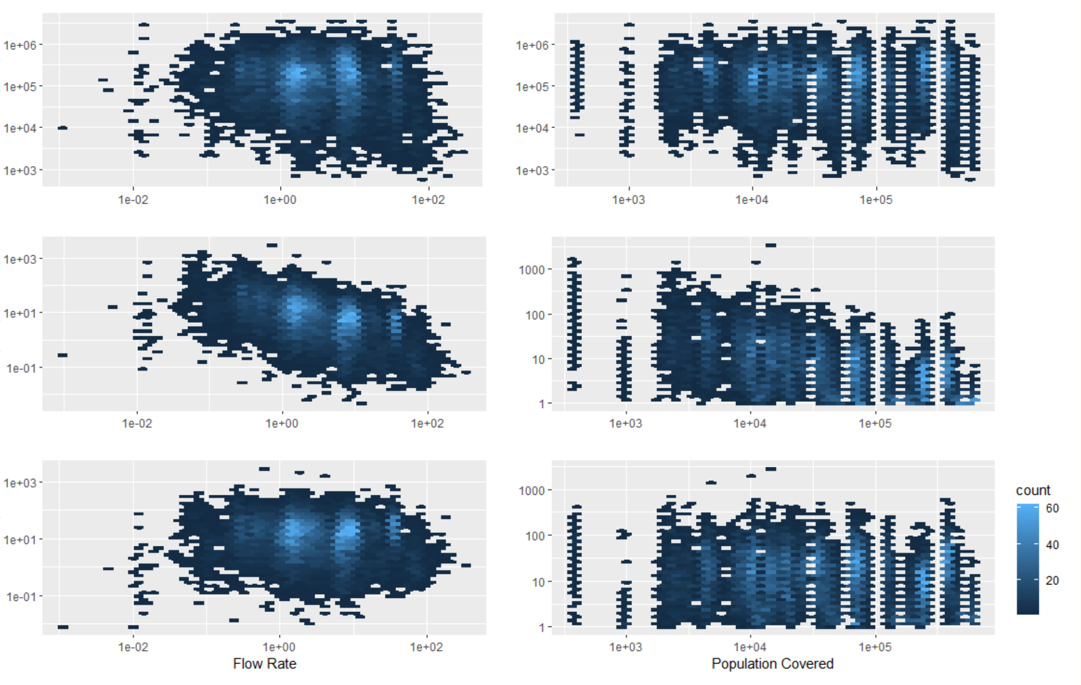
Conclusion:

