El Paso Wasterwater Surveillance Time Series Analysis

May 09, 2022

Contents

Analysis Objective: Process, analyze the time series data from WW Surveillance collected from El Paso, Texas

1. Process the data from multiple excel sheets to a combined dataset (long format)
2. Exploratory analysis of time series figures with WW concentration
3. Exploratory analysis of time series figures with COVID reported case data
4. Cross-correlation between the case and WW concentration - see this link for reading https://online.stat.psu.edu/stat510/lesson/8/8.2

Data processing

Read in raw data and export an organized dataset. Time framework is between 6/22/2020 and 3/2/2022.

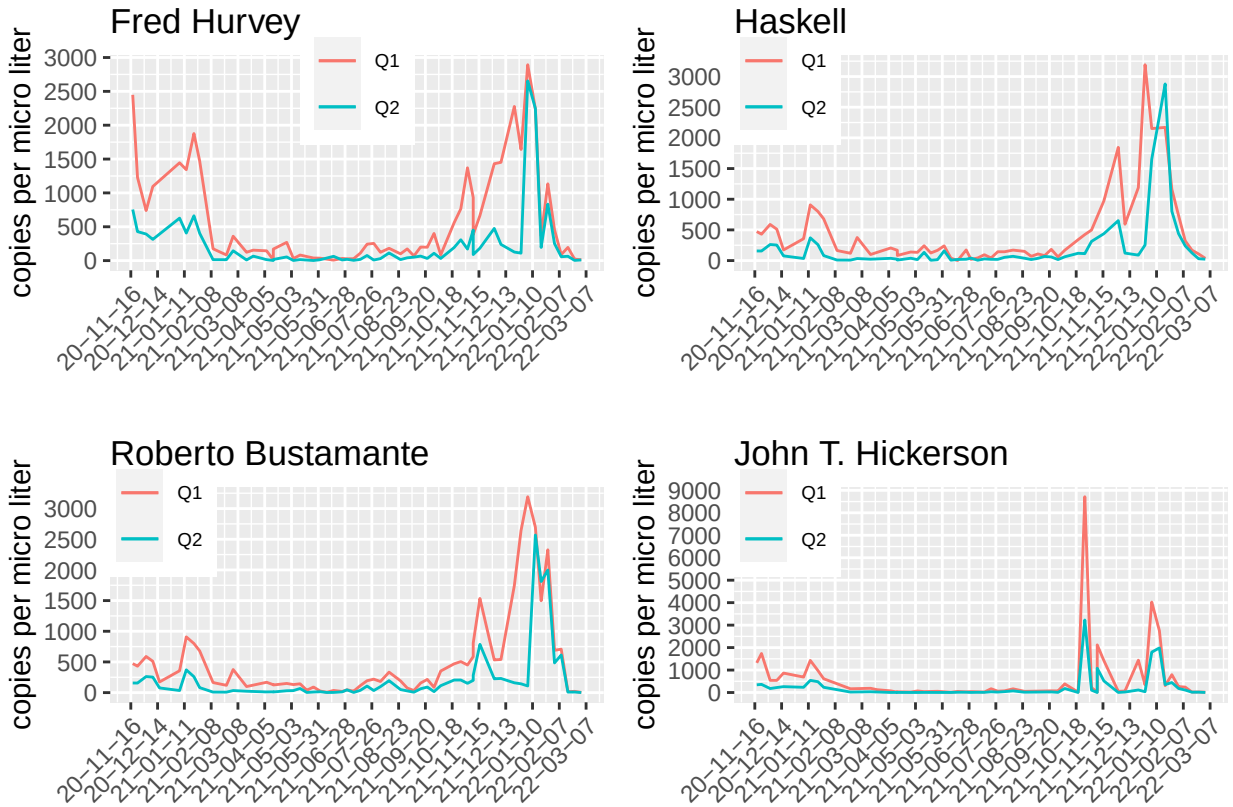
WWTP Site	Date Collected	Sample ID	Sample Name	N1 CT	Quantity for 10 ul volume into PCR (N1)	Copies per L waste water (N1)	N2 CT	Quantity for 10 ul volume into PCR (N2)	Copies per L waste water (N2)	AVG Flowrate (MGD)
Roberto Bustamante	2022-01-13	RB 22-01311	RB 01/13/22	29.3	2696.2	1797466.67	29.6	2568.6	1712400.00	26.52
Fred Hurvey	2020-12-02	NA	FH 12/02/20	31.0	739.1	492733.33	32.3	392.6	261733.33	6.05
Roberto Bustamante	2021-07-14	RB 21-20173	RB 7/14/21	33.5	114.4	76266.67	35.5	37.3	24866.67	26.52
Haskell	2020-07-15	HS 20-24431	HS 07/15/20	NA	NA	NA	NA	NA	NA	NA
John T. Hickerson	2020-06-23	JT 20-21360	JT 06/23/20	NA	NA	NA	NA	NA	NA	NA
Roberto Bustamante	2021-01-06	NA	HS 01/06/21	32.0	356.2	237466.67	35.6	34.2	22800.00	28.75
John T. Hickerson	2022-02-23	JT 22-04126	JT 02/23/22	35.5	25.1	16733.33	36.8	15.3	10200.00	7.07
Roberto Bustamante	2021-10-20	RB 21-31196	RB 10/20/21	31.6	468.8	312533.33	33.2	205.1	136733.33	26.52
Fred Hurvey	2022-01-19	FH 22-01939	FH 01/19/22	31.9	365.8	243866.67	33.2	194.5	129666.67	5.86
Roberto Bustamante	2021-03-17	NA	HS 03/17/21	33.7	99.1	66066.67	36.2	22.4	14933.33	28.75

- Sample extraction volume: 300ul; RNA elution: 100ul; PCR input volume: 10ul
- Conversion: $50ml$ –> $1ml$ then $300ul$ –> $100ul$, so the conversion factor is 150.
- copies per L waste is calculated from quantities for $10ul$ volume into PCR (PCR input volume): $((copies/10ulRT-PCR)*100,000)/150$
- Copies per MGD is calculated using copies per L waste water and average flow rate

- What does Ct mean? In a real time PCR assay a positive reaction is detected by accumulation of a fluorescent signal. The Ct (cycle threshold) is defined as the number of cycles required for the fluorescent signal to cross the threshold (ie exceeds background level). **Ct levels are inversely proportional to the amount of target nucleic acid in the sample (ie the lower the Ct level the greater the amount of target nucleic acid in the sample)**
  - Cts < 29 are strong positive reactions indicative of abundant target nucleic acid in the sample
  - Cts of 30-37 are positive reactions indicative of moderate amounts of target nucleic acid
  - Cts of 38-40 are weak reactions indicative of minimal amounts of target nucleic acid which could represent an infection state or environmental contamination.

Explorary analysis with the the WW concentration (Copies per micro liter waste water)

For the figures below I modified the following 1) legends shortened to "Q1" and "Q2" and 2) no need to plot all dates - I modified to print every 4 weeks in the label.



