Biostatistical Consulting
Wastewater-based Epidemiology of COVID-19 in Athens, GA, USA 2020-2021

Cody Dailey[1] & Megan Lott[2]

Erin Lipp [2] & Stephen Rathbun [1]

03 May 2021

[1] Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, USA

[2] Department of Environmental Health, University of Georgia, Athens, GA, USA

Preface

This document will serve as a report among collaborators working on the analysis of wastewater surveillance data as part of a biostatistical consulting course (BIOS8200) instructed by Dr. Stephen Rathbun. The wastewater surveillance project was designed and implemented by Dr. Erin Lipp and her environmental health doctoral students, Megan Lott and William Norfolk, and other members of the Lipp lab. Megan Lott acts as the primary "client" for correspondence with lead "consultant" Cody Dailey, under the supervision of Dr. Stephen Rathbun and with meaningful discussion among other course consultants: Nicholas Mallis, Amanda Skarlupka, Morgan Taylor, and Adrianna Westbrook.

The report is structured to highlight the analytic workflow from raw data management through analysis techniques. Particular attention is given to analytic options and decision-making to expand on reproducibility. This report does *not* mimic the structure found in scientific manuscripts. Rather and analogously, it will be more detailed and comprehensive *methods* and *results* sections with commentary and coding descriptions.

Common Abbreviations

SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2 (a virus)
COVID-19 = coronavirus disease 2019 (a disease)
RT-qPCR = reverse transcription quantitative polymerase chain reaction
LOD = limit of detection
LOQ = limit of quantification

Table of Contents

1 Background

1.1 SARS-CoV-2 and COVID-19 Pandemic

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the etiological agent of coronavirus disease 2019 (COVID-19). This virus stormed the modern world with a global pandemic that has imposed a gobsmacking burden on society. Direct impacts of deaths, disease, and disability have been met with indirect effects in societal upheaval and economic stress.

Governments, public health professionals, and academics alike were neither expecting of nor prepared for a global emergency of this magnitude. Consequently, there was an obvious dearth of public health and medical capacity. Important to our context, a valid and reliable surveillance system and methodology was missing. We initially had no sound approaches to track the extent of the pandemic, that is, how many and who were infected.

Even as diagnostic tests were developed and anthropocentric surveillance of cases yielded situation reports, there remained obscurity in the underlying pandemic processes. At best, infections are only partially observed (and reported) and potential biases from diagnostic accuracy and human behavior loom over our understanding of outbreaks as they happen. Retrospective serological studies can address some of the lapses in surveillance but do little to address concerns during an ongoing outbreak.

Dr. Erin Lipp and peers devised ways to monitor the pandemic from another perspective.

1.2 Wastewater-based Epidemiological Surveillance

When infected with pathogenic organisms, people will often and unknowingly release or shed the microbes into their environment. This is fundamental to understanding the transmission of communicable infectious disease (e.g., aerosol spread from a cough or sneeze). However, pathogen shedding is not limited to the more dominant / obvious modes of transmission.

Fecal shedding is a well-known and relatively common characteristic of viral infections and serves as the base rationale for Dr. Lipp's research into SARS-CoV-2 and COVID-19 surveillance in wastewater. **Compared to traditional case reporting systems, the detection and quantification of SARS-CoV-2 ribonucleic acid (RNA) in wastewater offers a more timely and comprehensive assessment of the extent of an outbreak.** Ultimately in this analysis, we aim to investigate the veracity of this statement.

1.3 Objectives

With this analysis, we hope to characterize the capabilities of wastewater-based epidemiological surveillance for SARS-CoV-2. In doing so, we outline the following objectives:

1. Explore viral load quantification methodologies
2. Address sources of bias in COVID-19 case reports

3. Develop a predictive framework using wastewater sampling to inform COVID-19 surveillance

1.4 Document Structure / Flow

The rest of this document will be organized in the following way (with terrible puns throughout). First in Section 2, I give a brief overview of the study design with respect to the collection of wastewater samples and the subsequent analysis. In Section 3, I describe the data to be used in the analysis. Section 4 outlines the sampling frequencies with resulting observations, quantifies the extent of data missingness, or rather null results, from the wastewater samples' analyses, and explores limits of detection and quantification. Section 5 covers the calibrations of reaction efficiencies, conversions from assay results to viral loads, and missing data replacements using estimated limits. Section 6 switches focus to COVID-19 incidence and estimation of a deconvoluted incidence curve from a time series of case counts by dates of report. Finally, Section 7 explores the associations between wastewater assays and COVID-19 case counts and works towards a prediction framework for using wastewater-based epidemiology to inform surveillance. Then, Section 8 discusses the entirety of the project, results, implications, and future directions.

All analyses were carried out using R (version 4.0.4 (2021-02-15), Lost Library Book;)[1] and RStudio (Windows 10 Desktop Version 1.4.1106, Tiger Daylily;).[2] This document was prepared using the `bookdown`[3] and `rmarkdown`[4] packages within the RStudio IDE.

2 From Feces to Species

2.1 Overview of Wastewater Sampling, Processing, and Analysis with Resulting Data

As part of the wastewater surveillance, three Athens-Clarke County (ACC) water reclamation facilities yield water samples. The sampling frequency varies across the study. Initially, samples were taken weekly, but later the sampling frequency was twice weekly (on Mondays and Wednesdays). However, there remained some irregularity in the sampling frequency (further discussed in Section 4).

Overall, the *three* wastewater reclamation facilities yield approximately *three* 24-hour composite water samples for a single sampling time. The three biological replicates each serve as a source for RT-qPCR experiments. The experiments use amplification primers for *two* viral sequence targets, N1 and N2, and for each primer *three* replicate experiments are conducted. So, for any given sampling time, there should be approximately 3x3x2x3=54 RT-qPCR results.

Initially, two water samples for each facility were taken at each sampling time, but the study design shifted to collecting three samples for most of the study duration; these redundant samples are referred to as biological replicates.

Each wastewater sample was processed and analyzed via a reverse transcription quantitative polymerase chain reaction (RT-qPCR) -based laboratory workflow required for the enumeration of SARS-CoV-2 in the wastewater samples. Briefly, viral genetic material (i.e., ribonucleic acid (RNA)) was extracted from the wastewater samples and used as a template to generate complementary deoxyribonucleic acid (DNA) which was then serially amplified to determine the concentration of viral genetic material from the original sample. This RT-qPCR workflow is done both in duplicate for two separate genetic sequence targets/primers (N1 and N2) and, for each sequence target, in triplicate (at least) for additional redundancies; these triplicate experiments are referred to as technical replicates.

More comprehensive documentation on data and methodology is found at the University of Georgia (UGA) Center for the Ecology of Infectious Disease (CEID) COVID-19 Portal: Wastewater Surveillance for SARS-CoV-2 in Athens, GA and at the Lipp Laboratory Protocol for Wastewater Surveillance of SARS-CoV-2.

3 Raw (sewage) Data

This section outlines the datasets used in these analyses.

3.1 RT-qPCR Generated Data

Megan Lott conducts the RT-qPCR procedures and aggregates raw data output using Microsoft Excel workbooks. These data contain unique identifiers for biological and technical replicates, dates of water sampling and RT-qPCR runs, and the cycle thresholds (ct; i.e., the number of PCR cycles that were needed to detect the fluorescent markers incorporated into the amplified DNA).

Data from RT-qPCR runs using sequence primers for N1 and N2 were imported into R as follows:

```
n1 <- read_csv("./data/raw_data/n1_all_cleaned2.csv")
n2 <- read_csv("./data/raw_data/n2_all_cleaned2.csv")
```

3.2 Reaction Calibration Data

In addition to the RT-qPCR runs of the water samples with unknown viral loads, positive controls are run as a means by which to calibrate the conversion equations for the RT-qPCRs. That is, water samples are "spiked" with known concentrations of viral sequence targets (N1 or N2) or a related sequence (orthologous or paralogous?) such as Bovine Coronavirus genes (not included in the current analyses). These data are needed to estimate reaction efficiency and in the calculations that convert RT-qPCR output Ct values to estimations of viral load (e.g., number of viral genetic copies per volume of water sample)

The following two lines of code were used to read in the quality-control data required for the standard curves:

```
qc <- read_csv("./data/raw_data/QC/all_curves.csv")
qc2 <- read_xlsx("./data/raw_data/QC/sarscov2_rna_control.xlsx")
```

3.3 Wastewater Reclamation Facilities

The wastewater reclamation facilities that serve as sources for samples provide data on total influent water / sewage volume (i.e., how much water flows through the facility) and total suspended solids. These characteristics may impact both the water sample and the downstream analysis.

The following line of code was used to import the facility data required in the conversion of viral load concentrations to an absolute count of viral copies in the wastewater:

```
plant <- read_csv("./data/raw_data/plant_data.csv")
```

3.4 COVID-19 Surveillance Reports

The purpose of these efforts to quantify the viral load in wastewater is to complement the underlying SARS-CoV-2 infections or COVID-19 epidemic curve. As such, we will also incorporate the COVID-19 case reports from the Georgia Department of Public Health (GaDPH). The data are available as two datasets of COVID-19 case reports. These two datasets are distinguished subtly based on how any particular case report is tied to a specific date. The first dataset contains simple COVID-19 case frequencies for the dates at which the cases were reported. The second dataset contains COVID-19 case frequencies corresponding to the date of symptom onset; however, if a date of symptom onset is not available for a record, then the date is replaced with the date of sample specimen collection.

Although the symptom onset dataset is imperfect, it is likely a useful addition to the analysis at hand. Viral shedding in feces occurs post symptom onset, but whether it occurs before symptom onset is unclear; however, some have noted similar viral loads and shedding in SARS-CoV-2 infections, regardless of symptomaticity. Still, it may be necessary to explore relations in the data with respect to the timing of cases in the population and viral loads in wastewater.

In addition, we have data on diagnostic test frequencies and positivity that may shed additional light onto the hidden processes. These data are also given for two date schema, report and specimen collection dates.

The following lines of code read in these data, subset to the records for Athens-Clarke County, and extract the variables pertinent to our analyses:

```
covid <- read_csv("./consult/01-data/ga_covid_data/epicurve_symptom_date.csv"
) %>%
            filter(county=="Clarke") %>%
            select(symptom.date=`symptom date`,
                   cases, moving_avg_cases)

covid.report <- read_csv("./consult/01-data/ga_covid_data/epicurve_rpt_date.c
sv") %>%
            filter(county=="Clarke") %>%
            select(report_date,
                   cases,
                   moving_avg_cases)

covid.testing <- read_csv("./consult/01-data/ga_covid_data/pcr_positives_col.
csv") %>%
            filter(county=="Clarke") %>%
            select(collection_date = collection_dt,
                   pcr_tests = `ALL PCR tests performed`,
                   pcr_pos = `All PCR positive tests`)
```

Appendix 10 contains a more detailed description of the raw data used in the analyses, the contents of the data, and, briefly, their management.

4 Plunging into Data Interrogation

In this section, I describe the sampling frequencies employed and the resulting number of records from wastewater processing, explore the extent and patterns of *missingness* within the wastewater samples data, and approach estimations of detection and quantification limits for the RT-qPCR analyses.

4.1 Sampling Frequencies

As mentioned in the study design overview, the sampling frequency varied across the study period. Figure 4.1 shows the sampling frequency as the number of days since the previous sample. Most of the observations in the dataset are weekly; however, there are many from the more recent Monday-Wednesday sampling design. There are some irregularities in the sampling frequency, such as in mid-November and mid-January, where there are observations sampled a single day apart.

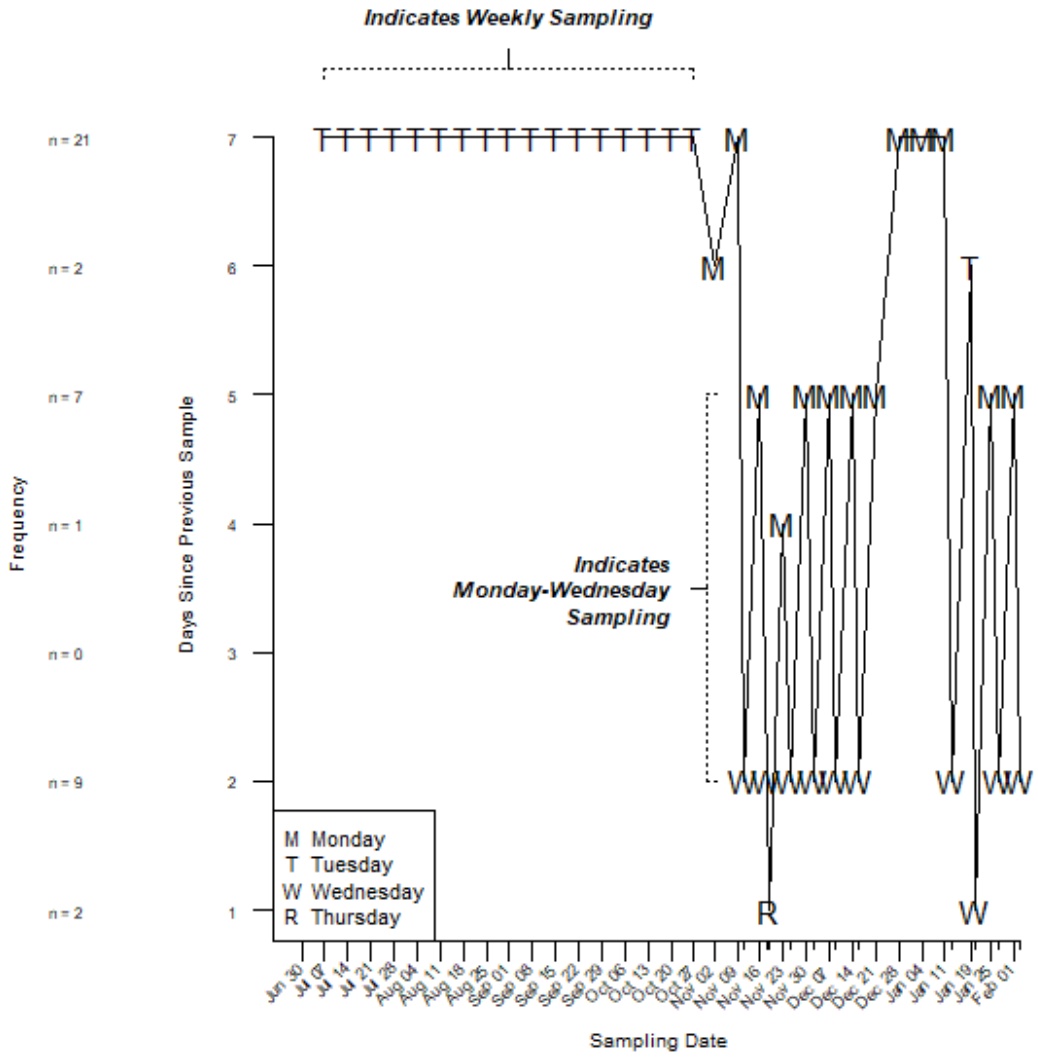


Figure 4.1: Sampling Frequency of Wastewater: Number of Days Between Samples

Figure 4.2 shows the number of 24-hour composite wastewater samples taken from each reclamation facility across the study period. Despite a few irregularities, the data show a transition from 2 samples taken to 3 samples taken in early October. There is a single date, 14 July 2020, where no samples were taken from the Cedar Creek facility.

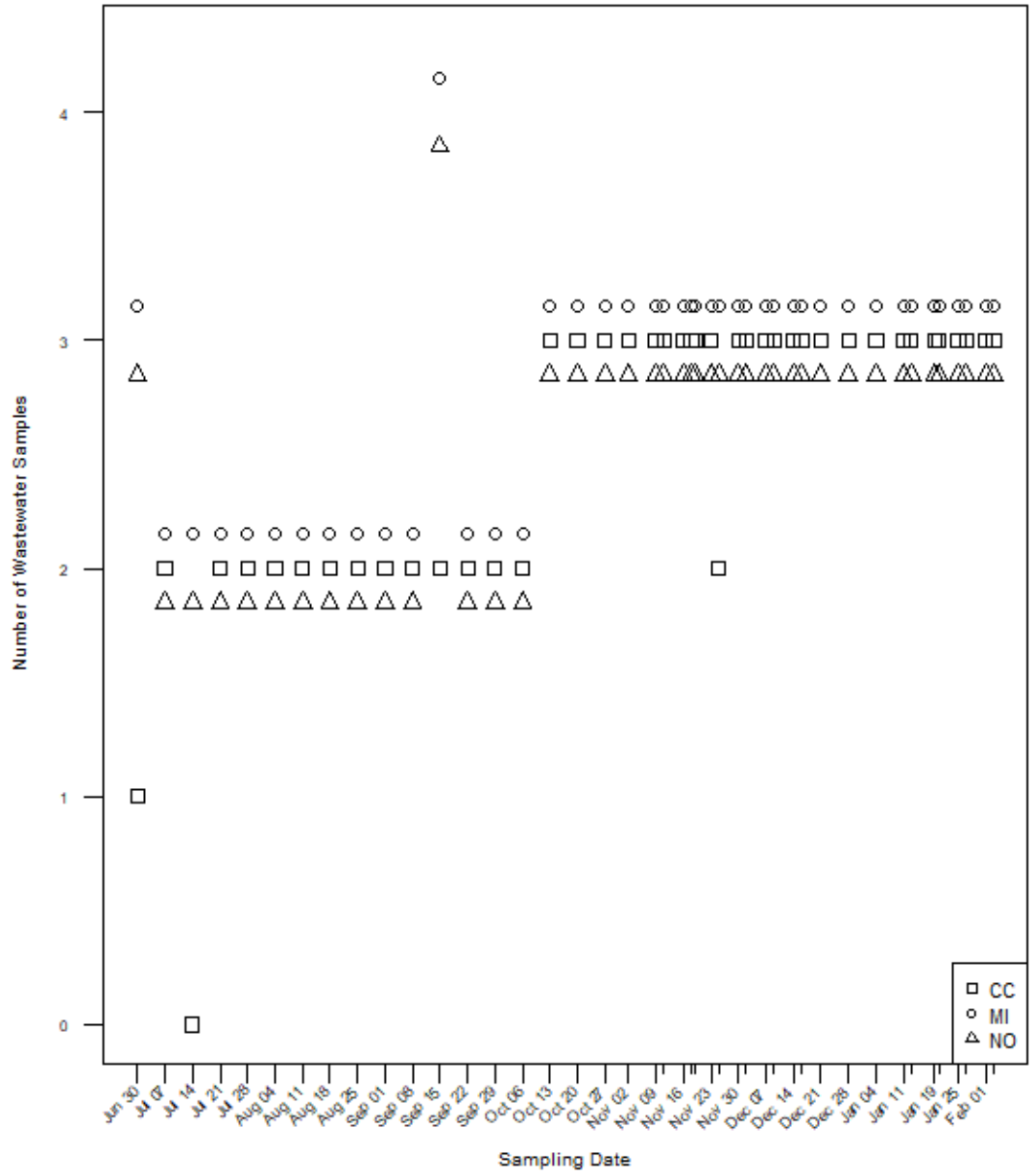


Figure 4.2: Number of 24-hour Composite Wastewater Samples from Reclamation Facilities across the Study

4.2 "Missingness" Evaluation

The RT-qPCR workflow can sometimes produce false negative results. That is, despite having a known concentration of some genetic material in a given sample, it is possible that the RT-qPCR technique can fail to detect any genetic material in the sample. A false negative may occur when the initial concentration of the genetic material in the sample is too small and, despite serial amplifications, this small concentration is never detected. This

scenario refers to the concept of a limit of detection; below this limit, the RT-qPCR technique is insensitive and unable to detect the presence of substrate. As such, it is possible for a given sample to return a null result when processed. From a data perspective, we may consider this as *missing* data, however, from a statistical perspective, a more accurate description would be aligned with the concepts of censoring or truncation. It is important to consider that, for those observations with null or "missing" results, all we know is that the viral load falls somewhere between zero and the limit of detection (given no spurious results). Therefore, it is essential to investigate the extent of this *missingness* within the wastewater samples' PCR results.

