Estimating Undetected COVID-19 Infections

Gregory Brown*, Eric Ghysels†, Lu Yi‡

January 8, 2021

Abstract

We specify and estimate a time-varying Markov model of COVID-19 cases for the US in 2020. We find that the estimated level of undetected infections spiked in March and remained elevated through May. However, since late April estimated undetected infections have generally declined though it was not until June or July that detected cases exceeded the estimated number of undetected cases. Our results suggest that the substantial increase in testing capacity in the US has identified a higher percentage of infections. However, these findings also indicate that much of the increase in the number of positive tests since spring represents a true increase in new cases as opposed to an increase resulting from more testing. According to our estimation, more than 20% of the US population has been infected by the Covid-19 virus which is consistent with other published estimates. One shortcoming of our analysis is that we are not able to condition our estimates on the age of people infected or hospitalized which may cause us to underestimate the current number of undetected cases.

*Prof. Gegory W. Brown, Sarah Graham Kenan Distinguished Professor of Finance and Executive Director, Frank H. Kenan Institute of Private Enterprise, The University of North Carolina at Chapel Hill. Corresponding Author: gregwbrown@unc.edu

†Prof. Eric Ghysels, Edward M. Bernstein Distinguished Professor of Economics,Professor of Finance, Kenan-Flager Business School and Faculty Research Director, Rethinc.Labs - Frank H. Kenan Institute of Private Enterprise, The University of North Carolina at Chapel Hill.

‡Lu Yi, Ph.D. student in Economics, The University of North Carolina at Chapel Hill.

1 Introduction and Summary of Findings

Having an accurate estimate of total infections can help planners make decisions about testing policy and economic openness, let business leaders better understand risks to their workers and customers, and inform economic projections. However, one of the challenges facing policymakers, business leaders, and the general public in understanding the spread of COVID-19 is the fact that many cases go undetected because of testing shortages or infected individuals not seeking a test, for example, asymptomatic individuals may not even consider the need for a test (Wu et al. (2020)). Unfortunately, the number of undetected cases, while hard to estimate, is much larger than the confirmed cases due to the vast amount of asymptomatic patients, which significantly undermines estimations of the total number of cases. As shown in the study by Friedman et al. (2020), most models predict the total number to be at least two to three times larger than the confirmed cases.

In fact, at the early stage of the pandemic, the number of positive tests in the US grew steadily faster than the number of hospitalizations. Likewise, hospitalizations have grown more quickly than deaths attributed to COVID-19. A very simple way to understand the disconnect between deaths and reported new cases is to estimate the total number of cases nationwide using lagged data on the number of deaths and recent estimates for infection fatality rates (see Meyerowitz-Katz and Merone (2020)). Figure 1 shows that these "death-implied" estimates suggest that the number of new cases in the US rose rapidly in March, then levelled off and started to decrease in April. This is obviously at odds with the number of new positive tests which was quite low comparing to the "death-implied" estimates until late June.

As vaccines become increasingly available, having accurate estimates of the cumulative number of cases is important because it could affect how vaccines are distributed. Specifically, in the context of a vaccination program, those already effected may already have immunity (if only temporarily) and so-called "herd immunity" could be achieved more quickly by deferring vaccination of those already infected (Randolph and Barreiro (2020)). Likewise, knowing the cumulative number of total infections may suggest how many people need to be vaccinated to achieve herd immunity. Therefore, estimation of the dynamics of total infections may provide better information on population immunity and thus enable planners to better distribute vaccines, make more informed public health decisions, and ultimately optimize social

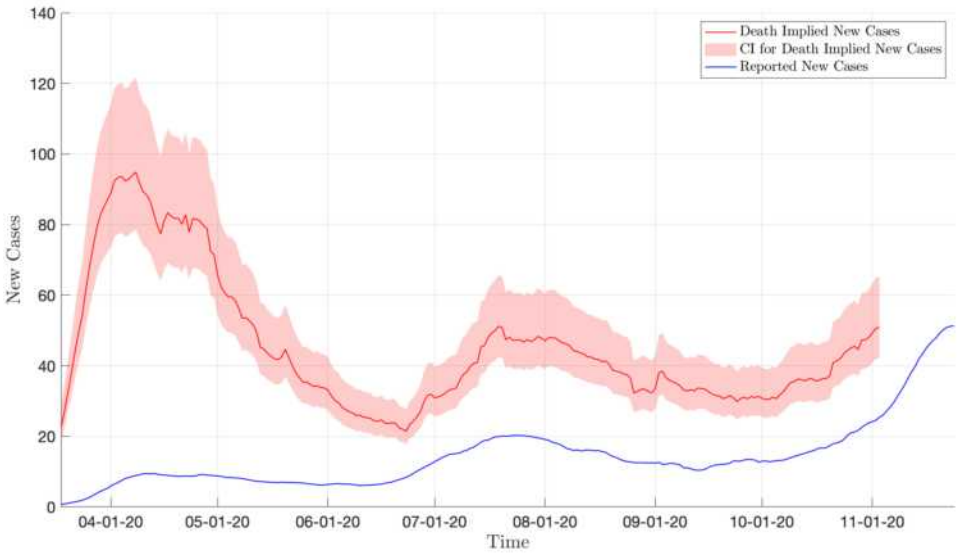


Figure 1: **New Cases in the US (per 100,000, 7-day moving average).**
The red line in the figure shows the death implied new cases, which are calculated using the 7-day average of new reported deaths in the US lagged by 14 days (to reflect the average time between contracting COVID-19 and death) divided by the infection fatality rate of 0.68% estimated by Meyerowitz-Katz and Merone (2020). The confidence band is calculated using the 7-day average of new reported deaths in the US lagged by 14 days divided by the 95% confidence interval of the estimated infection fatality rate. The blue line in the figure shows the reported new cases.

and economic well-being of the country.

To estimate the gap between observed and total cases, we use a variant of a standard time-varying Markov model to infer the number of undetected cases using easily observable data on reported cases, hospitalizations and deaths at the state and national level. Of course, other models have been proposed for estimating the number of undetected infections and we compare our results to some of these. Our model has the advantage of simplicity and ease of estimation. Specifically, in our analysis, we examine a standard 5-state time-varying Markov model based on Gourieroux and Jasiak (2020) (and cites therein) and apply it to data in the US and nine individual states. In our model the population is either susceptible (S), infected and undetected (IU), infected and detected (ID), hospitalized (H), or deceased (D). Recovered cases re-enter the susceptible pool. States are mutually exclusive so we track hospitalized separately from infected and detected. As conditioning variables in our analysis we include both the testing positivity rate and the intensity of testing (i.e., tests conducted per 100,000) and find that these are important factors in the estimation

with intuitive relations to infection probabilities. The model provides estimates of undetected infections that are plausible and the model fits observed levels of positive cases, hospitalizations and deaths well. We examine two different versions of the model and obtain similar results from both.

We find that the estimated IU was high in March and April. Since then, estimated IU has declined substantially but according to our estimates it was not until late-June or early-July that the number of detected cases exceeded the number of undetected cases. Our results suggest that the substantial increase in testing capacity has been successful in identifying a much higher percentage of infections. However, it also suggests that much of the increase in the number of positive tests since October represents a true increase in new cases as opposed to an increase resulting from more testing. According to our estimation, by the end of November about 23% of the US population has been infected by the Covid-19 virus, which suggests that we are still a long way from herd immunity but that vaccinations. Yet, if vaccinations could be prioritized based on past infection, this represents a substantial base of the population which may already be immune. One concern about our analysis is that we are not able to condition on the age of those with detected cases or who are hospitalized, and consequently, we may underestimate undetected cases if the age of those infected is declining on average. Our estimates could also underestimate cases if the quality of care has improved over time and reduced hospitalization and death rates in a way the model does not capture.

2 Model

The latent individual history variable $Y_{i,t}$, for individual $i = 1, ..., N$ at time $t = 1, ..., T$, is qualitative polytomous with J alternatives denoted by $j = 1, ..., J$. As in Gourieroux and Jasiak (2020), we assume that $Y_{i,t}$ have the same marginal distribution for all individuals $i = 1, ..., N$ at t fixed, which can be summarized by the J-dimensional vector $p(t)$. The j-th component of the marginal distribution is

$$p_j(t) = P(Y_{i,t} = j).$$

In addition, the individual history variable follows a Markov process with time-varying transition matrix $P\left[p(t-1);\theta\right]$, which gives

$$p(t) = P[p(t-1);\theta]'p(t-1), t = 2, ..., T,$$

with θ being a vector of parameters.

