Sensitivity of Wastewater Viral Loads for Detection Based on Technology Limits – Correlation Analysis

Viral loads (VL) in wastewater have gained importance in the recent years as a disease surveillance signal. It has been considered as a possible early warning signal for outbreaks [1]. However, it is noisy and there has been a lot of effort to extract signal from noisy observations. The level of correlation between wastewater detections and COVID-19 surveillance signals like reported cases and hospitalizations is not fully understood, and continues to be explored extensively [1, 2]. In addition, Wastewater surveillance has been considered an early indicator that the number of people with COVID-19 in a community is increasing or decreasing.

Aim: To analyze the correlation of SARS-COV2 viral loads in wastewater and the COVID-19 surveillance signals like hospitalizations. We also investigate the leading indicator characteristics of the viral load to cases and hospitalizations.

Data:

Wastewater data: We employ the wastewater data provided by Virginia Department of Health [2]. In Virginia, wastewater is collected each week from 36 WW treatment plants with 13 sites sampling twice weekly and 23 sites sampling once a week. The geographic distribution of the plants and population sizes served are shown in Figure 1. These samples help us track the amount of SARS-CoV-2, which is the virus that causes COVID-19, in the wastewater. VDH analyzes the amount of SARS-CoV-2 virus pieces found and using the total daily flow at the treatment plant, calculates the "viral load" (the total amount of viral pieces entering the wastewater treatment plant that day). Considering that people infected with COVID-19 can shed the virus in their feces (which then gets flushed down the toilet), VDH can track if COVID-19 infections are increasing or decreasing in the community served by a wastewater treatment plant. This community that feeds into the wastewater treatment plant is known as a "sewershed". The data should be interpreted with this limitation in mind and should be used together with other data points.

Hospitalizations data: The data is provided by HealthData.gov and is publicly available at https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh/about_data. The dataset consists of state-aggregated data for hospital utilization **in a timeseries format**. The hospital utilizations are derived from reports with facility-level granularity across three main sources: (1) HHS TeleTracking, (2) reporting provided directly to HHS Protect by state/territorial health departments on behalf of their healthcare facilities and (3) National Healthcare Safety Network (before July 15).

Methods: We employ Spearman's rank correlation to determine the linear relationship between VL and cases time series. In order to determine the leading indicator and lagging indicator we perform correlation between the hospitalizations signal and multiple shifted versions of the VL signal. We shift the VL signal to the right (+ve shift) by up to two weeks and compute its correlation with the hospitalization signal. Similary, we shift the VL signal to the left (-ve shift) by up to two weeks and correlate it with hospitalizations. Hence, for each shift, we obtain a correlation values, and the shift that yields the highest correlation is considered the *best shift.*

Results and Observations:

- The number of sites with significant correlation varies across time (cf. Figures 1).

- **During a surge the number of sites with significant correlation increase**
    - VLs from a significant proportion of these sites are leading indicators during a surge (cf. Figures 1)
- **Spatio-temporal trends for VDH regions based on relation to hospitalizations data**:
    - The distribution of correlation and shifts vary across regions.
    - Correlation distribution for Northern regions skewed towards higher correlation values, Far SW towards lower values (cf. Figures 2a)
    - Shifts distribution for several of these sites skewed towards +ve shift values (cf. Figures 2b)
- Typically, sites with higher population tend to have higher correlation (cf. Figure 3a)
    - The best shifts are also skewed towards positive shift values which indicates that the VLs are typically leading indicators to hospitalizations time series (cf. Figures 3b).