Due to the hierarchical structure in the data, we can explore the incomplete data records at various levels of replication.

Figures 4.3, 4.4, and 4.5 show the missing data profiles of RT-qPCR results for all experimental replicates, stratified by viral sequence target, and stratified by wastewater reclamation facility, respectively. Note that the tick marks on the x-axis represent unique sampling times of wastewater.

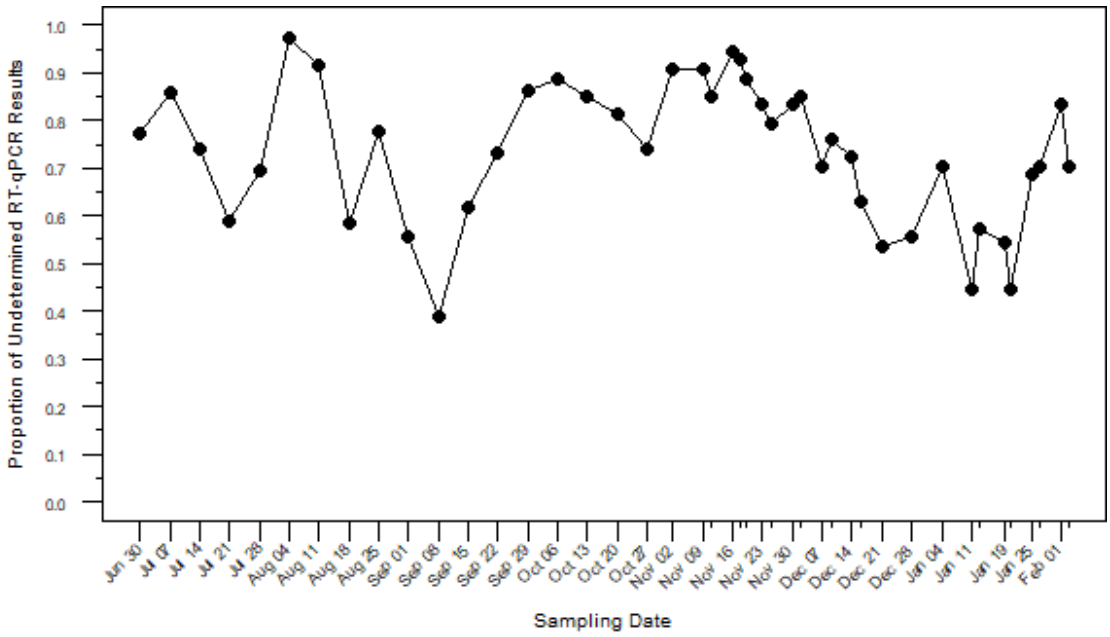


Figure 4.3: Overall Profile of Undetermined RT-qPCR Results

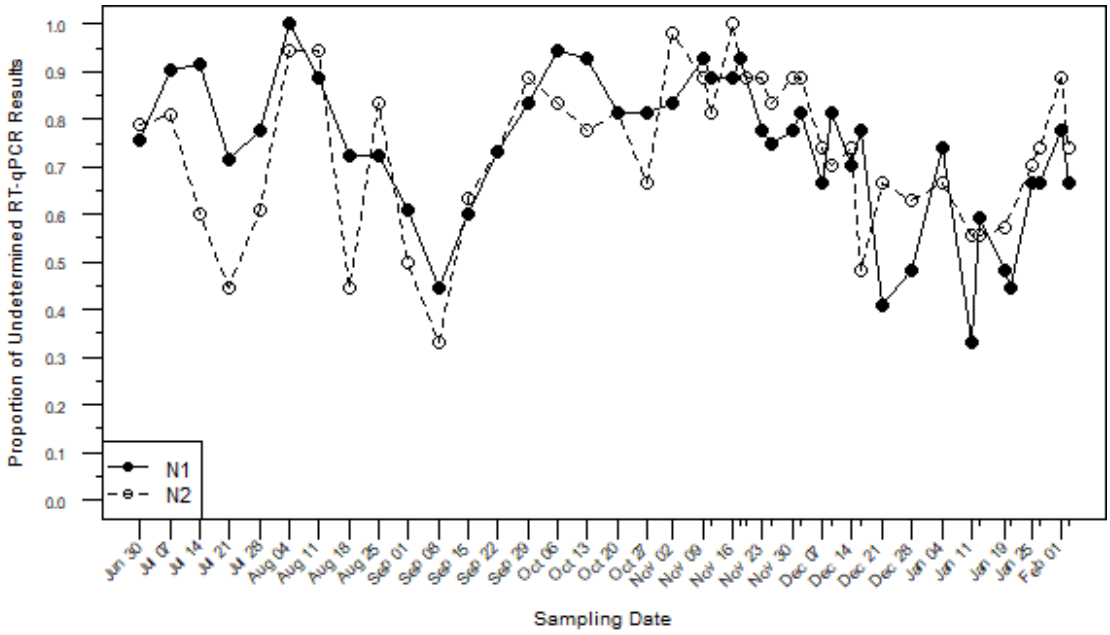


Figure 4.4: Profile of Undetermined RT-qPCR Results by Viral Sequence Target

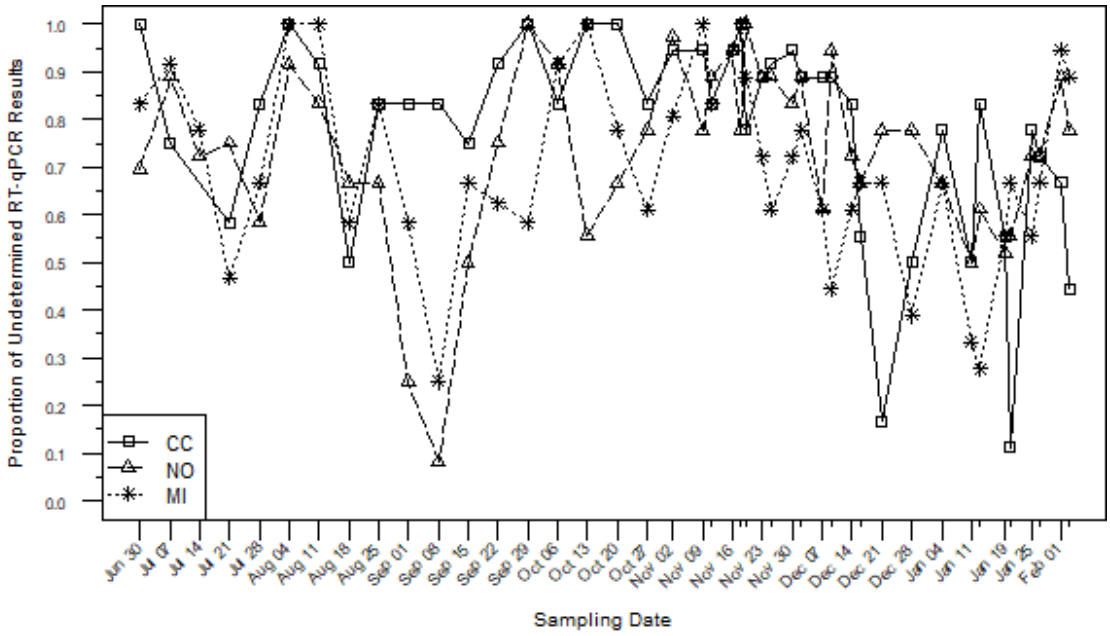


Figure 4.5: Profile of Undetermined RT-qPCR Results by Wastewater Reclamation Facility

Missing data are quite extensive. Of the total 2124 records, 1565 (73.7%) had undetermined values for the cycle threshold, i.e., effectively missing or censored. On the other hand, Figure 4.6 shows the total number of samples that RT-qPCR detected as positive. There are an average of 13 positive samples for each day across the study; over 80% of the sampling dates have more than 6 positive results.[3]

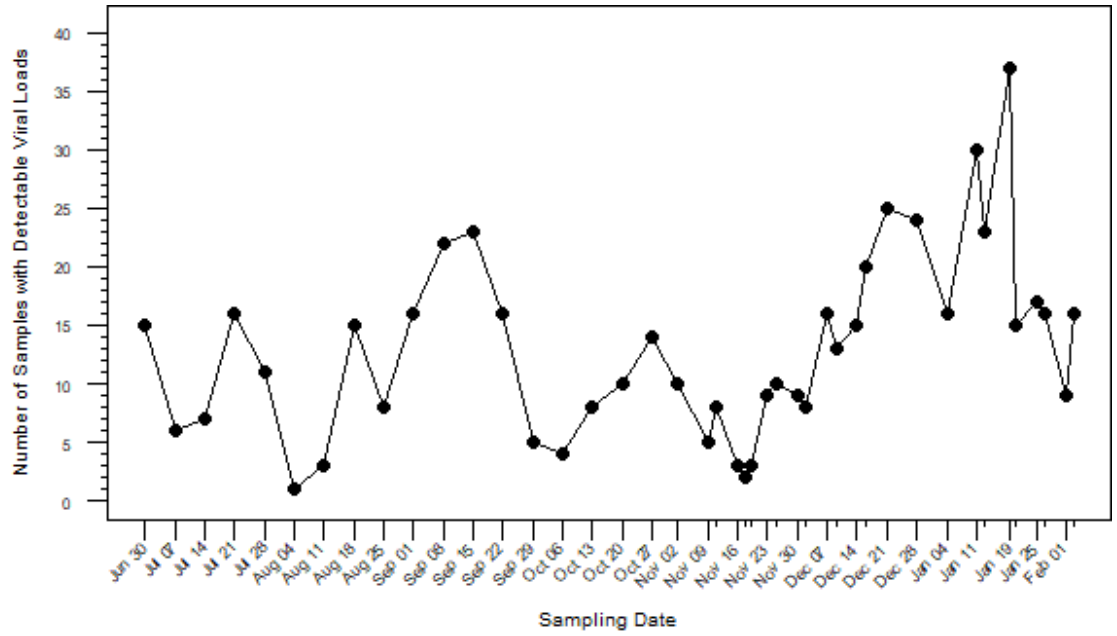


Figure 4.6: Overall Profile of Positive Samples from RT-qPCR

Given the hierarchical structure of the data, it is useful to explore the undetermined values at different levels. Table 4.1 and Figure 4.7 show the percentages / proportions of undetermined values stratified by wastewater reclamation facility and viral sequence target; these data are aggregated across the study duration. The three columns of percentages (with frequencies in parentheses) are given to show the three levels of organization within the data.

[3] The number of experiments ran / technical replicates was not strictly uniform across the study.

Table 4.1: Extent of Undetermined RT-qPCR Values at Various Levels of the Data Hierarchy

Wastewater Reclamation Facility	Viral Sequence Target	Percentage of Sampling Dates in which All Results were Undetermined	Percentage of Biological Replicates in which All Results were Undetermined	Percentage of Technical Replicates with Undetermined Results
CC	N1	22 (9 / 41)	52.3 (56 / 107)	77 (254 / 330)
	N2	30 (12 / 40)	52.9 (55 / 104)	78.5 (259 / 330)
MI	N1	14.3 (6 / 42)	31.9 (36 / 113)	67.5 (243 / 360)
	N2	22 (9 / 41)	41.4 (46 / 111)	72.2 (260 / 360)
NO	N1	16.7 (7 / 42)	50 (57 / 114)	75.5 (281 / 372)
	N2	14.6 (6 / 41)	41.4 (46 / 111)	72 (268 / 372)

The first column of percentages labeled "Percentage of Sampling Dates in which All Results were Undetermined" in Table 4.1 shows the frequencies of undetermined RT-qPCR values with respect to the date of sampling. To explain further, of the 43 distinct sampling dates in the data[4], there are some sampling dates where ***every*** RT-qPCR experiment for that day's samples yielded undetermined results. For example, the samples from Cedar Creek on 30 June 2020 all yielded undetermined results for all biologic and technical replicates for both viral sequence targets, N1 and N2, in the RT-qPCR experiments.

The next column, labeled "Percentage of Biological Replicates in which All Results were Undetermined" goes to the next smaller level in the data hierarchy: the biological replicates. There are a total of 110, 117, and 117 biological replicates from the Cedar Creek, Middle Oconee, and North Oconee Wastewater Reclamation Facilities, respectively.[5] So, the

[4] There are some implicitly missing data within the datasets with respect to sampling dates, facilities, and sequence targets. That is, there are some instances where no RT-qPCR results exist in the data for a given sampling date, facility, and viral sequence target. For example, there are no results Cedar Creek on 14 July 2020 for either viral sequence target, N1 or N2.

[5] Upon closer examination to the data, I may have uncovered a potential data entry error. For the Cedar Creek samples, there are uncrossed biological replicates with respect to viral sequence targets. For example, there are 3 technical replicates for each of 3 biological replicates for Cedar Creek on 18 November 2020; however, these records only exist for the N1 sequence target. Similarly, there are 9 records for Cedar Creek on 19 November 2020, but these are only for the N2 sequence target. I suspect these experimental units are from the same biological samples collected on the same day and that the 18 v 19 date is a simple entry error. There may be other entry errors, though. Closer inspection of the x-axes for Figures 4.3, 4.4, and 4.5 may shed some light on the matter.

percentages in this column refer to the frequencies in which *all* the technical replicate experiments among the biological replicates yielded undetermined results.

The last column in Table 4.1 shows the percentage of all technical replicates (i.e., the smallest experimental unit) that yielded an undetermined result from the RT-qPCR. Unlike the other columns, this one is a simple frequency calculation of the raw data. In fact, if you sum the denominators of each row in the last column, then you will get 2124 which is the total number of rows / experimental units in the RT-qPCR dataset.

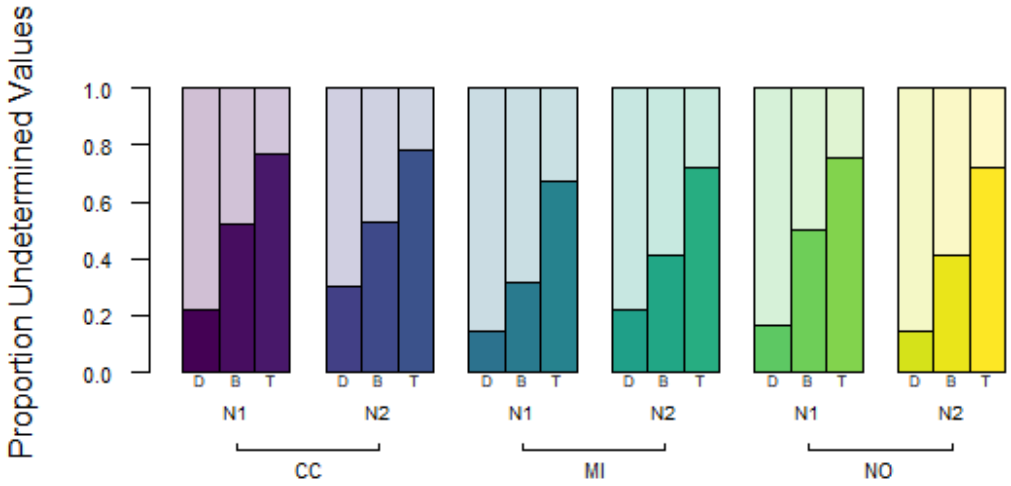


Figure 4.7: Barplot of Proportions Undetermined. Here, D, B, & T refer to sampling dates, biological replicates, and technical replicates, respectively. N1 & N2 refer to the viral sequence targets. CC, MI, & NO refer to the wastewater reclamation facilities Cedar Creek, Middle Oconee, and North Oconee, respectively.

To further explore the missingness and observed differences with respect to the viral sequence target and wastewater reclamation facility, logistic regression models were fit predicting the missing-data indicator from viral sequence target and wastewater reclamation facility.[6] Table 4.2 shows the results of these logistic regression models.

---

[6] The logistic regression models used the cbind() designation of a binomial for the response variable. The models were fit on the data used to generate Table 4.1. For example, the frequency of undetermined results aggregated to the sampling date for Cedar Creek and the N1 sequence target is 9 out of 41 total sampling dates. So, the model used the "missing" values and "non-missing" values to form the response, e.g., `cbind(9, 32)` for Cedar Creek and the N1 sequence target. Similarly, for all combinations of wastewater reclamation

Table 4.2: Logistic Regression for Undetermined RT-qPCR Results on Viral Sequence Target and Wastewater Reclamation Facility. A gives probability estimates from the logistic regression models for the missing frequencies; B shows the results of analyses of deviance; and C gives estimated odds ratios.

