Heuristic exploration of the relationship between cases and viral load

Marlin Lee

1/28/2022

This report looks is an update to the analysis shown on 1/14/2022. Most steps are the same with tweaking due to different data and outlier procedure

at exploring the relationship between wastewater and cases. There are four components to this analysis.

1) Removing putative outliers
2) Binning analysis
3) Smoothing signal
4) Statistical analysis

This report does not present any final answers but presents some very convincing heuristics.

"data Used from DSIWastewater package"

Data: The first look

The two data sets used in this analysis are the Madison case data sourced from the Wisconsin DHS and wastewater concentration data produced by the Wisconsin State Laboratory of Hygiene. This wastewater data has entries every couple of days from 15 September 2020 to 19 April 2022.

```
##    min(Date)  max(Date)
## 1 2020-09-15 2022-04-19
```

Date	Site	Cases	MACases	sars_cov2_adj_load	MAWaste
2020-09-15	Madison	65	191.00000	26.763985	NA
2020-09-19	Madison	129	142.14286	6.236600	NA
2020-09-22	Madison	65	120.00000	9.722190	NA
2020-09-23	Madison	122	106.28571	8.244951	NA
2020-09-24	Madison	90	94.85714	4.703074	NA
2020-09-25	Madison	92	87.42857	3.176360	NA

The case data has a strong weekend effect so for this section we look at a seven day smoothing of cases. The simple display of the data shows the core components of this story. First, wastewater data is noisy. And that there is a clear relationship between the two signals.

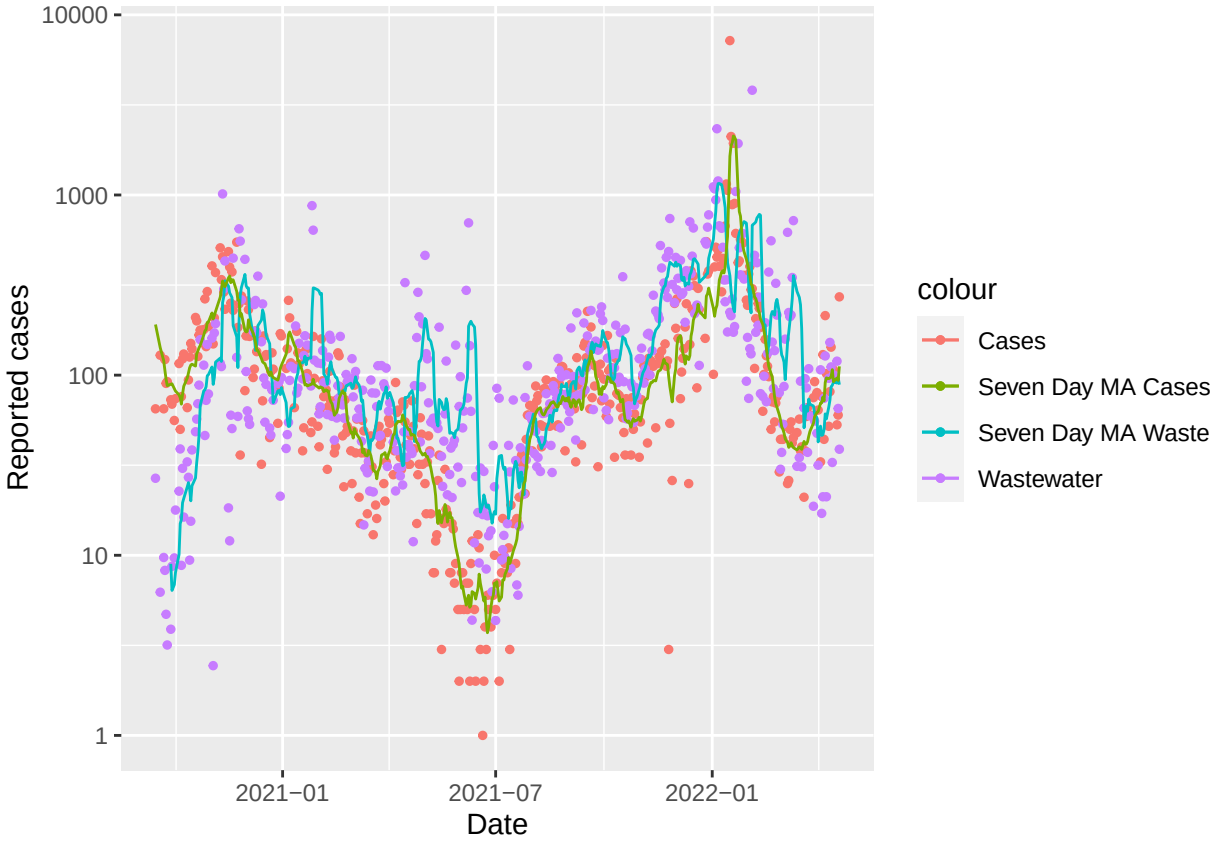

Figure 1: Wastewater concentration and daily Covid-19 case data for Madison. A seven day moving average of cases is used to reduce a day of the week effect.

Removing potential outliers

Looking at the wastewater measurements we observe there were some points many times larger than adjacent values hinting at them being outliers. We used the adjacent 10 values on each side and marked points 2.5 standard deviations away from the group mean as outliers.