Time series analysis with the reported number of cases

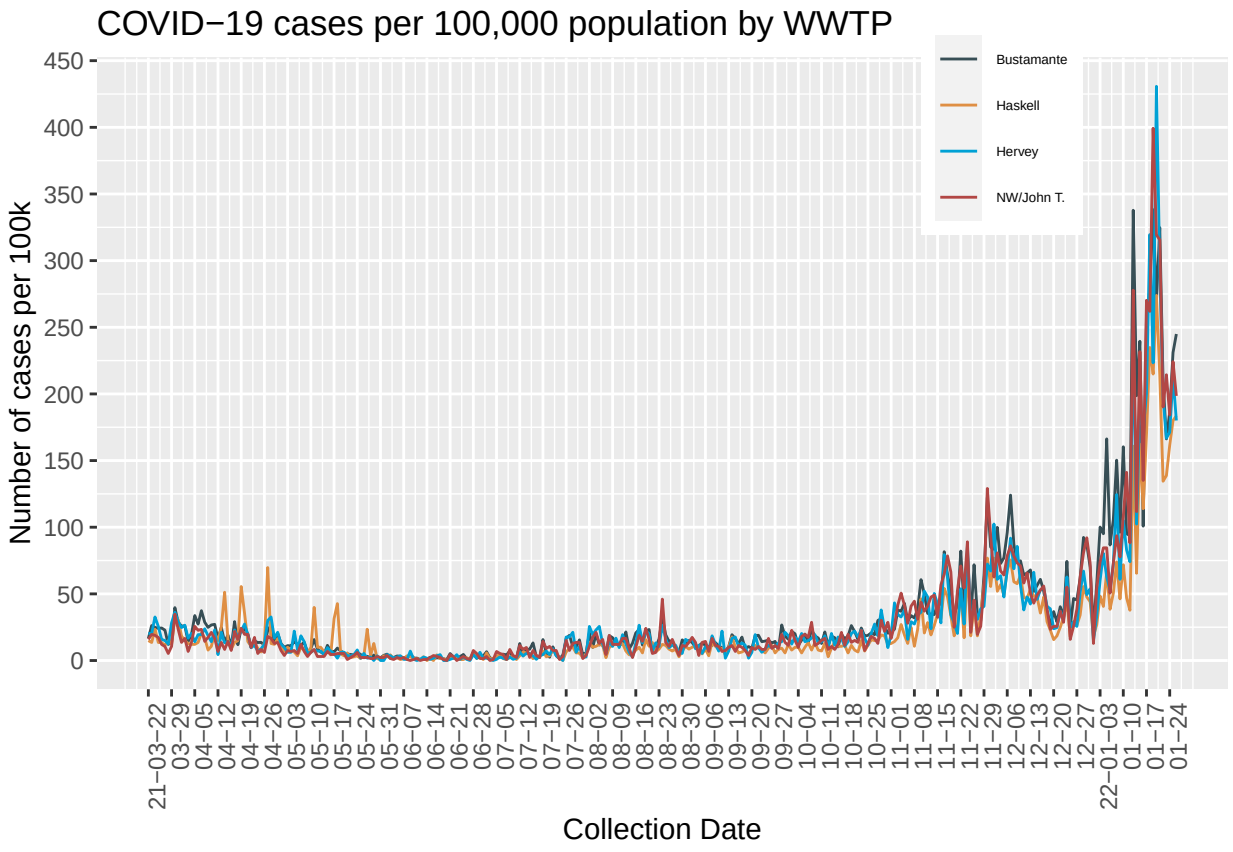
Processed COVID-19 case data (Time period: 3/22/2021 - 1/26/2022)

Date	WWTP	CaseRate
2021-07-13	Haskell	6.404737
2021-09-21	Haskell	11.386199

Date	WWTP	CaseRate
2021-05-18	Haskell	42.698244
2021-09-10	Hervey	7.061399
2021-07-26	Hervey	17.653497
2021-12-08	Haskell	59.065900
2021-04-13	Hervey	21.184200
2021-09-26	Hervey	9.709423
2021-12-23	Haskell	32.735320
2021-04-04	Bustamante	19.386650

For the figures, can we over lay the WW time series and the case data (perhaps case data y-axis on the right), to see how the trends correlate? Looks like we need to do this for each facility (i.e., Fred Hurvey, Haskell etc)

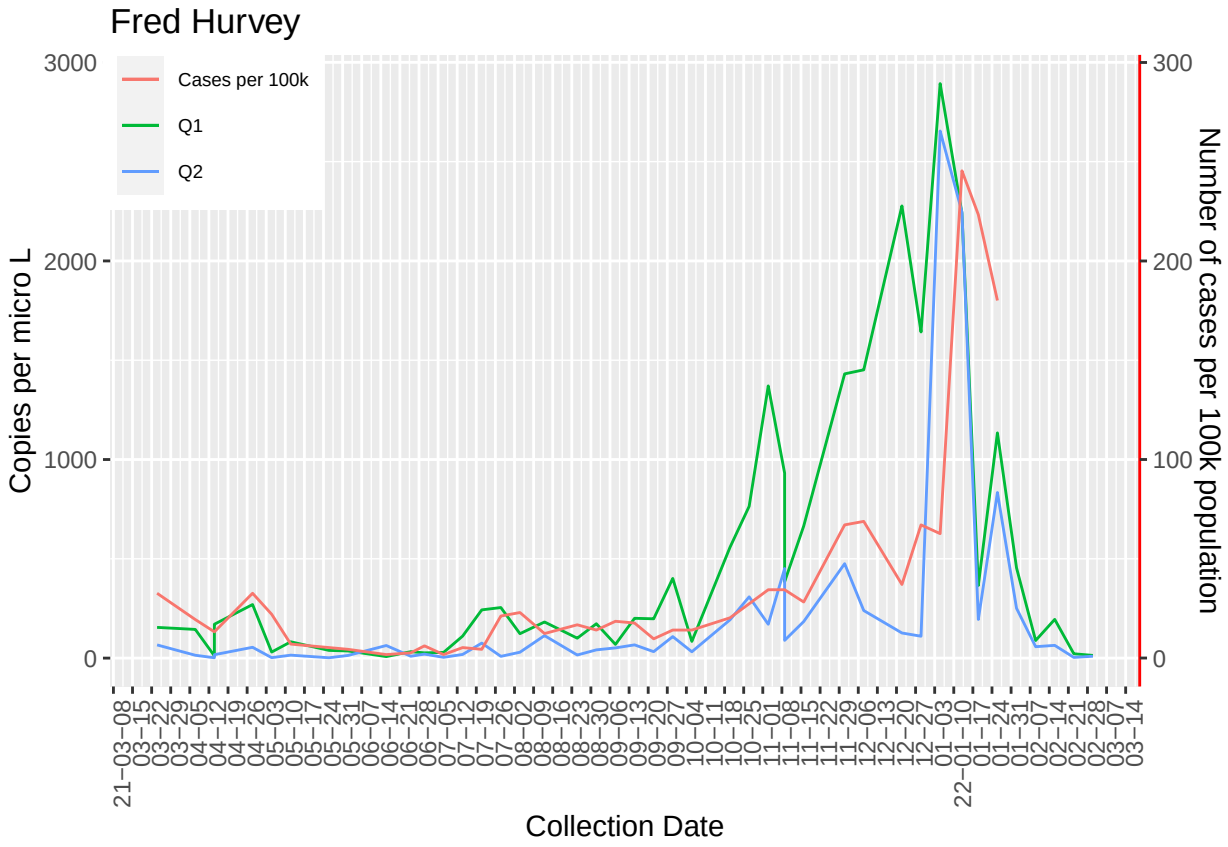
COVID-19 infection rate (per 100k population) by four waste water treatment facilities