A

Wastewater Reclamation Facility	Viral Sequence Target	$\hat{P}$ of Sampling Days in which All Results were Undetermined	$\hat{P}$ of Biological Replicates in which All Results were Undetermined	$\hat{P}$ of Technical Replicates with All Undetermined Results
CC	N1	0.523 (0.446, 0.601)	0.523 (0.446, 0.601)	0.773 (0.737, 0.81)
	N2	0.529 (0.451, 0.607)	0.529 (0.451, 0.607)	0.781 (0.746, 0.817)
MI	N1	0.364 (0.291, 0.436)	0.364 (0.291, 0.436)	0.694 (0.655, 0.733)
	N2	0.369 (0.296, 0.441)	0.369 (0.296, 0.441)	0.703 (0.664, 0.742)
NO	N1	0.455 (0.38, 0.53)	0.455 (0.38, 0.53)	0.734 (0.697, 0.771)
	N2	0.46 (0.385, 0.536)	0.46 (0.385, 0.536)	0.742 (0.706, 0.779)

B

Model	Effect	LR Chisq	Df	Pr(>Chisq)
Days	target	0.808	1	0.369
	facility	2.904	2	0.234
Bios	target	0.018	1	0.894
	facility	11.433	2	**0.003**
Techs	target	0.198	1	0.657
	facility	11.043	2	**0.004**

C

Model	Parameter	OR Estimate
Days	N2 v N1	1.335 (0.711, 2.531)
	MI v CC	0.629 (0.293, 1.324)

facility and viral sequence target, the variables for the frequency of missing and non-missing observations were combined using cbind() which served as the response variable in the models. This grouped binomial designation is equivalent in estimation as the more common 0 / 1 coding in logistic regression.

	NO v CC	0.529 (0.239, 1.136)
Bios	N2 v N1	1.021 (0.749, 1.392)
	MI v CC	**0.52 (0.354, 0.762)**
	NO v CC	0.761 (0.521, 1.108)
Techs	N2 v N1	1.045 (0.861, 1.268)
	MI v CC	**0.664 (0.52, 0.846)**
	NO v CC	*0.807 (0.63, 1.031)*

Interestingly, the facility was determined to have had some effect on the frequencies of undetermined results for the data at the biological replicate level (χ^2 = 11.4, df = 2, p = 0.003) and the technical replicate level (χ^2 = 11, df = 2, p = 0.004) after controlling for sequence target. No significant differences were observed for the data aggregated at the sampling date level (χ^2 = 2.9, df = 2, p = 0.234), but this is not surprising as the artificial reduction in sample size likely reduced the power of the analysis. The viral sequence target does not exhibit evidence of having any impact on "missing" data (p = 0.369, 0.894, and 0.657 for analyses of sampling dates, biological replicates, and technical replicates, respectively); that is, the frequencies of undetermined experimental values are indistinguishable between N1 and N2 at each level while controlling for facility. The differences among facilities appears to be mostly the result of lower frequencies (lower estimated odds) of undetermined values for the Middle Oconee wastewater reclamation facility when compared with Cedar Creek for both analyses at the biological and technical replicate levels. Notably, the estimated odds ratio comparing undetermined values in North Oconee to Cedar Creek is *marginally* significant for the analyses of technical replicates (i.e., its confidence interval just barely includes the null value so it is inconclusive whether North Oconee's missingness differs from that of Cedar Creek). It seems that Cedar Creek samples are more likely to yield undetermined results from the RT-qPCR experiments.

Regardless of the ***patterns*** of missingness within the data, it may be important to explore approaches to use any information of a missing / undetermined result. To accomplish this, we explore limits of detection in the next subsection.

4.3 Exploring Limits of Detection and Quantification from the Data Perspective

The limit of detection refers to a threshold where a positive sample (i.e., one that **does** contain targeted substrate) is indistinguishable from a negative sample (i.e., one that **does not** contain any targeted substrate) given some technique. Therefore, some of the undetermined values within the data could be truly positive, but, since the substrate is at

such a low concentration, the experimental approach fails to detect anything (i.e., a false negative).[7]

Theoretically, it may be possible to calculate a limit of detection through considerations of the experimental procedure alone.[8] The RT-qPCR workflow begins with the water / sewage sample from which 280 microliters, μL, is used in RNA extraction. Following, the RNA is eluted into 60 μL of a buffer solution. Three μL of the 60 μL of RNA-in-buffer is added to a reverse transcription reaction of 25 μL. Finally, 2 μL of the reverse transcription reaction product is transferred to 20 μL wells. These 25 μL product of the reverse transcription reaction is distributed in $\sim$2 μL quantities for three technical replicates for each viral sequence target; the remainder is used in controls.[9]

Altogether, these quantities are important for understanding the dilution scheme imposed by the lab procedures which can then be used in calculating the theoretical limit of detection (Equation (4.1)). In words, if there was a ***single*** viral sequence copy within our sample, then we can track the copy through the dilution to calculate the concentration that is used in the PCR amplifications:

$$Theoretical\ LOD = \frac{1\ viral\ copy}{60\mu L} \times \frac{3\mu L}{25\mu L} \times \frac{2\mu L}{20\mu L} = 0.0002\ copies\ \mu L^{-1}\ of\ reaction. \qquad (4.1)$$

Although a useful quantity / concentration to know, this limit of detection is rather optimistic. So, in lieu of an experimental approach, we investigated some data analysis approaches to quantify the limit.

4.3.1 Observed Distributions of Cycle Thresholds and Normality

The cycle threshold values for detected samples range from approximately 32.16 to 39.03; a histogram of these data is shown in Figure 4.8. There appears to be a stacking of observations around Ct = 37. This potential "boundary" in the data distribution may indicate both limits of detection (37+) and quantification (37-).

---

[7] It may be important to acknowledge that these experimental methods in themselves are not without flaw nor devoid of other potential sources of measurement error. So, even given a sample with substrate concentrations above the limit of detection, it is still possible and not unlikely that an experiment could have an undetermined / negative result (i.e., spurious).

[8] This theoretical limit of detection may be more aligned with an absolute lower bound for procedure as it would assume perfectly efficient lab techniques and reactions.

[9] I am not confident in my verbose description of the lab procedure nor essential verbage.

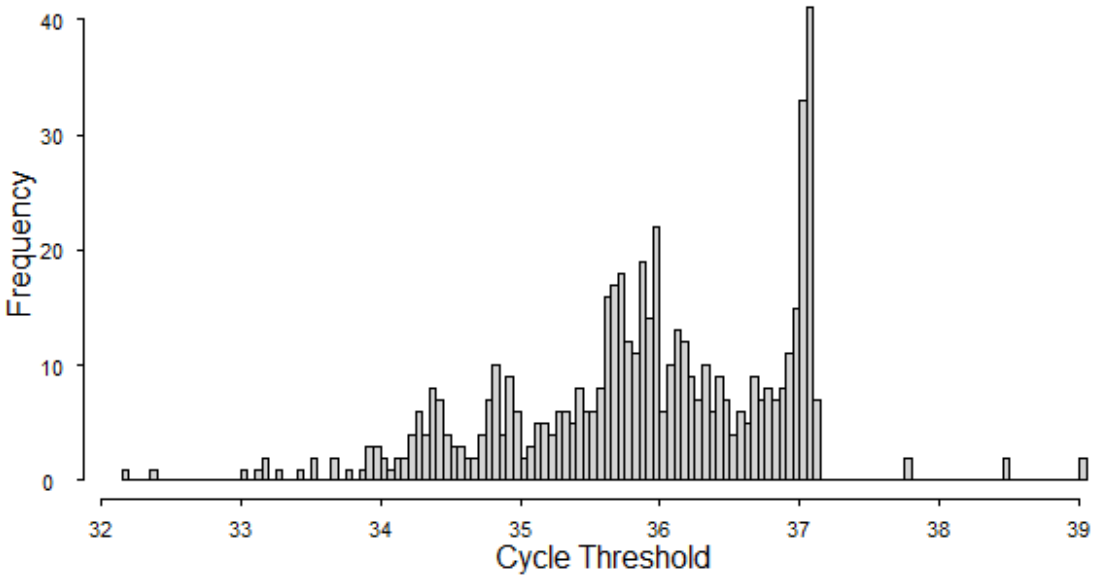


Figure 4.8: Histogram of Cycle Thresholds

Figure 4.9 shows the normal quantile plots for the cycle thresholds of the two viral sequence targets, N1 and N2, separately. If the data were approximately normally distributed, we would expect a roughly straight line (shown by the gray diagonal line). However, there appears to be some deviations from normality at the upper tail for higher Ct values. From this normal quantile plots, we can identify two points of interest. The first point of interest would be the initial inflection point (scanning from left to right or low Ct values to higher Ct values) where the data deviate from a normal distribution tail, i.e., begin stacking in the histogram. The second point of interest would approximate to the point of truncation seen in Figure 4.8 and in Figure 4.9 as the right-hand side of the flatter portion of the "curve" of data points.[10]

[10] To estimate the ct values associated with these points of interest, I programmatically scanned the qqplot and the slopes for each pair of points (scanning right to left with respect to the qqplots [artifact of first writing program for copy number scanning]). When the slopes were successively below 1, this indicated the first point of interest (i.e., the limit of detection). Following, when the slopes then approximated 1 again, this indicated the second point of interest (i.e., the limit of quantification).

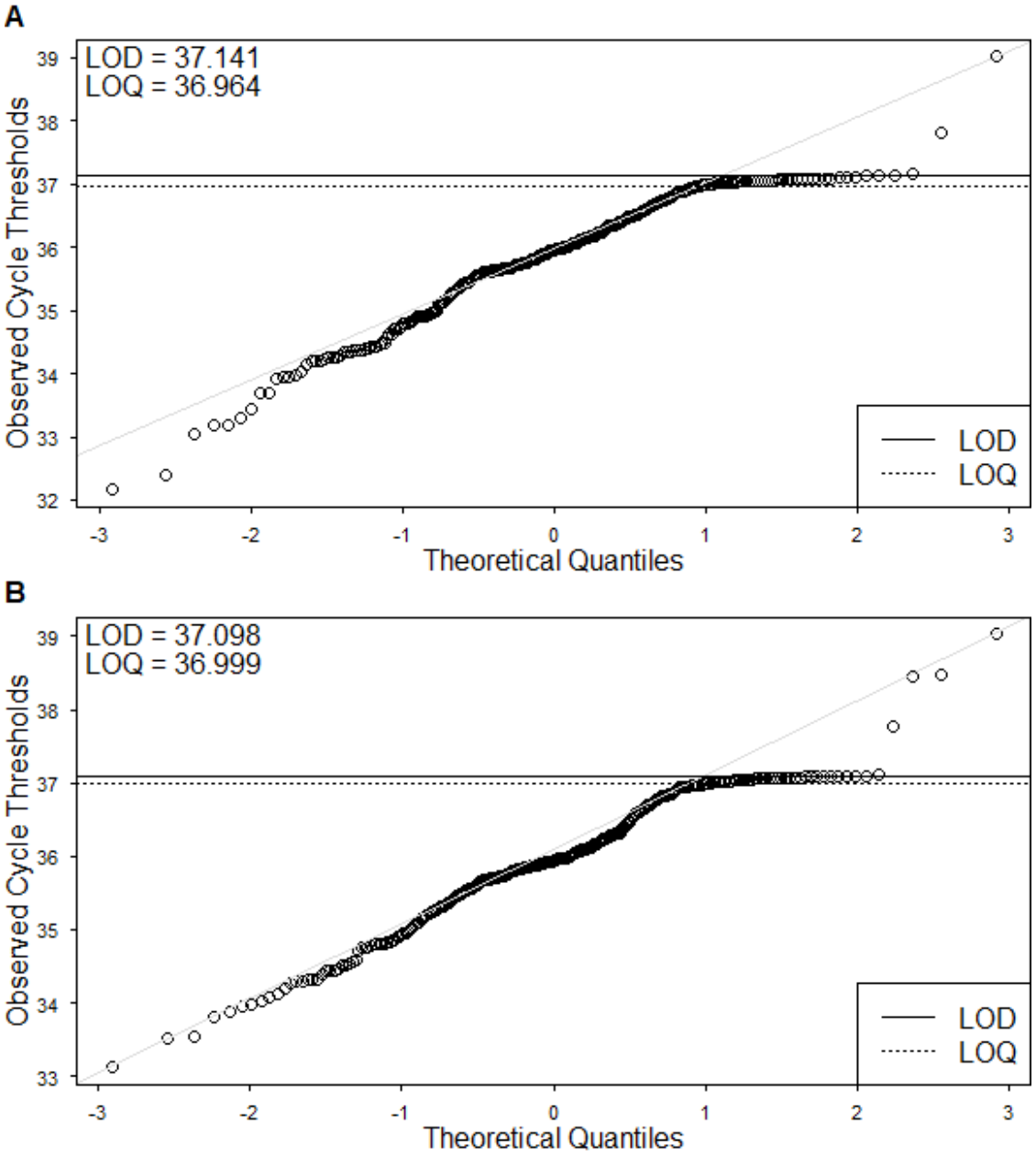


Figure 4.9: Normal Quantile-Quantile Plots of Observed Cycle Thresholds for Viral Sequence Targets N1 and N2

So, with careful consideration, we can rationalize the meaning of these points. The point of truncation within the histogram and the larger point of interest determined from the normal quantile plots may serve as a proxy for the limit of detection. There are few values more extreme than this cutoff point and it seemingly acts as a boundary in the Ct values distribution.

Also, the point of inflection where the data deviate from normality may serve as a proxy for a limit of quantification. Similar to a limit of detection, a limit of quantification represents a lower boundary of concentrations / quantities beyond which the experimental procedure performs poorly. The limit of quantification is defined for higher concentrations / larger quantities of substrate when compared to the limit of detection. This means that the experimental procedure would be able to detect a sample as positive as it is above the detection limit. However, for those samples with concentrations above the limit of detection yet below the limit of quantification, the experiment fails to accurately estimate and distinguish the concentrations of substrate.

Table 4.3 explores where the cycle threshold data fall with respect to these newly defined limits. A more detailed table can be found in Appendix 11.

Table 4.3: Frequencies and Percentages of Technical Replicates whose Cycle Thresholds are Undetermined, Above the LOD, Between the LOD and LOQ, and Below the LOQ

Wastewater Reclamation Facility	Viral Sequence Target	n	Undetermined	Above LOD	Between LOD and LOQ	Below LOQ
CC	N1	330	254 (77)	2 (0.61)	10 (3)	64 (19.4)
	N2	330	259 (78.5)	2 (0.61)	8 (2.4)	61 (18.5)
MI	N1	360	243 (67.5)	0 (0)	26 (7.2)	91 (25.3)
	N2	360	260 (72.2)	2 (0.56)	15 (4.2)	83 (23.1)
NO	N1	372	281 (75.5)	1 (0.27)	10 (2.7)	80 (21.5)
	N2	372	268 (72)	1 (0.27)	16 (4.3)	87 (23.4)

Table 4.3 has some overlap in presented data with Table 4.1; specifically, the frequencies for undetermined results. However, the additional columns show further breakdowns of the cycle threshold values with respect to the LOD and LOQ.

Only 8 technical replicates with Ct values were above the newly specified limit of detection (Table 4.3).[11]

[11] Normally, values *below* the limit of detection would be of undetectable. However, since we are referring to the cycle thresholds, the values *above* the limit of detection are of interest. These higher Ct values correspond to lower concentrations of substrate or genetic material within the sample.

So, after identifying these potential limitation boundaries within the data, we have a value that may be useful in managing the “missingness” / undetermined results: the limit of detection.

5 Waste (Data) Management

5.1 Standard Curves and Reaction Efficiency

Output of RT-qPCR is the cycle threshold (Ct) or the number of amplification cycles at which fluorescence intensity is detectable. This value in its own is not directly related to the absolute concentration / quantity of genetic material within the original sample; rather, it is a semi-quantitative value that corresponds to the relative concentrations following serial amplification. In order to convert these cycle threshold values to estimates of the absolute concentration of genetic material within the original sample, we need to calibrate conversion values to represent the implemented RT-qPCR reaction.

5.1.1 Algebraic Foundations of Amplification

As an aside (and more to walk myself through the understanding), the following equations from Kralik and Ricchi (2017) link the original sample concentration (N_0) to the concentration after n rounds of amplification (N_n):

$$N_n = N_0 \times (1 + E)^n. \qquad (5.1)$$

The parameter E corresponds to the efficiency of the reaction: for a perfectly efficient reaction, $E = 1$ which simplifies the term $(1 + E)^n = 2^n$ to reflect the doubling of genetic material for each of n rounds of amplification (Equation (5.2)); that is,

$$N_n/N_0 = 2^n. \qquad (5.2)$$

Typically, the samples with known concentrations in ten-fold serial dilutions are used to estimate calibration or standard curves for the reactions; The following represents a simplification of Equation (5.2) using two ten-fold dilutions:

$$10 = 2^n. \qquad (5.3)$$

Solving Equation (5.3) yields

$$log_2(10) = n \approx 3.322, \qquad (5.4)$$

showing that under a perfectly efficient doubling reaction, the difference in the number of amplification rounds between any two ten-fold dilutions should be approximately 3.322.

This known value can thusly be used to calculate the reaction's efficiency

$$E = 10^{-\left(\frac{1}{n}\right)} - 1. \qquad (5.5)$$

Altogether, this framework can be used to convert RT-qPCR output Ct values to an absolute concentration of genetic material.

5.1.2 Fitting Standard Curves

The standard curves are fit to the data using linear regression of cycle threshold on base 10 logarithm of the known quantity of substrate:

$$Ct = \beta_0 + \beta_1 log_{10}(quantity) + \epsilon. \qquad (5.6)$$

Figure 5.1 shows fit lines for each repetition of the standard curve calibration experiments, highlighting the mean of all fits (grey) and the single results chosen to represent the reaction (black line for fit line, black circles for data points) for both viral sequence targets, N1 (A) and N2 (B). Appendix 12 has a more extensive description of each of the standard curve fits from the many experimental repetitions (denoted with CN=collection number).