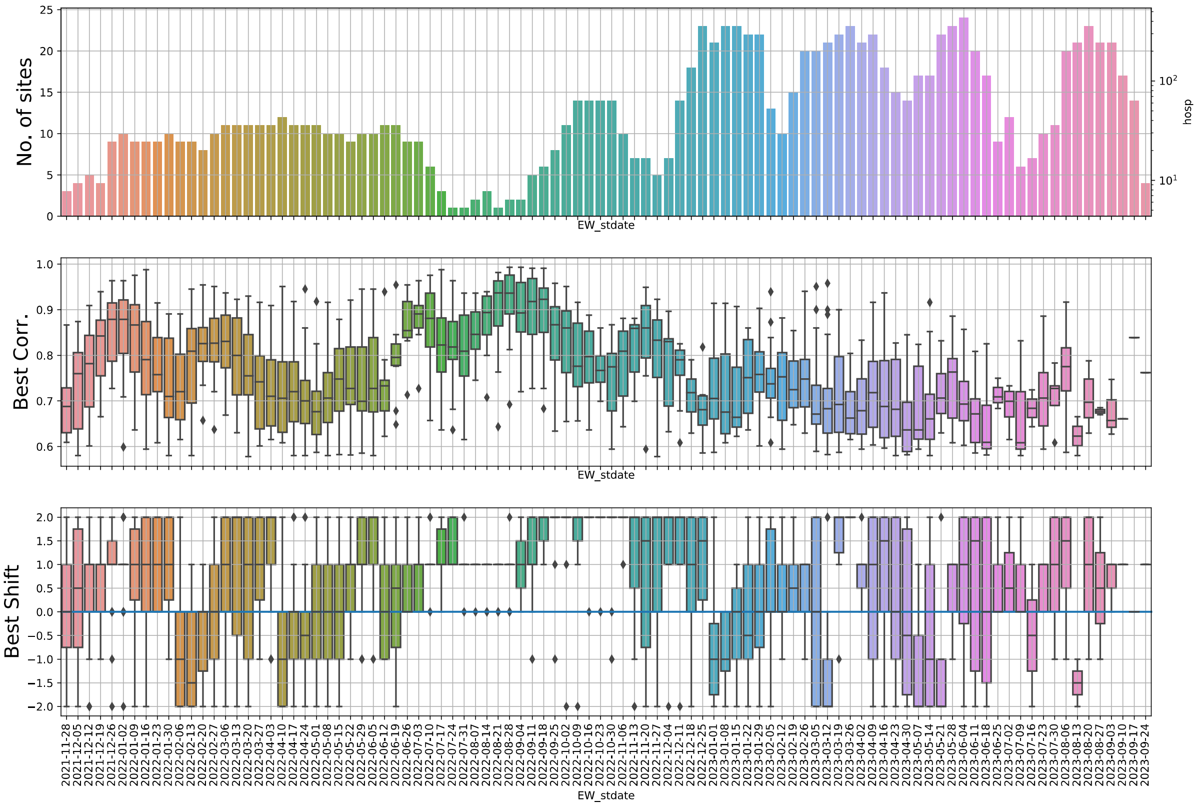


Figure 1: Correlation results for hospitalizations. Row1 – number of sites in a particular week with significant correlation, Row2 -- the distribution of correlation values in a particular week computed across all sites. We observe that during a surge the distribution is skewed towards higher correlation values. Row3 – The distribution of shifts in a particular week computed across all sites. Again, during a surge in hospitalizations, we observe that the best shift distribution is skewed towards positive shifts, indicating that the VL signal tends to lead the hospitalizations signal, mostly.