The data of individual histories may not be available in practice. With the assumptions of independent individual histories and homogeneous population of risks, the J-dimensional cross-sectional frequency vector $f(t)$, where $f_j(t)$ is the state j frequency of the population, can be seen as the sample counterpart of $p(t)$. However, the cross-sectional frequencies are only partially observed. A state aggregation matrix A is used to account for the unobserved states and the observations are $\hat{A}_t = Af(t)$ for $t = 1, ..., T$, where A is a $K \times J$ matrix of full rank K. The parameters of interest, θ and the sequence of the unobserved component of $p(t)$, can then be estimated by solving the following optimization problem,

$$\begin{aligned}(\hat{p}(1), ..., \hat{p}(T), \hat{\theta}) = \text{argmin} \sum_{t=2}^{T} \|p(t) - P\left[p(t-1), \theta\right]' p(t-1)\|_2^2 \\ \text{s.t. } Ap(t) = Af(t) = \hat{A}_t, t = 1, ..., T,\end{aligned} \tag{1}$$

where $\|.\|_2$ denotes the Euclidean norm.

To model the COVID-19 propagation, we consider a Markov process with 5 states: $1 = S$, for susceptible, $2 = IU$, for Infected and Undetected, $3 = ID$, for Infected and Detected, $4 = H$ for Hospitalized, and $5 = D$ for Deceased. The sum of the frequencies across all the five states equals to the size of the population. For simplicity, we assume no immunity in our estimation, hence the recovered cases re-enter the susceptible pool. This assumption lets us avoid having an unobservable recovered state but will have little impact on estimation for low levels of overall infection.

The transition matrix $P[p(t-1);\theta]$ of the Markov process is defined as

$$
\begin{array}{c} \\ \text{S} \\ \text{IU} \\ \text{ID} \\ \text{H} \\ \text{D} \end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{S} & \text{IU} & \text{ID} & \text{H} & \text{D} \end{array} \\
\begin{bmatrix}
1-p_i & p_i(1-p_d) & p_i p_d & 0 & 0 \\
p_{21} & (1-p_{21}-p_{24})(1-p_d) & (1-p_{21}-p_{24})p_d & p_{24} & 0 \\
p_{31} & 0 & 1-p_{31}-p_{34} & p_{34} & 0 \\
p_{41} & 0 & 0 & 1-p_{41}-p_{45} & p_{45} \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
$$

with

$$
\begin{aligned}
p_i &= \text{logist}\left(a_1 + a_2\left(p_2(t-1) + p_3(t-1)\right)\right) + a_3 x_t), \\
p_d &= \text{logist}\left(b_1 + b_2 y_t\right),
\end{aligned}
$$

where $\text{logist}(x) = 1/[1+\exp(-x)]$ is the logistic function, i.e. the inverse of the logit function. The probability of infected p_i follows a multinomial logit model for the competing propagation driven by lagged IU and lagged ID, and it also depends on the testing positivity rate x_t. Conditioning on being infected, the probability of being detected p_d is a function of testing intensity y_t. Each row of the transition matrix sums to one by construction. The structure of zeros indicates that one cannot go backward from ID to IU, patients who died are hospitalized before death, the hospitalized patients will stay in hospital until they recover or die, and death is considered an absorbing state.

In addition, we consider two model specifications for the transition probabilities from state IU and ID to state H. The basic specification assumes constant transition probabilities p_{24} and p_{34}. In this model, there are 11 parameters in $\theta = [a_1, a_2, a_3, b_1, b_2, p_{21}, p_{24}, p_{31}, p_{34}, p_{41}, p_{45}]'$. The full specification assumes time-varying transition probabilities driven by the lagged frequency of the corresponding state with

$$
\begin{aligned}
p_{24} &= \text{logist}\left(c_1 + c_2 p_2(t-1)\right), \\
p_{34} &= \text{logist}\left(d_1 + d_2 p_3(t-1)\right),
\end{aligned}
$$

in which $\theta = [a_1, a_2, a_3, b_1, b_2, c_1, c_2, d_1, d_2, p_{21}, p_{31}, p_{41}, p_{45}]'$ has 13 parameters. The results from these two versions of the model are very similar so we only report the results from the basic model (but results from the full model are available on request).

Empirically, $IU(t)$ and $ID(t)$ represent the state of currently infected excluding those hospitalized. The frequency of $ID(t)$ is observable by assumption, while $IU(t)$

is the unobserved state of unidentified infections and will be considered as additional quantities of interest to be estimated jointly. Also, the frequencies of $H(t)$ and $D(t)$ are both observable. Therefore, we have the state aggregation matrix A expressed as

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

3 Data, Estimation and Results

We estimate the two specifications of the time-varying Markov model on the national Covid-19 propagation data of the US over the period of 271 days between March 4 to November 29, 2020. We use the daily data reported by The Covid Tracking Project. The frequency of $ID(t)$ is measured by the rolling 2-week sum of the new positive tests in the US, which assumes that a person with positive test will either be hospitalized or recover within 14 days. The frequency of $H(t)$ is the actual number of hospitalized in US on any given date and the frequency of the absorbing state $D(t)$ is measured by the cumulative deaths caused by COVID-19 in the US. In constructing the cross-sectional frequency vector $f(t)$, we do everything in per $100,000$ population to facilitate interpretation as well as comparison to estimated infection rates across geographies.

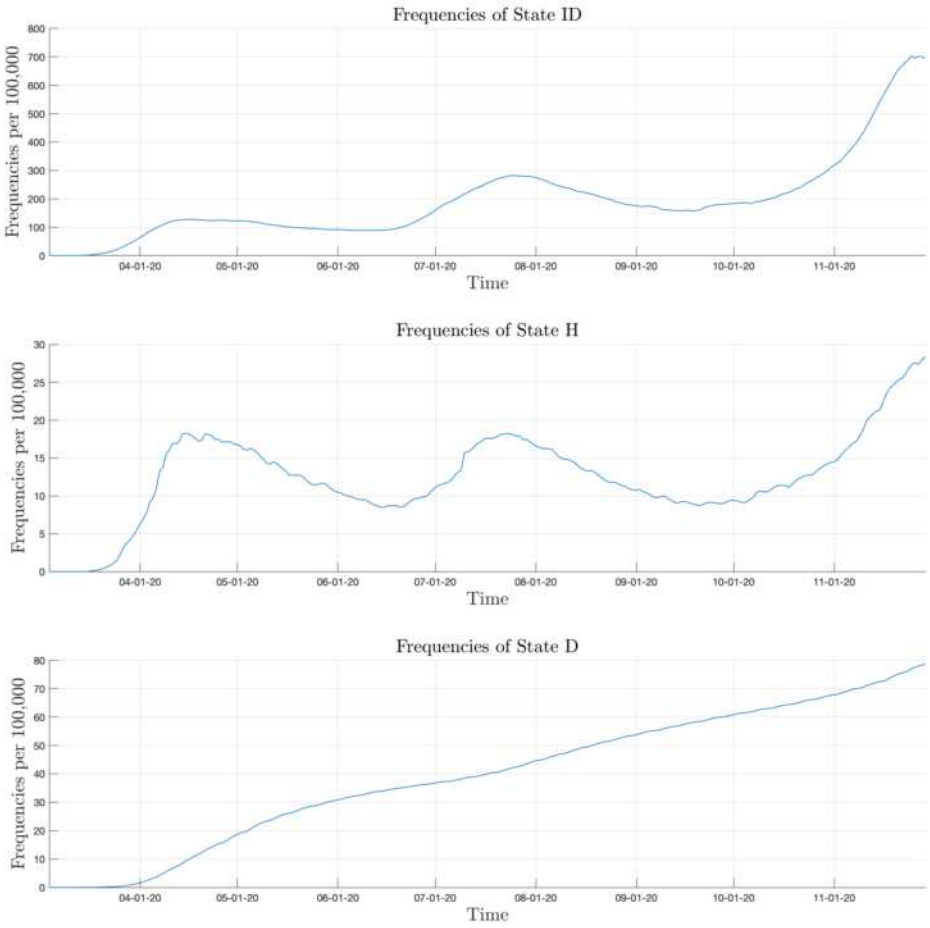


Figure 2: **The Frequencies of the Observed States in the US (per 100,000).** The top panel shows the time series data for the frequency of state $ID(t)$, which are measured by the rolling 2-week sum of the new positive tests in the US. The middle panel shows the time series data for the frequencies of state $H(t)$, which is measured by the actual number of hospitalized in the US at date t. The bottom panel shows the time series data for the frequencies of state $D(t)$, which is measured by the sum of deaths caused by COVID-19 in the US up until date t.

The daily evolutions of the observed components of $f(t)$ for the US are displayed in Figure 2. For the two conditioning variables, the test positivity rate x_t is measured by the weekly moving average of the testing positivity rate (i.e., out of all tests) and the test intensity y_t is measured by the rolling 7-day average of tests per day per $100,000$ population as of date t. Figure 3 shows the plots of these two conditioning variables calculated from the US data.

The initial frequency is set equal to $100,000$ for state $S(0)$ and 0 for all other states. The model parameters θ and the series of frequencies of the unobserved state $IU(t)$ are then estimated by solving the optimization problem in Equation (1) numerically using the $fminsearch$ function in Matlab. The estimates of the parameters for the basic model on US data are provided in Table 1. The mean fitted values are within 2.36% of observed values. The comparisons of fitted and observed frequencies for state ID, state H and state D are shown in Figure 6 in an appendix. We see that the estimated frequencies track the observations closely.

