Estimating Unobserved COVID-19 Infections in the United States

Quinn White

Submitted to the Department of Statistical and Data Sciences
of Smith College
in partial fulfillment
of the requirements for the degree of
Bachelor of Arts

Ben Baumer, Primary Faculty Advisor
Nicholas Reich, Secondary Faculty Advisor

May 2023

Acknowledgements

Will add

Table of Contents

Abstract

As we have navigated the COVID-19 pandemic, case counts have been a central source of information for understanding transmission dynamics and the effect of public health interventions. However, because the number of cases we observe is limited by the testing effort in a given location, the case counts presented on local or national dashboards are only a fraction of the true infections. Variations in testing rate by time and location impacts the number of cases that go unobserved, which can cloud our understanding of the true COVID-19 incidence at a given time point and can create biases in downstream analyses. Additionally, the number of cases we observe is impacted by the sensitivity and specificity of the diagnostic test. To quantify the number of true infections given incomplete testing and diagnostic test inaccuracy, this work implements probabilistic bias analysis at a biweekly time scale from January 1, 2021 through February 2022. In doing so, we can estimate a range of possible true infections for a given time interval and location. This approach can be applied at the state level across the United States, as well as in some counties where the needed data are available.

Dedication

You can have a dedication here if you wish.

Chapter 1

Motivation

Placeholder

Chapter 2

Background

Placeholder

2.1 Probabalistic Bias Analysis

2.2 Background for the Approach

2.2.1 Simple Discrete Example

2.2.2 General Solution for the Discrete Case

2.2.3 General Solution for the Continuous Case

2.2.4 Implementation through the Sampling-Importance-Resampling Algorithm

2.3 Bayesian Melding Applied to COVID-19 Misclassification

2.3.1 Distribution of $\theta = \{\alpha, \beta, P(S_1|\textbf{untested})\}$

2.3.2 Direct Prior and Induced Prior Distributions for $P(S_0|\textbf{test}_+, \textbf{untested})$

2.3.3 Pooling

2.3.4 Derivation of M

2.4 Sampling-Importance-Resampling Algorithm

2.4.1 Overview

2.4.2 Proof that Algorithm Obtains Approximate Sample from Target Distribution

Example 1:

Example 2:

2.4.3 Obtaining Logarithmic Pooled Distribution with the Sampling-Importance-Resampling Algorithm

2.4.4 Implications of the Sample Size and Resample Size

2.5 LOESS Smoothing

2.5.1 Introduction

2.5.2 Fitting the LOESS Curve

2.6 Kernel Density Estimation

2.6.1 Overview

Chapter 3

Definition of Prior Distributions for Bias Parameters

Placeholder

3.1 Background on the Beta Distribution

3.2 Background on the Gamma Distribution

3.3 Definition of Prior Distributions for Incomplete Testing Correction

3.3.1 Defining $P(S_1|Untested)$

3.3.2 Defining α

3.3.3 Defining β

3.3.4 Defining $P(S_0|test+, untested)$

3.4 Definition of Priors for Test Inaccuracy Correction

3.4.1 Defining Test Sensitivity (S_e)

3.5 Defining Test Specificity (S_p)

3.6 Summary Table of Bias Parameter Distributions

3.7 Correction for Incomplete Testing

3.8 Correction for Diagnostic Test Inaccuracy

3.8.1 Derivation of Formula for Correction for Diagnostic Test Inaccuracy

Chapter 4

Details of Implementation

- Describe each step here
- Mention reproducible workflow with make

4.1 Reproducible Workflow

Chapter 5

Results

5.1 Comparison to the Covidestim Model

5.1.1 Overview

One challenge in correcting for biases in general is that although we may have some information about the influence of possible biases, we do not have a ground truth for comparison. However, one approach to handle the fact that the true cases are unobserved is comparing our estimates to those from other approaches seeking to estimate a similar quantity. In particular, if other approaches make different assumptions and come to a similar result, this can give us more confidence in our estimates.

One notable project seeking to estimate the true infection burden at the county-level over time is the COVIDestim project. In this work, Chitwood et al. proposed a mechanistic model that includes states for asymptomatic/pre-symptomatic infection, symptomatic but mild infection, severe COVID-19 presentations, and death. This approach also enables the estimation of R_t, the number of secondary infections a single infected individual causes at time t. This is a useful quantity to estimate, but is sensitive to reporting delays and changes in testing practices (**pitzer2021a?**).

5.1.2 The Covidestim Model

Chitwood *et al.* propose a Bayesian evidence synthesis model to correct for reporting delays and time varying case ascertainment testing rate in the estimation of incident infections and R_t.

To estimate the expected cases and deaths at a particular point in time, the model uses a convolution of the time series of observed cases and deaths and reporting delay distributions that are specific to the health state categories. This enables the model to account for the fact that reporting delay is different For any health state, for example, asymptomatic, the individual can either transition to the next health state (symptomatic) or recover. Thus, with each transition between a defined health state, for example, asymptomatic, there is a probability of transitioning to the next health state (in this case, asymptomatic to symptomatic); the complement of this probability is the probability of recovery.

Each of these transitions is defined by a delay distribution. For example, the distribution for moving from asymptomatic to symptomatic represents the probability an individual moves to the symptomatic state at a point in time. The probabilities asymptomatic to symptomatic and symptomatic to severe are modeled as not varying with time. Meanwhile, the probability of transitioning from severe to death was defined to be higher in 2020 due to higher case fatalities early in the pandemic. The infection fatality rates, adjusted to be specific to a given state or county based on age distributions and the prevalence of risk factors for COVID-19, are used to inform the probability of moving from the severe category to the death category.

The change in daily infections from the previous day (i.e., the new infections) is calculated as a function of the estimated effective reproductive number R_t and the mean serial interval, where serial interval is the time from the onset of infection of a primary case to the time of onset of infection in the secondary case. R_t is estimated using a log-transformed cubic spline, under the assumption individuals can only be infected once.

They also defined a distribution for the delay to diagnosis, which was distinct by health state category to reflect differences in diagnosis delays that occur depending on the disease severity. The probability of diagnosis among different health states was allowed to vary by time to reflect changing testing rates throughout the pandemic.

A separate distribution models the reporting delay to correct the total number of diagnoses on a given day for the fact that these diagnoses correspond to past infections.

The observed cases and death data for each state to the model were fitted using negative binomial likelihood functions.

5.1.3 Assumptions

This approach relies on infection fatality ratios and death counts to estimate the true case counts. Thus, it is sensitive to estimates of infection fatality rate, with higher infection fatality ratio estimates resulting in lower estimated infections. The infection fatality ratio is defined as the proportion of COVID-19 infections that lead to death, which means there is uncertainty in estimating both the numerator and the denominator of the ratio. The true cumulative incidence depends on the same uncertainties in estimating the true case burden at any point in time. Estimating the infection fatality ratio itself is a challenging task.

The COVIDestim model uses age-specific estimates of IFR produced by (**odriscoll2021?**). This group used national-level age-stratified, and when possible sex-stratified, COVID-19 death counts and cumulative infection estimates from seroprevalence studies. Of note, the estimates of infection fatality ratio are assumed to be constant over time, which may not be the case due to improving treatments (e.g., Paxlovid), different variants leading to less severe presentations, or changes in the demographics of individuals being infected. However, reliable estimates of infection fatality ratio that vary with time difficult to acquire; COVIDestim assumed a higher case fatality in 2020 given the novelty of the virus and consequent lack of available treatments.

5.1.4 The Model

Chitwood et al. propose a Bayesian evidence synthesis model. To estimate the expected cases and deaths at a particular point in time, the model uses a convolution of the time series of observed cases and deaths and reporting delay distributions that are specific to the health state categories. This enables the model to account for the fact that reporting delay is different For any health state, for example, asymptomatic, the individual can either transition to the next health state (symptomatic) or recover. Thus, with each transition between a defined health state, for example, asymptomatic, there is a probability of transitioning to the next health state (in this case, asymptomatic → symptomatic); the complement of this probability is the probability of recovery.

