Stat 416 Assignment 1 Due Monday, September 30 at 11:59:59PM

Julia Schedler

A paper I worked on as a research scientist considered the time series of the concentration (measured as $\log_{10}$ copies per Liter) of the SARS-CoV-2 virus from 5 different locations in the City of Houston, visualized in parts (c)-(g) of the figure below.

The goal of this study was to see whether the information gleaned from sampling the lift stations, which represent smaller populations, was different than the information gleaned from sampling only the larger wastewater treatment plant. In other words, one research question was to determine whether the WWTP (dark blue) time series has different dynamics (behavior) than those that represent the lift stations.

The methods in this paper are touched on in chapter 8 of our textbook. For this assignment, we will use the wastewater data as an example and practice our plotting and time series data science skills.

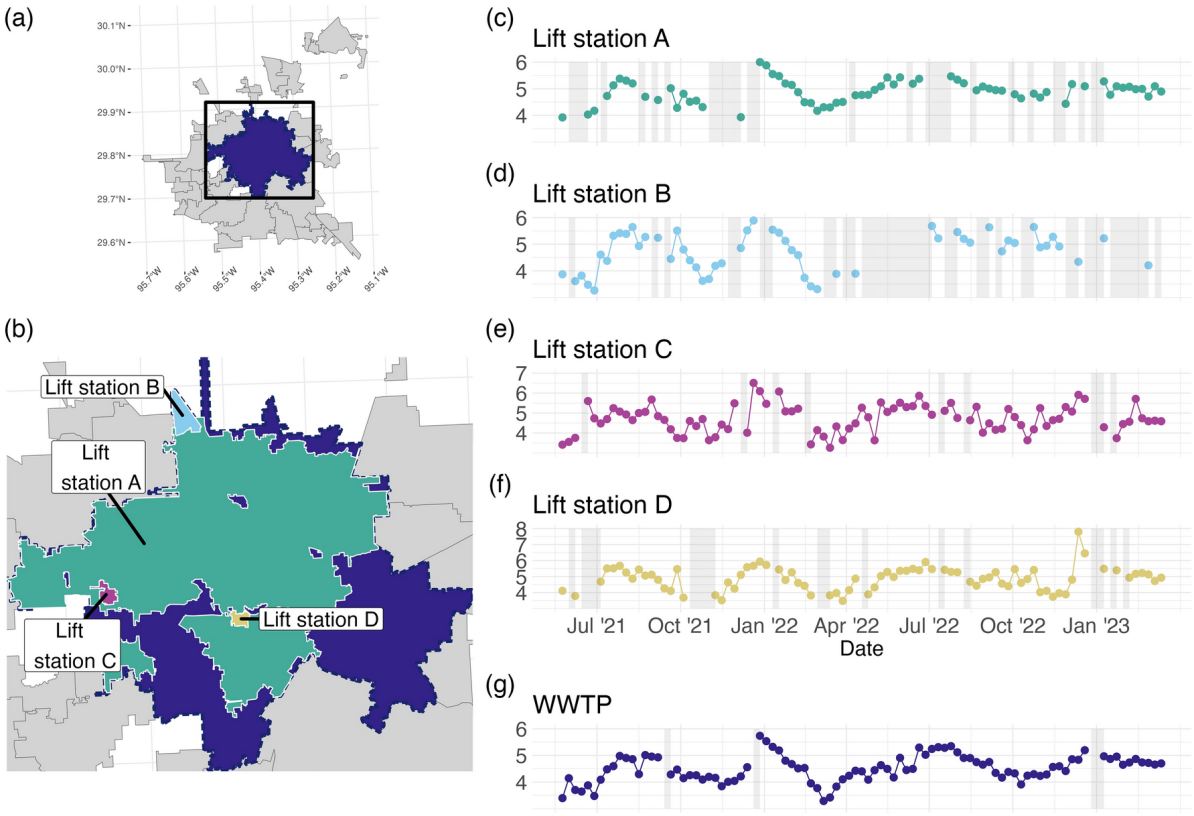

Figure 1: (a) The WWTP catchment areas for the City of Houston, with the WWTP of focus shaded. The box shows the extent of (b), the map showing the 4 lift stations considered in the analysis. (c–g) Plot the time series of Log10 Copies/L for the WWTP and the 4 lift station facilities, referred to as Lift Station A–D, with periods of missing values indicated by grey rectangles.

1. Which of the time series has the most missing data? Which appears to have the most variability? Does the overall behavior of the series seem to be similar?
2. Load the (synthetic) wastewater data from https://raw.githubusercontent.com/hou-wastewater-epi-org/online_trend_estimation/refs/heads/main/Data/synthetic_ww_time_series.csv using the `read.csv` function
3. Inspect the data. Verify that each of the series from the map above are included in the .csv (hint: what are the unique values of the `name` field?)
4. Convert the date field to a Date format using the function as.Date.
5. Install and load the `tidyverse` package.
6. We will work with just the WWTP series for now. Use `dplyr::filter` to extract the values for just the WWTP series.

```
ww <- read.csv(#your code here)

ww$dates <- as.Date(# your code here)

#install.packages("tidyverse")
library(tidyverse)
ww_WWTP <- ww %>% dplyr::filter(#your code here)
```

7. What is the time interval between the observations? How do you know?
8. Use the `tsplot` function from the `astsa` package to plot the `WWTP` series.

   Make sure to use the `dates` field for the x-axis and specify good axis and plot labels using the `xlab`/`ylab`, and `main` arguments. (see the documentation `?tsplot` for more)
9. Apply a moving average filter with 3 time points using the stats::filter function and save the result in a vector called `ww_ma_3`. You can choose the order of the moving average. (Similar to the final part of problem 1.1, see here in Lecture Notes).

```
ww_ma_3 <- stats::filter(#your code here)
```

10. Plot the moving average you computed on top of the tsplot in a different color using the lines function (see linked Problem 1.1 above). In the call to the lines function, also use `type = l` and `lwd = 2`.

```
tsplot(# your code here)
lines(# your code here)
```

11. Apply the moving average filter again, but this time use 5 time points, call it `ww_ma_5`. Plot just the `WWTP` series data and the `ww_ma_5` you just computed, and use a different color for this MA process than you used in question 10.
12. Inspect the plot you generated in questions 10 and 11. Which MA process looks "smoother"?
13. Describe the different way that the missing data in the WWTP series impacts the moving average estimates for the case of 3 time points vs. 5 time points.