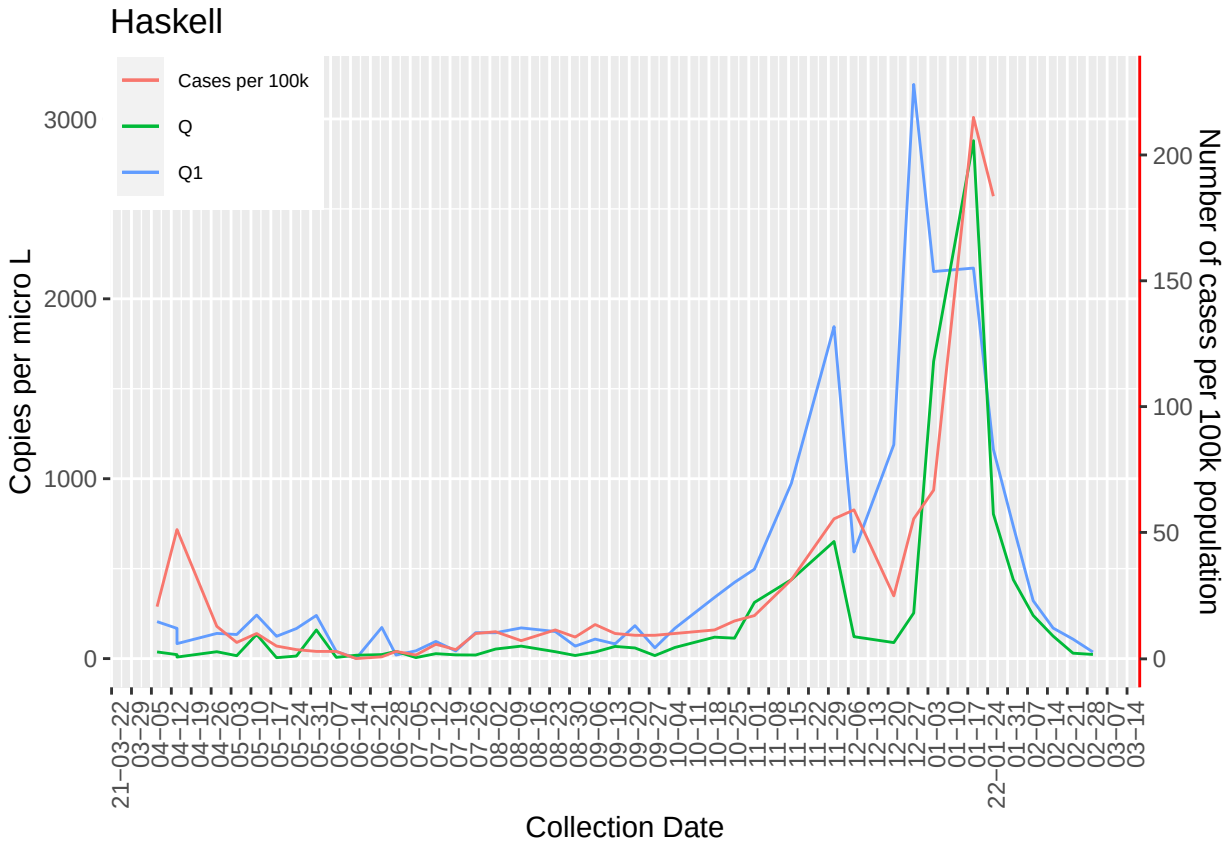
Overlay cases with WW time series plot

Time overlap: 4/7/2021 - 1/26/2022

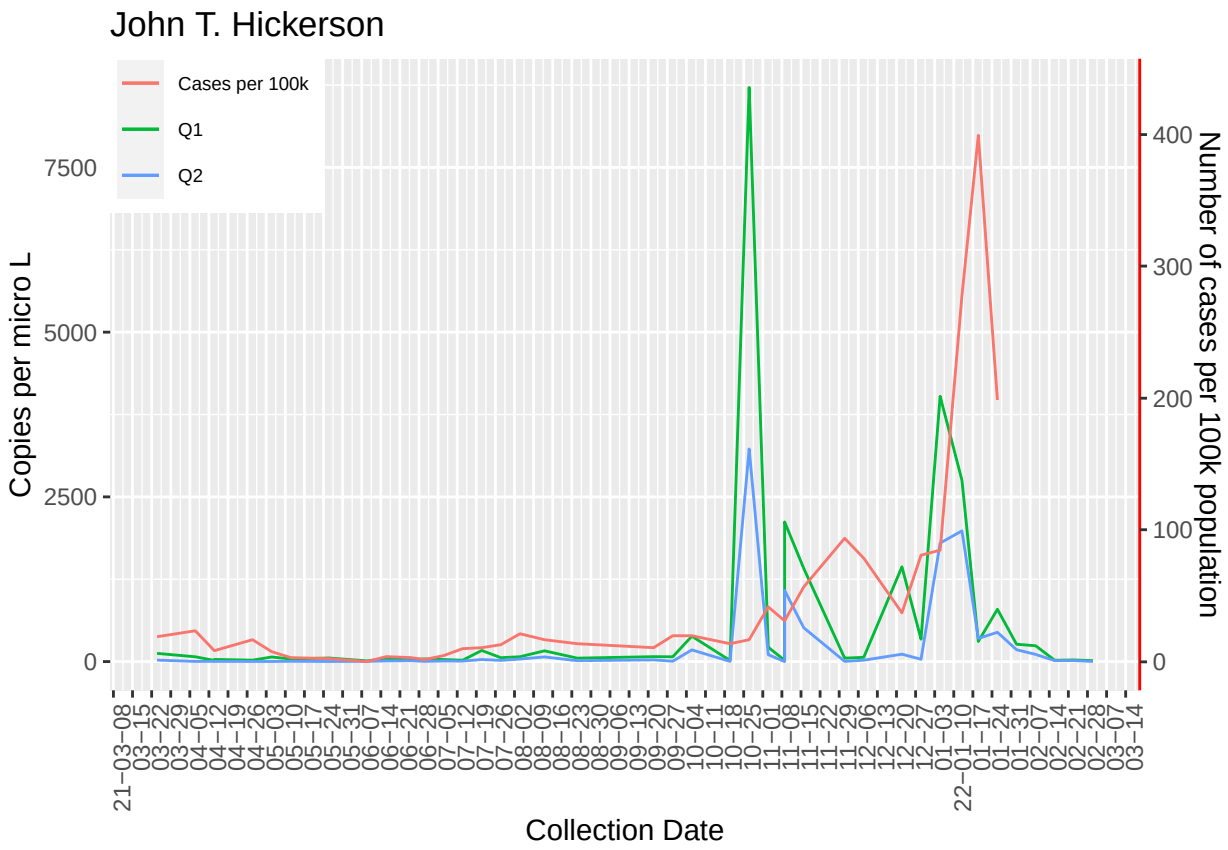
Fred Hurvey


Fred Hurvey
Cases per 100k
Q1
Q2
Copies per micro L
Number of cases per 100k population
Collection Date
3000
2000
1000
0
300
200
100
0
21–03–08
03–15
03–22
03–29
04–05
04–12
04–19
04–26
05–03
05–10
05–17
05–24
05–31
06–07
06–14
06–21
06–28
07–05
07–12
07–19
07–26
08–02
08–09
08–16
08–23
08–30
09–06
09–13
09–20
09–27
10–04
10–11
10–18
10–25
11–01
11–08
11–15
11–22
11–29
12–06
12–13
12–20
12–27
01–03
22–01–10
01–17
01–24
01–31
02–07
02–14
02–21
02–28
03–07
03–14