Each of these transitions is defined by a delay distribution. For example, the distribution for moving from asymptomatic to symptomatic represents the probability an individual moves to the symptomatic state at a point in time. The probabilities asymptomatic to symptomatic and symptomatic to severe are modeled as not varying with time. Meanwhile, the probability of transitioning from severe to death was defined to be higher in 2020 due to higher case fatalities early in the pandemic. The infection fatality rates, adjusted to be specific to a given state or county based on age distributions and the prevalence of risk factors for COVID-19, are used to inform the probability of moving from the severe category to the death category.

The change in daily infections from the previous day (i.e., the new infections) is calculated as a function of the estimated effective reproductive number R_t and the mean serial interval, where serial interval is the time from the onset of infection of a primary case to the time of onset of infection in the secondary case. R_t is estimated using a log-transformed cubic spline, under the assumption individuals can only be infected once.

They also defined a distribution for the delay to diagnosis, which was distinct by health state category to reflect differences in diagnosis delays that occur depending on the disease severity. The probability of diagnosis among different health states was allowed to vary by time to reflect changing testing rates throughout the pandemic.

A separate distribution models the reporting delay to correct the total number of diagnoses on a given day for the fact that these diagnoses correspond to past infections.

The observed cases and death data for each state to the model were fitted using negative binomial likelihood functions.

5.1.5 Comparison to Serological Data

There are known issues with seroprevalence estimates. For one, these samples are drawn from a convenience (i.e. nonrandom) sample of individuals with blood specimens taken for purposes other than COVID-19 antibody detection (**zotero-1003?**). Secondly, while a positive serological test is evidence for infection, a negative serological test is less clear to interpret. The person may have been infected but not yet have developed antibodies, or their immune system may not have produced antibodies at a detectable level [(**cdc2020?**).

Indeed, Chitwood et al. found limited concordance between their estimates and seroprevalence data. However, there was a stronger correlation between estimates of cumulative infection and cumulative hospitalizations and cumulative deaths [1].

5.1.6 Limitations of this Comparison

At this point in the pandemic, there is no true gold standard to compare to. Covidestim is one model, among many, that makes key assumptions about aspects of the virus. Another note is that estimates from the Covidestim model are reported on the daily timescale for counties, while the probabilistic approach we implemented here is at the biweekly time scale.

To ensure the comparisons are on the same time scale, we sum the reported 95% credible intervals for the days in each 2-week interval. These intervals do not represent a 95% credible interval for the 2-week interval, and while such an interval would be ideal for the comparison, computation of a 95% credible interval for the two-week interval is not feasible because of the model structure. Due to the correlation between observations for each day for a given location, summing the intervals yields an estimate that is likely to be more conservative than a true 95% credible interval for the two-week interval would be. More detail on this assumption is in the appendix.

5.2 County-level Results

We performed county-level probabilistic bias analysis for Michigan and Massachusetts. This work can be expanded to consider other states as well where the needed data is available. In particular, we need both county-level positive PCR tests and county-level total PCR tests. Because the assumptions of the bias correction are related to test positivity, it does not make sense to apply the method to a positive cases count that includes positive PCR tests lumped together with probable cases. In some states, this is the only value reported.

5.2.1 Massachusetts

Figure 5.1 shows the bias corrected estimates for each implementation, as well as the observed cases. We note that the lower bounds of the bias corrected estimates are

[1] The correlation employed here is the Spearman rank correlation, which measures the strength of the monotonic relationship rather than the strength of the linear relationship, in which case the Pearson correlation coefficient is the usual choice. The Spearman rank correlation is equivalent to the Pearson correlation of the rank values rather than the values themselves.

always above the observed cases because adding (unobserved) infections among the untested population to the observed positives among the population never results in a decrease in the estimated infections. In theory such a decrease could be possible since we do correct for differences due to imperfect test accuracy, and if the false positive rate was high enough, we might estimate the lower bound of cases as lower than the observed cases. However, the false positive rate of the COVID-19 PCR test is so low that in practice we do not see lower bounds lower than the number of infections.[2]

We can see although the trends are broadly similar between versions for each county, centering the distribution at the empirical value of β leads to peaks not present in the version where priors do not vary by county and date. However, only centering $P(S_1|\text{untested})$ at the empirical value leads to a distribution that is highly similar to the version where priors do not vary by county and date.

These results make sense when we consider that this analysis is much more sensitive to the choice of β than $P(S_1|\text{untested})$. This follows from the fact we compute the number of positive infections among those who are untested and *asymptomatic* as

$$\begin{aligned} N^{+}_{untested,S_0} &= P(\text{test}_{+}|S_0,\text{untested})(N_{S_0,\text{untested}}) \\ &= \Big(\beta P(test_{+}|tested)\Big) N_{\text{untested}}(1 - P(S_1|\text{untested})) \end{aligned}$$

and the number of positive infections among those who are untested and *symptomatic* as

$$\begin{aligned} N^{+}_{\text{untested},S_1} &= P(\text{test}_{+}|S_1,\text{untested})(N_{S_1,\text{untested}}) \\ &= \Big(\alpha P(\text{test}_{+}|\text{tested})\Big) N_{\text{untested}}(P(S_1|\text{untested})). \end{aligned}$$

Since $N_{S_1,\text{untested}}$ is so much larger than $N_{S_1,\text{untested}}$ for any of the specified values of $P(S_1|\text{untested})$ (since the bulk of this distribution is less than 5%), β has a larger impact on the number of estimated infections.

[2] The false positive rate differs by platform and laboratory, but multiple analyses estimated that it is less than 0.10% (**chandler2021?**).