Gene * Flow Rate / Population has the lowest covariation

Copies / liter

Copies / Liter / Person

Copies / Person



Normalization of Reported Cases

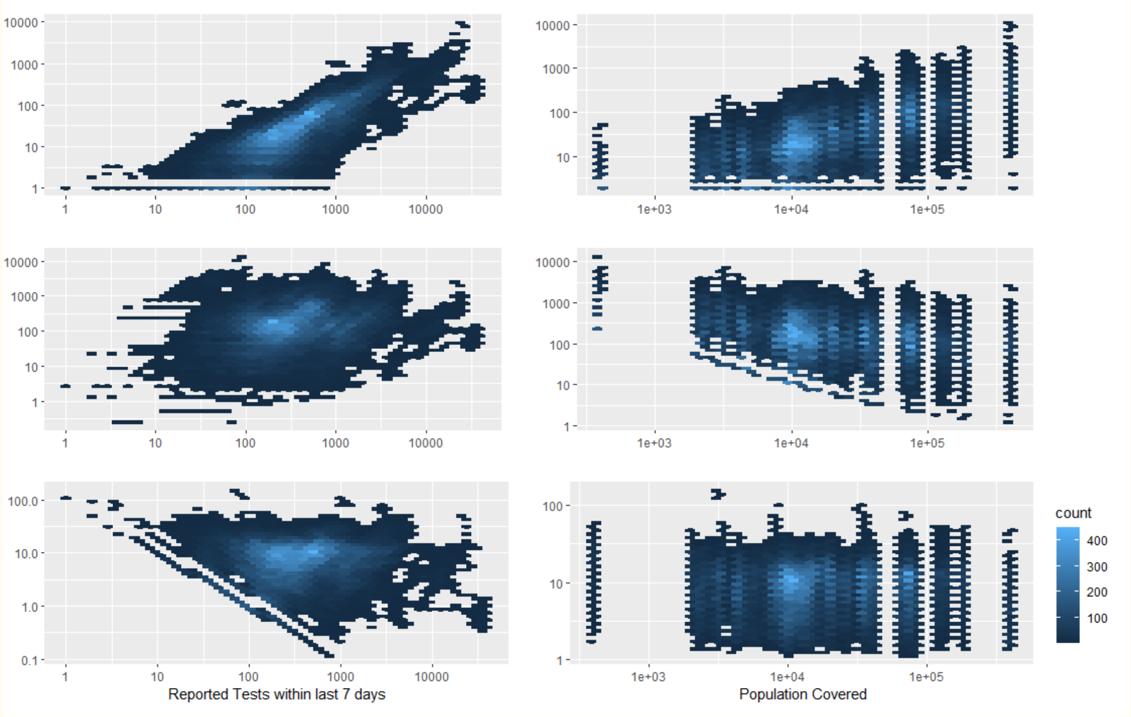
Conclusion:

Percent positive has the lowest covariation

Cases

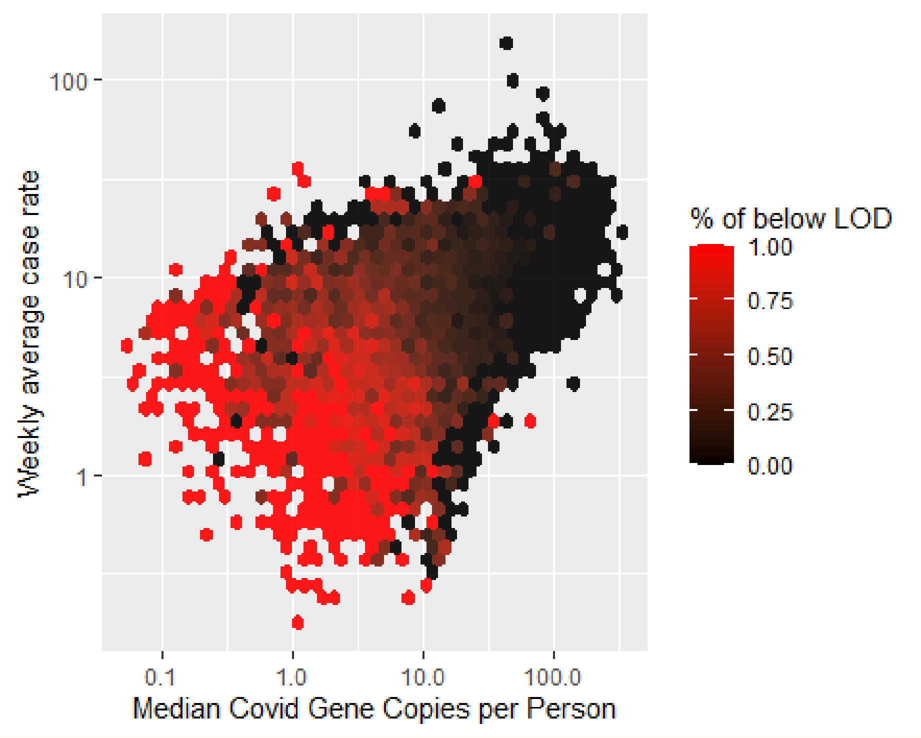
Cases / Person

Cases / Tests



Basic Level of Detection (LOD) Filtering

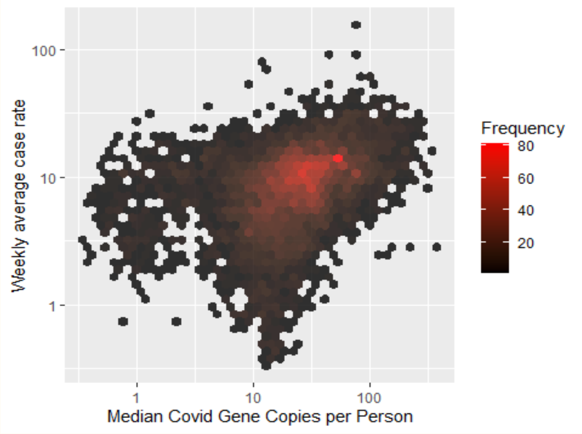
- Reported LOD changed over:
  - Time
  - Location
  - Method
- The occurrence of low case rates appears to be correlated with being below the LOD, independent of variations in gene copies.



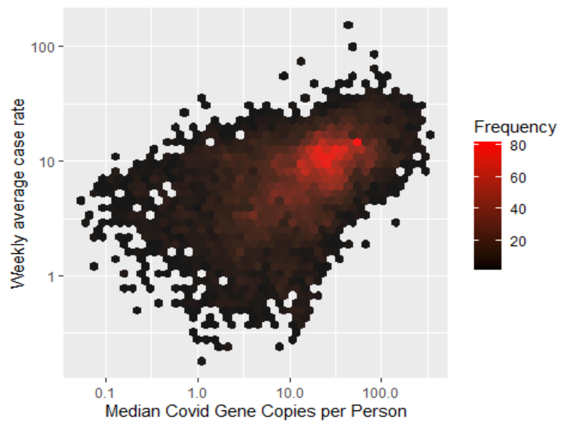
Level of Detection Regulating

- Controlling for values below the LOD threshold helps to reveal a more distinct trend.
- There is a decently strong connection between cases and N1/N2 concentration discounting time.