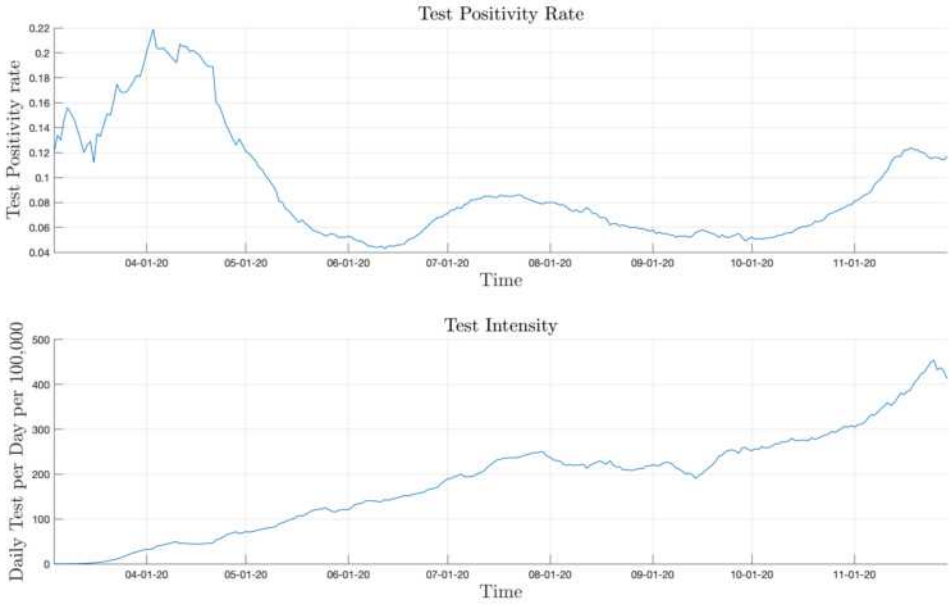


Figure 3: **The Conditioning Variables (x_t and y_t).**
The top panel shows the time series data of the test positivity rates x_t, which is measured by the weekly moving average of the rates of positivity in testing (i.e., out of all tests). The bottom panel shows the time series data of test intensity y_t, which is measured by the rolling 7-day average of tests per day per $100,000$ population as of date t.

In Table 1, $p_{21} = 0.2979$, which corresponds to a less than 1 week average recovery time of for state IU, and $p_{31} = 0.0461$, which represents an average recovery time around 20 days for state ID. The model estimates that it takes longer for a patient in the detected state to recover, which is reasonable considering it is more likely that patients with severe cases will get tested (and be detected) thus the overall health condition of state ID is worse than state IU. This is also consistent with the estimated transition probabilities to state H. The probability of transition to state H is 0.0057

from state ID, which is higher than the probability of 0.0013 from state IU. The estimates of p_{33} is 0.9482, which means that people stay in the state ID for an average around 18 days and are then either hospitalized or recover. This is roughly consistent with how we construct the variable representing state ID (i.e. rolling 2-week sum of the positive tests). The mortality rate conditional on being hospitalized is 2.25%, which is higher than the estimated value of 0.68% for the overall infection-fatality rate of COVID-19 in Meyerowitz-Katz and Merone (2020). This is not surprising considering that the severity of the illness is higher for the hospitalized patients than the average severity of all cases. We have a large positive estimate a_3 meaning that the probability of being infected, p_i, is increasing with higher positivity rate of tests. The estimate of b_2 is also positive, which means that if a person is infected, the probability of being detected, p_d, is increasing with the intensity of testing.

a_1	a_2	a_3	b_1	b_2
-8.4486	-0.0030	25.7573	-5.0047	0.0120

	$1 = S$	$2 = IU$	$3 = ID$	$4 = H$	$5 = D$
$2 = IU$	0.2979	Time-varying	Time-varying	0.0013	0
$3 = ID$	0.0461	0	0.9482	0.0057	0
$4 = H$	0.0805	0	0	0.8970	0.0225
$5 = D$	0	0	0	0	1

Table 1: Parameter Estimates of the Basic Model

The time series of the frequencies of state $IU(t)$ are the quantities of primary interest. Figure 4 shows the estimated frequencies of the state IU (dashed red line) and the observed frequencies of the state ID (dash-dot blue line). In addition, we calculate the total infections per 100,000 population at time t by the summation of $ID(t)$ and $IU(t)$. The solid black line in Figure 4 shows the time series of total infections per 100,000 population and the green line shows the evolution of the unidentified percentage of the total cases with corresponding values on the right y-axis. From Figure 4, we find that the estimated IU grew rapidly since mid-March until peaking in early-April. After that, estimated IU has declined substantially but according to our estimates it was not until around July 1st that the number of detected cases exceeded the number of undetected cases. Our results suggest that the substantial increase in testing capacity has been successful in identifying a much higher percentage of infections. However, it also suggests that much of the increase in the number of

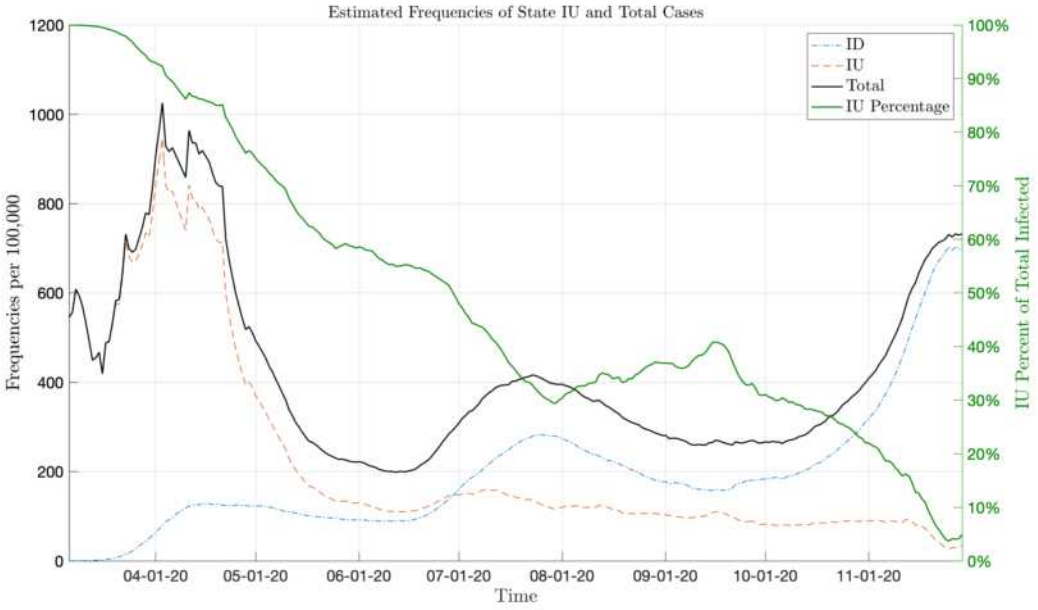


Figure 4: **The Estimated Frequencies of State IU and Total Infections.** The figure shows the estimated frequencies of state IU and total infections. The dash-dot blue line is the observed frequencies of state ID and the dashed red line is the estimated frequencies of state IU. The solid black line is the estimated frequencies of total cases (e.g. total frequencies of state IU and state ID). The green line with corresponding values on the right y-axis shows the evolution of unidentified percentage of the total cases.

positive tests since October is in fact an increase in new cases as opposed to an increase related to a higher number of tests. Our results are consistent with models discussed in Friedman et al. (2020).[1]

3.1 Estimates for individual states

We also estimate models for nine individual states including Arizona, California, Florida, Georgia, North Carolina, New Jersey, New York, Pennsylvania and Texas, of which the total residential population account for nearly half of the US population. We believe that these states constitute a good representation of the overall demographic not only because of their population, but also since they have experienced the pandemic in very different ways since March in terms of the trends of confirmed cases, hospitalizations, and deaths. Despite the similar recent surge in confirmed

[1]The estimates of total infections using these models are on https://ourworldindata.org/covid-models.

Covid-19 cases and hospitalizations, these data peaked in late-April for New York, New Jersey and Pennsylvania, while the other states have their peaks in mid-July.