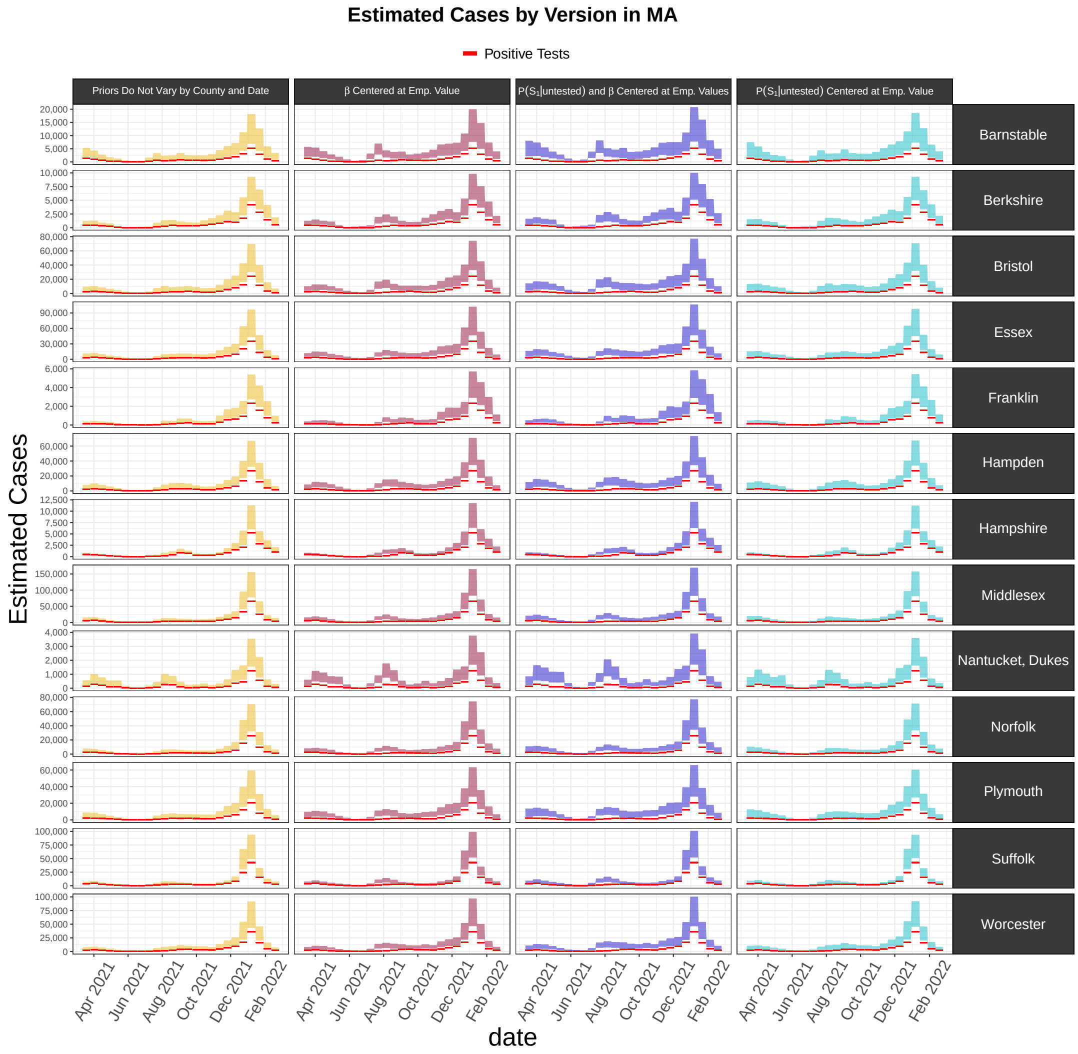


Figure 5.1

To better see the overlap between versions, in Figure 5.2 we can look at the versions together. This allows us to see more clearly how the version with both $P(S_1|\text{untested})$ and β centered at their empirical values is consistently the highest. Meanwhile, the version with only $P(S_1|untested)$ centered at its empirical value corresponds so closely to the version that does not vary by date or location that there is no part of the intervals for the version not varying by date or location that do not overlap with the $P(S_1|\text{untested})$ version.

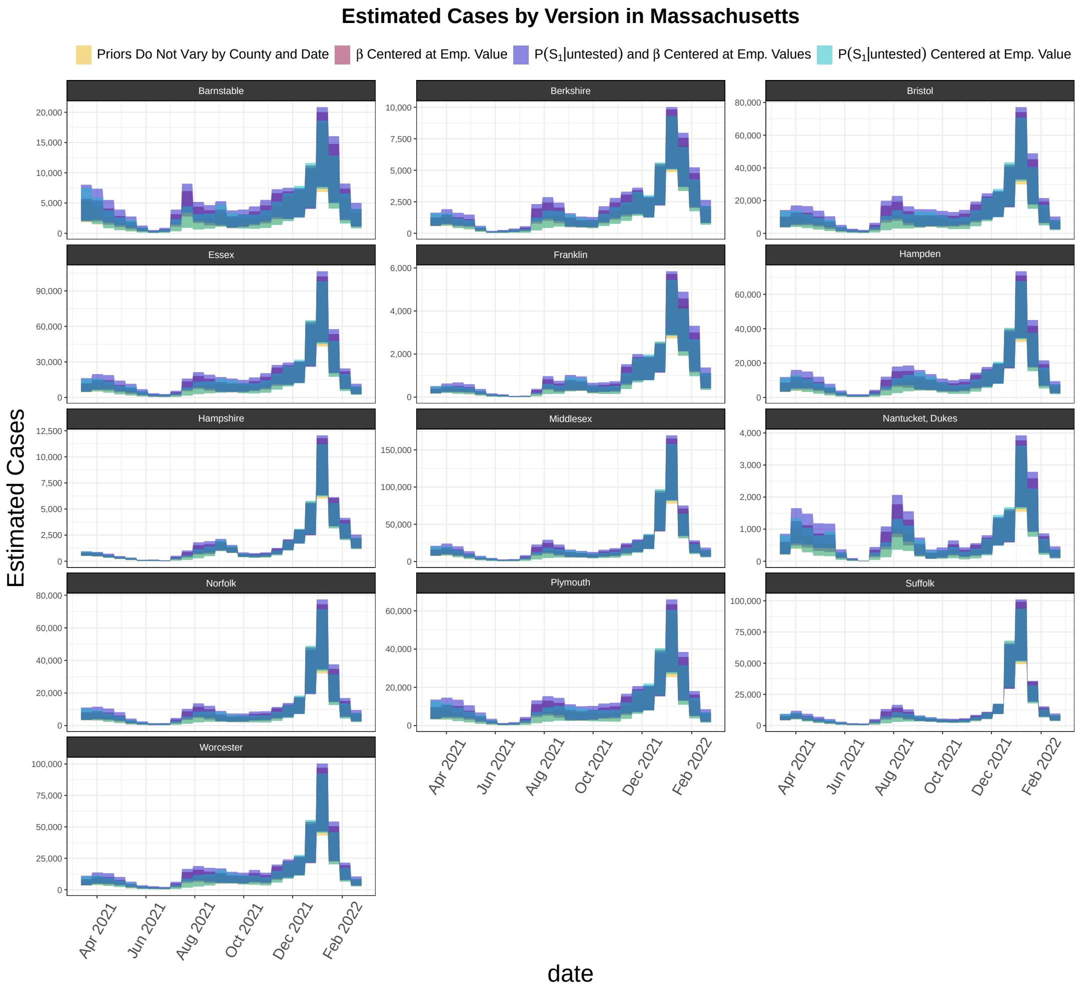

Estimated Cases by Version in Massachusetts
Priors Do Not Vary by County and Date
β Centered at Emp. Value
P(S$_1$|untested) and β Centered at Emp. Values
P(S$_1$|untested) Centered at Emp. Value
Barnstable
0
5,000
10,000
15,000
20,000
Berkshire
0
2,500
5,000
7,500
10,000
Bristol
0
20,000
40,000
60,000
80,000
Essex
0
30,000
60,000
90,000
Franklin
0
2,000
4,000
6,000
Hampden
0
20,000
40,000
60,000
Hampshire
0
2,500
5,000
7,500
10,000
12,500
Middlesex
0
50,000
100,000
150,000
Nantucket, Dukes
0
1,000
2,000
3,000
4,000
Norfolk
0
20,000
40,000
60,000
80,000
Plymouth
0
20,000
40,000
60,000
Suffolk
0
25,000
50,000
75,000
100,000
Worcester
0
25,000
50,000
75,000
100,000
Apr 2021
Jun 2021
Aug 2021
Oct 2021
Dec 2021
Feb 2022
Estimated Cases
date