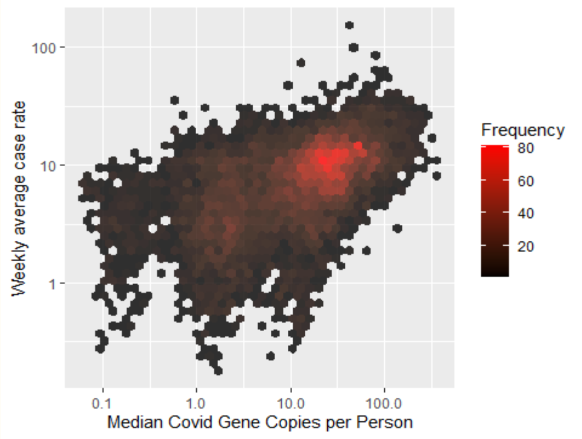
Drop values below LOD



No values removed

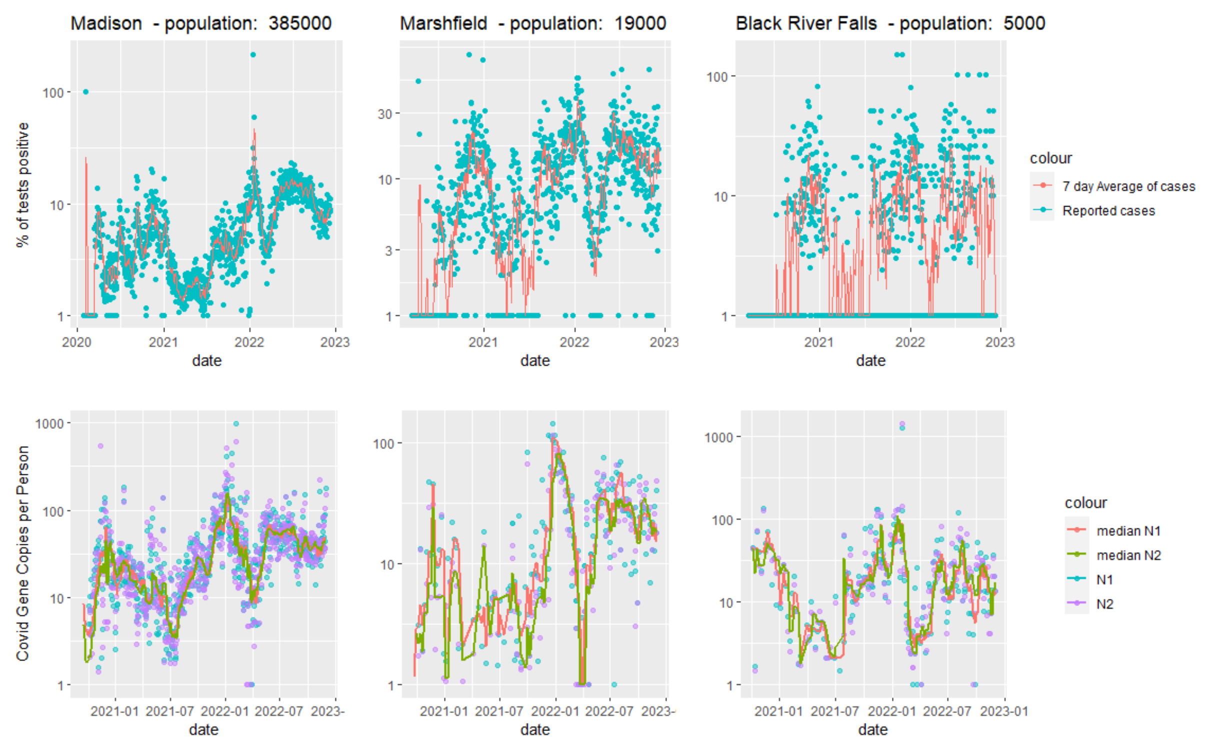


Set values below LOD to LOD / 2



Correlation Between Reported Cases and Gene Copies

- Reported cases (above) track with gene copies (below) over time.

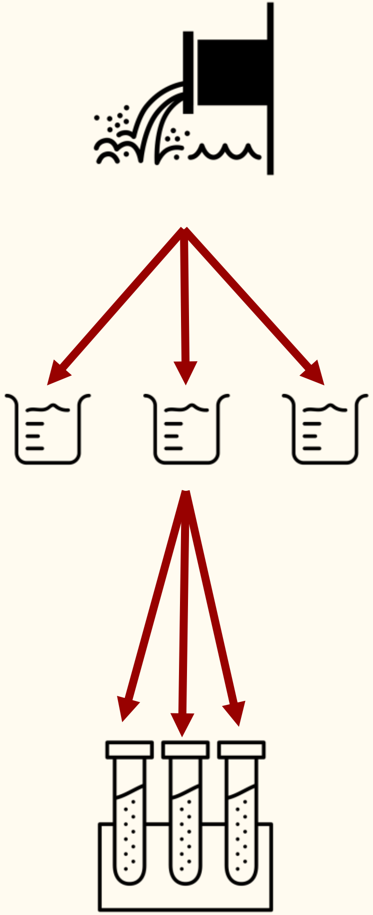


N1, N2 = gene markers

Variance Analysis Overview

Made easy:

- WSLH ran a high frequency testing period at 10 sites for 42 days
- Each day 3 filters were collected with 3 measurements each meaning 9 total measurements
- Used to calculate sources of error in each step of collection