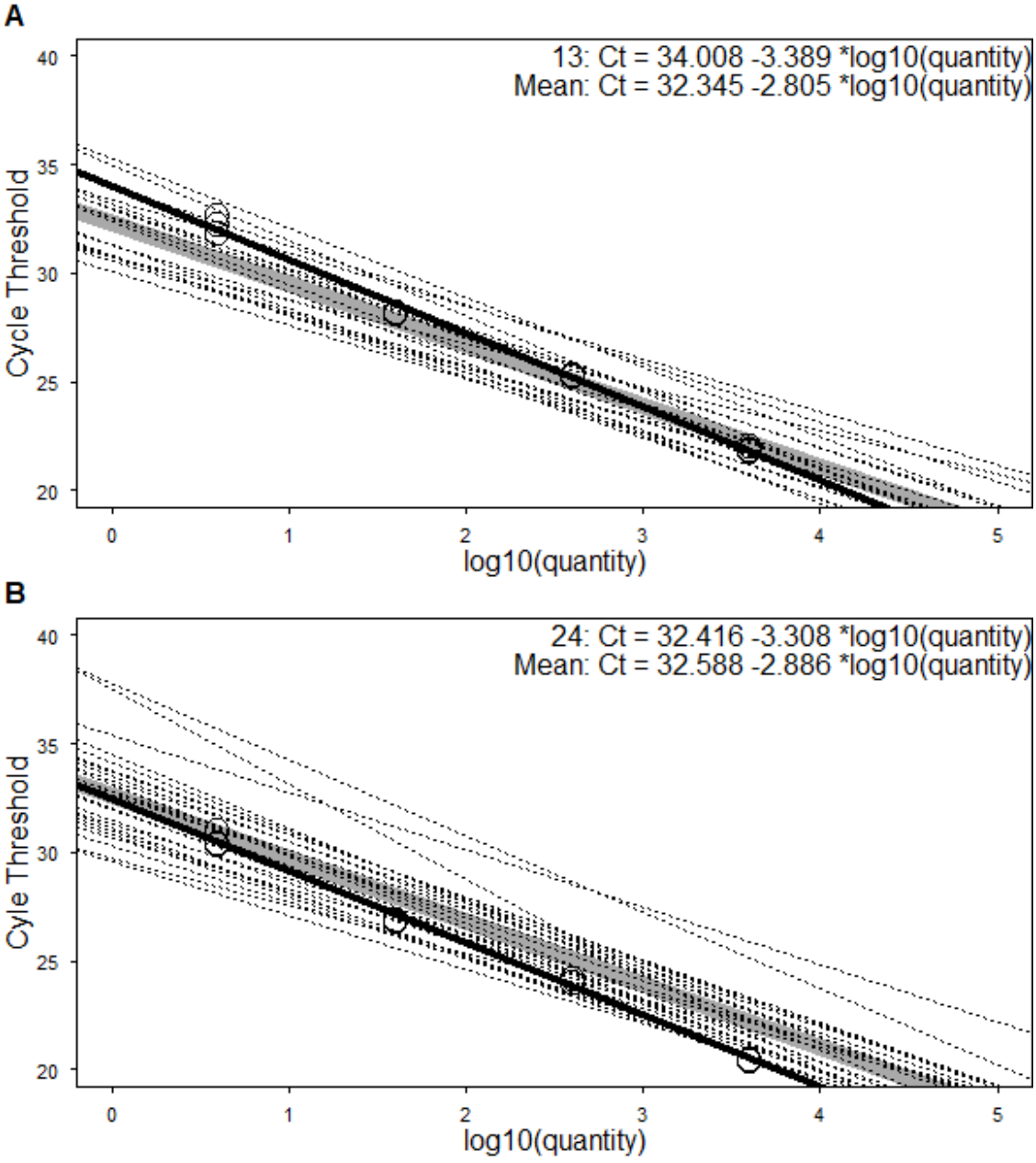


Figure 5.1: RT-qPCR Standard Curves for Viral Sequence Targets, N1 & N2

5.1.3 Calculating Sample Concentrations using Calibrated Standard Curves and Reaction Dilutions

Now with the fit lines to the standard curve data, we can use the values from the intercepts and the slopes to convert cycle threshold values to the original sample concentration:

$$Quantity = 10^{-\frac{Ct-\beta_0}{\beta_1}}, \qquad (5.7)$$

a rearrangement of Equation (5.6)). Substituting our fit estimates for the intercept and slopes, we have

$$\begin{aligned} Quantity_{N1} &= 10^{-\frac{Ct-34.008}{-3.3890}}, \\ Quantity_{N2} &= 10^{-\frac{Ct-32.416}{-3.3084}}, \end{aligned} \qquad (5.8)$$

for the viral sequence targets, N1 and N2, respectively. The concentrations yielded from Equation (5.8) would be in units $copies\ per\ \mu L\ of\ reaction$. We need to take into consideration the dilution scheme (Equation (4.1)) to have the units be simply $copies\ per\ \mu L$.

5.2 Data Adjustment using Limits of Detection and Quantification

It is relatively common to use the limit of detection or some fraction thereof to replace undetermined results within the data. For the purposes of this analysis we will use a generic approach to replace the missing results (or those that fall below the estimated LOD) with half of the LOD value (i.e., LOD / 2). Similarly, we will replace those values that fall between the LOD and LOQ with half of the LOQ value (i.e., LOQ / 2). Figure 5.2 shows the histograms of the natural logarithm of estimated copies per microliter before (panels A and B) and after (panels C and D) replacing the values as previously described. Similar to Figure 4.8, the histograms in panels A and B of Figure 5.2 show that the data deviate from a normal distribution, even after the normalizing natural logarithm transformation. Also similar, we can observe the points of interest within the distribution that indicate limits of detection and quantification (highlighted with the vertical bars). Panels C and D of Figure 5.2 show these same histograms after the data were adjusted using the LODs and LOQs. Note that, due to the extensive missingness, the replaced values dominate the distributions.

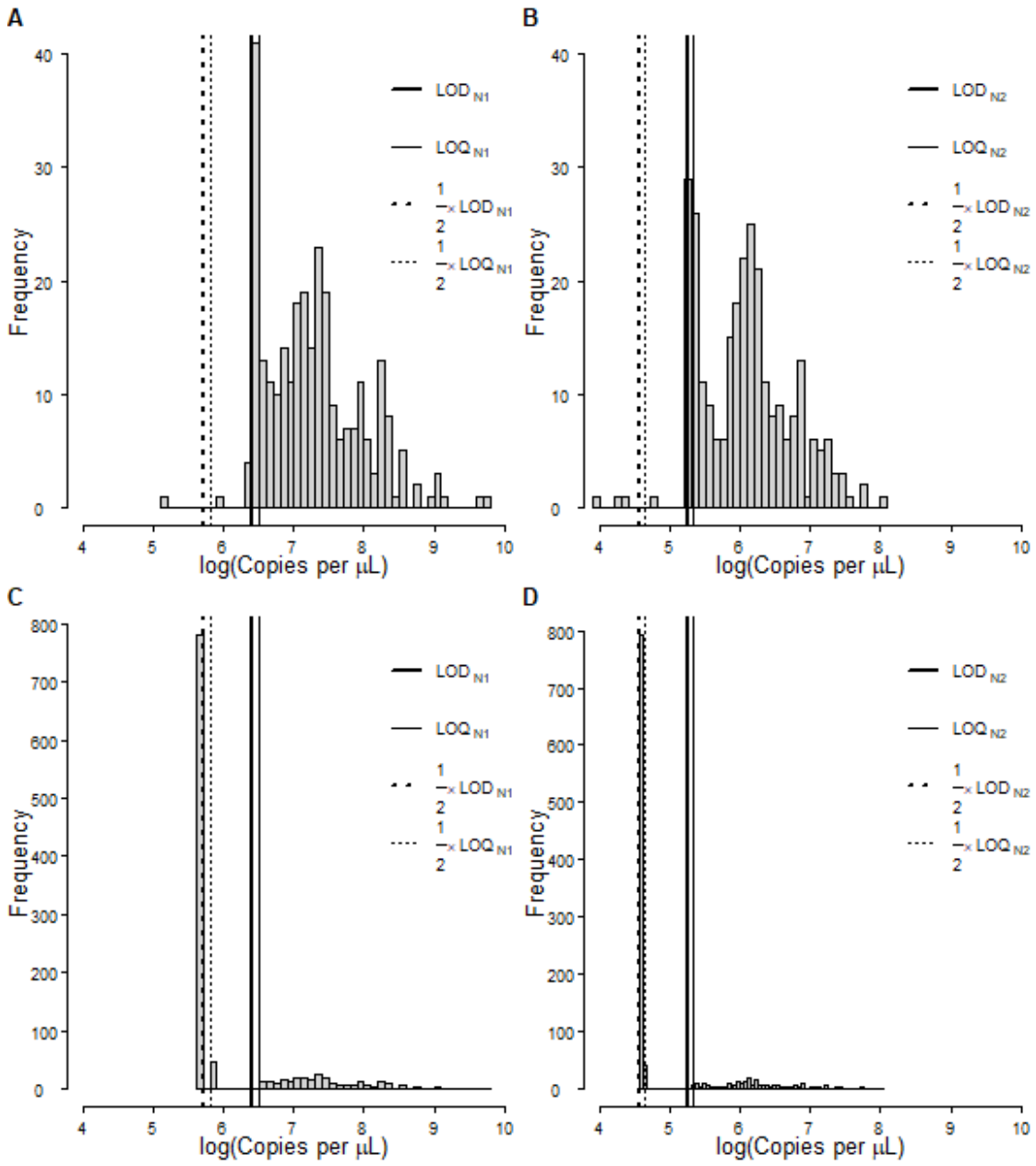


Figure 5.2: Histograms of the Natural Logarithm of Sample Viral Sequence Copy Concentrations. Panels A and B show the histograms for N1 and N2, respectively, before any adjustments were made. Panels C and D show the histograms for N1 and N2, respectively, after values which were either undetermined or fell below the LOD were replaced with half of the LOD and, similarly, values which fell between the LOD and LOQ were replaced with half of the LOQ.

6 Wiping away Convolution

There are certain inherent flaws within most, if not all, infectious disease surveillance systems. Perhaps the most well-known is the discrepancy between the times a case is reported through the surveillance system and the times that person is potentially infectious and transmitting. These discrepancies leave public health professionals with an imperfect understanding of the true epidemic curve.

The data available for COVID-19 cases for Georgia attempt to overcome this issue. In addition to the more widely available time series of cases by date of report, the Georgia Department of Public Health has also included time series of cases by the date of symptom onset. This subtle difference has a meaningful rationale as it would better demarcate the times at which the cases were potentially infectious and contributing to transmission. Similarly, this would give us a better glimpse at the times at which cases were shedding into their environment. However, the data documentation indicates that the dates correspond to the date of symptom onset only when that information is available and, otherwise, refer to dates of positive specimen collection. A date of positive specimen collection may still be preferable to report date, but it is likely there still exists some discrepancy with that of the true transmissible period.

Figure 6.1 shows the epidemic curve for Athens-Clarke County with cases by dates of report and symptom onset, and PCR positive tests by date of specimen collection. These data are complete from 01 February 2020 to 08 February 2021 and have records for cases totaling 11468, 11451, and 11821, respectively for cases by report date, symptom onset date, and date of specimen collection.

The difference in the curves is slight, but left shifts are notable. That is, the curve for cases by report date lags behind the curve for PCR positive cases by date of specimen collection which, in turn, lags behind the curve for cases by symptom onset date.

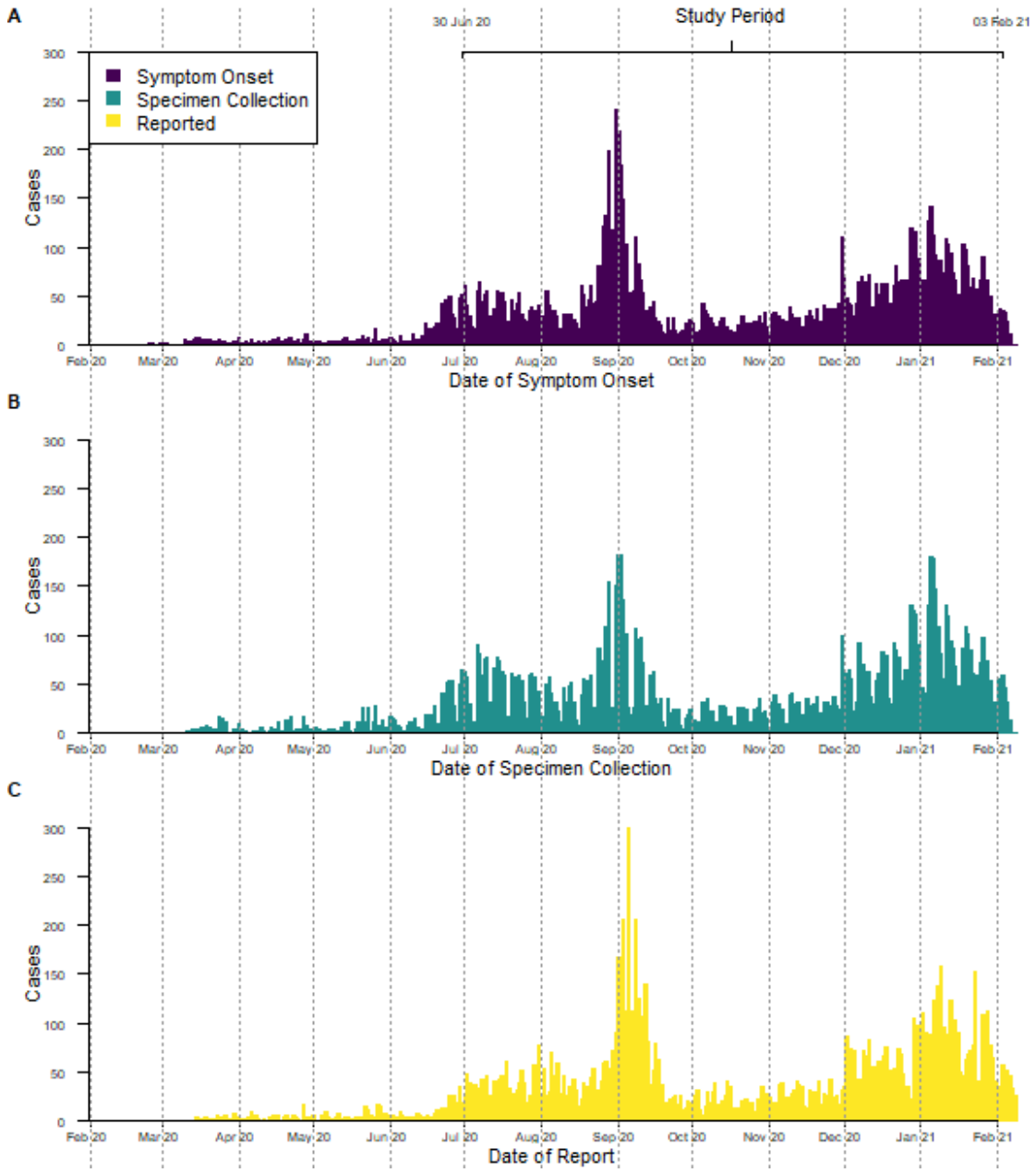


Figure 6.1: Epidemic Curve of COVID-19 in Athens-Clarke County, GA, USA

6.1 Comparing Deconvoluted Case Counts by Report Date to Case Counts By Symptom Onset Date

The case counts by report date are essentially a *convolution* of the underlying true epidemic curve and the distribution of times between true and reported infection dates. If we have an understanding of the distribution of these delays (i.e., how likely a delay of *n* amount of time is), then we can use a method of deconvolution to attempt to remove the delays from the data. The `incidental` package of R[5] uses an empirical Bayes estimation method to accomplish this. Furthermore, the package also has a given delay distribution for COVID-19 (Figure 6.2). This delay distribution was estimated using information on the distribution of the incubation period from Lauer et al, 2020 and reporting delays after symptom onset from Florida case listings. I was unable to find any associated peer-reviewed publications using this delay distribution.

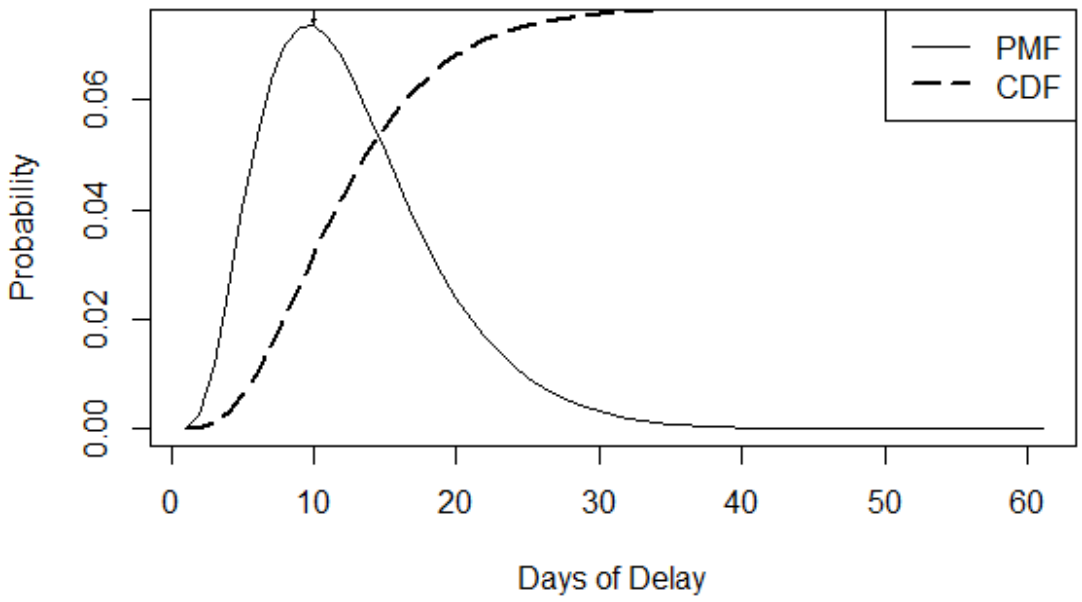


Figure 6.2: Incidental Package Delay Distribution for COVID-19

The estimation procedure for the deconvoluted incidence curve includes the fitting of splines to the convoluted data using Poisson basis functions for a regularized Poisson likelihood function. Therefore, it is necessary to specify the number of knots for the splines fits and a hyperparameter, λ, for the regularization. A vector of potential values is explored for both the numbers of knots and the hyperparameter and the algorithm used by the `incidental` package selects the best performing vales based on AIC for spline knots and validation likelihood for lambda. Additional information on the `incidental` package may be found on its CRAN webpage or within the built-in help documentation `?incidental`.