State	PA	CA	NJ	NY
a_1	−7.892	−8.624	−8.299	−8.333
a_2	−0.007	−0.003	−0.004	−0.003
a_3	27.728	31.092	58.942	38.874
b_1	−4.461	−5.297	−4.840	−4.909
b_2	0.018	0.011	0.007	0.006
p_{21}	0.409	0.483	0.121	0.222
p_{24}	0.003	0.001	0.002	0.003
p_{31}	0.078	0.039	0.124	0.052
p_{34}	0.005	0.005	0.003	0.013
p_{41}	0.088	0.053	0.097	0.146
p_{45}	0.022	0.020	0.037	0.051
%RMSE	4.68%	8.22%	6.06%	5.89%

State	NC	AZ	TX	GA	FL
a_1	−8.120	−7.705	−7.783	−8.696	−7.756
a_2	−0.005	−0.003	−0.004	−0.006	−0.003
a_3	37.847	15.302	25.073	30.044	17.278
b_1	−4.748	−3.935	−4.068	−4.775	−3.745
b_2	0.010	0.016	0.013	0.018	0.013
p_{21}	0.351	0.507	0.482	0.407	0.644
p_{24}	0.000	0.001	0.000	0.001	0.006
p_{31}	0.043	0.102	0.094	0.060	0.099
p_{34}	0.006	0.006	0.010	0.004	0.003
p_{41}	0.050	0.072	0.106	0.038	0.046
p_{45}	0.017	0.020	0.018	0.020	0.028
%RMSE	6.64%	15.15%	15.72%	4.77%	10.54%

Table 2: Model Parameters for Individual States

The historical data of Covid-19 for each state is from The Covid Tracking Project. The time series data used in model estimation is constructed in the same way as the US data. Figure 7 - 9 in the appendix show the daily evolutions of $ID(t)$, $H(t)$ and $D(t)$ for the nine states. The two conditioning variables, test positivity rate x_t and test intensity y_t for each state are shown in Figure 10 and 11 in the appendix. The estimates of model parameters and the goodness-of-fit measures (percentage root mean square error (%RMSE)) for the nine states are in Table 2. Table 3 shows the average of the estimated parameters across these individual states. We calculate the

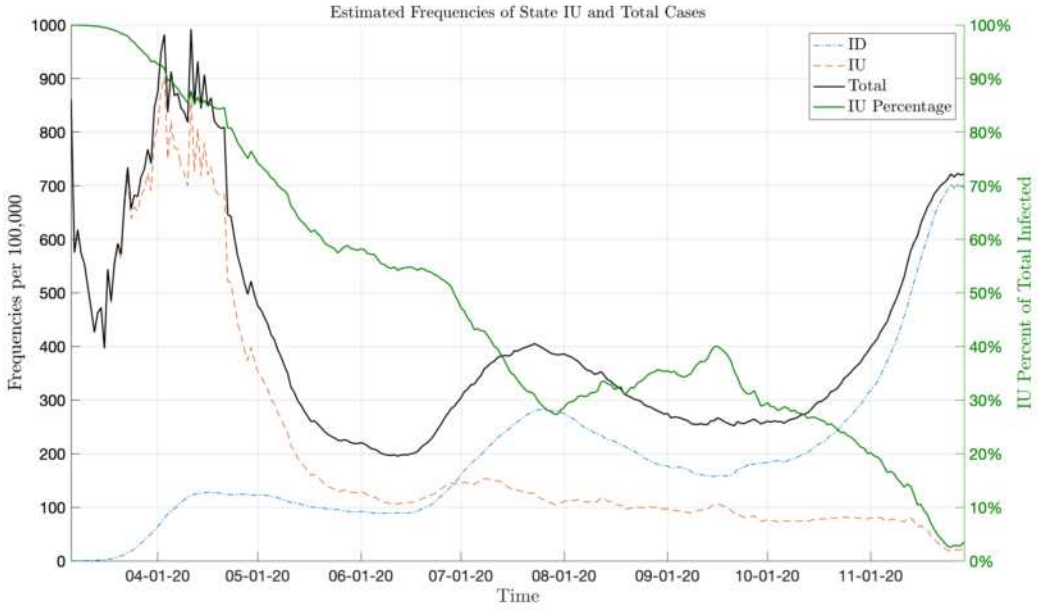


Figure 5: **The Estimated Frequencies of State IU and Total Infections for the US using Average State Parameters.**
See Figure 4 for details.

estimated frequencies for state $IU(t)$ and total infections for each state using both the state parameters and the average parameters. The results are shown in Figure 12 - 20.

a_1	a_2	a_3	b_1	b_2
-8.134	-0.004	31.353	-4.531	0.012

	$1 = S$	$2 = IU$	$3 = ID$	$4 = H$	$5 = D$
$2 = IU$	0.403	Time-varying	Time-varying	0.002	0
$3 = ID$	0.077	0	0.917	0.006	0
$4 = H$	0.077	0	0	0.8970	0.026
$5 = D$	0	0	0	0	1

Table 3: Average of Model Parameters

According to our estimation, all states in our model experienced a rapid increase in total infections during April, though some of these states only observed a mild spread of virus according to the confirmed cases. The surge in total cases during this period were mainly driven by the infected and unidentified cases, while the later increase were driven by confirmed cases. We also re-estimate the frequencies of state $IU(t)$

and total infections for the US using the average of state parameters. The results are shown in Figure 5 and we find that the overall pattern is similar comparing to the results in Figure 4.

3.2 Estimates for cumulative cases

Having accurate estimates of the cumulative infections is important as vaccines become increasingly available. It may suggest what percentage of the population is immune to the virus if those already infected are not susceptible, or likewise how many people need to get a vaccine before we reach herd immunity. We estimate the cumulative total cases based on the data of confirmed cases and our estimated infected and unidentified cases. The cumulative total cases $CI(T)$ up to date T is

$$CI(T) = CD(T) + CU(T),$$

where $CD(T)$ is the cumulative confirmed cases and $CU(T)$ is the cumulative undetected cases. Our data set has the cumulative number of positive tests and we use it as a measure of cumulative confirmed cases. Using our estimated model, we calculate the cumulative undetected cases as

$$CU(T) = \sum_{t=1}^{T} p_i(t)(1 - p_d(t))S(t-1) - \sum_{t=1}^{T}(1 - p_{21} - p_{24})p_d(t)IU(t-1). \quad (2)$$

The first summation in Equation (2) is the total number of daily new entrants to state IU, which measures the total number of patients who have been through the infected and unidentified state. According to our model specification, a proportion of patients in state IU transit to state ID at each t and they were "detected" and included into the cumulative number of positive tests, therefore, we subtract this portion of patients from the total number of patients who been in state IU to get the cumulative number of undetected cases.

The estimated cumulative infections as of November 29, 2020 are shown in Table 4. We also do the estimation using the state average parameters and the results are in Table 5. As shown in Table 4, our estimated percentage of undetected infections out of total infections is 82.34% for the US, which is similar to the results of 79.21%

Panel (a)	**Number of Infections**			
State	Confirmed	Undetected	Total	% Undetected Infections
AZ	325,995	1,606,040	1,932,035	83.13%
CA	1,198,934	7,674,817	8,873,751	86.49%
FL	976,944	5,534,536	6,511,480	84.99%
GA	420,601	1,893,227	2,313,828	81.82%
NC	361,778	1,748,355	2,110,133	82.86%
NJ	334,114	3,200,884	3,534,998	90.55%
NY	641,161	10,414,697	11,055,858	94.20%
PA	357,196	2,747,136	3,104,332	88.49%
TX	1,157,273	7,018,290	8,175,563	85.84%
State Sum	5,773,996	41,837,985	47,611,981	87.87%
US	13,188,777	61,488,270	74,677,047	82.34%

Panel (b)	**Infections per 100,000 Population**			
State	Confirmed	Undetected	Total	Population
AZ	4,477	22,061	26,538	7,280,000
CA	3,034	19,425	22,459	39,510,000
FL	4,548	25,766	30,314	21,480,000
GA	3,960	17,827	21,787	10,620,000
NC	3,445	16,651	20,096	10,500,000
NJ	3,762	36,046	39,808	8,880,000
NY	3,296	53,546	56,842	19,450,000
PA	2,790	21,462	24,252	12,800,000
TX	3,990	24,201	28,191	29,000,000
State Sum	3,619	26,227	29,847	159,520,000
US	4,018	18,735	22,753	328,200,000

Table 4: Estimated Cumulative Infections as of Nov 29, 2020 (state parameters)

and 74.23% in Gu (2020) and Friedman et al. (2020) respectively[2]. Based on our estimation using individual state parameters, the cumulative number of cases in the US up to November 29 is 74,677,047, which accounts for 22.75% of the US population. The infections per 100,000 population are quite different across states. New York and New Jersey, which are the early epicenters of the Covid-19 pandemic, have a much higher percentage of population being infected.

[2]Please see https://ourworldindata.org/covid-models for details.