Figure 5.2

5.2.2 Michigan

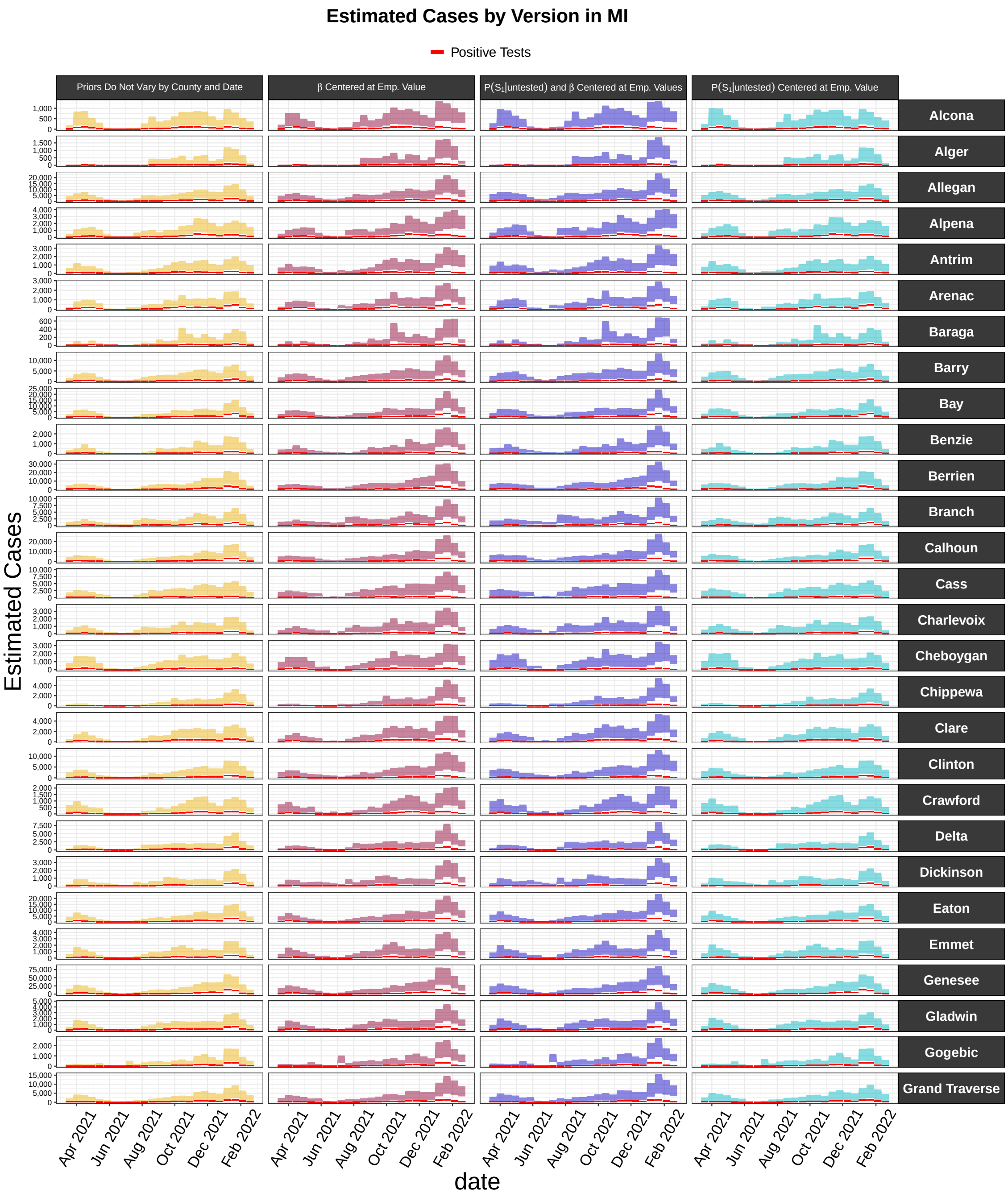


Figure 5.3

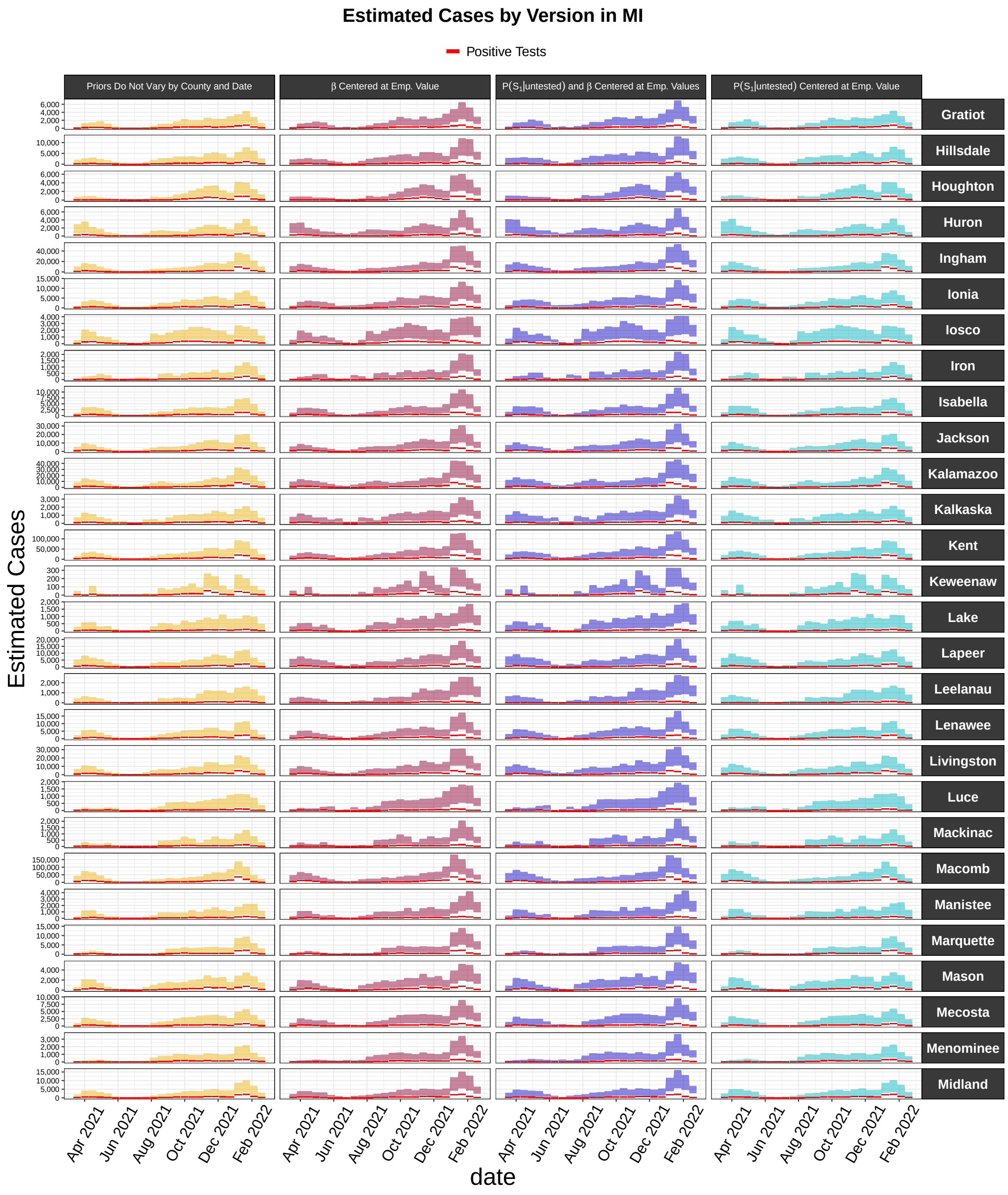

Estimated Cases by Version in MI
Positive Tests
Priors Do Not Vary by County and Date
β Centered at Emp. Value
P(S$_1$|untested) and β Centered at Emp. Values
P(S$_1$|untested) Centered at Emp. Value
Gratiot
Hillsdale
Houghton
Huron
Ingham
Ionia
Iosco
Iron
Isabella
Jackson
Kalamazoo
Kalkaska
Kent
Keweenaw
Lake
Lapeer
Leelanau
Lenawee
Livingston
Luce
Mackinac
Macomb
Manistee
Marquette
Mason
Mecosta
Menominee
Midland
Estimated Cases
6,000 4,000 2,000 0
10,000 5,000 0
6,000 4,000 2,000 0
6,000 4,000 2,000 0
40,000 20,000 0
15,000 10,000 5,000 0
4,000 3,000 2,000 1,000 0
2,000 1,500 1,000 500 0
10,000 7,500 5,000 2,500 0
30,000 20,000 10,000 0
40,000 30,000 20,000 10,000 0
3,000 2,000 1,000 0
100,000 50,000 0
300 200 100 0
2,000 1,500 1,000 500 0
20,000 15,000 10,000 5,000 0
2,000 1,000 0
15,000 10,000 5,000 0
30,000 20,000 10,000 0
2,000 1,500 1,000 500 0
2,000 1,500 1,000 500 0
150,000 100,000 50,000 0
4,000 3,000 2,000 1,000 0
15,000 10,000 5,000 0
4,000 2,000 0
10,000 7,500 5,000 2,500 0
3,000 2,000 1,000 0
15,000 10,000 5,000 0
Apr 2021
Jun 2021
Aug 2021
Oct 2021
Dec 2021
Feb 2022
date