Figure 6.3 shows the estimated incidence curves from the deconvolution. The deconvoluted incidence curve does seem fairly well matched with the epidemic curve for symptom onset. The deconvoluted incidence curve apprears to be shifted to the left of the symptom onset data, as expected. Peak levels in the deconvoluted curve are observed to fall below the peaks for the reported case counts. This could be attributed in part to the correlation between the mean and variance in case counts under the Poisson distribution. Moreover, the spline model will tend to smooth the data and, hence, lead to lower peak levels. Each dataset of case counts time series and the deconvolution estimates will be explored in its association with the RT-qPCR data.

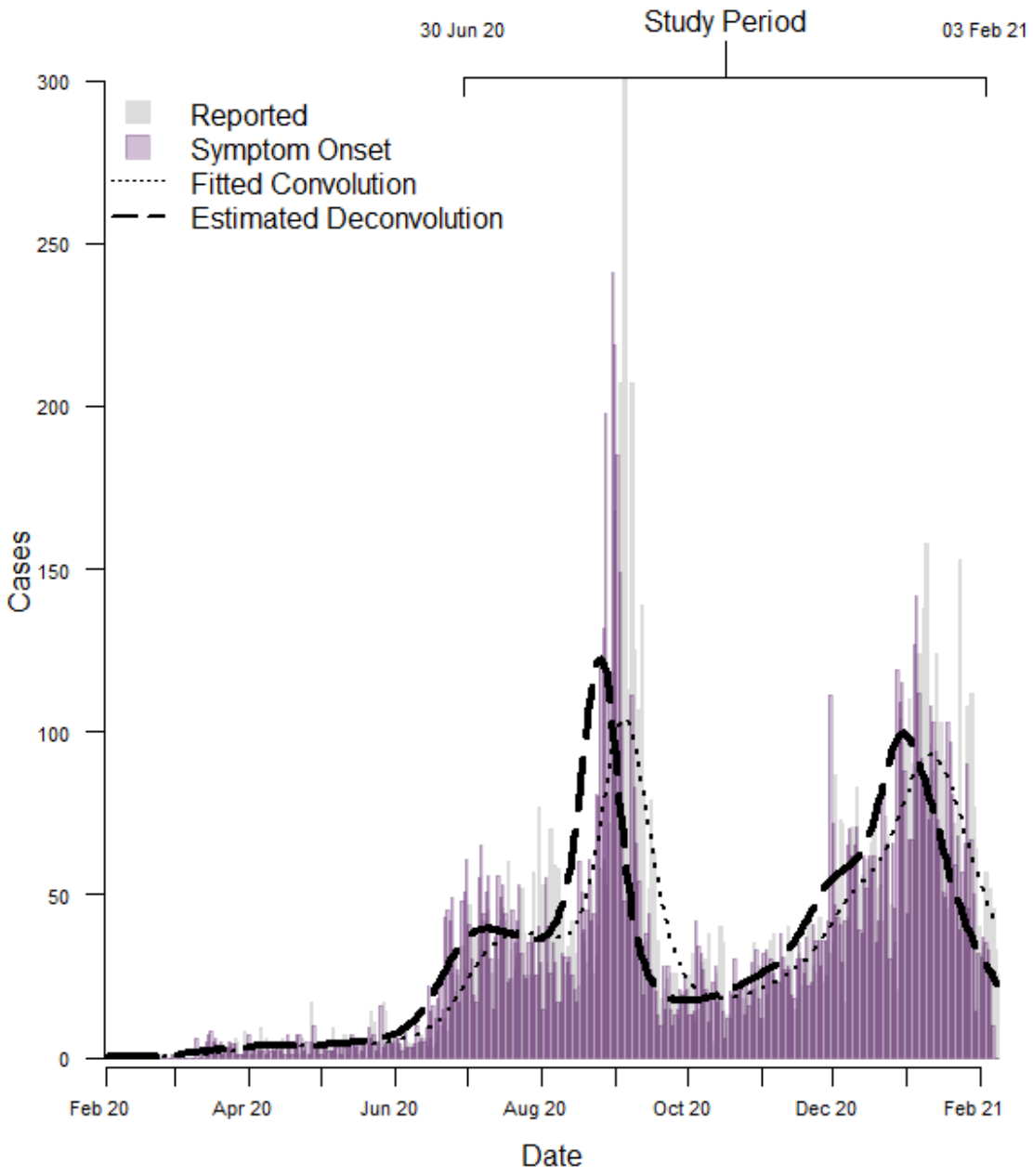


Figure 6.3: Comparison of Epidemic Curves from Dates of Report and Symptom Onset to a Deconvoluted Incidence Curve

7 It's All Clumping Together…

The ultimate goal of this project was to assess the predictability of COVID-19 cases given the wastewater-based epidemiological surveillance for SARS-CoV-2. So, we will explore the associations between the case counts and the viral loads in wastewater samples. Since we have a hierarchical dataset for the RT-qPCR results, we can take a couple different approaches with the data. We could use the data at any one of the hierarchy levels, but as the unit of analysis gets smaller, the model complexity increases where at the technical replicate level we would likely need to control for the correlation among replicates of the same sample and location. If we were to take a meaningful summary of RT-qPCR results for every given sampling time, then we would effectively remove the additional sources correlation among observations. The correlations among the different replicates will impact the uncertainty but will not impact averages. Furthermore, since the data are time series, we should address potential autocorrelation in the time of the samples. Of note, this approach of summarization to a single viral load estimate per sampling date could contribute to a loss of information from those lower levels of data, but this may be minimal given an appropriate summarization.

7.1 Determining an Appropriate Summarization Scheme

As can be seen in Figure 5.2, the distributions of the two viral sequence targets are quite different. To further show the discrepancies in N1 and N2 assays, Figure 7.1 shows the time-series of estimated copies in the wastewater for both N1 and N2 separately. The time-series show two stark contrasts: (1) the timing of the peaks and valleys and (2) the relative magnitudes. The differences in the relative magnitudes is due to a combination of the differences in the fits of the standard curves (Section 5) and the differences in the estimated limits of detection (Section 4). The differences in the timing of the peaks and valleys are a bit more difficult to rationalize. For any reason, it still stands that the results are substantially heterogeneous.

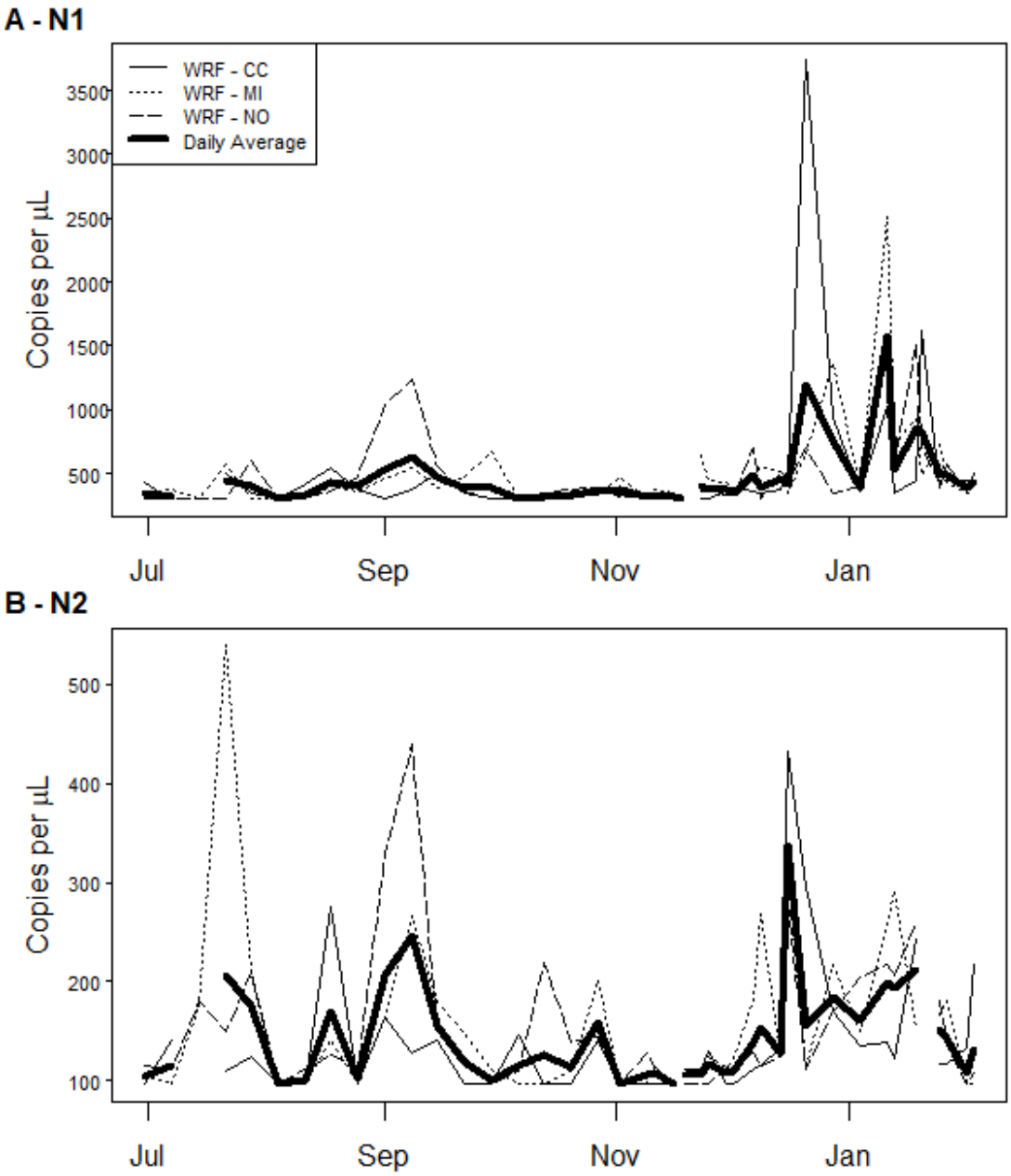


Figure 7.1: Profiles of Estimated Viral Loads in Wastewater for both Viral Sequence Targets used in RT-qPCRs

It is also evident in Figure 7.1 that the concentration time-series differ among the wastewater reclamation facilities. Therefore, it may be necessary to revisit the earlier analyses to address these discrepancies.

However, for the sake of simplicity and consistency, the data will be aggregated to a single measure of viral copies per sampling day in a joint-evaluation of their association with case incidence.

7.2 Averaging Technical and Biological Replicates

For any given sampling date and any given wastewater reclamation facility, there are typically three separate wastewater samples (i.e., biological replicates) each of which are used as a source for typically six separate runs (i.e., three technical replicates for both viral sequence targets) of the RT-qPCR. So, to summarize the results for samples from a single day at a single facility, we must average the technical replicates and biological replicates.

The choice of measure of centrality (read average) is not trivial and may have large implications in the analysis. This is explored a bit in Appendix 14. In the remainder of this report, geometric means will be used to summarize the technical replicates and, subsequently, the biological replicates.

Next, we can aggregate data across facilities so as to account for the total influent water / sewage volume and the total influent suspended solids concentrations. Here, we multiply the averaged viral loads (concentrations with units $viral\ copies\ per\ \mu\ L$) by the total volume of wastewater to estimate the total number of viral copies present within the wastewater (assuming a homogenous concentration). These total viral copies data for each wastewater reclamation facility are then summed across the three facilities to yield a single value for each sampling date.

7.3 Cross-correlations

Now, with a summarized dataset (i.e., a single observation for each of the 43 sampling dates), it is finally possible to explore the relationships between the viral loads detected in the wastewater samples and the case counts data for Athens-Clarke County. As mentioned previously, the relative timing of viral shedding in wastewater compared to case incidence (or case diagnosis or report) is not well-established. However, some studies have estimated that the RNA concentrations in wastewater ***sludge*** lead positive tests by date of specimen collection by 0-2 days and positive tests by report date by 6-8 days (Peccia et al, 2020).[12]

To similarly investigate these associations, I computed cross-correlations between total viral loads in wastewater (i.e., the summarized estimate of total viral copies for each sampling date) and case counts at various leads and lags. Instead of using the raw case counts, smoothed 7-day moving averages were used for the cases by report date, PCR-positive tests by date of specimen collection, cases by symptom onset date, and deconvoluted cases by report date were considered. The correlations were calculated using Spearman rank correlations. Figure 7.2 shows these estimated cross-correlations.

---

[12] Peccia et al (2020) used distributed lag Poisson regression models to estimate these lead times (i.e., they fit time-series models to the number of positive cases and included the wastewater sludge RNA concentrations at various lags as predictors and significant coefficient estimates were used to determine time ranges).

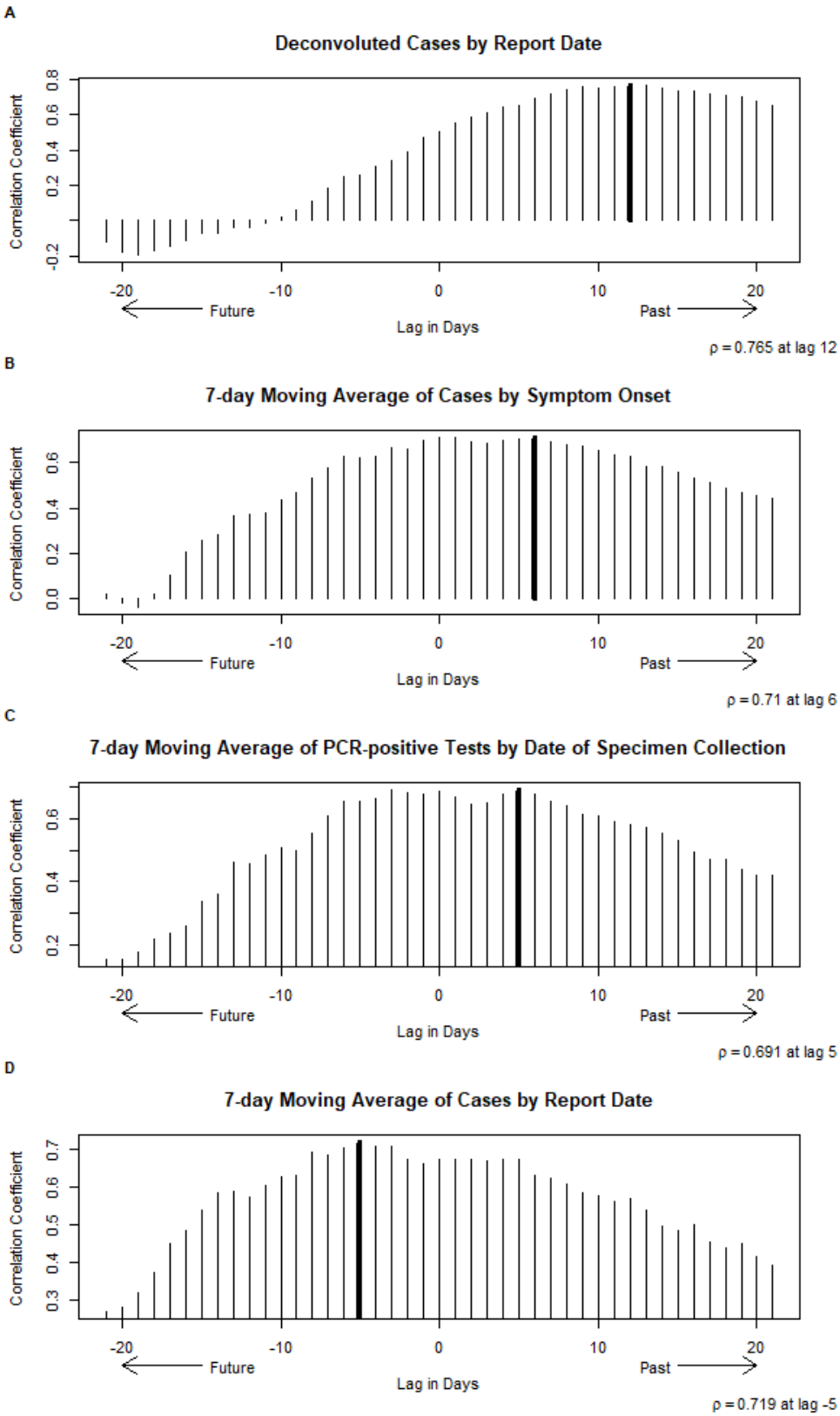


Figure 7.2: Cross-correlations Between the Total Viral Copies and COVID-19 Cases

From Figure 7.2 we can note the point at which the correlations are the strongest. The viral loads within wastewater seem to lead the 7-day moving average of cases by report date by about 5 days ($\hat{\rho}$ = 0.719, p < 0.001), but seem to lag behind the 7-day moving averages of PCR-positive test by date of specimen collection by approximately 5 days ($\hat{\rho}$ = 0.691, p < 0.001) and cases by symptom onset by about 6 days ($\hat{\rho}$ = 0.71, p < 0.001) and the deconvoluted cases by report date by about 12 days ($\hat{\rho}$ = 0.765, p < 0.001).

To assess any potential effects of the sampling interval, I calculated a spline function for the viral loads in waste and used the estimated spline curve to infer / interpolate the values of the wastewater viral loads missing due to the interval censoring. These new data were used similarly as before to estimate the cross-correlations at various leads and lags of COVID-19 cases (Figure 7.3). Additionally, I used another spline fit to smooth the wastewater viral loads data and, again, calculated cross-correlations (Figure 7.4, smoothing parameter = 0.56).