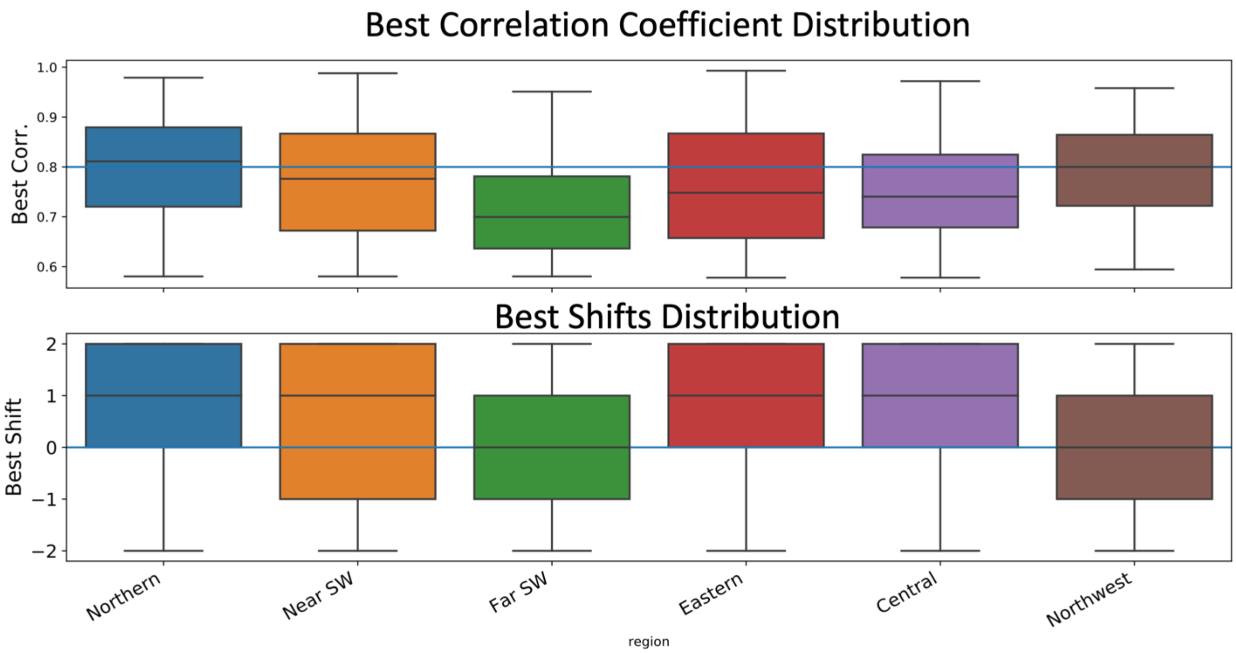


Figure 2: Top row Boxplot indicates the distribution of the correlation values across time for a particular region. These plots indicate that the distribution of best correlation values for regions like Northern are skewed towards higher correlation values compared to Far southwest and central regions. Bottom row Boxplot indicates the shift distribution across time for a particular region. This plot indicates that for most regions the boxplot is skewed towards positive shift values.