Figure 5.4

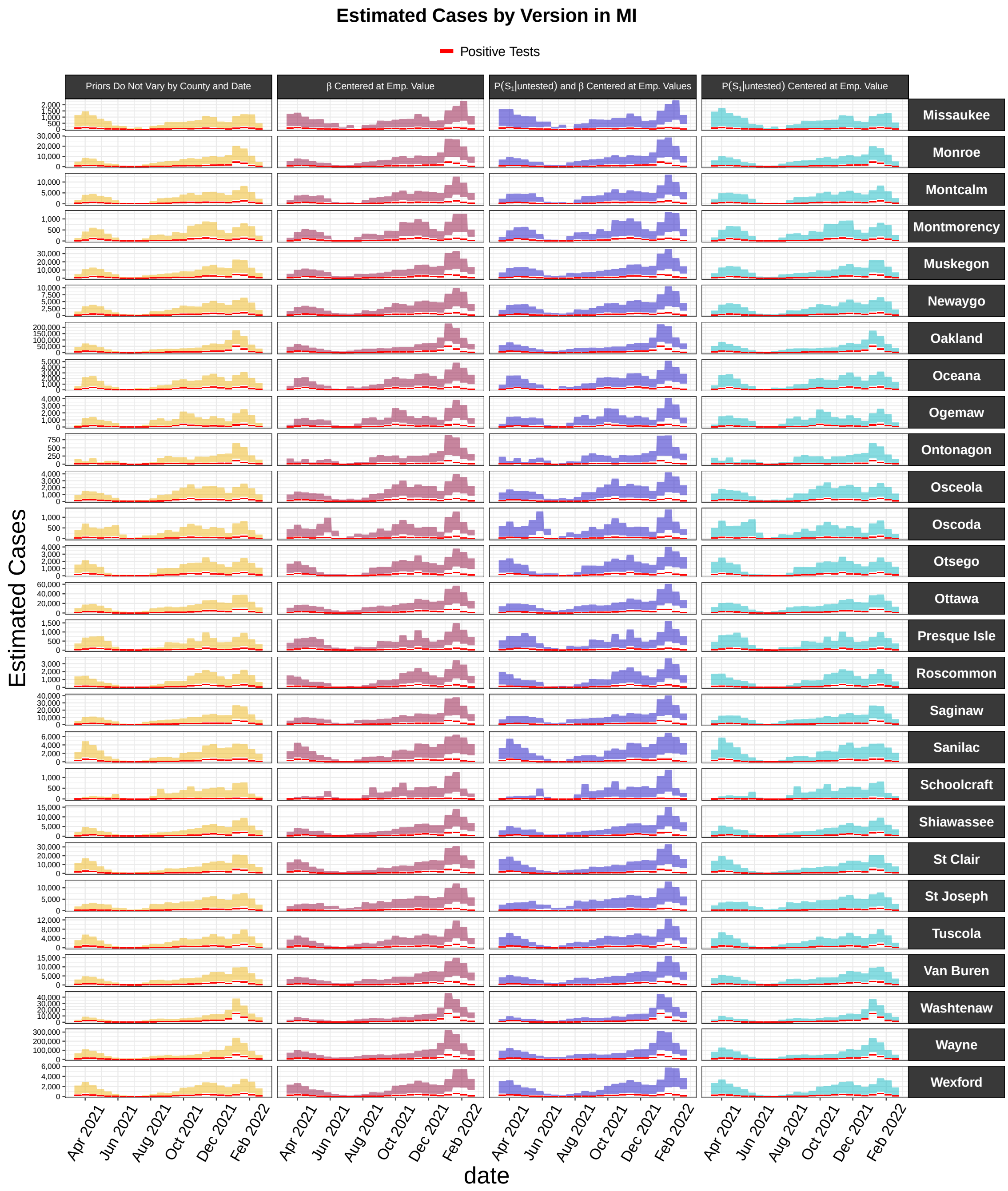


Figure 5.5

5.3 State-level Results

There is much more comprehensive testing data at the state-level than at the county level, as John Hopkins has tracked state-level testing data throughout the pandemic. As a result, we can apply probabilistic bias analysis across the United States at the state-level.

However, since versions 2-4 of the analysis utilize empirical estimates of $P(S_1|\text{untested})$ or β from the COVID-19 Trends and Impact Survey, these versions are only possible when there is sufficient data. As we see in Figure 5.6, we see that

For versions 2-4, we did not attempt the probabilistic bias analysis for states where more than 60% of observations were missing at the daily time step. Missing values for states with sufficient data were imputed before summarizing to the biweek level using a linear weighted moving average.

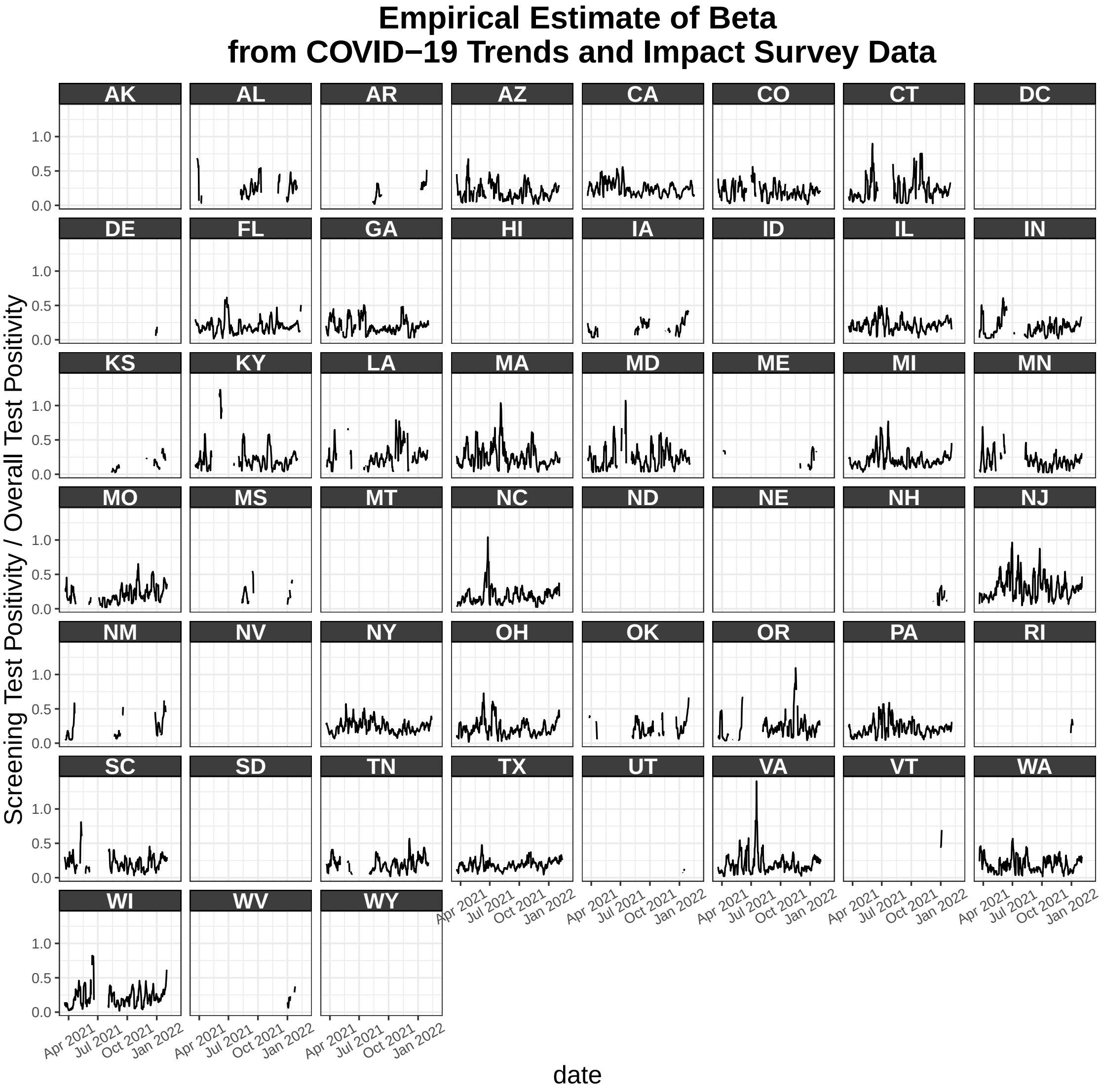


Figure 5.6

5.4 Cross Correlation Comparison

5.4.1 Background

An ongoing challenge for assessing the quality of the probabilistic bias intervals is that there is no ground truth to compare to. To broaden our comparison beyond Covidestim, we can look at wastewater data.