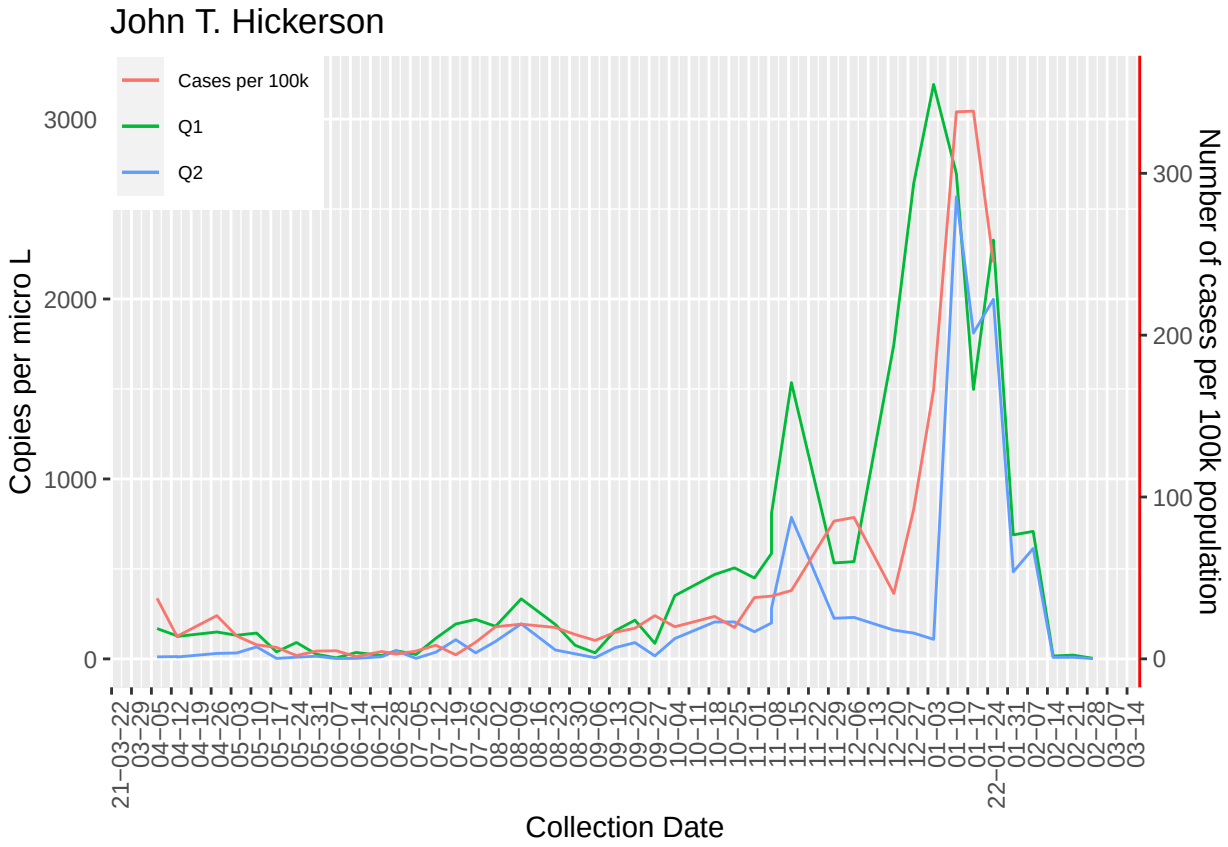
Haskell


Haskell
Cases per 100k
Q
Q1
Copies per micro L
Number of cases per 100k population
Collection Date

John T.


John T. Hickerson
Cases per 100k
Q1
Q2
Copies per micro L
Number of cases per 100k population
Collection Date
7500
5000
2500
0
400
300
200
100
0
21–03–08
03–15
03–22
03–29
04–05
04–12
04–19
04–26
05–03
05–10
05–17
05–24
05–31
06–07
06–14
06–21
06–28
07–05
07–12
07–19
07–26
08–02
08–09
08–16
08–23
08–30
09–06
09–13
09–20
09–27
10–04
10–11
10–18
10–25
11–01
11–08
11–15
11–22
11–29
12–06
12–13
12–20
12–27
22–01–03
01–10
01–17
01–24
01–31
02–07
02–14
02–21
02–28
03–07
03–14

Roberto Bustamante



Cross-correlation analysis between reported cases and WW concentration

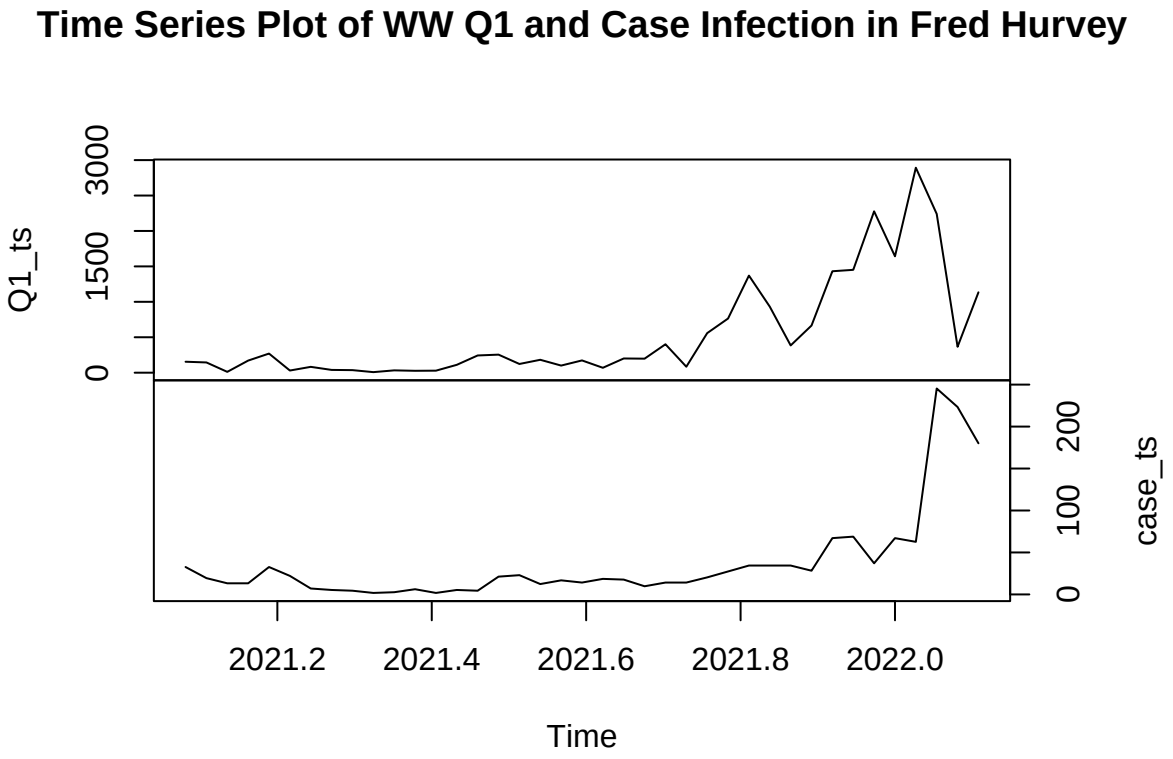
Cross-correlation may help us identify the leading time when virus in WW started to show a pikes before the cases were increasing.

The CCF is defined as the set of correlations (height of the vertical line segments in figures below between two time series $x_t + h$ (waste water concentration) and y_t (case rate) for lags $h = 0, \pm 1, \pm 2, \ldots$. A negative value for h represents a correlation between the x-series at a time before t and the y-series at time t.
The CCF helps to identify lags of x_t that could be predictors of the y_t series.