Figure 2: Wastewater concentration for Madison with potential outliers marked. Using a rolling symmetrical bin of 21 days as a sample we use 2.5 standard deviations of the bin as a metric to reject extreme points. This process is ran multiple times to get a robust process to select outliers.

Data smoothing

The goal in this section is to smooth the data to get a similar effect without losing resolution.

viral load smoothing

To get a good smoothing of the sars_cov2_adj_load measurement we employ loess smoothing. Loess smoothing takes a locally weighted sliding window using some number of points. we found the best smoothing when it uses data within approximately 0 weeks of both sides of the data. The displayed plot shows the visual power of this smoothing. We see in general that the smoothed N1 trails SLD. However loess is symmetric meaning that it can not be used in predictive modeling due to it using points from the future to smooth points.

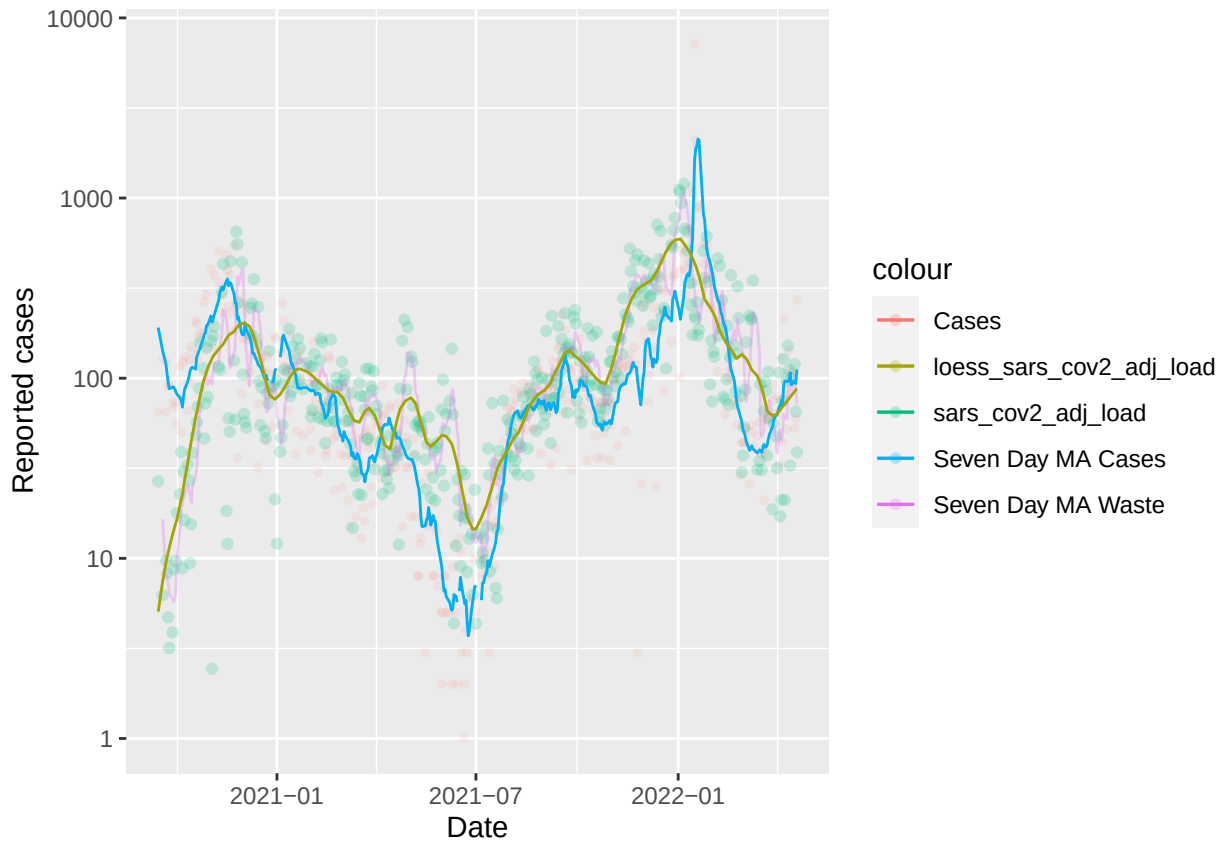

Figure 3: Loess smoothed N1 and SLD cases for Madison data. Using a Locally Weighted Scatterplot Smoothing process along with the previous figure SLD cases we get the most sophisticated relationship between the two signals discussed in this document.

Towards a formal analysis

Cross correlation and Granger Causality are key components to formalize this analysis. Cross correlation looks at the correlation at a range of time shifts and Granger analysis performs a test for predictive power.

	Max Cross Correlation	Lag of largest Cross correlation	P-value Wastewater predicts Cases	P-value Cases predicts wastewater
Section 1: Cases vs sars_cov2_adj_load	0.4673	18	0.0590	0.0000
Section 1: 7 Day MA Cases vs sars_cov2_adj_load	0.5231	9	0.0286	0.0000
Section 2: Cases vs sars_cov2_adj_load	0.5517	1	0.3790	0.0002
Section 4.3: 7 Day MA Cases vs Loess smoothing of sars_cov2_adj_load	0.7109	1	0.0234	0.0532