Wastewater data is a source of data that has been of rising interest throughout the pandemic, in part due to its cost effectiveness in assessing community-level burden, but also due to the fact it represents a much more unbiased sample than COVID-19 testing does.

That said, there are challenges in relating wastewater concentrations to the true number of infections, in part because of the same issue we face here of the lack of ground truth for the true number of infections in any location. The choice of normalization of the viral RNA concentrations of SARS-CoV-2 is important for understanding how these concentrations scale to the number of infections, since the concentration of virus (in genome copies per liter) in a sample will be influenced by various factor unrelated to the true prevalence COVID-19, such as processing differences between treatment plants or trends in water usage. One common choice is to normalize against the concentration of a virus that has a relatively stable population in wastewater, such as Pepper Mild Mottle Virus (PMMoV) (**zhan2022?**).

Wastewater testing has become increasingly widespread throughout the pandemic as the technology and analysis approaches have evolved, as well as the demand for a source of data on the presence of COVID-19 that is less reliant on access to tests (or symptoms strong enough to warrant a test, which differ by the variants circulating). A comprehensive source of wastewater data across the United States is provided by Biobot Analytics, which is the institution partnering with the CDC for the National Wastewater Surveillance System (NWSS) (**duvallet2022?**). Biobot Analytics provide wastewater concentrations aggregated at the county scale by using a weighted average of the concentrations at sampling locations within the county, weighted by the size of the corresponding sewershed populations. This data is publicly available on a public github repository.

Most notable for this work, several counties in Massachusetts have reported wastewater data for a substantial period throughout 2021 to 2022. This allows us to compare the bias-corrected estimates – as well as the Covidestim estimates – to the wastewater concentrations.

Wastewater concentrations are typically a leading indicator of observed cases, though there may be some variability in the lead time during different waves of the pandemic (**hopkins2023?**). In particular, the lead time was strongest in the earliest waves of the pandemic, and has since declined (**xiao2022?**). Various factors can create the changes we see in lead time over the course of the pandemic; for example, the lead time can be impacted by differences in viral shedding, diagnostic testing turnaround times, and testing capacity and behavior (**olesen2021?**).

Since the correlation between the time series as well as the lag at which the maximum correlation occurs are both of interest, we assessed the cross correlation between the series.

First, we define autocorrelation since the definition of cross correlation is very similar. The definition here uses the notation of (**shumway2011?**).

Definition: Autocorrelation

Denote the set of time points of a time series T. For any time series $(x_t)_{t \in T}$, we define the auto-correlation function (ACF) as

$$\rho_{XX}(\tau) = \frac{E[(X_{t+\tau} - \mu_{X_{t+\tau}})(X_t - \mu_{X_t})]}{sd(X_{t+\tau})sd(X_t)}.$$

Assuming second order stationarity[3], we have $\mu_{X_{t+\tau}} = \mu_{X_t}$ and similarly $\text{Var}(X_{t+\tau}) = \text{Var}(X_t)$, so we an simplify the expression for $\rho_{XX}(\tau)$ to yield

$$\rho_{XX}(\tau) = \frac{E[(X_{t+\tau} - \mu_X)(X_t - \mu_X)]}{Var(X)}.$$

The auto-correlation function $\rho_{XX}(\tau)$ measures the linear dependence between $X_{1+\tau}, \ldots X_n$ and $X_1, \ldots, X_{n-\tau}$, that is, the difference between the original time series and the time series shifted forward by τ time units.

We can extend this definition to quantify the linear relationship between distinct lagged time series $X_1, X_2, \ldots, X_t$ and $Y_1, Y_2, \ldots, Y_t$ by defining the cross correlation function. The function is only defined on two time series that are over the same time interval and sampled at the same frequency.

Definition: Cross-Correlation

We compute the cross-correlation function (CCF) as

$$\rho_{XY}(s,t) = \frac{E[(X_s - \mu_{X_s})(Y_t - \mu_{Y_t})]}{\sqrt{\text{Var}(X_s)\text{Var}(Y_t)}}.$$

Again assuming the series satisfy second-order stationarity, we have

$$\rho_{XY}(s,t) = \frac{E[(X_s - \mu_X)(Y_t - \mu_Y)]}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

The implementation of the cross correlation in base R (`stats::ccf`) assumes second order stationarity (**venables2002?**).

Looking at cross correlation can be useful in the sense that we can both consider the strength of correlation and the lag at which the correlation is maximized. Before presenting the cross correlation results of the county level time series, we can consider a more concrete example, where the lag is known.

[3] Second order stationarity is also referred to as weak stationarity, and implies that the mean, variance are constant over time and the autocovariance function depends only on the difference between time points.

In Figure 5.7, we consider simulated data where (Z_t) is (Y_t) lagged by 3 time units with noise added. We can see that Z_t and Y_t are not second-order stationary since the mean clearly is not constant over time. However, to stabilize the mean, we can apply first order differencing, where we take the differences between consecutive observations.

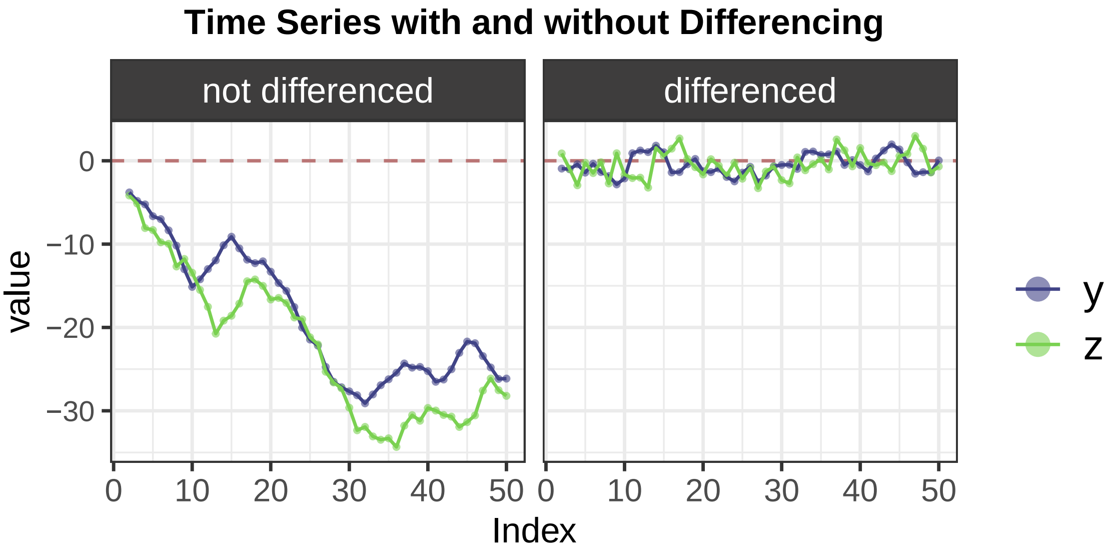
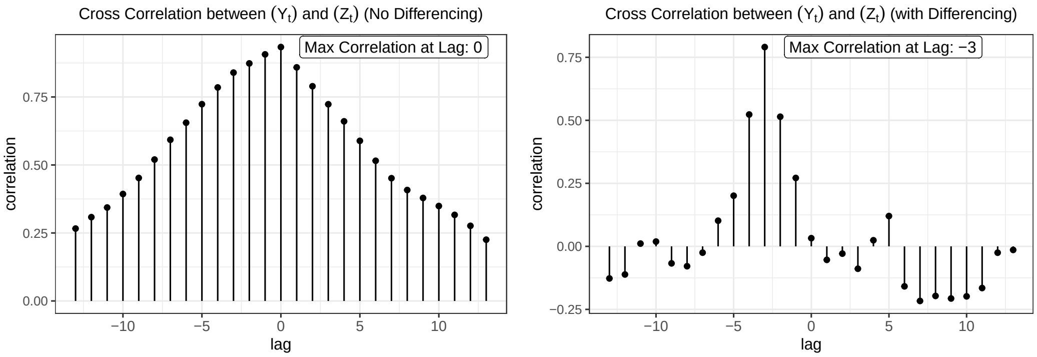