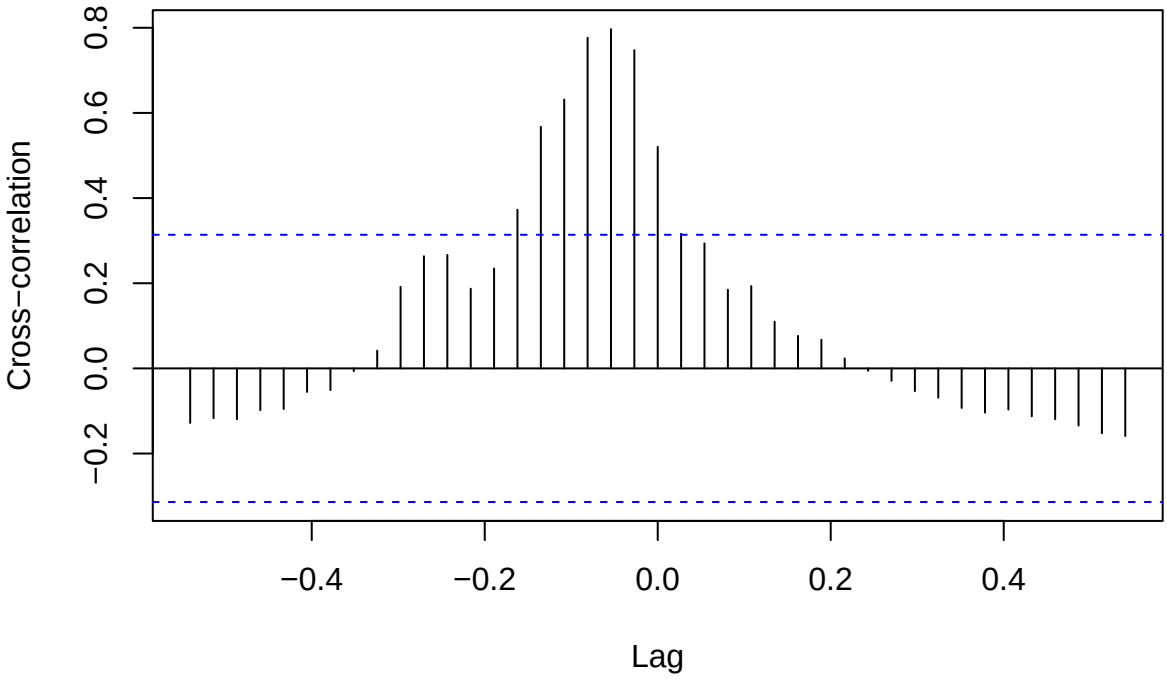
1. When $h < 0$, x leads y.