Panel (a)	**Number of Infections**			
State	Confirmed	Undetected	Total	% Undetected Infections
AZ	325,995	2,562,268	2,888,263	88.71%
CA	1,198,934	6,026,460	7,225,394	83.41%
FL	976,944	6,602,092	7,579,036	87.11%
GA	420,601	3,427,286	3,847,887	89.06%
NC	361,778	1,573,530	1,935,308	81.31%
NJ	334,114	3,471,724	3,805,838	91.22%
NY	641,161	7,054,515	7,695,676	91.67%
PA	357,196	4,623,488	4,980,684	92.83%
TX	1,157,273	7,221,290	8,378,563	86.18%
State Sum	5,773,996	42,562,656	48,336,652	88.05%
US	13,188,777	80,582,946	93,771,723	85.93%

Panel (b)	**Infections per 100,000 Population**			
State	Confirmed	Undetected	Total	Population
AZ	4,477	35,196	39,673	7,280,000
CA	3,034	15,253	18,287	39,510,000
FL	4,548	30,736	35,284	21,480,000
GA	3,960	32,272	36,232	10,620,000
NC	3,445	14,986	18,431	10,500,000
NJ	3,762	39,096	42,858	8,880,000
NY	3,296	36,270	39,566	19,450,000
PA	2,790	36,121	38,911	12,800,000
TX	3,990	24,901	28,891	29,000,000
State Sum	3,619	26,681	30,301	159,520,000
US	4,018	24,553	28,571	328,200,000

Table 5: Estimated Cumulative Infections as of Nov 29, 2929 (average parameters)

4 Conclusion

We estimate a model of COVID-19 infections, hospitalizations, recoveries, and deaths. The results of the estimation are intuitive and indicate a high percentage of undetected cases early in our sample period followed by a decline to a much lower percentage of undetected cases by July. Taken at face value, our results suggest that reported cases in the US increasingly reflect the true number of infections. We also estimate the model using data from nine individual states, which gives similar estimates on total infections as well as the proportion of undetected cases. Nonetheless, our model is fairly simple. Given anecdotal evidence that age of detected cases is changing

through time, the estimation is also likely to benefit by conditioning estimates on other variables such as the average age of hospitalized patients or the average age of those testing positive.

References

Friedman, J., P. Liu, C. E. Troeger, A. Carter, R. C. Reiner, R. M. Barber, J. Collins, S. S. Lim, D. M. Pigott, T. Vos, S. I. Hay, C. J. Murray, and E. Gakidou (2020). Predictive performance of international COVID-19 mortality forecasting models. *medRxiv*.

Gourieroux, C. and J. Jasiak (2020). Time varying markov process with partially observed aggregate data: An application to coronavirus.

Gu, Y. (2020). Estimating True Infections Revisited: A Simple Nowcasting Model to Estimate Prevalent Cases in the US. `https://covid19-projections.com/estimating-true-infections-revisited/`. Accessed: 2020-11-30.

Meyerowitz-Katz, G. and L. Merone (2020). A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. *medRxiv*.

Randolph, H. E. and L. B. Barreiro (2020). Herd immunity: Understanding COVID-19. *Immunity 52*(5), 737 – 741.

Wu, S., A. Mertens, Y. Crider, A. Nguyen, N. Pokpongkiat, S. Djajadi, A. Seth, M. Hsiang, J. Colford, A. Reingold, B. Arnold, A. Hubbard, and J. Benjamin-Chung (2020, 09). Substantial underestimation of SARS-CoV-2 infection in the united states. *Nature communications 11*, 4507.

Appendix

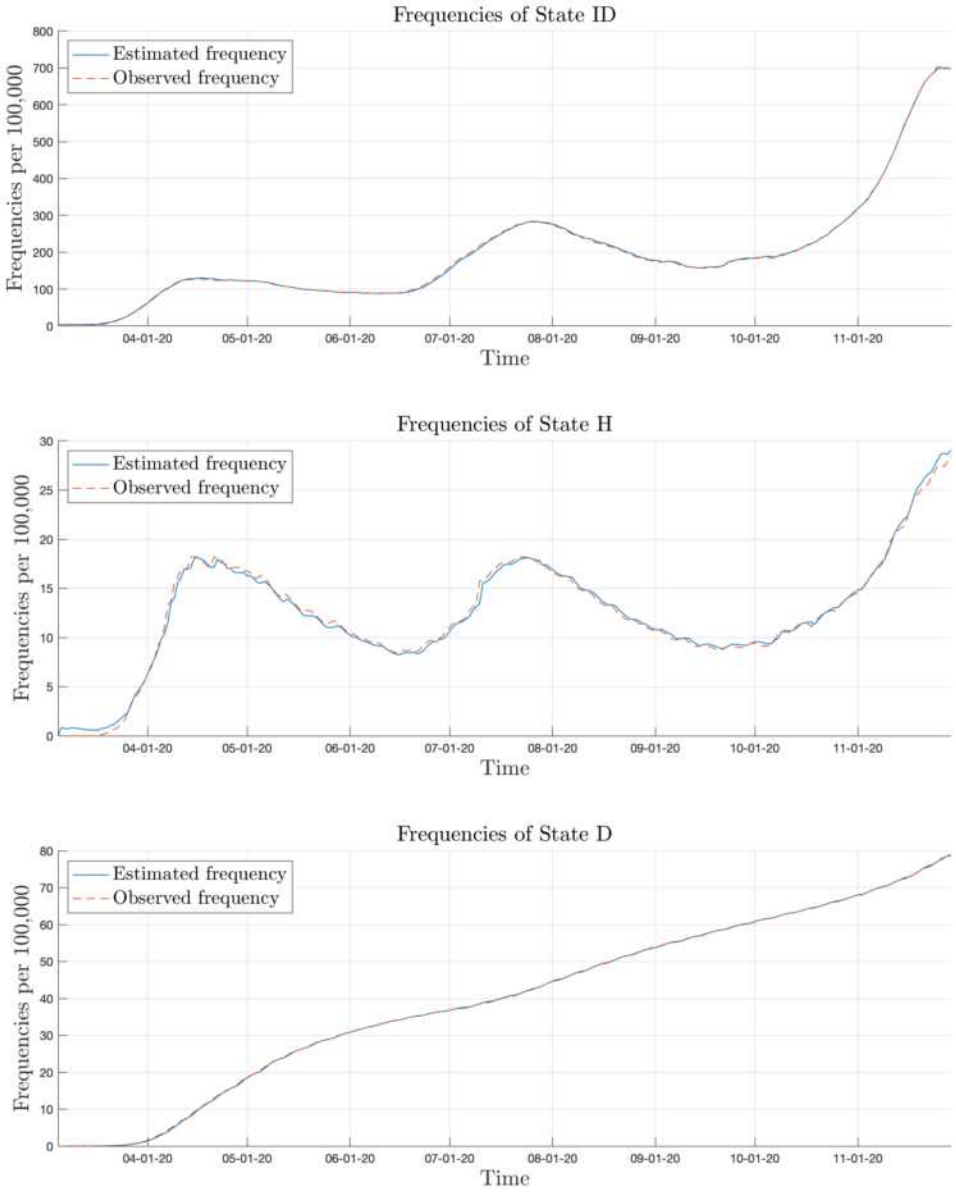


Figure 6: **The Observed and Estimated Frequencies.**
The figure compares the observed frequencies with the estimated frequencies for state ID (top panel), state H (middle panel) and state D (bottom panel). The dashed red line is the observed frequencies and the solid blue line is the estimated frequencies.