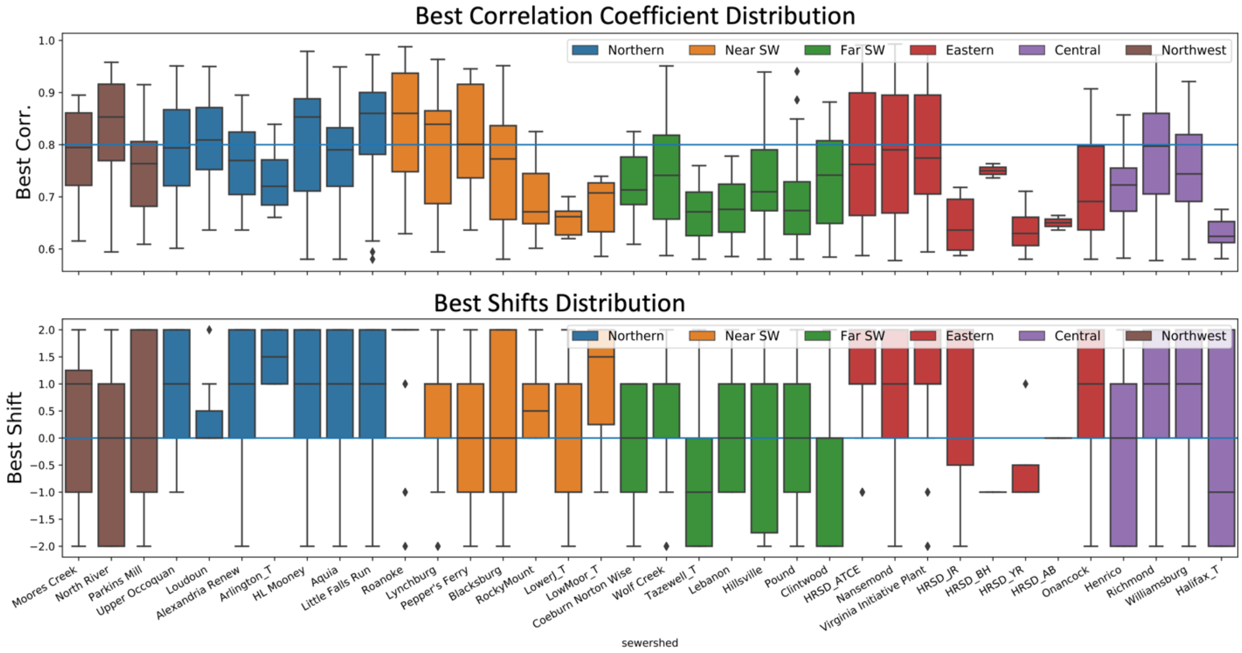
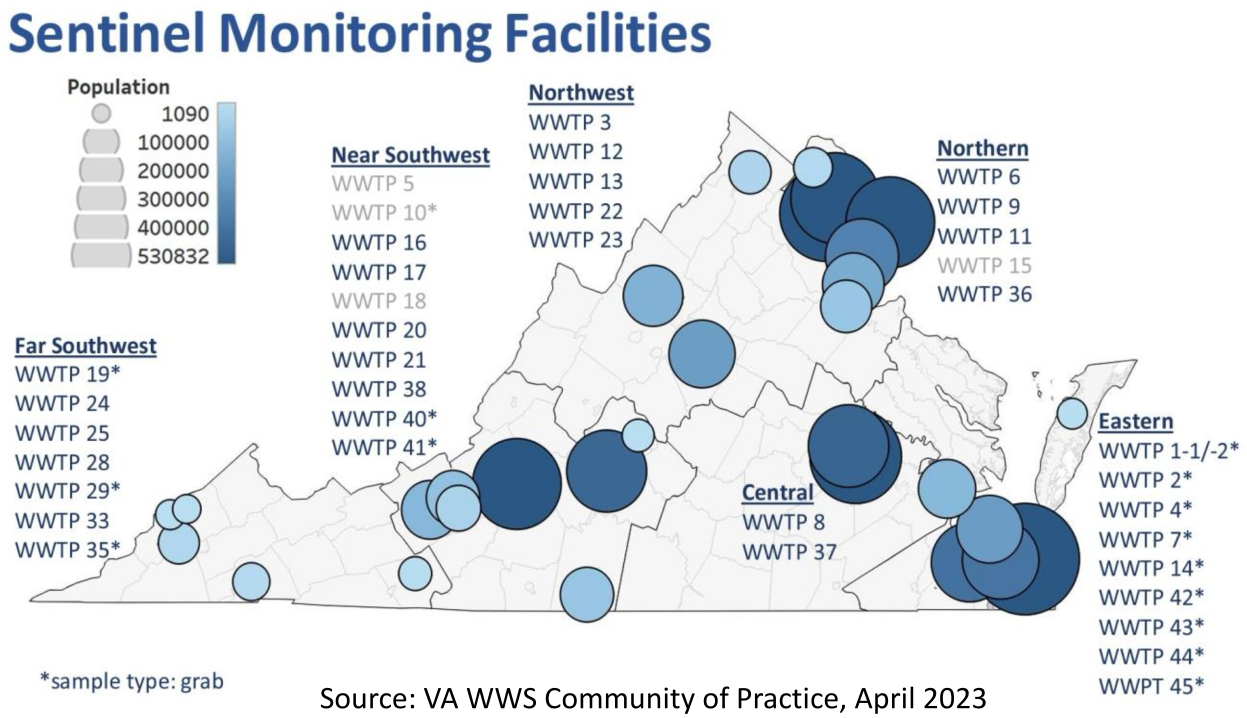


Figure 3: Top row Boxplot indicates the distribution of the correlation values across time for the various sewersheds. Bottom row Boxplot indicates the shift distribution across time for a particular sewershed. Further, the sewershed within a region are ordered from the highest population site to the lowest population site. Typically we observe that the distribution of the higher population sites tend to have higher correlation values.