Figure 5.7

We can see the effect of applying differencing to the time series when we compute the cross correlations of (Z_t) and (Y_t), as shown in Figure 5.8. The true lag of -3 time units was recovered when considering the differenced time series, but not when we considered the original time series. In what follows, because the time series we are considering are not stationary, we consider the cross correlation between the differenced time series.


Cross Correlation between (Y_t) and (Z_t) (No Differencing)

Cross Correlation between (Y_t) and (Z_t) (with Differencing)

Figure 5.8

5.4.2 Cross Correlation Results Comparing Bias Corrected Counts, Covidestim Estimates, and Wastewater Concentrations

Because wastewater data is reported at the weekly time scale while the bias corrected estimates are at the 2-week time scale, we take a mean of the effective concentration for each 2 week interval, such that the time series are sampled at the same frequency.[4]

Since the effective concentration of SARS-CoV-2 in wastewater samples reported by Biobot is in genome copies per liter and is not directly comparable to estimates of infections, we place the wastewater concentration on a separate scale.

Looking at the counties in Figure 5.9, we see that, with the exception of Barnstable, MA, the wastewater trends are highly similar to trends captured by the bias corrected infection counts. We also see that the trends are similar both with regard to shape but also with regard to time, with little visible lag between the series. This is expected because although wastewater cases do in general lead cases, lead times generally are not on the order of 2 weeks. This means that since we are summarizing to 2-week intervals we would expect the lag to be very small, if present at all.

[4] We cannot interpret the cross correlation if the time steps are different.

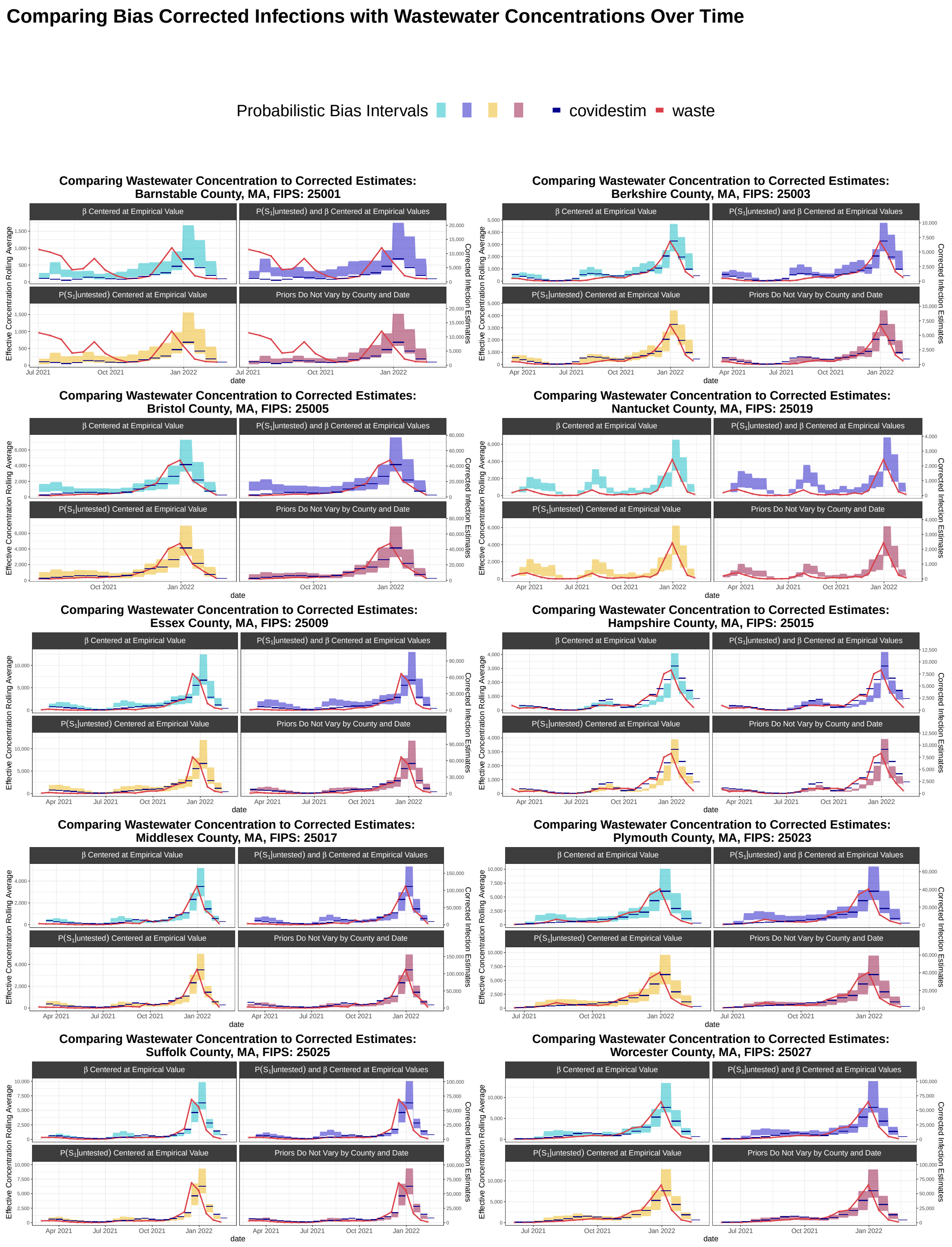

Comparing Bias Corrected Infections with Wastewater Concentrations Over Time
Probabilistic Bias Intervals
covidestim
waste
Comparing Wastewater Concentration to Corrected Estimates: Barnstable County, MA, FIPS: 25001
Comparing Wastewater Concentration to Corrected Estimates: Berkshire County, MA, FIPS: 25003
Comparing Wastewater Concentration to Corrected Estimates: Bristol County, MA, FIPS: 25005
Comparing Wastewater Concentration to Corrected Estimates: Nantucket County, MA, FIPS: 25019
Comparing Wastewater Concentration to Corrected Estimates: Essex County, MA, FIPS: 25009
Comparing Wastewater Concentration to Corrected Estimates: Hampshire County, MA, FIPS: 25015
Comparing Wastewater Concentration to Corrected Estimates: Middlesex County, MA, FIPS: 25017
Comparing Wastewater Concentration to Corrected Estimates: Plymouth County, MA, FIPS: 25023
Comparing Wastewater Concentration to Corrected Estimates: Suffolk County, MA, FIPS: 25025
Comparing Wastewater Concentration to Corrected Estimates: Worcester County, MA, FIPS: 25027
β Centered at Empirical Value
P(S1|untested) and β Centered at Empirical Values
P(S1|untested) Centered at Empirical Value
Priors Do Not Vary by County and Date
Effective Concentration Rolling Average
Corrected Infection Estimates
date
Apr 2021
Jul 2021
Oct 2021
Jan 2022

Figure 5.9

Comparison Between Implementations of Probabilistic Bias Analysis

In Figure 5.10, we see that, in general, infections were highly correlated with the wastewater effective concentrations, which was true across all implementations of probabilistic bias analysis. In most cases, the implementation where priors did not vary by state or date were the most highly correlated with the wastewater concentrations. Exceptions to this were Barnstable County (25001), where the implementation with the prior for β centered at the empirical value was the most highly correlated, and Worcester County (25027), where the implementation with the prior for $P(S_1|\text{untested})$ centered at empirical value was the most highly correlated. In all counties except for Barnstable, the lag at which the maximum correlation was obtained was 0 units, while for Barnstable it was -1, indicating that wastewater concentrations led infections by one two-week interval.