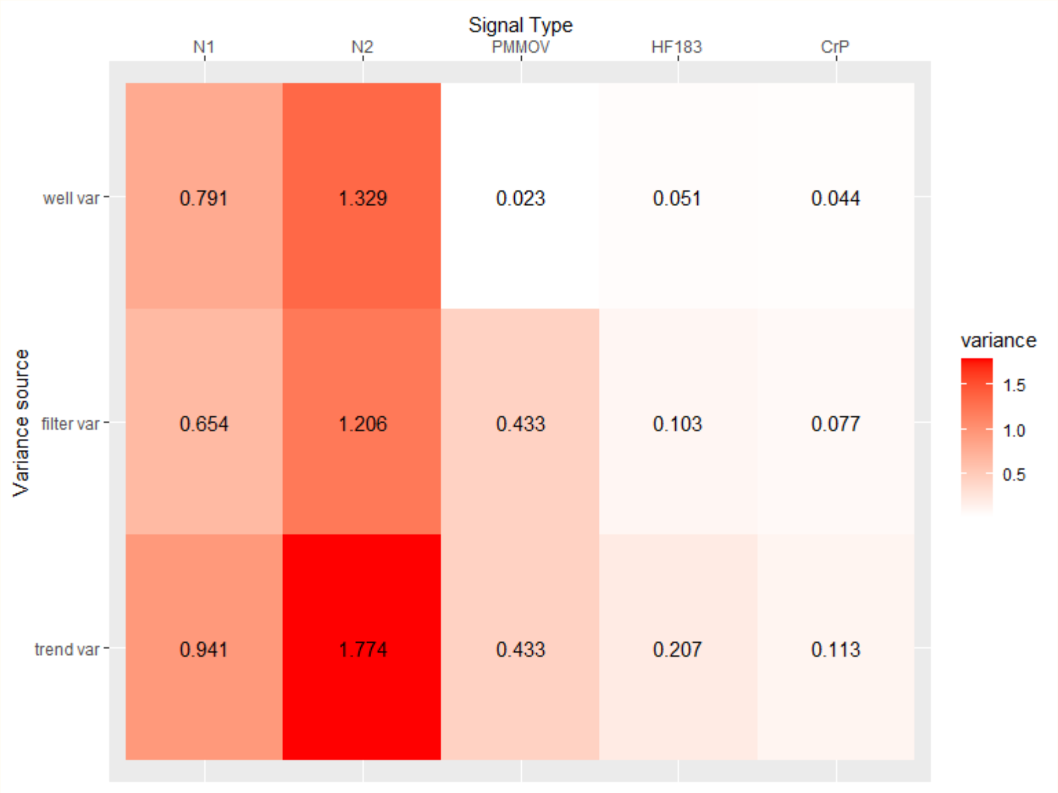


Variance Analysis Results

- Normalizers (HF183, CrP, and PMMoV) have lower systemic variance than N1 and N2.

- PMMoV has higher systemic variance than other normalizers

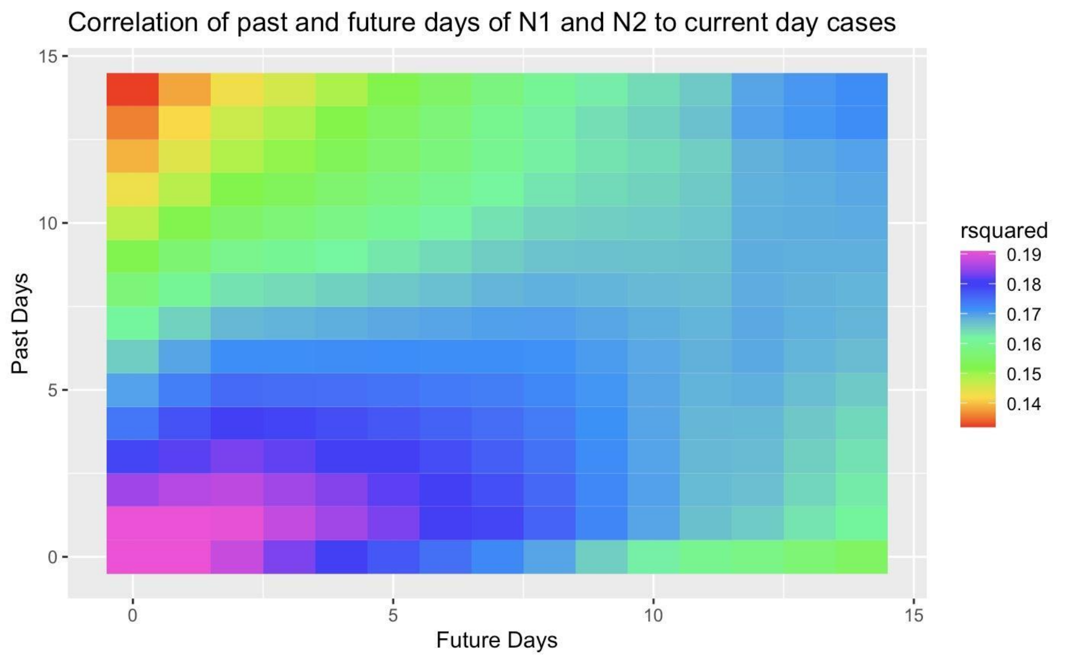
- N2 has higher variance at ever level then N1