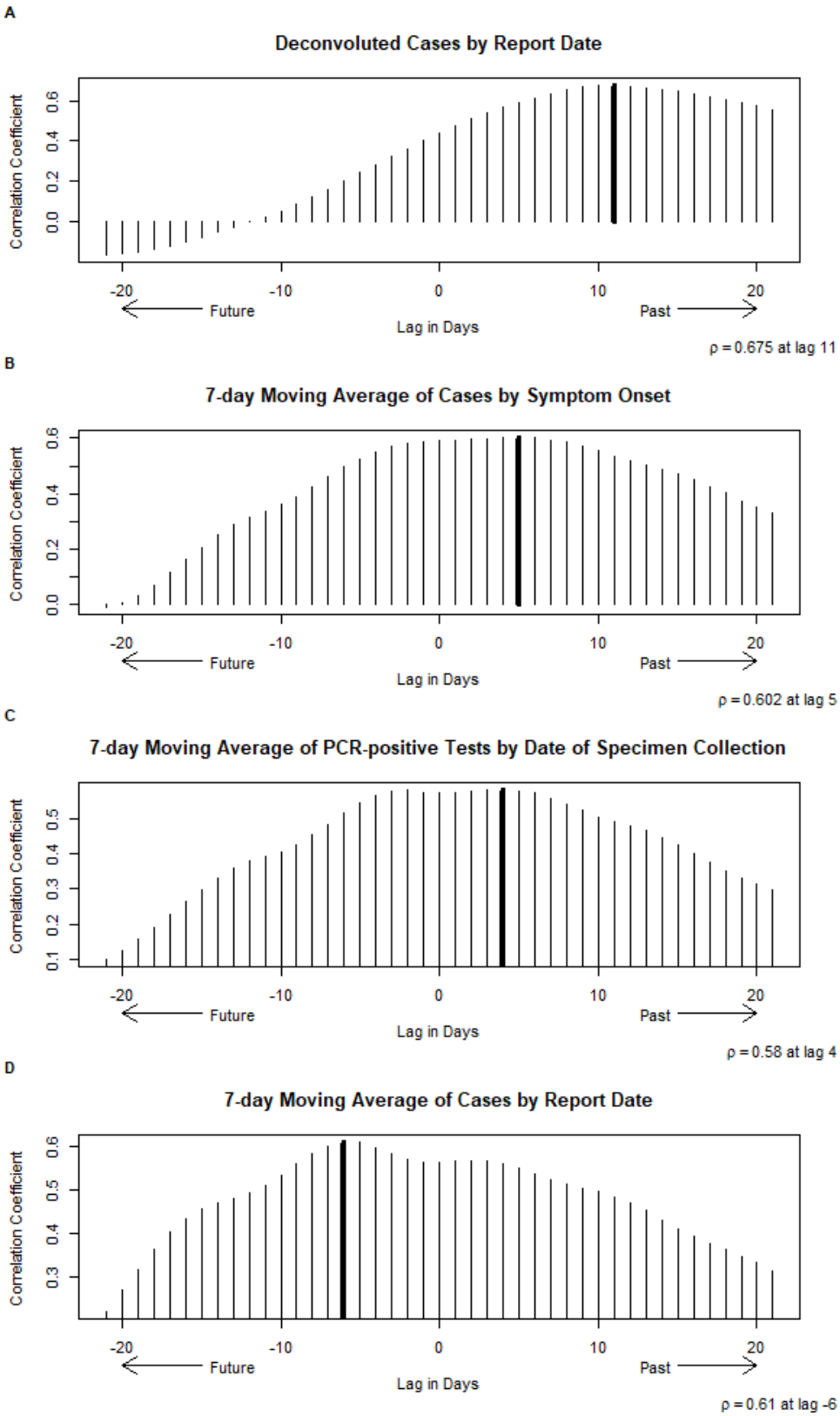


Figure 7.3: Cross-correlations Between the Interpolated Total Viral Copies and COVID-19 Cases

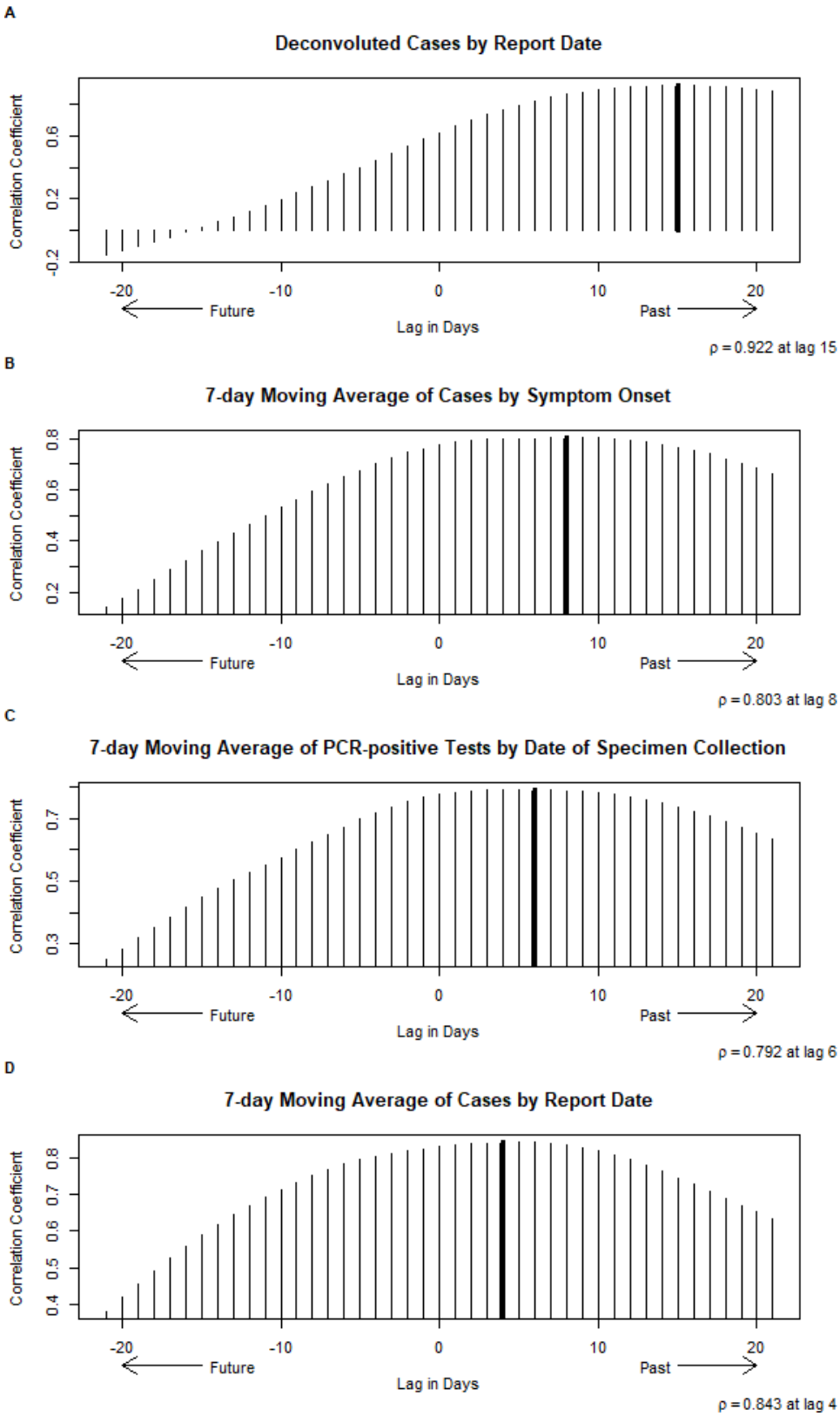


Figure 7.4: Cross-correlations Between the Smoothed, Interpolated Total Viral Copies and COVID-19 Cases

For the spline interpolation (Figure 7.3), the pattern holds consistent with wastewater samples leading cases by report date and lagging behind cases by symptom onset date and the deconvoluted case reports. On the other hand, the smoothing spline (Figure 7.4) indicates that the wastewater samples lag behind all case incidence indicators. However, the extent of the smoothing seems to have a large effect on this observed pattern, as can be seen in Figure 7.5 where the data were smoothed to a lesser extent (smoothing parameter = 0.3) and the pattern reverts to as previously observed.

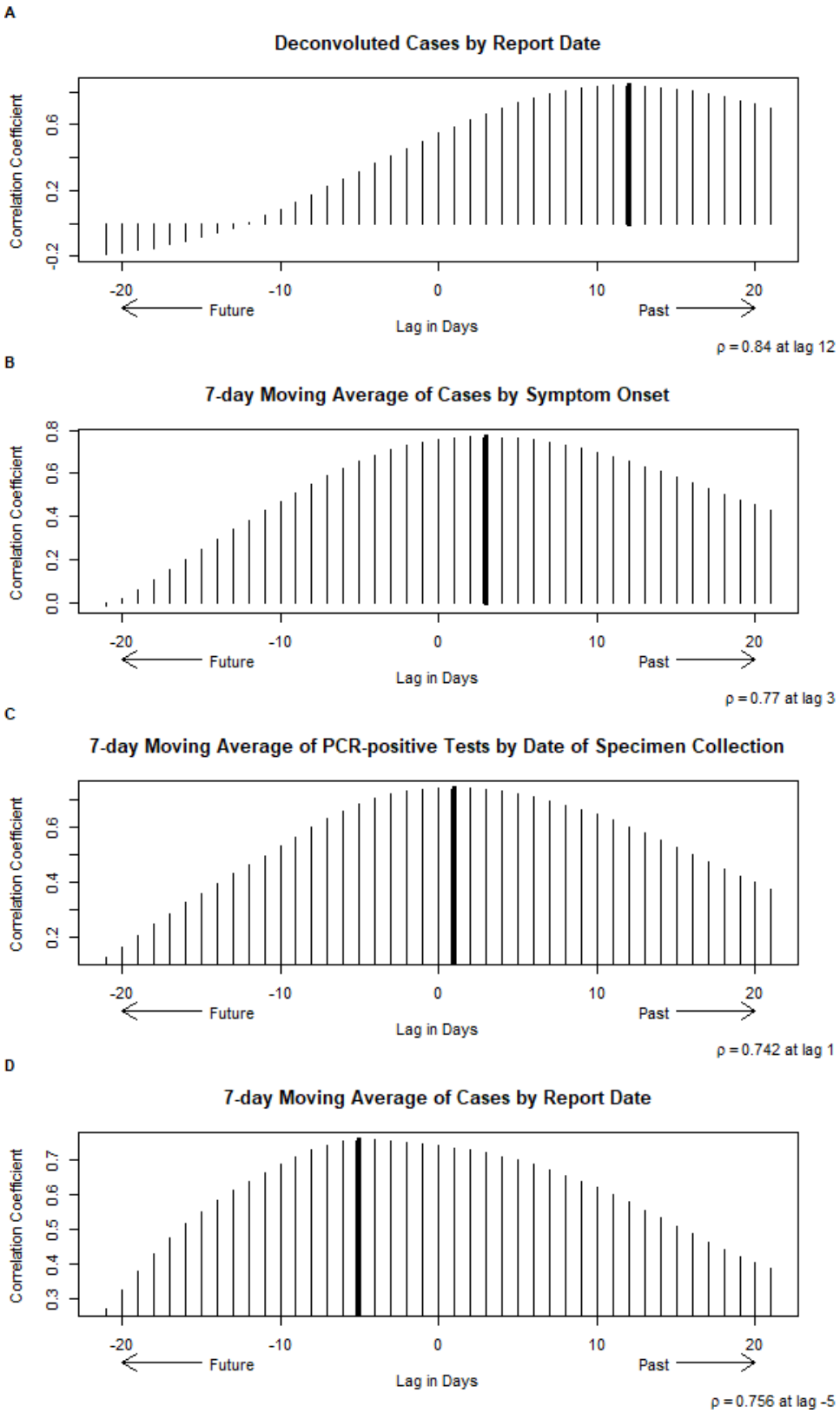


Figure 7.5: Cross-correlations Between the Less Smoothed, Interpolated Total Viral Copies and COVID-19 Cases

So, it seems that our wastewater samples lead the COVID-19 case reports and lag behind the PCR-positive tests by date of specimen collection, cases by symptom onset date, and deconvoluted case reports. However, we may interpret rather a range of days where these correlations are the strongest to be a 3-8 day lead for case reports, a 3-day lead to 5-day lag of PCR-positive tests by date of specimen collection, a 1-day lead to 7-day lag for cases by symptom onset date, and a 8-14-day lag for deconvoluted case reports.

8 Conclusions

This report covers an analysis for the utility of COVID-19 wastewater epidemiological surveillance. We: investigated the occurence of undetermined / missing data from the RT-qPCR analyses; estimated detection and quantification limits; elaborated on the conversion schemes from RT-qPCR results to viral load quantities in wastewater samples; employed our estimates of detection and quantification limits to overcome / mitigate the missingness in the data; explored a method of deconvolution for the COVID-19 epidemic curve; characterized the temporal associations between wastewater samples and population-level COVID-19 case indicators; and, ultimately, conclude that the wastewater samples may serve as useful and informative complements to traditional anthropocentric case surveillance.

Figure 8.1 shows the time-series of the wastewater samples viral loads along with the four case incidence indicators used in comparisons. We can see how closely the wastewater sample profiles approximate the curves for the cases by symptom onset date and cases by report date and how it seemingly lags behind the deconvoluted case reports curve. Biologically, this makes sense as the deconvoluted case reports curve would give us our best estimate of the *onset* of infections / exposures in the population. The actual viral shedding likely doesn't initiate so proximal to exposure, but rather it would take some time for the infection to develop; perhaps, it would better approximate the onset of infectiousness.

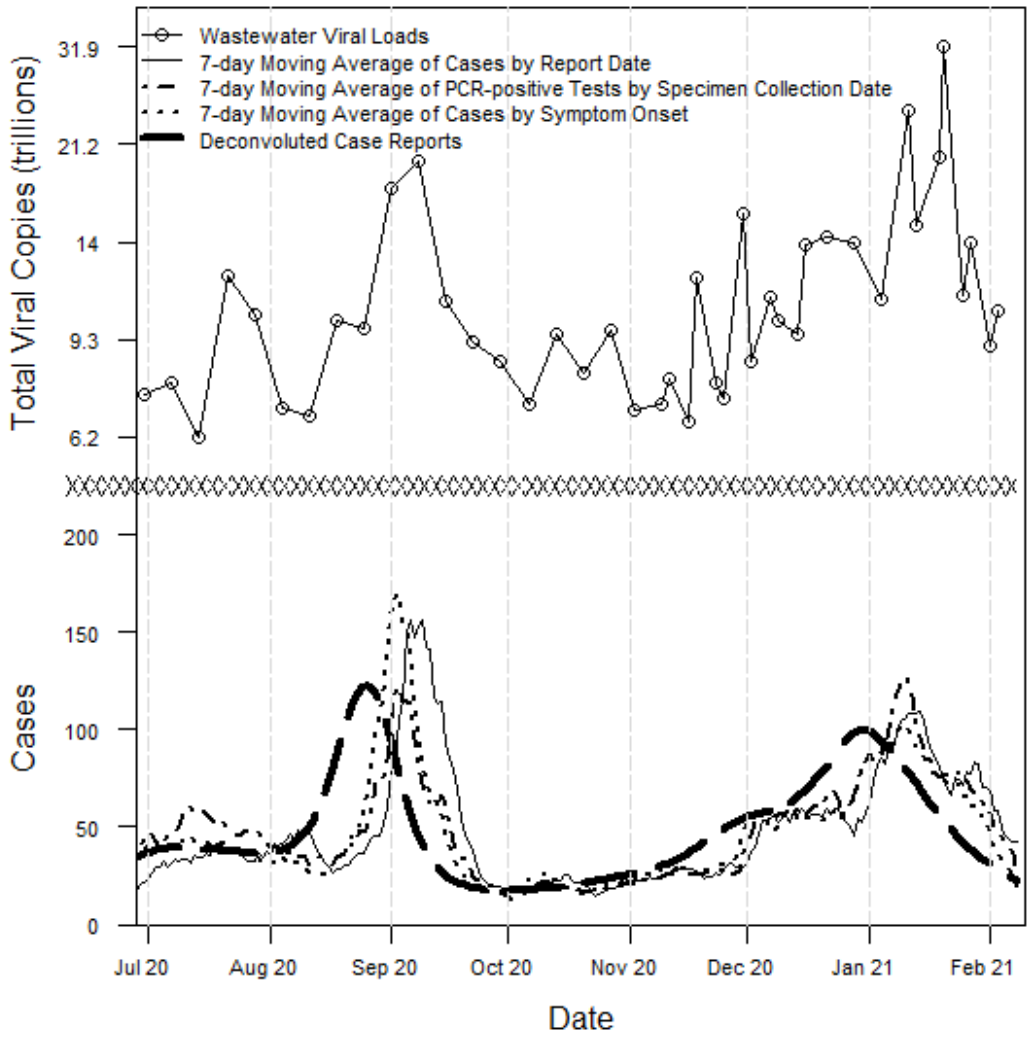


Figure 8.1: Comparison of Epidemic Curves from Dates of Report and Symptom Onset to a Deconvoluted Incidence Curve

This was a multi-step analysis and decisions made at earlier steps may have impacts on the downstream results. Therefore, it is critical to evaluate the sensitivity of our results to these analytical decisions and potential alternatives. Here, I will note on a select few of these decisions that I feel may warrant further attention.

First, the conversion calculations from RT-qPCR results using the standard curves and positive controls could use further attention. Due to the variability imposed by the

standard stock solution (e.g., batch decay?) and lab techniques (e.g., pipetting), the repeated runs of the standard curves were not utilized in this analysis. Instead, single, well-performing runs were considered characteristic of the reaction and used in the calculations. However, the standard curves chosen for the two sequence targets, N1 and N2, differ substantially and their difference can be seen in the downstream analyses. So, it may be worth considering the biological implications of such a difference in standard curves and, perhaps, reassess which values to use in conversions.

Perhaps the single most impactful step in this analysis was the data management using replacement values for the limits of detection and quantification. Although we took a practical (and common) approach using half of these limits in replacement, it may be possible to devise more complex schemes, and, in Appendix Section 13, I toyed around with different ways to replace undetermined values within the data. For example, I thought it may be informative to replace undetermined results with values proportional to the limit of detection that scale with the extent of the missingness of for a given sampling frame (e.g., proportion of technical replicates missing on a given day at a given facility from a given biological replicate and for a given sequence target). The impacts of these considerations have yet to be explored.