2. When $h > 0$, x lags y.

Fred Hurvey



Cross-correlogram plot

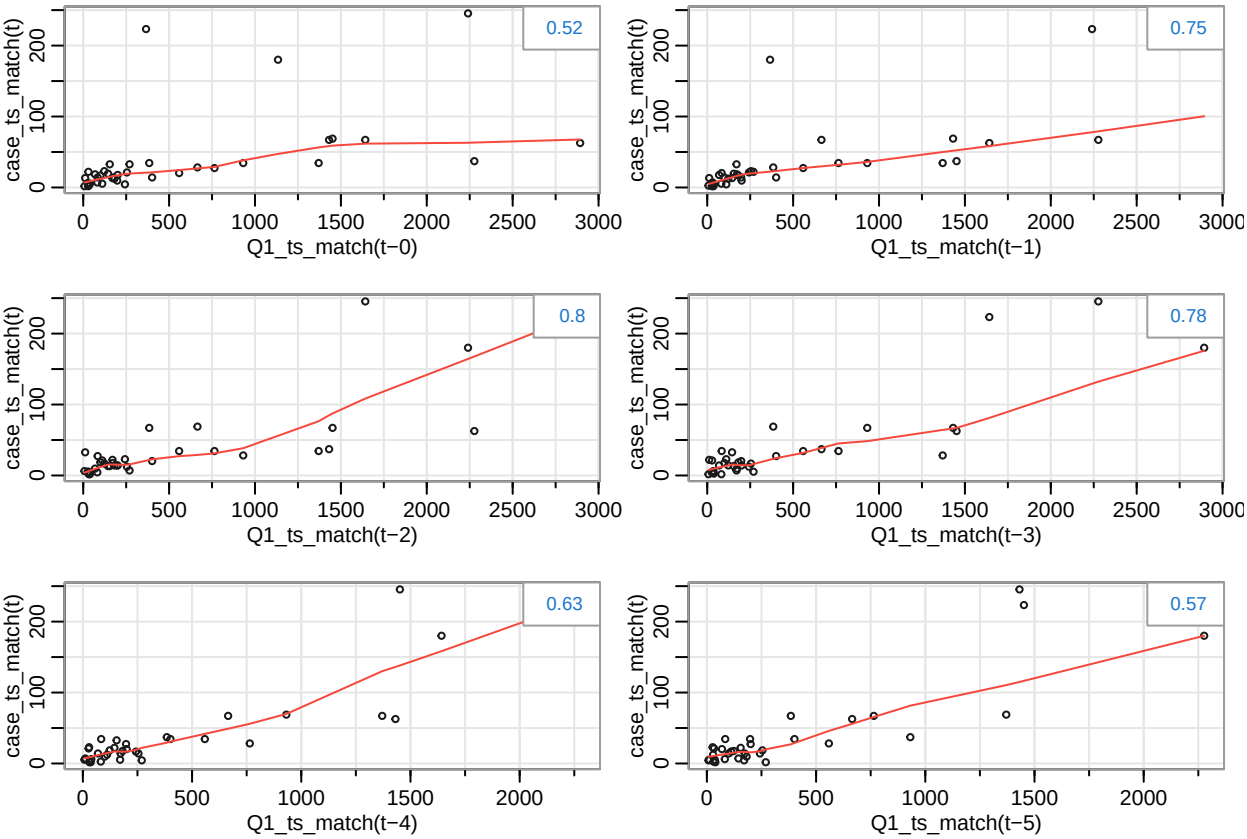



```
## [1] -2
```

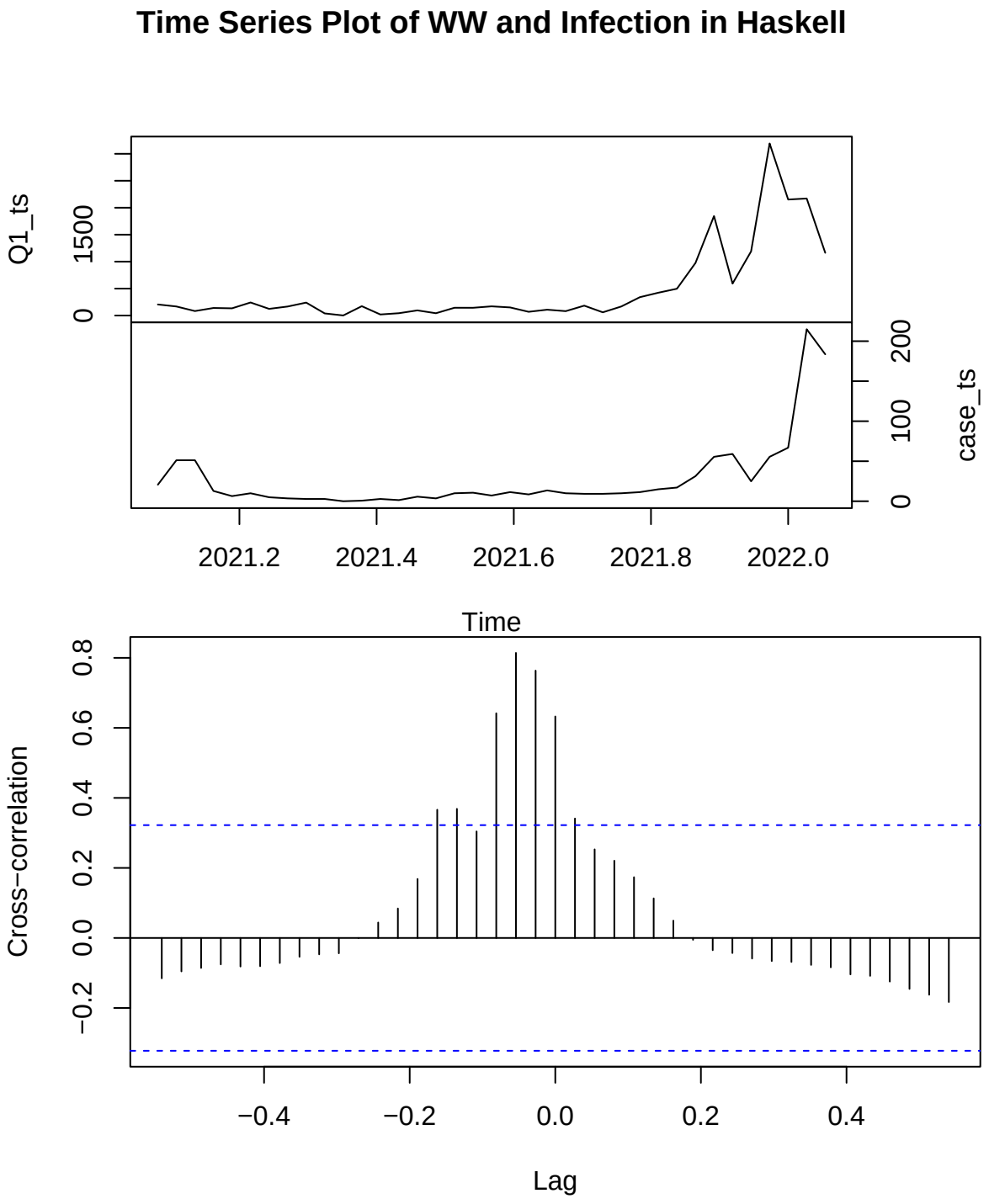
Here one lag is 37 days, so the maximum correlation is at -2 days - Infection rates are relatively high around 2 days after high WW concentration

Scatterplots of y_t versus x_{t-h} for negative lags from 0 back to a lag -5. In each plot, (WW Q1 variable) is on the vertical and a past lag of infection rate is on the horizontal. Correlation values are given on each plot.

- Correlation on the top right is the cross-correlation between WW Q1 at time t and Infection rate at time t-h (h=0,1...5)
- Red line in the plot is the lowess fit.