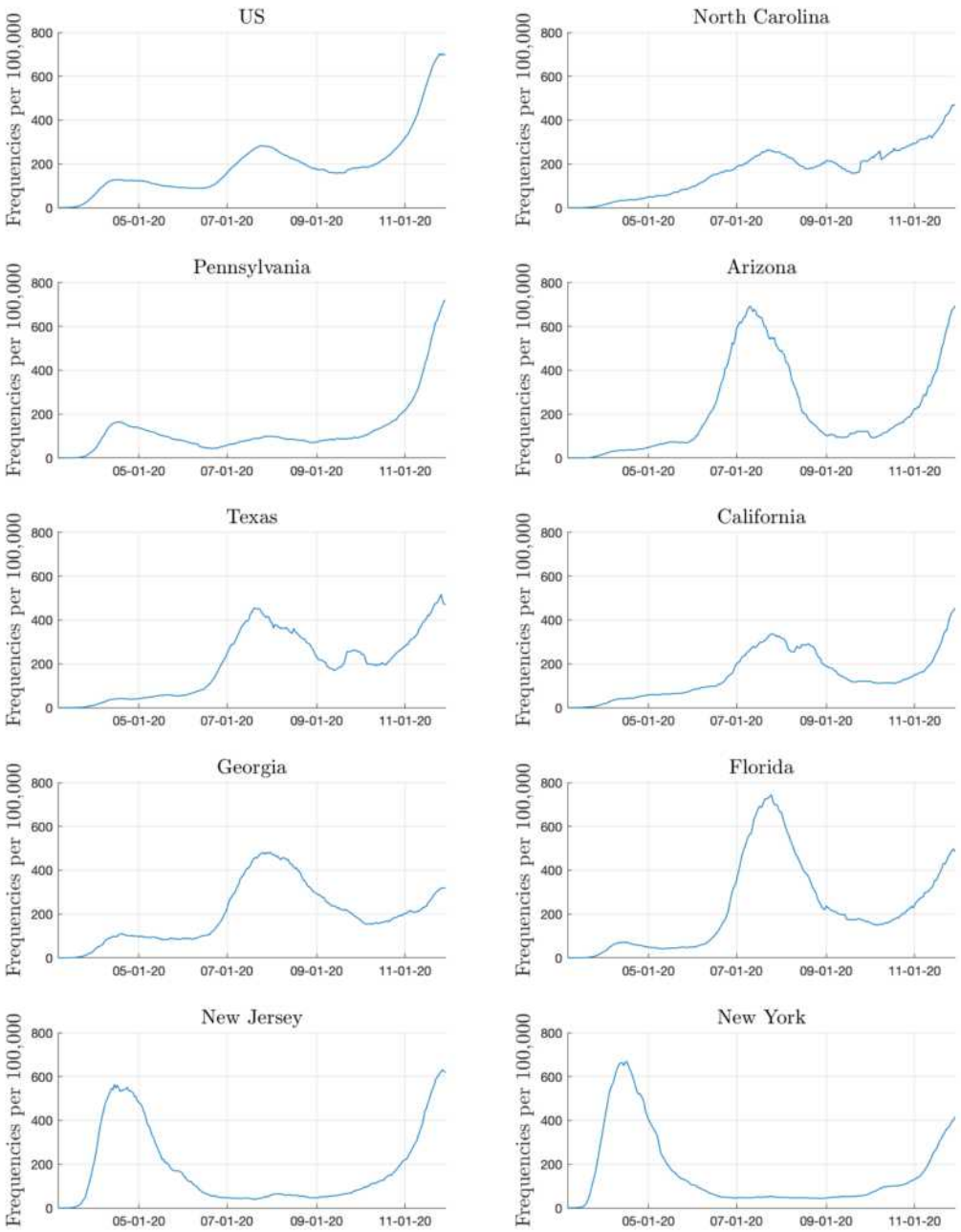


Figure 7: **The Frequencies of State $ID(t)$ for Individual States.**
The figure shows the observed frequencies of state $ID(t)$ of nine individual states. The time series of $ID(t)$ for the US is included for a better comparison, the same plot is also shown in the top panel of Figure 2.

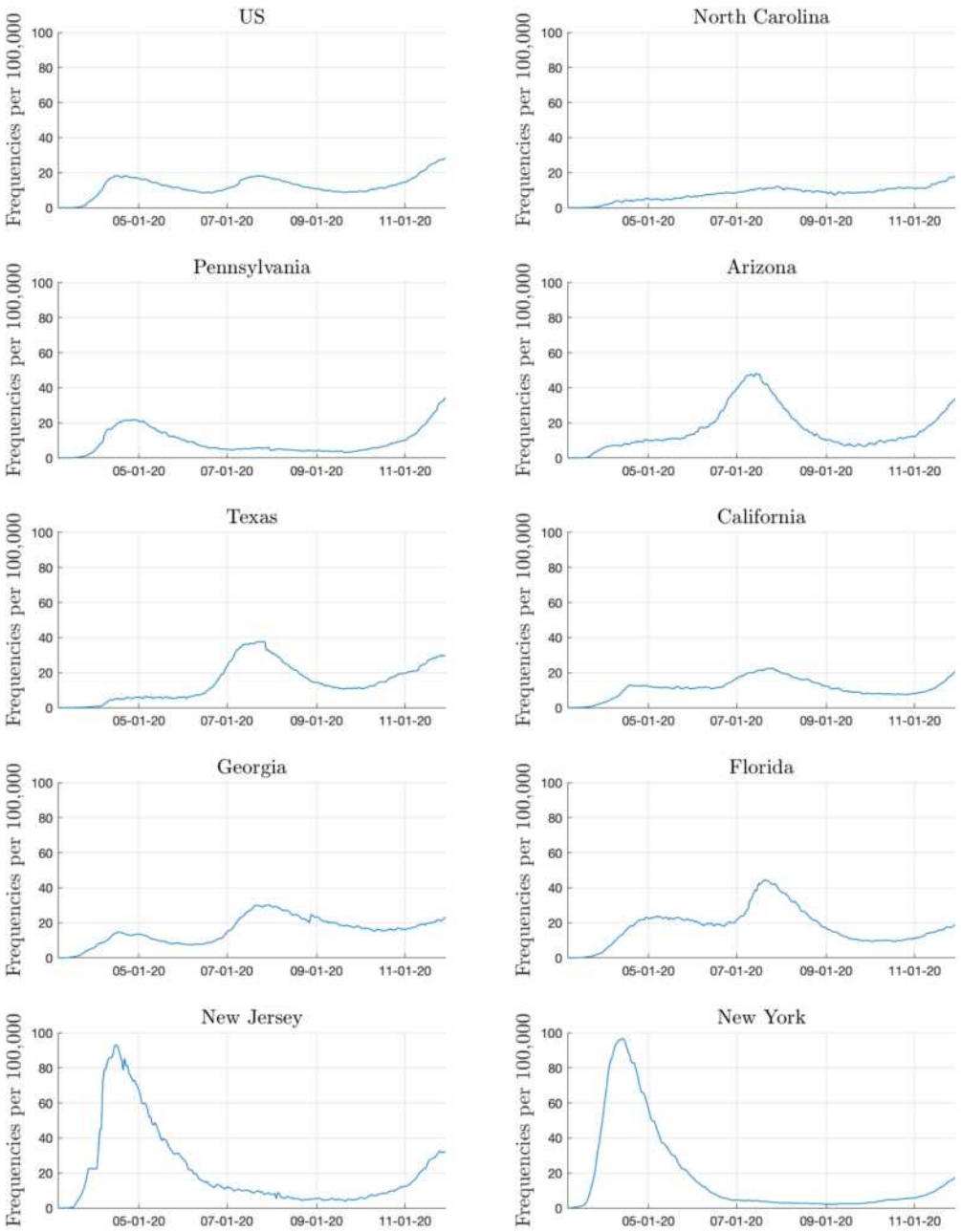


Figure 8: **The Frequencies of State $H(t)$ for Individual States.**
The figure shows the observed frequencies of state $H(t)$ of nine individual states. The time series of $H(t)$ for the US is included for a better comparison, the same plot is also shown in the middle panel of Figure 2.

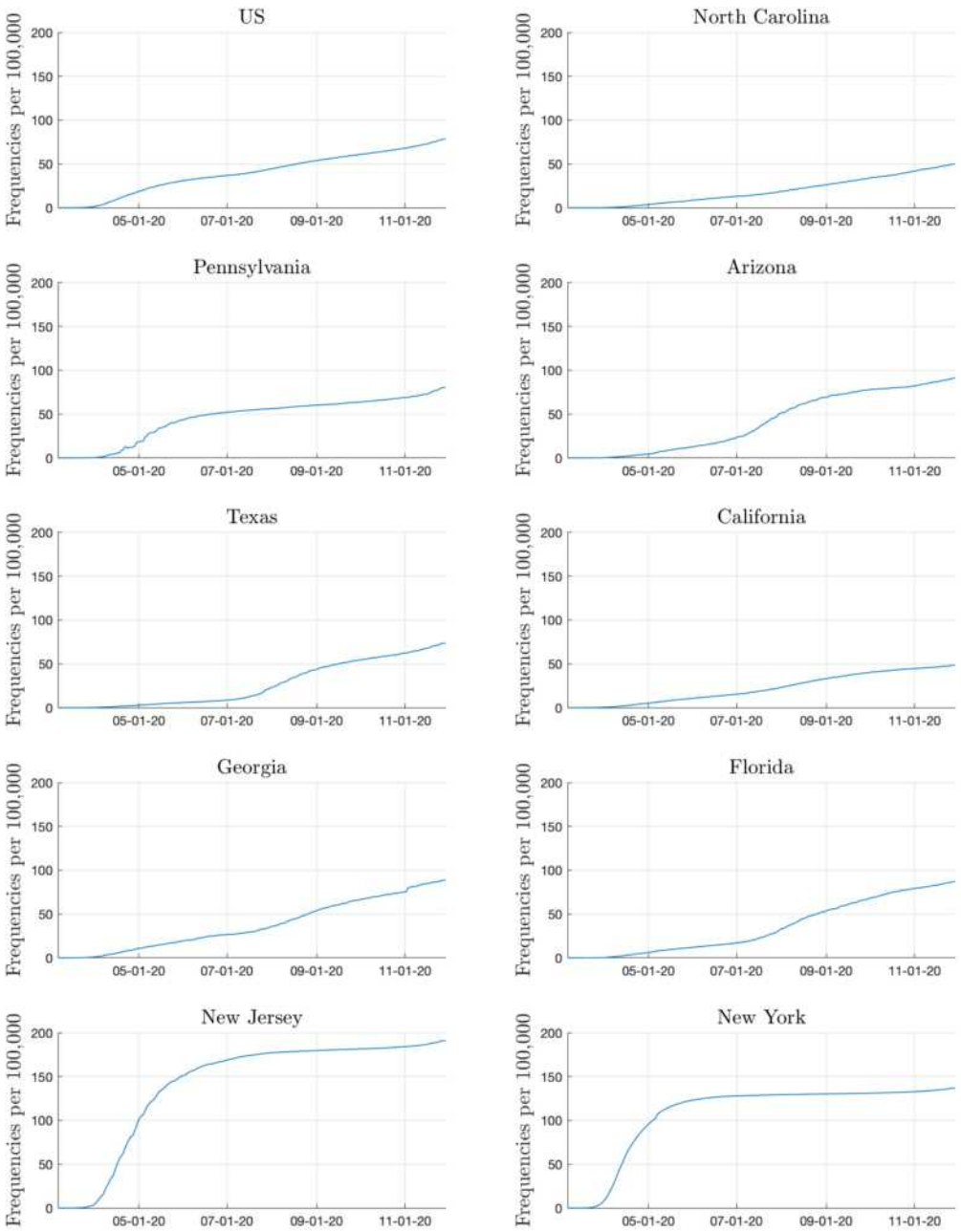


Figure 9: **The Frequencies of State $D(t)$ for Individual States.**
The figure shows the observed frequencies of state $D(t)$ of nine individual states. The time series of $D(t)$ for the US is included for a better comparison, the same plot is also shown in the bottom panel of Figure 2.