Given the small size of Barnstable relative to other counties and high variability in its early estimates in 2021 (as seen in Figure 5.9), it is possible that there were still aspects of the SARS-CoV-2 detection process that took time to refine. Another possibility is that the way Biobot aggregated wastewater concentrations by county failed to capture the infection dynamics in this county, since wastewater catchments are not contained within county lines. This is a central challenge in relating cases to wastewater concentrations, since these values are recorded for distinct geographic units.

Comparing the maximum correlations obtained the observed cases, only in Hampshire County were the observed cases more correlated with the wastewater concentrations than all implementations of probabilistic bias analysis. We also see again that in most cases the maximum correlation is obtained at zero lag in observed cases; however, for Barnstable, the correlation is highest when wastewater concentrations lead infections by two biweeks, and for Hampshire the correlation is highest when wastewater concentrations lead infections by 1 biweek.

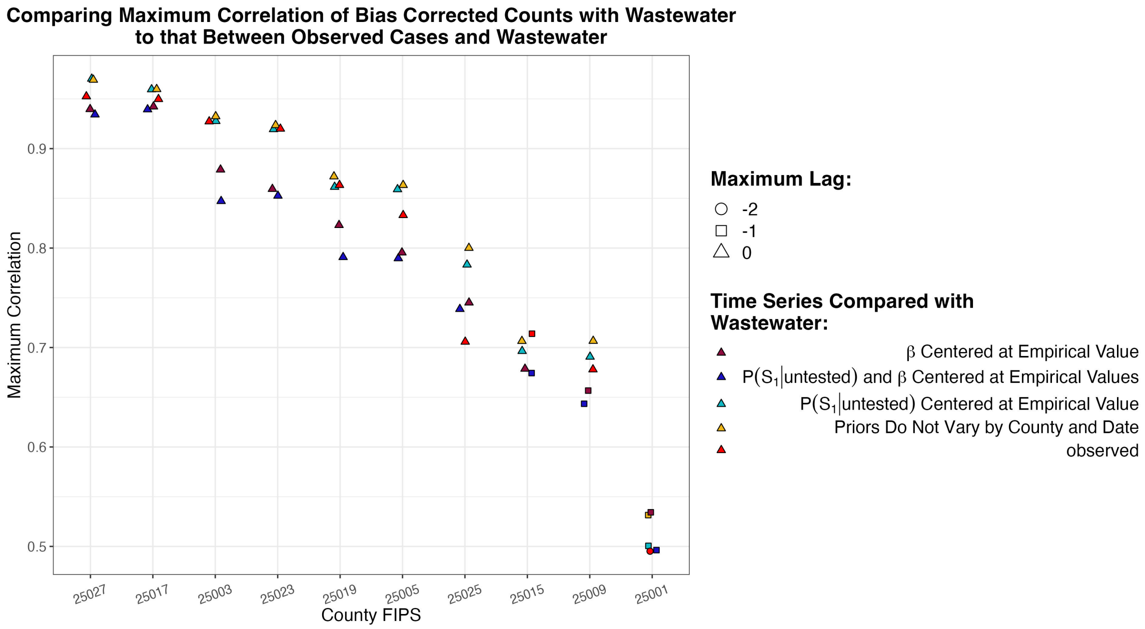


Figure 5.10

Comparison Between Covidestim, Observed Cases, and Bias Corrected Counts

In Figure 5.11, we also compare the Covidestim estimates to the wastewater concentrations. In general, both Covidestim and bias-corrected counts are more correlated with wastewater concentrations than observed infections. Of note, Nantucket County (25019) is not included here because Covidestimdoes not report estimates are not reported for Nantucket.[5]

[5] In reporting of COVID-19 data, Nantucket values are grouped with Dukes County, which is likely why Covidestim does not try to estimate the grouped counts.

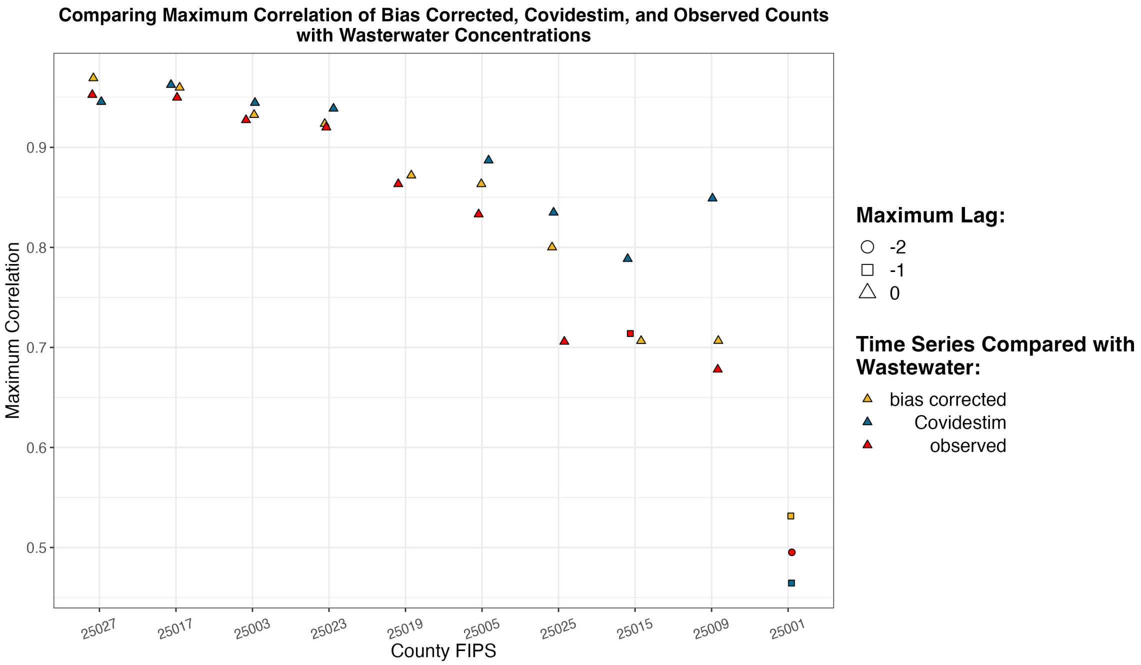


Figure 5.11

Takeaways

The aim of the cross correlation analysis was to add another source of comparison for the county-level counts from an entirely different source of data – in particular, a source of data that is less impacted by access to testing or test behavior. We see that in most counties considered here, there is high agreement between the time series. An avenue for future exploration would be to consider this analysis among a broader set of counties to see which time series tends to be most highly correlated with wastewater concentrations, a question that we cannot confidently address here when looking only at counties in Massachusetts.

Chapter 6

Appendix

Placeholder

6.1 Smoothing Span

6.1.1 Changing SPAN for LOESS Smoothing of β

6.2 Changing Mean and Variance for Prior Distribution Specifications

6.3 Relationship Between $(X + Y)_\alpha$ and X_alpha +Y_α for Dependent Variables X, Y

6.3.1 Simulation: Bivariate Normal

6.3.2 Derivation of the Distribution of X+Y for Bivariate Normal

References

Placeholder