Time Series Analysis

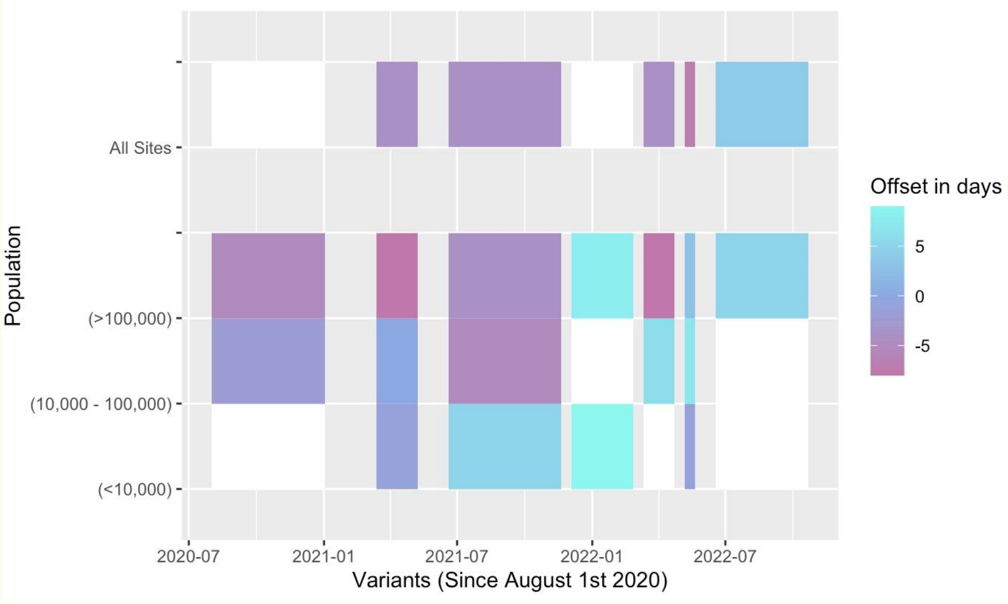
- Goal:
  Determine temporal offset between cases and wastewater (N1/N2) measurements.

- Results:
  Future days more important



Offset Analysis Results

- Offset between wastewater and case data varies by variant.
- Offset is sometimes lagging and sometimes leading case data.
- Variant and population size both affect offset.
- Reasons for differences between variants are unknown.