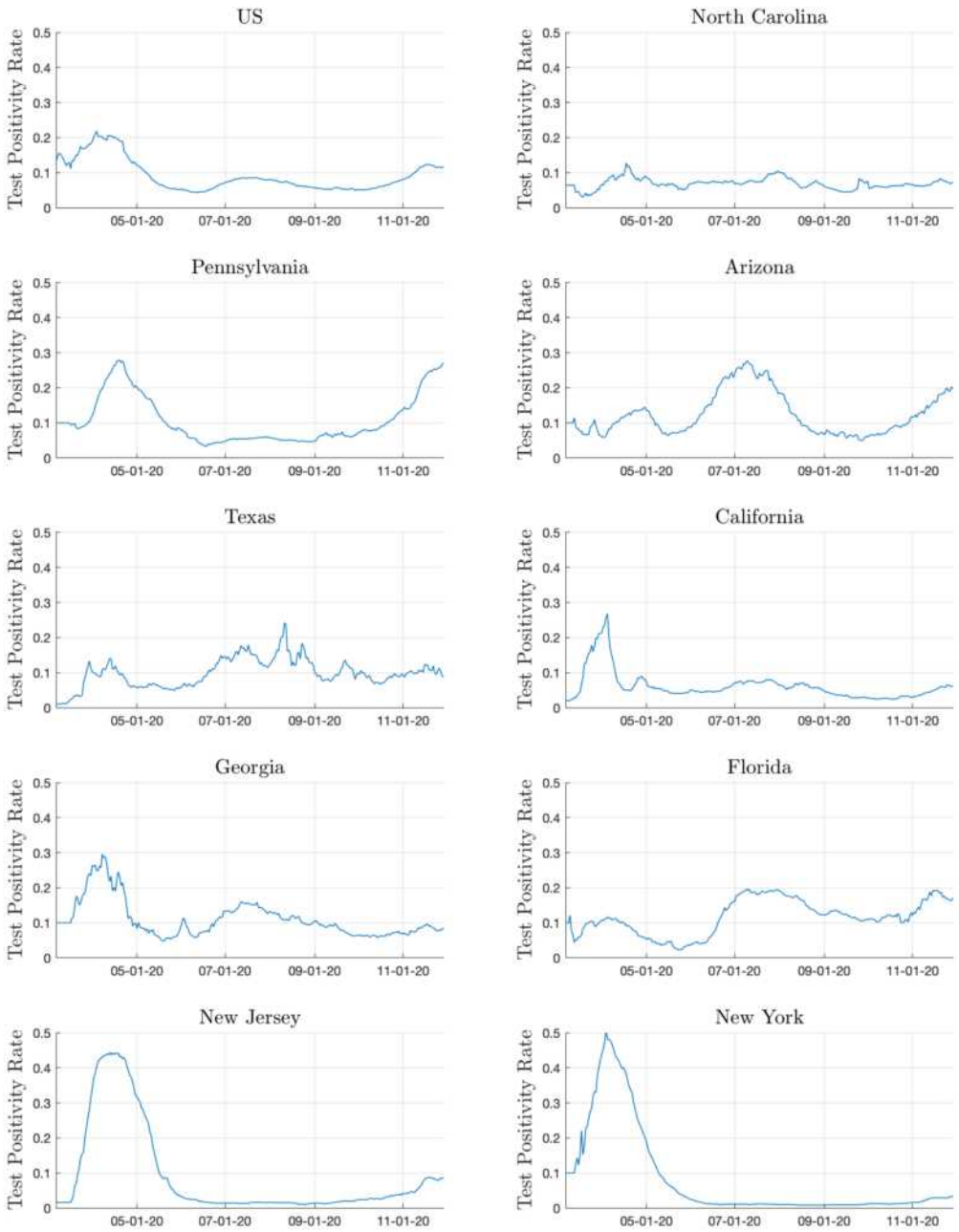


Figure 10: **The Test Positivity Rates x_t of Individual States.**
The figure shows the time series data of the test positivity rates x_t of nine individual states. The time series of x_t for the US is included for a better comparison, the same plot is also shown in the top panel of Figure 3.