[1] https://www.cdc.gov/nwss/wastewater-surveillance.html
[2] https://www.vdh.virginia.gov/coronavirus/see-the-numbers/covid-19-data-insights/sars-cov-2-in-wastewater/

Appendix



Source: VA WWS Community of Practice, April 2023

Figure 4: The geographic placement of the WWTPs across Virginia along with the population sizes served. In total, there are 36 WWTPs with 13 sites sampling twice weekly and 23 sites sampling once a week. Approximately, 50% of the Virginia's population is monitored through these sites.

Correlation Analysis: We employ Spearman's rank correlation to determine the linear relationship between logarithm-transformed viral signal and the cases time series. The correlation is computed on a rolling-window basis with a window size of 12 weeks. We shift the VL signal to the right (+ve shift) by up to two two weeks and to the left (-ve shift) by up to two weeks and find the shift with the best correlation. If the shift with the best correlation is positive then we conclude that VL is a leading indicator for that window period, if the shift with the best correlation is negative then we conclude that VL is a lagging indicator to the cases time series. See figures A, B and C for reference. The

Corr(VL(t+shift), cases(t)), *where* $\boldsymbol{shift} \in \{-\mathbf{2}, -\mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{2}\}$

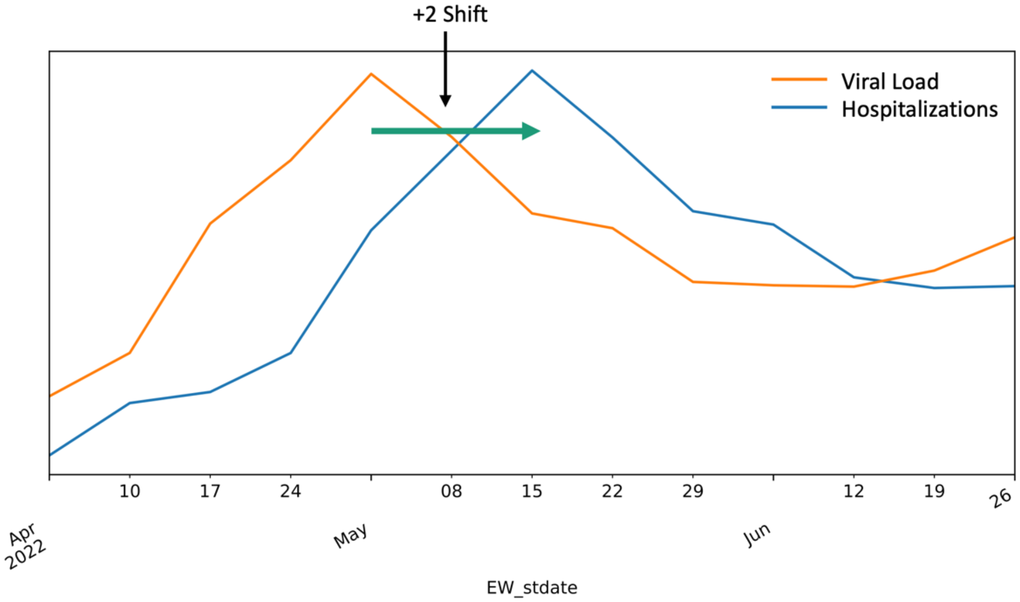


Figure 5: An example of the correlation analysis showing an instance where VL (orange) is a leading indicator of the hospitalizations (blue). Shifting the VL forward by two weeks (positive shift) and computing the correlation yields the highest correlation.