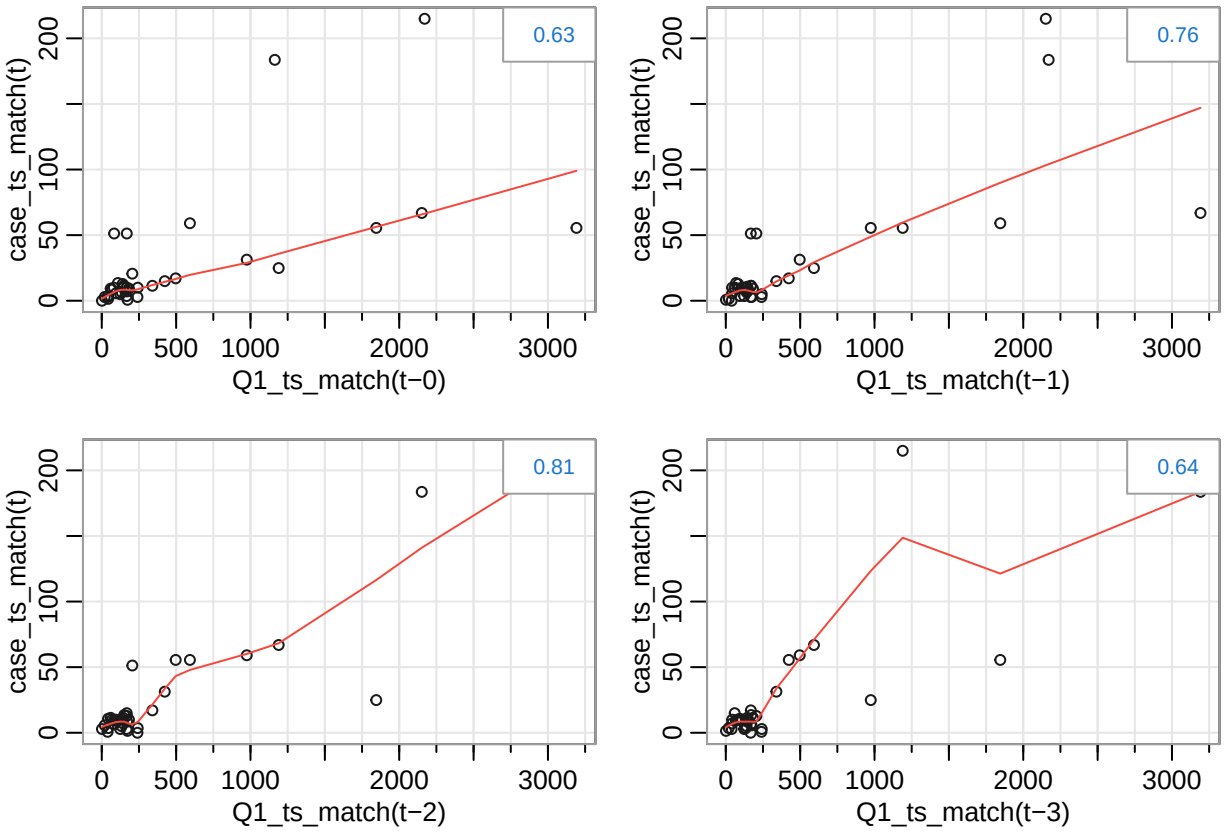


Haskell

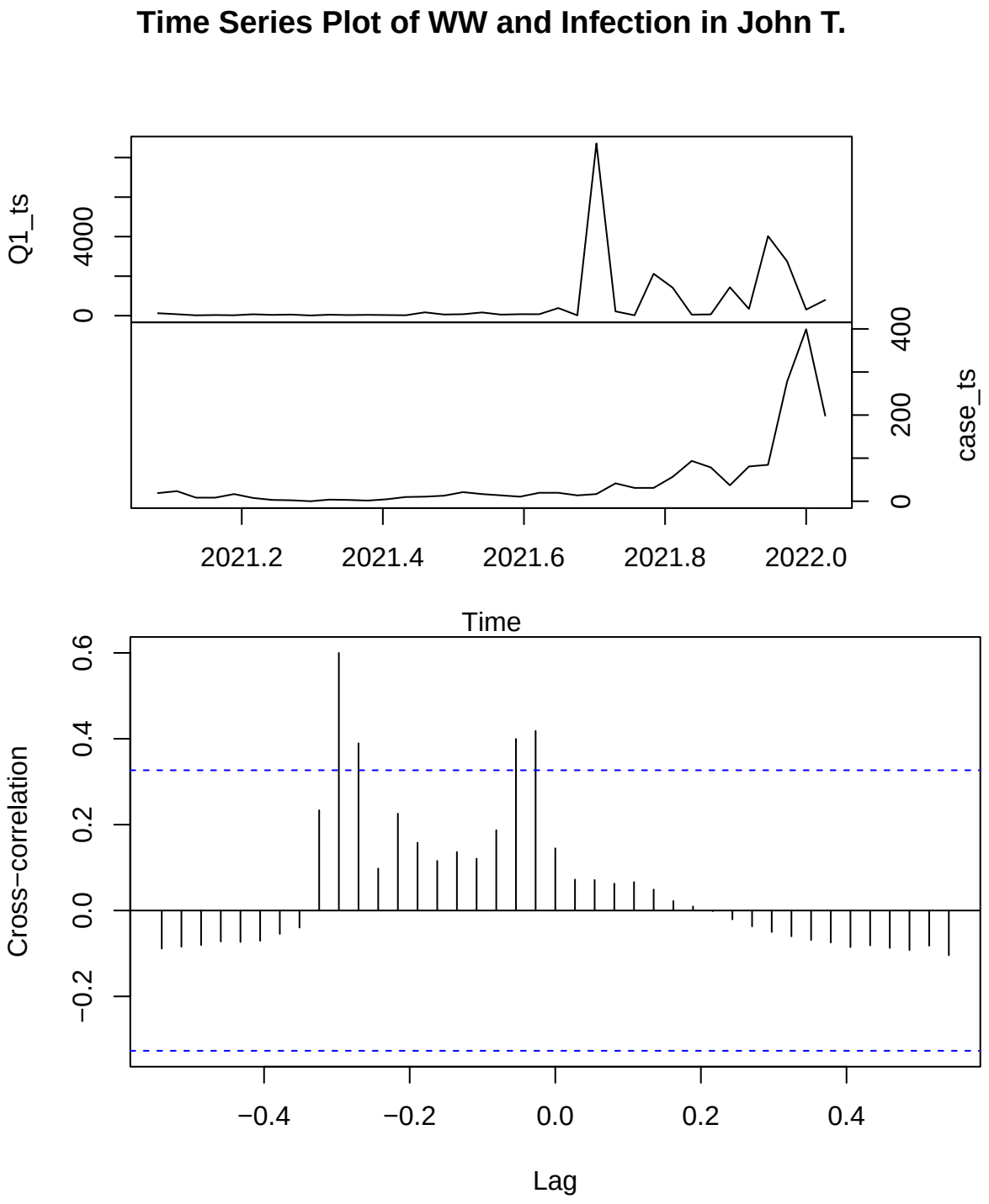



```
## [1] -2
```

Infection rates are relatively high around 2 days after high WW concentration

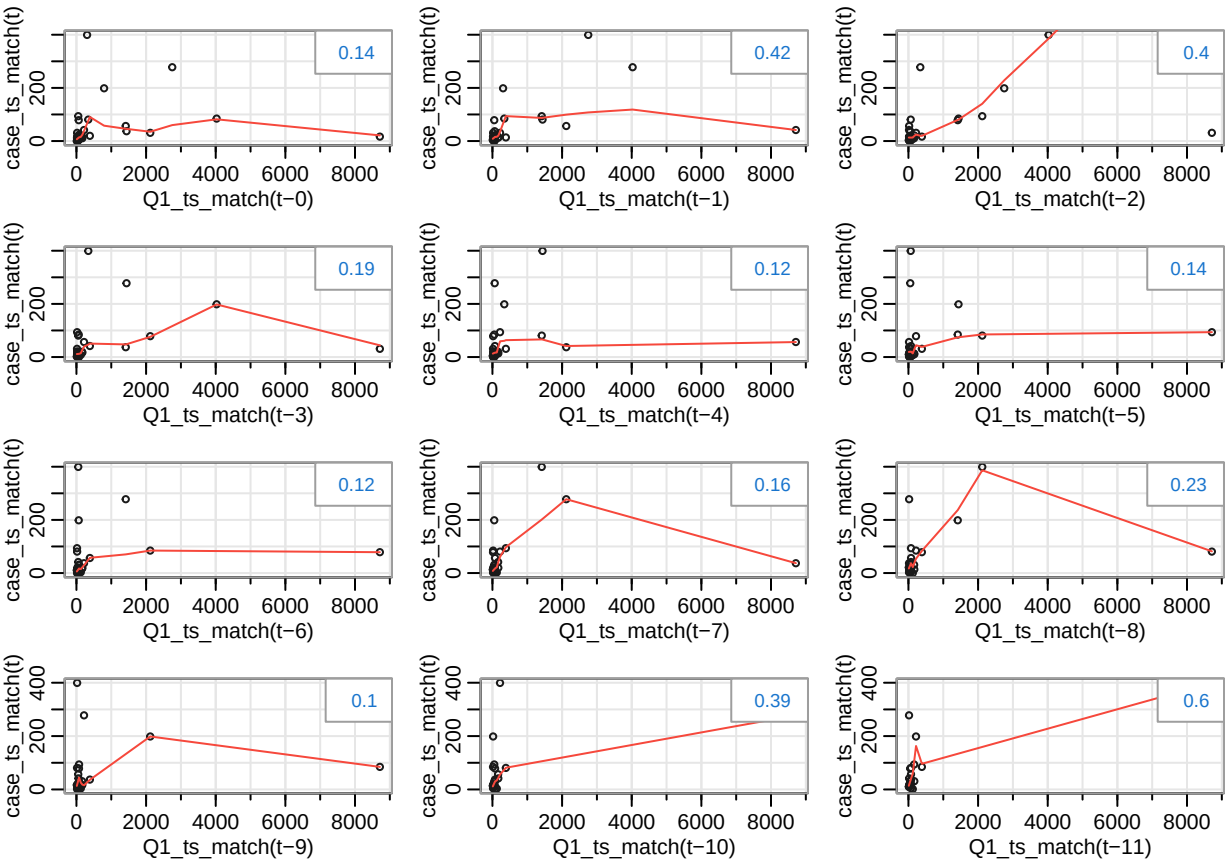

0.63
case_ts_match(t)
Q1_ts_match(t−0)
0.76
Q1_ts_match(t−1)
0.81
Q1_ts_match(t−2)
0.64
Q1_ts_match(t−3)
0
50
100
200
500
1000
2000
3000