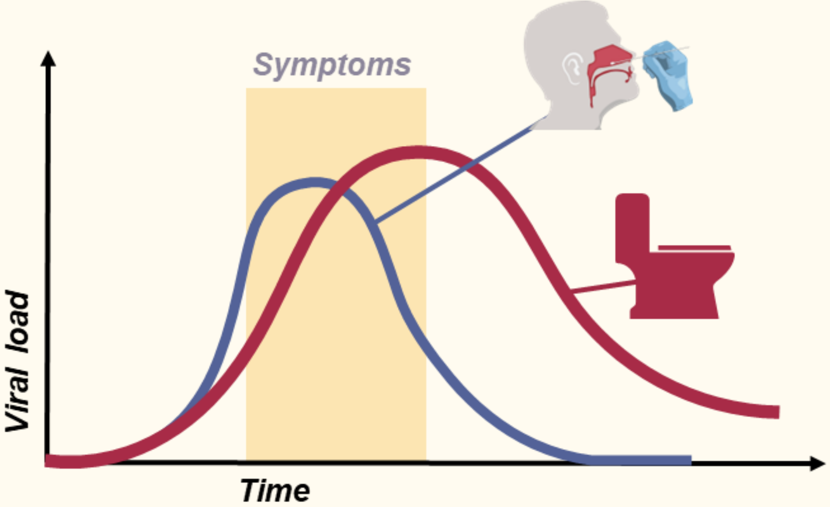
□ Offset out of range >10 or <-10

Conclusions

- Outlier detection helps fix local (short term) issues but does nothing for the wider trend.
- Smoothing works but creates lagging issues.
- Normalization has huge impact on signal comparison.
- A more comprehensive breakdown of the sources of variance allows for better, more predictive models.

Future Directions

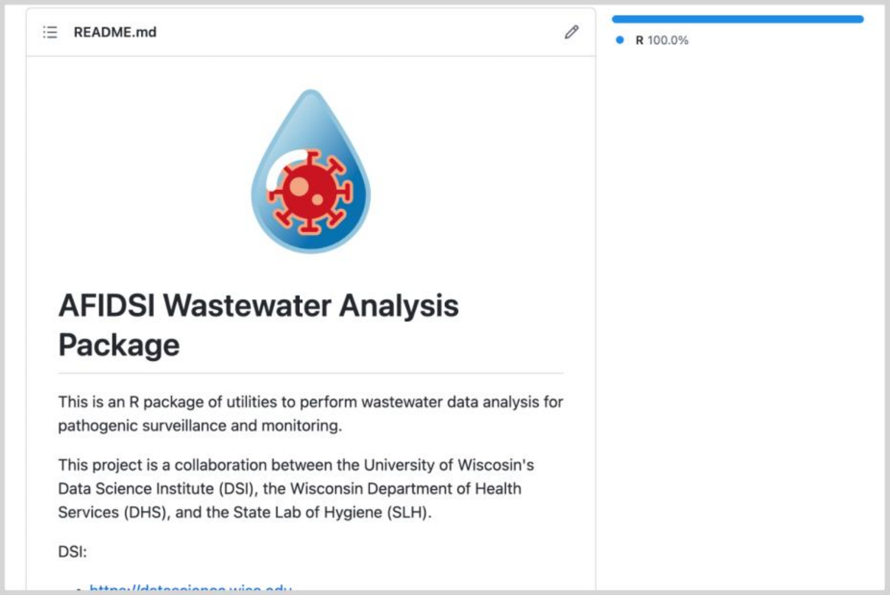
- Achieve a better understanding of variance in order to create more comprehensive model.
- Improve estimates of shedding and case testing viral load distributions.
- Continue optimizing current methods.



Final Products

- Presented two poster sessions
- GitHub repository
- Contact us if you’d like more information.

Contact:
Marlin Lee
mrlee6@wisc.edu



AFIDSI Wastewater Analysis Package

This is an R package of utilities to perform wastewater data analysis for pathogenic surveillance and monitoring.

This project is a collaboration between the University of Wisconsin's Data Science Institute (DSI), the Wisconsin Department of Health Services (DHS), and the State Lab of Hygiene (SLH).

DSI:

- https://datascience.wisc.edu

https://github.com/AFIDSI/Covid19-Wastewater-Analysis