Similarly with respect to our *confidence* in the observations made in a sampling frame, the summarization schemes (read averaging) of technical and biological replicates could be done in a variety of ways. Appendix Section 14 illustrates various mean calculations and where a particular measure of centrality falls in relation to other (e.g., geometric mean < arithmetic mean < quadratic mean). With this in mind, I thought that it may be possible to use the extent of the missingness to help choose our averaging calculation (e.g., scale the power mean parameter p with the proportion missing). In this analysis, we used exclusively geometric means in our summarizations. Geometric means tend to bias against ***larger*** values in a distribution. This property may be desired given sparse results where very few yielded a detectable result. However, in a situation with a single missing result, we may instead wish to bias against ***smaller*** values, especially when noting the replacement values based on the limit of detection are considerably smaller than other values in the distributions; in this case, we may prefer a quadradic (or higher power) mean.

Again in consideration of the extent of the "missingness" in the data due to undetermined values, all aforementioned effects may manifest in our cross-correlation analyses. If we were able to capture this degree of "uncertainty" within a particular summarized value, it could be informative in establishing the lead / lag times of associated case incidence indicators. For example, if we could assign some range of values (or simply a measure of dispersion) or some weight to a particular observation, we may prioritize the alignment of more *certain* wastewater samples with case indicators (i.e., disregarding lack of fit for more *uncertain* wastewater samples). This idea is a little less developed with respect to actual implementation and it may be unnecessary given the aforementioned, upstream considerations for accounting for the missingness.

As of now, I consider this analysis incomplete. I can envision more involved development of modeling strategies to better characterize the data. For example, Peccia et al (2020) fit distributed lag models to their data by regressing case indicators on lagged wastewater

samples to characterize the temporal associations within the data. These methods are new to me and I may not well understand them, but I think a distributed lag model could be very useful in our analyses. However, I think it would be more well-suited to regress the wastewater samples on led / lagged case indicators to be more respectful to the biology of the system (e.g., cases are shedding into wastewater instead of wastewater exposure generating cases (disproven fecal transmission?)). Furthermore, the characterization of this environmental shedding may be directly incorporated to developing mathematical infectious disease models (e.g., as the process of shedding). I think that an infectious disease model could be very useful in establishing the utility and practicality of wastewater epidemiological surveillance. From a specialized model, we would not only be able to characterize the complex interplay of mechanisms, but we would also be able to fit the model to data for parameter estimation (e.g., average shedding, decay, *true* incidence) and we would establish a forecasting framework.

9 References

1. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/

2. RStudio Team. *RStudio: Integrated Development Environment for r*. RStudio, PBC; 2021. http://www.rstudio.com/

3. Xie Y. *Bookdown: Authoring Books and Technical Documents with r Markdown*.; 2020. https://github.com/rstudio/bookdown

4. Allaire J, Xie Y, McPherson J, et al. *Rmarkdown: Dynamic Documents for r*.; 2021. https://CRAN.R-project.org/package=rmarkdown

5. Miller A, Hannah L, Foti N, Futoma J. *Incidental: Implements Empirical Bayes Incidence Curves*.; 2020. https://CRAN.R-project.org/package=incidental

6. Public Health GD of. Daily status report: Georgia COVID-19 daily status of cases and hospitalizations with interactive charts and graphs. Published online 2021. Downloaded from https://dph.georgia.gov/covid-19-daily-status-report

7. *Real-Time PCR: Understanding C_t*. ThermoFisher Scientific, appliedbiosystems; 2016. https://www.thermofisher.com/us/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learning-center/real-time-pcr-basics/real-time-pcr-understanding-ct.html

8. Kralik P, Ricchi M. A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. *Frontiers in Microbiology*. 2017;8:108. doi:10.3389/fmicb.2017.00108

9. Demirhan H. *dLagM: Time Series Regression Models with Distributed Lag Models*.; 2020. https://CRAN.R-project.org/package=dLagM

10. Wickham H, François R, Henry L, Müller K. *Dplyr: A Grammar of Data Manipulation*.; 2021. https://CRAN.R-project.org/package=dplyr

11. Gohel D. *Flextable: Functions for Tabular Reporting*.; 2021. https://CRAN.R-project.org/package=flextable

12. Xie Y. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*.; 2021. https://yihui.org/knitr/

13. Bache SM, Wickham H. *Magrittr: A Forward-Pipe Operator for r*.; 2020. https://CRAN.R-project.org/package=magrittr

14. Schauberger P, Walker A. *Openxlsx: Read, Write and Edit Xlsx Files*.; 2020. https://CRAN.R-project.org/package=openxlsx

15. Wickham H, Hester J. *Readr: Read Rectangular Text Data*.; 2020. https://CRAN.R-project.org/package=readr

16. Wickham H, Bryan J. *Readxl: Read Excel Files*.; 2019. https://CRAN.R-project.org/package=readxl

17. Wickham H. *Tidyr: Tidy Messy Data*.; 2020. https://CRAN.R-project.org/package=tidyr

18. Zeileis A, Grothendieck G, Ryan JA. *Zoo: S3 Infrastructure for Regular and Irregular Time Series (z's Ordered Observations)*.; 2021. https://zoo.R-Forge.R-project.org/

19. Xie Y. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman; Hall/CRC; 2016. https://github.com/rstudio/bookdown

20. Demirhan H. dLagM: An R package for distributed lag models and ARDL bounds testing. *PLoS ONE*. 2020;15(2):e0228812. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0228812

21. Xie Y. *Dynamic Documents with R and Knitr*. 2nd ed. Chapman; Hall/CRC; 2015. https://yihui.org/knitr/

22. Xie Y. Knitr: A comprehensive tool for reproducible research in R. In: Stodden V, Leisch F, Peng RD, eds. *Implementing Reproducible Computational Research*. Chapman; Hall/CRC; 2014. http://www.crcpress.com/product/isbn/9781466561595

23. Xie Y, Allaire JJ, Grolemund G. *R Markdown: The Definitive Guide*. Chapman; Hall/CRC; 2018. https://bookdown.org/yihui/rmarkdown

24. Xie Y, Dervieux C, Riederer E. *R Markdown Cookbook*. Chapman; Hall/CRC; 2020. https://bookdown.org/yihui/rmarkdown-cookbook

25. Zeileis A, Grothendieck G. Zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*. 2005;14(6):1-27. doi:10.18637/jss.v014.i06

Appendix

10 Data Descriptions

10.1 Brief Item Analysis

Table 10.1 gives a brief description of the data used in this analysis. Aside from some minor manipulations upon import (e.g., COVID-19 case reports were subset to only include records for Athens Clarke County), the descriptions are for the raw data. From this, we can see some initial data formatting may be necessary. For example, many of the date variables are being read as character strings so they need to be parsed as dates to be useful for analysis. Eventually, the data represented in Table 10.1 will be aggregated / condensed to a single dataframe object used in final analyses.

Table 10.1: Description of Raw Datasets used in Analyses

Dataframe Object	Dataframe Description	Number of Observations	Number of Variables	Variable Names	Variable Classes	Variable Labels
n1	RT-qPCR Results for N1	1,062	10	run_date, run_num, sample_date, collection_num, sample_id, target, ct, slope, y-intercept, copy_num_uL_rxn	character, numeric, character, numeric, character, character, character, numeric, numeric, numeric	PCR Run Date, PCR Run ID Number, Sample Collection Date, Sample Collection Number, Sample ID, Sequence Target, Cycle Threshold, Slope of Standard Curve, Y-intercept of Standard Curve, Copy Number per Microliter of Reaction
n2	RT-qPCR Results for N2	1,062	10	run_date, run_num, sample_date, collection_num, sample_id, target, ct, slope, y-intercept, copy_num_uL_rxn	character, numeric, character, numeric, character, character, character, numeric, numeric, numeric	PCR Run Date, PCR Run ID Number, Sample Collection Date, Sample Collection Number, Sample ID, Sequence Target, Cycle Threshold, Slope of Standard Curve, Y-intercept of Standard Curve, Copy Number per Microliter of Reaction
qc	Standard Curve Results	557	5	collection_num, target, ct, quantity, log_quant	numeric, character, numeric, numeric, numeric	Sample Collection Number, Sequence Target, Cycle Threshold, Concentration Quantity Spiked in Sample, Logarithm Base 10 of Concentration Quantity
qc2	Positive Controls for SARS-CoV-2	231	3	collection_num, target, ct_value	numeric, character, numeric	Sample Collection Number, Sequence Target, Cycle Threshold
plant	Wastewater Reclamation Facility Processing Data	147	4	date, wrf, influent_flow_mg, influent_tss_mg_l	character, character, numeric, numeric	Date, Wastewater Reclamation Facility ID, Volume of Influent Flow in Millions of Gallons, Total Suspended Solids Concentration in Milligrams per Liter

Dataframe Object	Dataframe Description	Number of Observations	Number of Variables	Variable Names	Variable Classes	Variable Labels
covid	COVID-19 Cases by Symptom Onset Date	374	3	symptom.date, cases, moving_avg_cases	Date, numeric, numeric	Date of Symptom Onset, Number of Cases, Average Number of Cases in Previous 7 Days
covid.report	COVID-19 Cases by Report Date	374	3	report_date, cases, moving_avg_cases	Date, numeric, numeric	Date of Report, Number of Cases, Average Number of Cases in Previous 7 Days
covid.testing	COVID-19 PCR Diagnostics Data	374	3	collection_date, pcr_tests, pcr_pos	Date, numeric, numeric	Date of Specimen Collection, Total PCR Tests Reported or Collected, Total Positive PCR Tests

10.2 Initial Data Management and Cleaning (Formatting)

The separate datasets for the N1 and N2 RT-qPCR results are aggregated to a single dataframe to represent all experimental instances. Similarly, the COVID-19 surveillance data are aggregated to a single dataframe as are the positive controls. Each of the variables across all datasets are appropriately formatted and some additional variables are created to aid in structuring (e.g., the wastewater facility IDs are parsed out of the sample IDs). Additionally, some variables are dropped / deleted. These variables will be recreated in later analyses. Finally, the influent flow data from the wastewater reclamation facilities has a unit of a million gallons (assumed US liquid gallon). The influent flow is converted to liters using the unit conversion formula outlined in Equation (10.1).

$$1\ US\ liquid\ gallon \times \frac{231\ in^3}{1\ US\ liquid\ gallon} \times \frac{0.0254^3\ m^3}{1\ in^3} \times \frac{1000\ L}{1\ m^3} \qquad (10.1)$$

Dataframe Object	Dataframe Description	Number of Observations	Number of Variables	Variable Names	Variable Classes	Variable Labels
wbe	RT-qPCR Results for All Experiments	2,124	8	sample_date, facility, target, biological_replicate, collection_num, run_date, run_num, ct	Date, character, character, character, numeric, Date, numeric, numeric	Sample Collection Date, Wastewater Reclamation Facility, Sequence Target, Biological Replicate ID, Sample Collection Number, PCR Run Date, PCR Run ID Number, Cycle Threshold
qc	Standard Curves and Positive Controls for SARS-CoV-2	788	5	collection_num, target, ct, quantity, df	numeric, character, numeric, numeric, character	Sample Collection Number, Sequence Target, Cycle Threshold, Concentration Quantity Spiked in Sample, Dataframe Indicator (two separate before merge)
plant	Wastewater Reclamation Facility Processing Data	147	4	date, wrf, influent_flow_L, influent_tss_mg_l	Date, character, numeric, numeric	Date, Wastewater Reclamation Facility ID, Volume of Influent Flow in Liters, Total Suspended Solids Concentration in Milligrams per Liter

Dataframe Object	Dataframe Description	Number of Observations	Number of Variables	Variable Names	Variable Classes	Variable Labels
covid	COVID-19 Cases by Symptom Onset Date, Cases by Report Date, and PCR Diagnostics Data	374	5	date, cases.symptom.onset, cases.reported, pcr_tests, pcr_pos	Date, numeric, numeric, numeric, numeric	Date, Number of Cases with Symptom Onset on the Date, Number of Cases Reported on the Date, Total PCR Tests Reported or Collected on the Date, Total Positive PCR Tests from those Reported or Collected on the Date

11 Limits of Detection and Quantification

Table 4.3 explores where the cycle threshold data fall with respect to these newly defined limits. This table is three separate tables (A, B, C) stacked on top of one another showing summaries at the different hierarchy levels within the data: A = technical replicates; B = biological replicates; and, C = sampling days.

Table 11.1: Distributions of Cycle Thresholds at Each Hierarchy

A

Wastewater Reclamation Facility	Viral Sequence Target	Tech: Undetermined	Tech: Above LOD	Tech: Below LOD & Above LOQ	Tech: Below LOQ
CC	N1	254 (77)	2 (0.61)	10 (3)	64 (19.4)
	N2	259 (78.5)	2 (0.61)	8 (2.4)	61 (18.5)
MI	N1	243 (67.5)	0 (0)	26 (7.2)	91 (25.3)
	N2	260 (72.2)	2 (0.56)	15 (4.2)	83 (23.1)
NO	N1	281 (75.5)	1 (0.27)	10 (2.7)	80 (21.5)
	N2	268 (72)	1 (0.27)	16 (4.3)	87 (23.4)

B

Wastewater Reclamation Facility	Viral Sequence Target	Bio: All TR Undetermined	Bio: All TR Above LOD	Bio: All TR Above LOQ	Bio: Others
CC	N1	52.3 (56 / 107)	52.3 (56 / 107)	57.9 (62 / 107)	42.1 (45 / 107)
	N2	52.9 (55 / 104)	52.9 (55 / 104)	58.7 (61 / 104)	41.3 (43 / 104)
MI	N1	31.9 (36 / 113)	31.9 (36 / 113)	44.2 (50 / 113)	55.8 (63 / 113)
	N2	41.4 (46 / 111)	42.3 (47 / 111)	46.8 (52 / 111)	53.2 (59 / 111)
NO	N1	50 (57 / 114)	50 (57 / 114)	53.5 (61 / 114)	46.5 (53 / 114)

	N2	41.4 (46 / 111)	42.3 (47 / 111)	48.6 (54 / 111)	51.4 (57 / 111)

C

Wastewater Reclamation Facility	Viral Sequence Target	Days: All BR Undetermined	Days: All BR Above LOD	Days: All BR Above LOQ	Days: Others
CC	N1	22 (9 / 41)	22 (9 / 41)	31.7 (13 / 41)	68.3 (28 / 41)
	N2	30 (12 / 40)	30 (12 / 40)	35 (14 / 40)	65 (26 / 40)
MI	N1	14.3 (6 / 42)	14.3 (6 / 42)	19 (8 / 42)	81 (34 / 42)
	N2	22 (9 / 41)	22 (9 / 41)	26.8 (11 / 41)	73.2 (30 / 41)
NO	N1	16.7 (7 / 42)	16.7 (7 / 42)	23.8 (10 / 42)	76.2 (32 / 42)
	N2	14.6 (6 / 41)	14.6 (6 / 41)	19.5 (8 / 41)	80.5 (33 / 41)

Table 11.1 has some overlap in presented data with Table 4.1; specifically, the frequencies for undetermined results. However, the additional columns show other breakdowns of the cycle threshold values within the data.

There are only 8 technical replicates with Ct values that are above the newly specified limit of detection (Part A, Table 11.1).[13] As such, the columns `Bio: All TR Above LOD` (Part B) and `Days: All BR Above LOD` (Part C) are nearly identical to their respective columns for undetermined results.

The Part A column for `Tech: Below LOD & Above LOQ` shows the frequency of ct values falling between the limits of detection and quantification, a gray area with respect to distinguishing concentrations of genetic material. The Part A column for `Tech: Below LOQ` shows the frequency of ct values falling below the limit of quantification, a favorable range. The similarly names columns within Parts B and C have a bit different approach in their frequency calculations and, consequently, their interpretations. For Part B `Bio: All TR Above LOQ`, the frequencies refer to biological replicates where ***all*** technical replicates ct values fell above the limit of detection. The `Bio: Non-Miss Below LOQ` column gives the frequencies of biological replicates that have at least one technical replicate with a ct value below the limit of quantification. The Part C columns similarly give aggregations of biological replicates instead of technical replicates as in Part B. Note that in Parts B and C, the columns from left to right, or from all undetermined to all above LOQ correspond to increasingly inclusive thresholds so that the frequencies for all above LOQ are always

[13] Normally, values *below* the limit of detection would be of undetectable. However, since we are referring to the cycle thresholds, the values *above* the limit of detection are of interest. These higher Ct values correspond to lower concentrations of substrate or genetic material within the sample.

greater or equal to the other two columns. Also, for Parts B and C the frequencies in the All above LOQ and Others columns sum to account for the total number of replicates or days.

11.1 An Alternative Summary Table

The Part A column for `Tech: Below LOD & Above LOQ` shows the frequency of ct values falling between the limits of detection and quantification, a gray area with respect to distinguishing concentrations of genetic material. The Part A column for `Tech: Below LOQ` shows the frequency of ct values falling below the limit of quantification, a favorable range. The similarly names columns within Parts B and C have a bit different approach in their frequency calculations and, consequently, their interpretations. For Part B `Bio: Non-Miss Below LOD & Above LOQ`, the frequencies refer to biological replicates where ***all*** technical replicates ct values fell between the limits. The denominators given for the relative frequencies correspond to the number of biological replicates excluding the ones where all technical replicates yielded undetermined results. The `Bio: Non-Miss Below LOQ` column gives the frequencies of biological replicates for which all technical replicates yielded ct values below the limit of quantification. Note that these two instances of values either purely between the limits or below the LOQ do not account for all the biological replicates excluding the ones where all technical replicates yielded undetermined results. The Part C columns similarly give aggregations of biological replicates instead of technical replicates as in Part B.