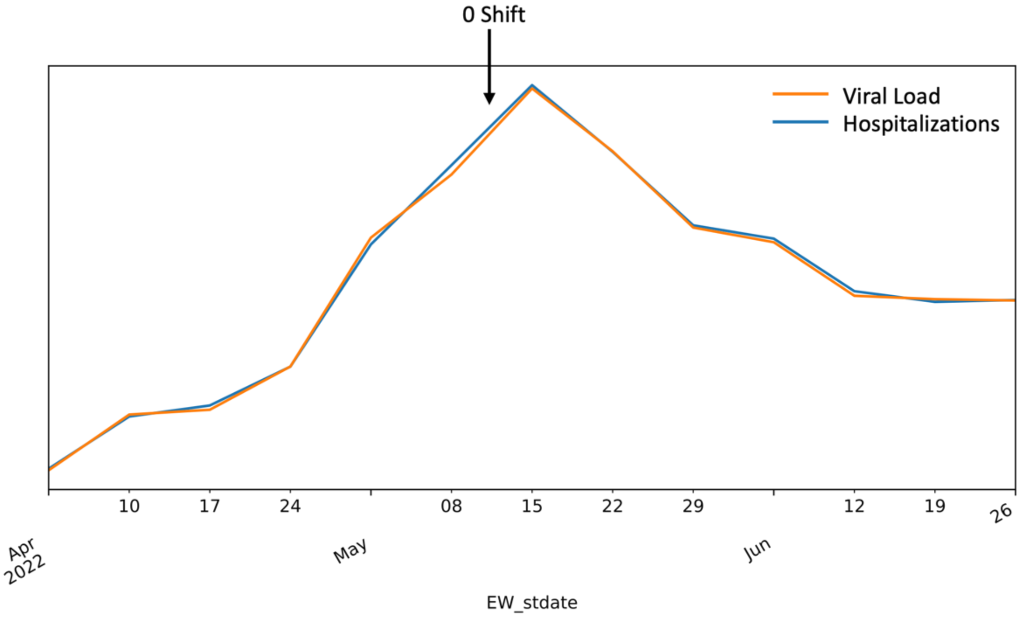


Figure 6: An example of the correlation analysis showing an instance where VL (orange) is neither a leading or a lagging indicator of the hospitalizations (blue). The VL and hospitalizations time series are aligned.

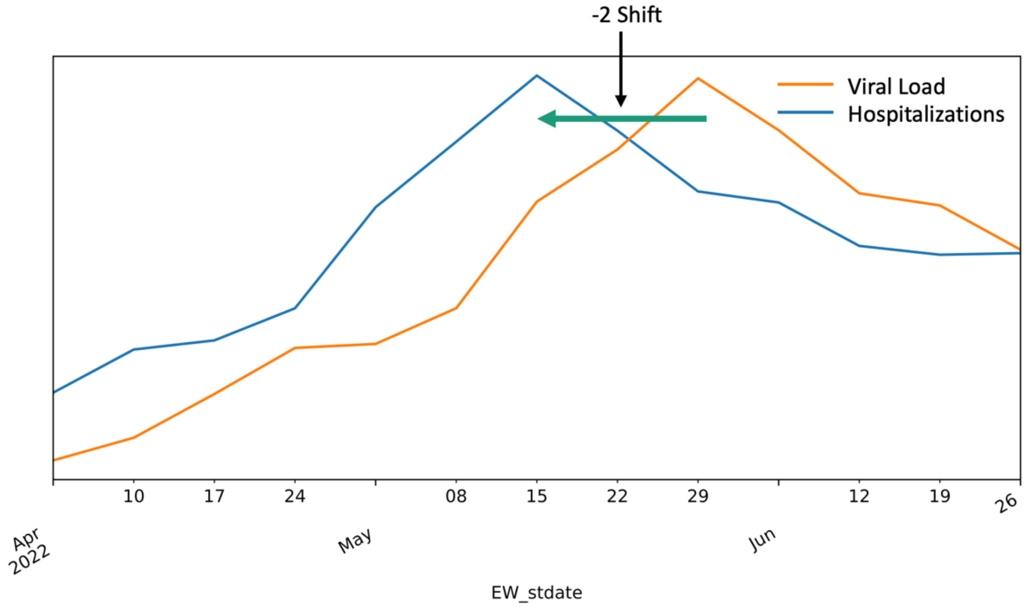


Figure 7: An example of the correlation analysis showing an instance where VL timeseries(orange) is a lagging indicator of the hospitalizations time series (blue). Shifting the VL behind by two weeks (negative shift) and computing the correlation yields the high.