John T

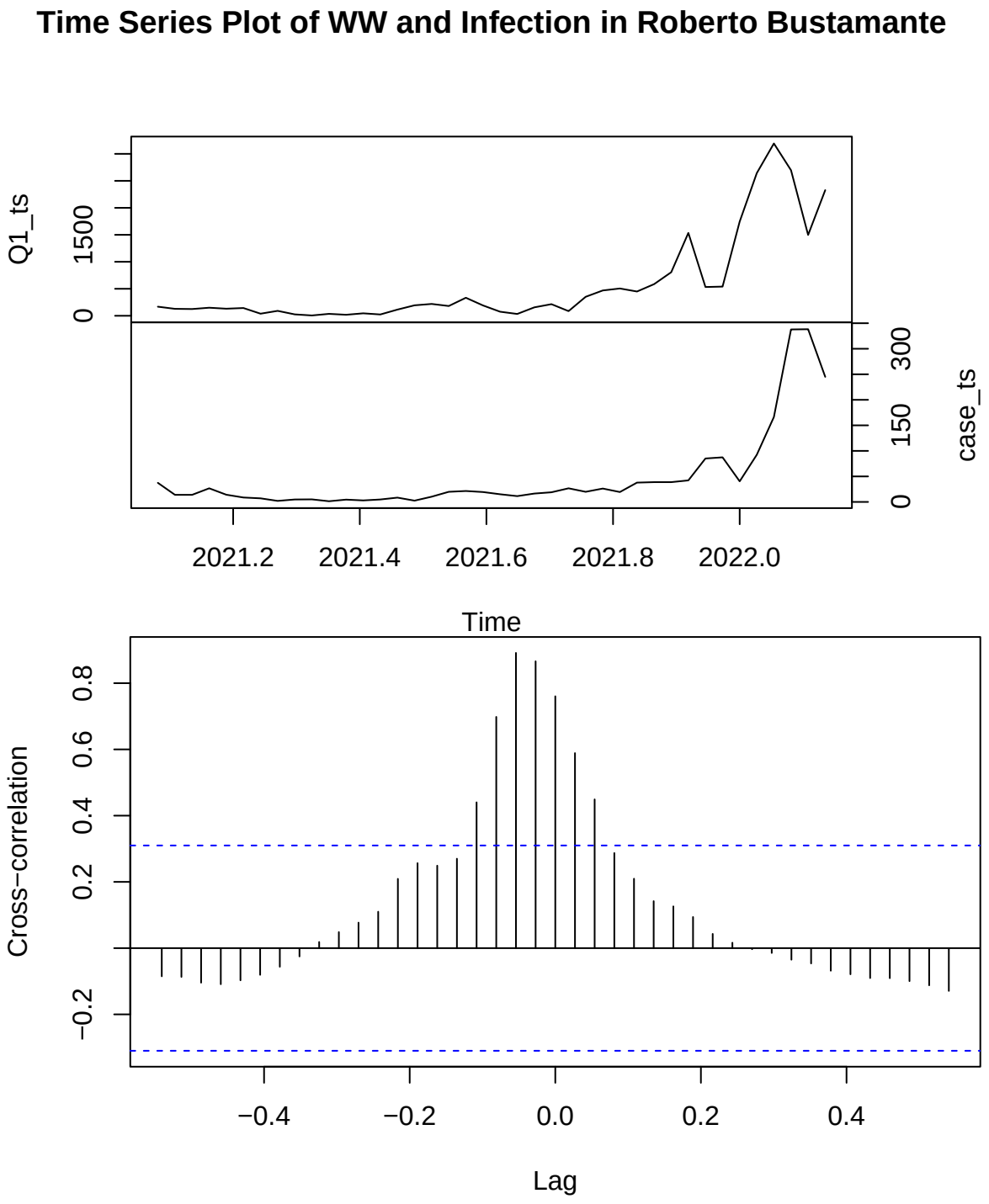



```
## [1] -11
```

Infection rates are relatively high around 11 days after high WW concentration

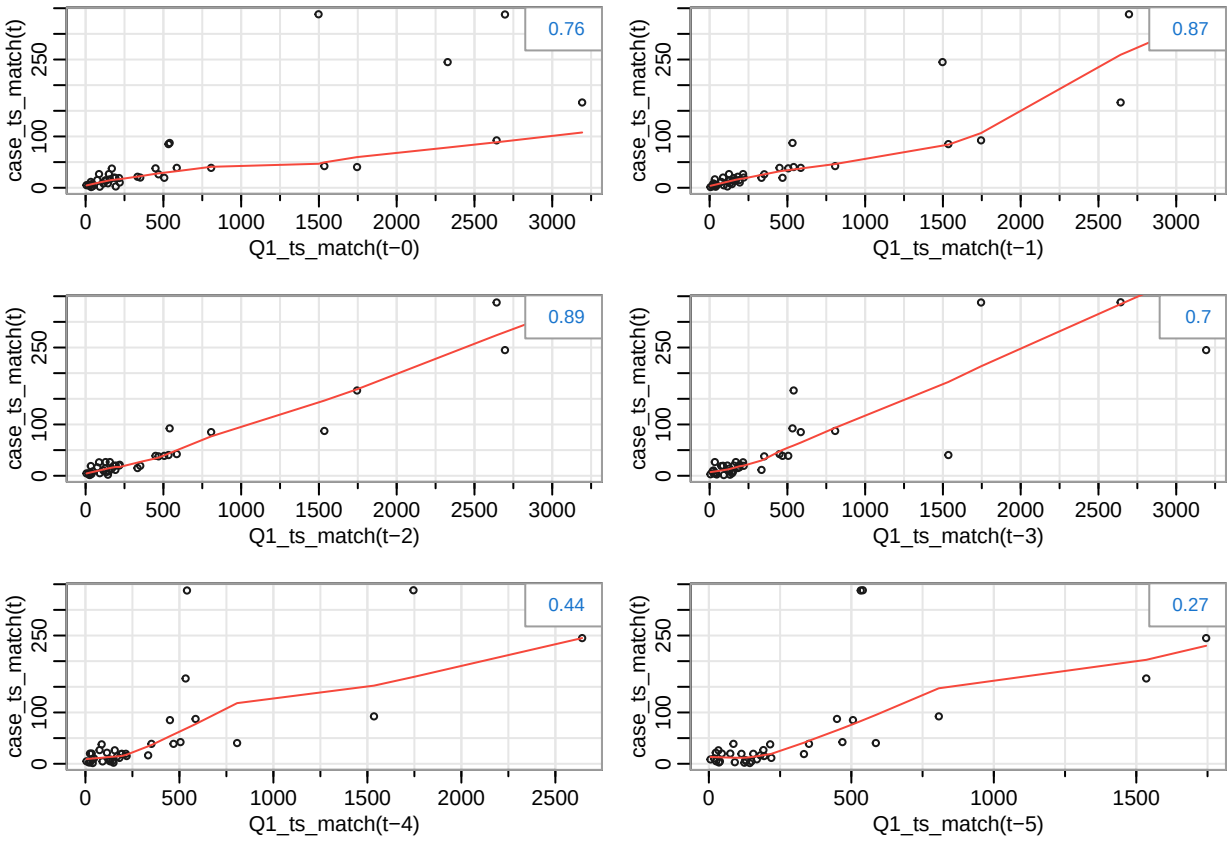

case_ts_match(t)
Q1_ts_match(t−0)
0.14
Q1_ts_match(t−1)
0.42
Q1_ts_match(t−2)
0.4
Q1_ts_match(t−3)
0.19
Q1_ts_match(t−4)
0.12
Q1_ts_match(t−5)
0.14
Q1_ts_match(t−6)
0.12
Q1_ts_match(t−7)
0.16
Q1_ts_match(t−8)
0.23
Q1_ts_match(t−9)
0.1
Q1_ts_match(t−10)
0.39
Q1_ts_match(t−11)
0.6

Roberto Bustamante




```
## [1] -2
```

Infection rates are relatively high around 2 days after high WW concentration


0.76
case_ts_match(t)
0
100
250
0
500
1000
1500
2000
2500
3000
Q1_ts_match(t−0)
0.87
case_ts_match(t)
0
100
250
0
500
1000
1500
2000
2500
3000
Q1_ts_match(t−1)
0.89
case_ts_match(t)
0
100
250
0
500
1000
1500
2000
2500
3000
Q1_ts_match(t−2)
0.7
case_ts_match(t)
0
100
250
0
500
1000
1500
2000
2500
3000
Q1_ts_match(t−3)
0.44
case_ts_match(t)
0
100
250
0
500
1000
1500
2000
2500
Q1_ts_match(t−4)
0.27
case_ts_match(t)
0
100
250
0
500
1000
1500
Q1_ts_match(t−5)