12 Fitting of Standard Curves

Table 12.1: Standard Curve Linear Regression Fits

target	CN	(Intercept)	log10(quantity)	r2	Efficiency
N1					
	10	32.623	-2.659	0.691	1.378
	12	30.855	-2.816	0.994	1.266
	13	**34.008**	**-3.389**	**0.993**	**0.973**
	14	30.606	-2.472	0.977	1.538
	15	33.367	-2.435	0.981	1.575
	16	35.258	-3.204	0.947	1.052
	20	32.509	-3.039	0.965	1.133
	21	31.266	-2.506	0.994	1.506
	22	33.2	-3.049	0.936	1.128
	23	33.975	-2.71	0.987	1.339

target	CN	(Intercept)	log10(quantity)	r2	Efficiency
	24	32.384	-3.25	0.992	1.031
	29	35.005	-3.483	0.989	0.937
	30	30.721	-1.994	0.7	2.173
	31	33.013	-2.765	0.986	1.3
	33	30.691	-2.493	0.973	1.519
	34	32.963	-2.981	0.996	1.165
	35	31.295	-2.928	0.999	1.196
	36	30.038	-2.458	0.972	1.551
	37	30.78	-2.673	0.997	1.367
	MEAN (SD)	32.345 (1.57)	-2.805 (0.379)	0.951 (0.092)	1.322 (0.29)
N2					
	10	34.134	-3.154	0.982	1.075
	11	33.089	-3.312	0.998	1.004
	12	31.46	-2.882	0.983	1.223
	13	33.477	-2.992	0.945	1.159
	14	33.32	-2.813	0.961	1.267
	15	35.373	-2.63	0.968	1.4
	16	32.82	-2.708	0.984	1.341
	17	32.53	-3.138	0.992	1.083
	18	37.496	-4.384	0.976	0.691
	19	33.097	-3.195	0.999	1.056
	20	32.783	-2.901	0.961	1.212
	21	30.683	-2.367	0.991	1.646
	22	31.951	-2.889	0.953	1.219
	23	33.795	-2.932	0.975	1.193
	24	**32.416**	**-3.308**	**0.994**	**1.006**

target	CN	(Intercept)	log10(quantity)	r2	Efficiency
	25	31.315	-2.515	0.984	1.498
	26	32.678	-2.74	0.974	1.317
	27	33.7	-2.877	0.98	1.227
	28	33.7	-2.877	0.98	1.227
	29	37.782	-3.501	0.999	0.93
	30	29.694	-2.168	0.745	1.892
	31	31.109	-2.973	0.933	1.169
	32	34.535	-3.402	0.973	0.968
	33	30.317	-2.586	0.965	1.436
	34	31.984	-2.999	0.995	1.155
	35	30.986	-2.869	0.997	1.231
	36	29.573	-2.485	0.963	1.526
	37	30.836	-2.874	0.997	1.228
	MEAN (SD)	32.737 (2)	-2.946 (0.418)	0.969 (0.047)	1.228 (0.236)

The rows of Table 12.1 for target N1, collection number 13 and target N2, collection number 24 are bold as they were selected as the ideal candidates for the calibration. Briefly, these single runs were chosen due to the variability in the standard curves as a whole. The variability in the data is not representative of the process itself, rather due mainly to the quality of the samples used in the experiment; the samples with known concentrations of substrate are known to degrade over time, yielding inconsistent results.

13 Data Adjustment using Limits of Detection and Quantification

It is relatively common to use the limit of detection to replace undetermined results within the data. So, I have concocted a few scenarios and created a few datasets varying on the replacement of undetermined values within the dataset. Briefly, scenario 1 uses the data as is without replacing undetermined values. Scenario 2 uses half of the limit of detection to replace undetermined values and half the limit of quantification to replace the data with concentrations below that limit. Scenario 3 again uses half of the limit of detection for undetermined, but also scales values between the LOD and LOQ to stretch from the LOQ to the LOD/2. Scenario 4 replaces undetermined values with the LOD scaled to incorporate the relative frequency of undetermined results and, then, scales the values between the LOD and LOQ to the missing frequency-scaled LOD. Scenario 5 does not replace

undetermined values, but does scale the values below the LOQ to half of the LOD. Figure 13.1 shows histograms of each of the datasets.

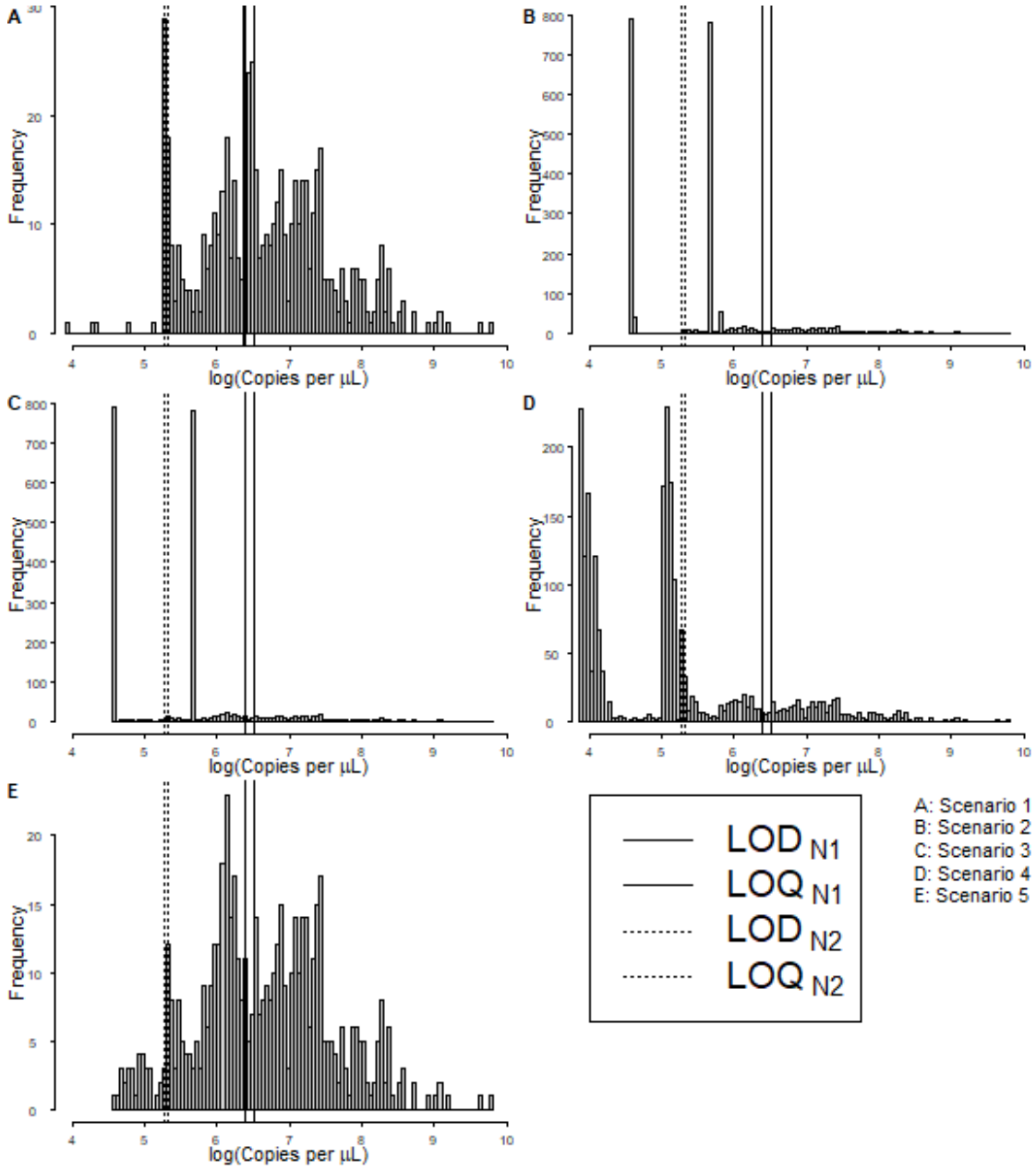
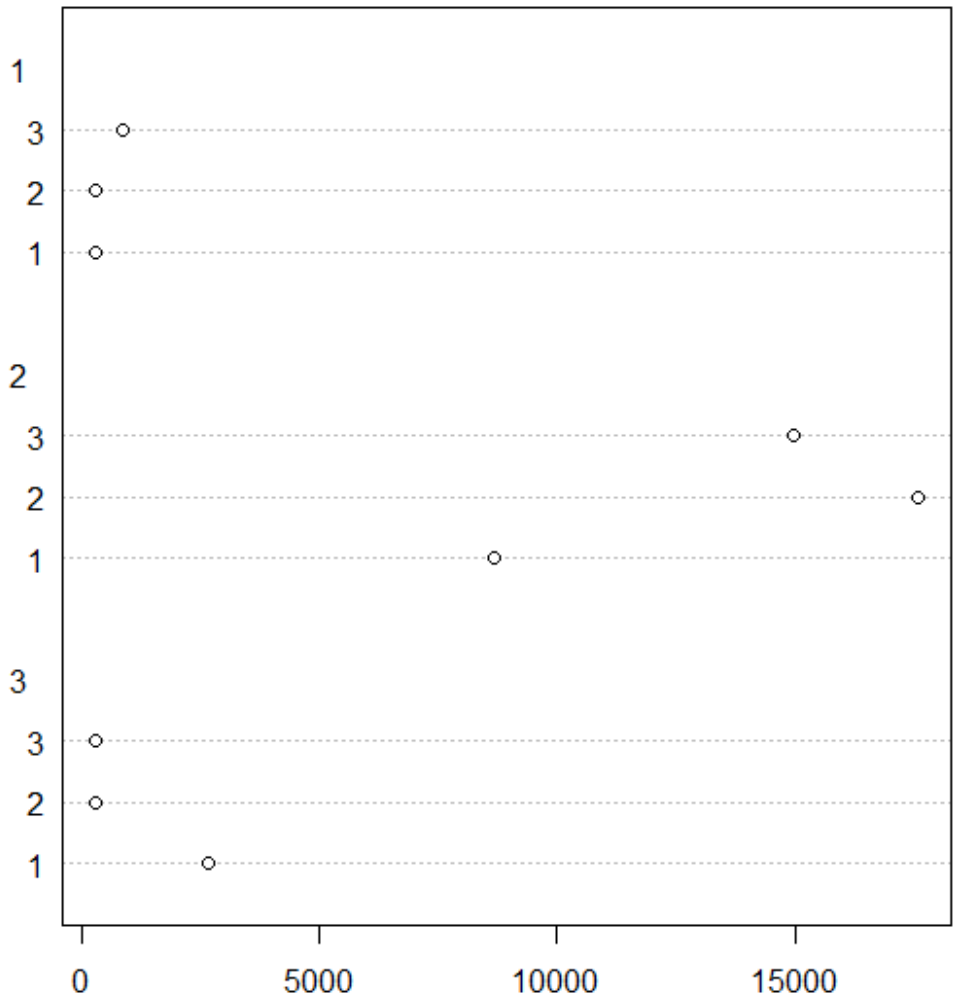
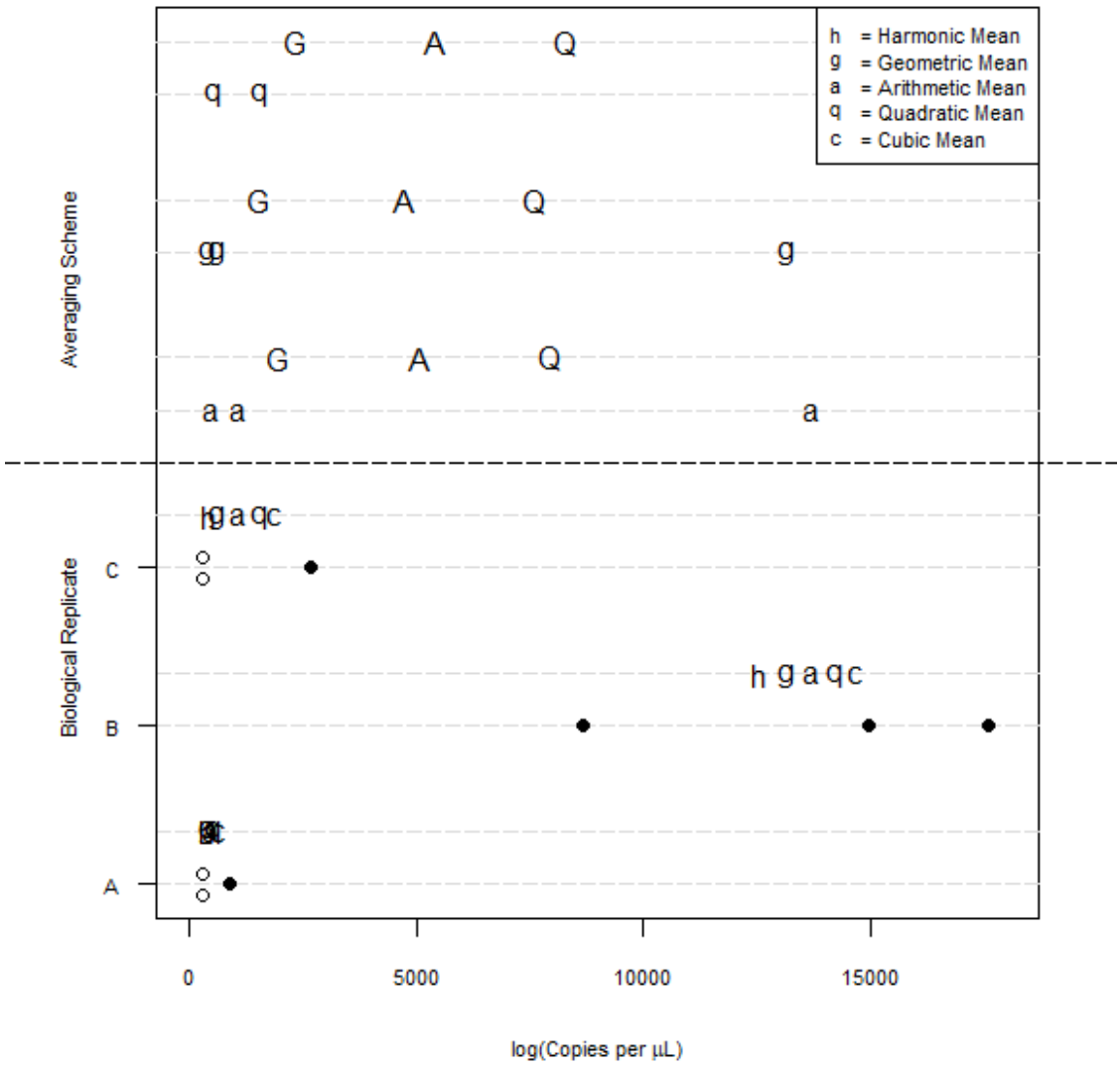


Figure 13.1: Histograms of the Natural Logarithm of Sample Viral Sequence Copy Concentrations

14 Averages Calculations


1
3
2
1
2
3
2
1
3
3
2
1
0
5000
10000
15000


h = Harmonic Mean
g = Geometric Mean
a = Arithmetic Mean
q = Quadratic Mean
c = Cubic Mean
Averaging Scheme
Biological Replicate
C
B
A
0
5000
10000
15000
log(Copies per µL)

15 Extra Code

15.1 Copies and Normality

15.2 Autocorrelation

15.3 Scatterplots and Correlations

15.4 Extensive Plotting

15.5 All the Plots

These many plots vary with respect to their x-variables (either copy concentration, or copy number), y-variables (raw case counts, 7-day moving averages, deconvoluted case counts), and the level of the dataset (facility level or sampling day level).

15.6 Exploring Model Fits

15.7 Distributed Lag (Lead) Models

To further investigate the temporal association of wastewater viral loads and case incidence, a distributed lag (lead) model will be fit to the data in a regression of cases on wastewater viral loads at various leads (Equation (15.1) ; adapted from Peccia et al, 2020).

$$Cases = \beta_0 + \sum_{j=-1}^{d} \beta_j \, VL_{t-j} \qquad (15.1)$$

where *Cases* refers to the case counts, *d* refers to the lead times, and *VL* refers to he wastewater viral loads at time *t*.

Since the smoothed case counts are not discrete, a Poisson regression may not be appropriate. Instead, I fit a generalized linear model for a Gamma distribution with a natural logarithm link function.

15.8 Ramblings

15.8.1 Concepts of Sensitivity

Similar to Bayesian Framework discussed by Kyle Curtis during the Wastewater-based Epidemiology Researchers Collaboration Network Webinar on 31 March 2021.

$$Pr(+Sample \mid +RT-qPCR) = \frac{Pr(+RT-qPCR \mid +Sample) \times Pr(+Sample)}{[Pr(+RT-qPCR \mid +Sample) \times Pr(+Sample)] + [Pr(+Rt-qPCR \mid -Sample) \times Pr(-Sample)]}$$

It is important that we consider the sources of these water samples and how heterogeneity in these sources may impact downstream analyses. The wastewater reclamation facilities, as sources of water samples, may differ in at least two potentially meaningful ways: the facility itself (e.g., size, processing, infrastructure, …) and the origin of the water / sewage

that is processed. Admittedly, the demarcation between facility and water characteristics may be impertinent as they are likely interrelated (e.g., size may reflect volume totals of influent water / sewage). However, considerations may still be warranted for their potential impacts and artifacts on data.

Ultimately, what characteristics of the wastewater reclamation facilities and service areas ***could*** affect water samples and, subsequently, their analysis?

Well, before considering the differences among facilities, let us first explore the sampling of water / sewage. Ideally, the collected samples would be representative of the water / sewage as a whole. What could prevent a sample from being representative? Using sparse chemistry knowledge and definitions, the water / sewage is likely more a mixture than a solution and, as such, the mixture may not be homogeneous (i.e., well mixed). So, a small sample may not be representative of the whole. The samples are, however, composites from a 24-hour period which may alleviate this as well as concerns from a cross-sectional perspective.

With these sample and procedural impacts highlighted, the potential differences among facilities could potentially exacerbate these differences. For example, if Facility A collects larger volumes of water / sewage than Facility B, then we may expect the samples from Facility A to be drawn from a more homogeneous (well-mixed) source (I'm picturing a river with a waterfall versus a creek). There may be additional sources of heterogeneity due to a facility itself (e.g., water processing and point at which samples are taken), but at this point it is more of a thought experiment than empirical.

Perhaps more profound in their effects on sample heterogeneity, these facilities have different service areas. The facilities effectively divide the county in three with respect to the service boundaries and the sewer networks. It is not much of a stretch to assume that the service areas differ in served population size and characteristics. So, perhaps the water samples are representative of these "subpopulations" and data analysis should consider them explicitly. Furthermore, due to potential differences in landscapes and built environments of service areas, weather events such as rainfall may be important to consider as sources of "error" in measurement.