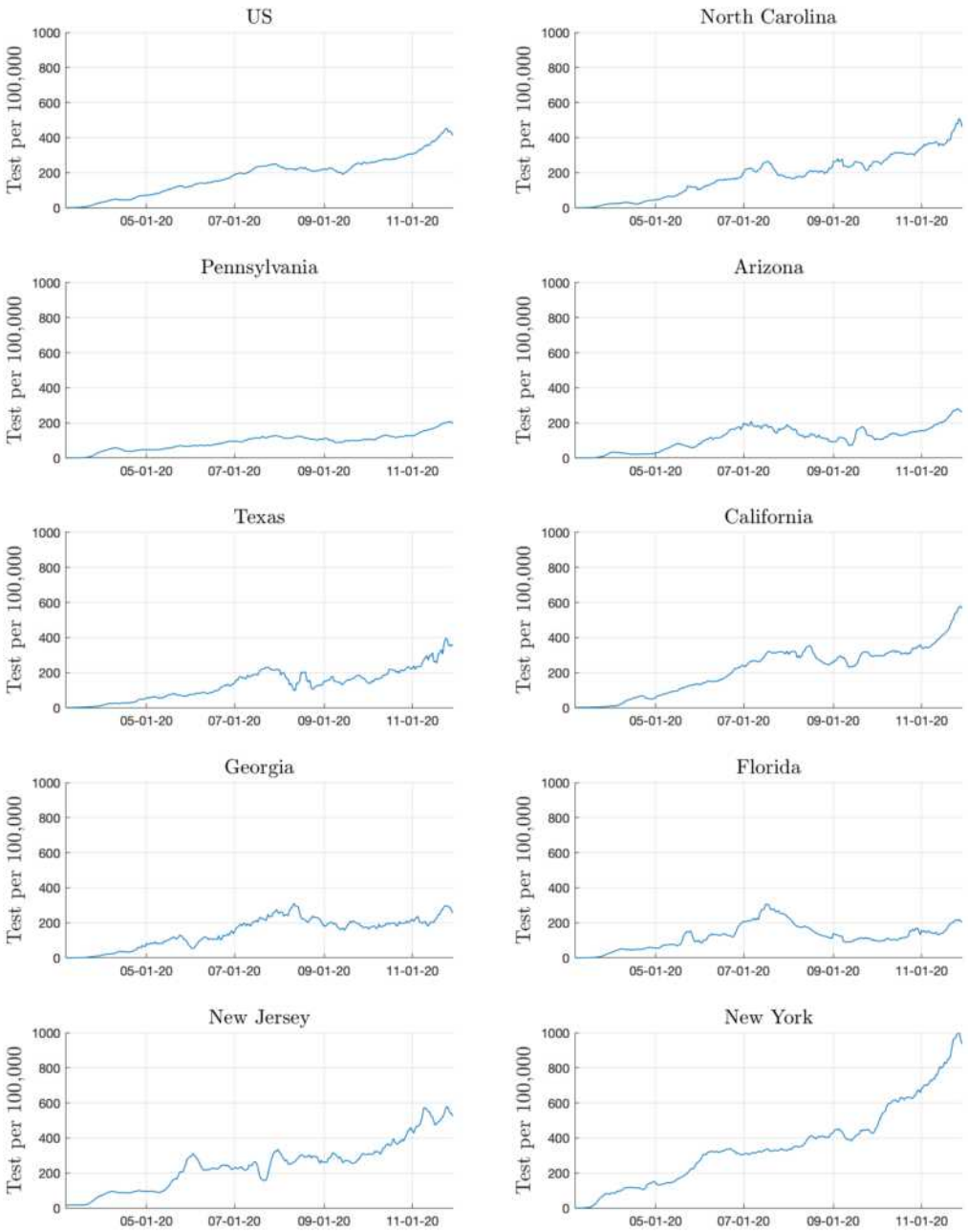
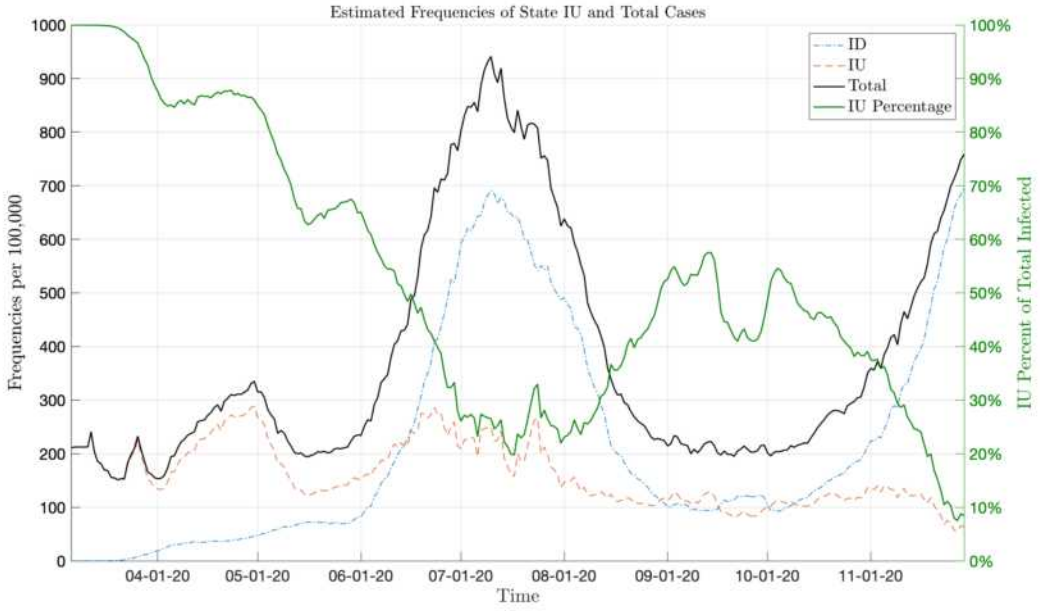
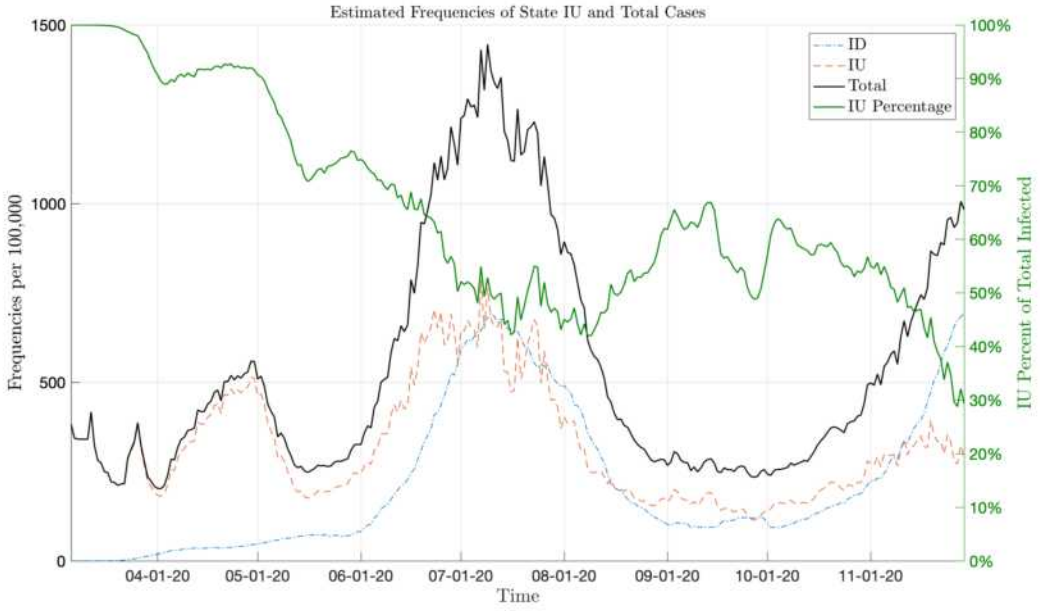


Figure 11: **The Test Intensity y_t of Individual States.**
The figure shows the time series data of the test intensity y_t of nine individual states. The time series of y_t for the US is included for a better comparison, the same plot is also shown in the bottom panel of Figure 3.

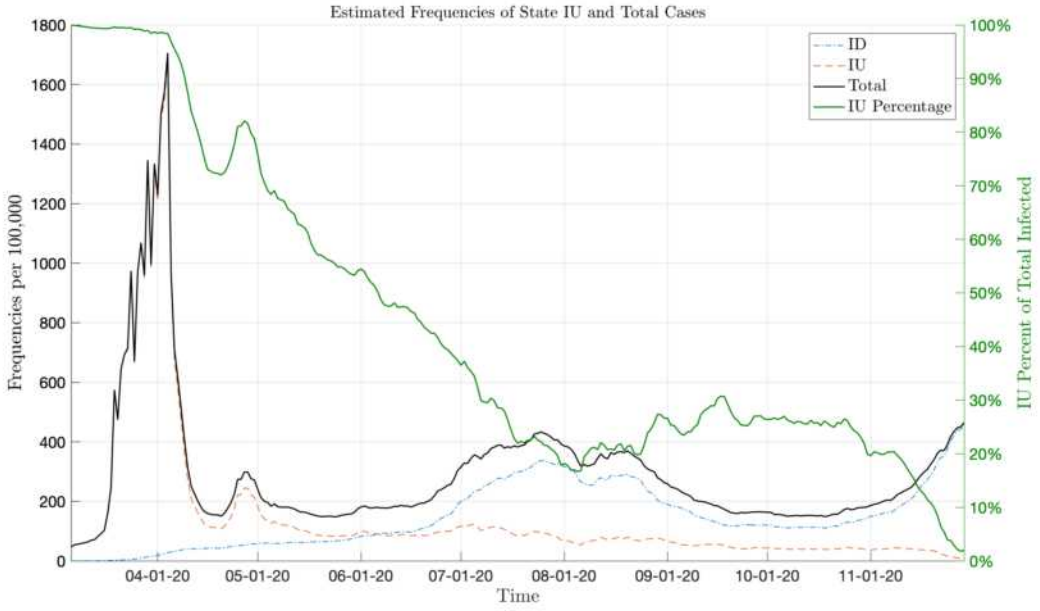


(a) **State Parameters**

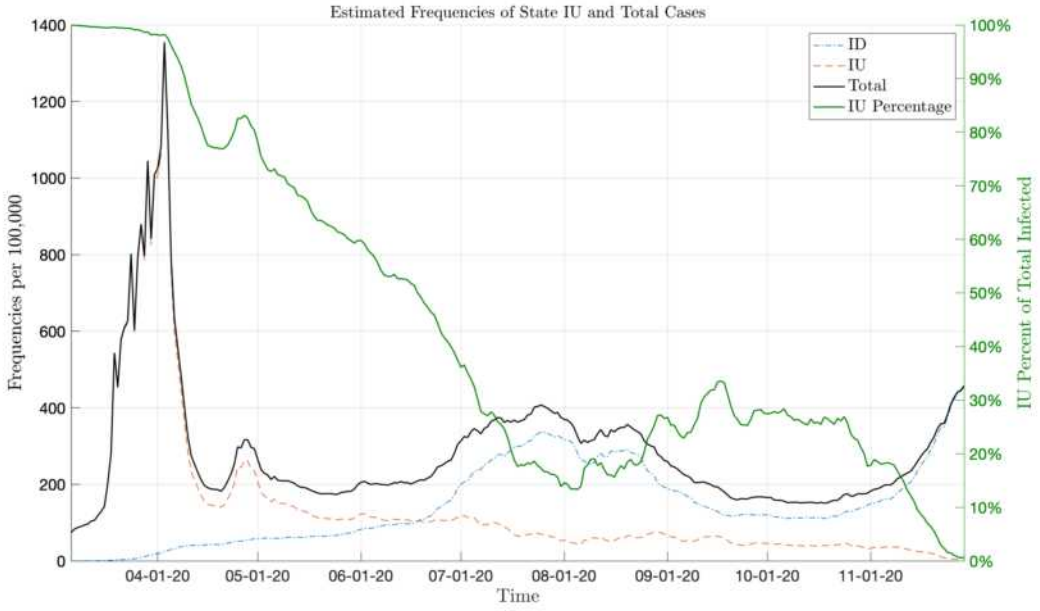


(b) **Average Parameters**

Figure 12: **Estimated Frequencies of State IU and Total Infections (AZ)**
See Figure 4 for more details.

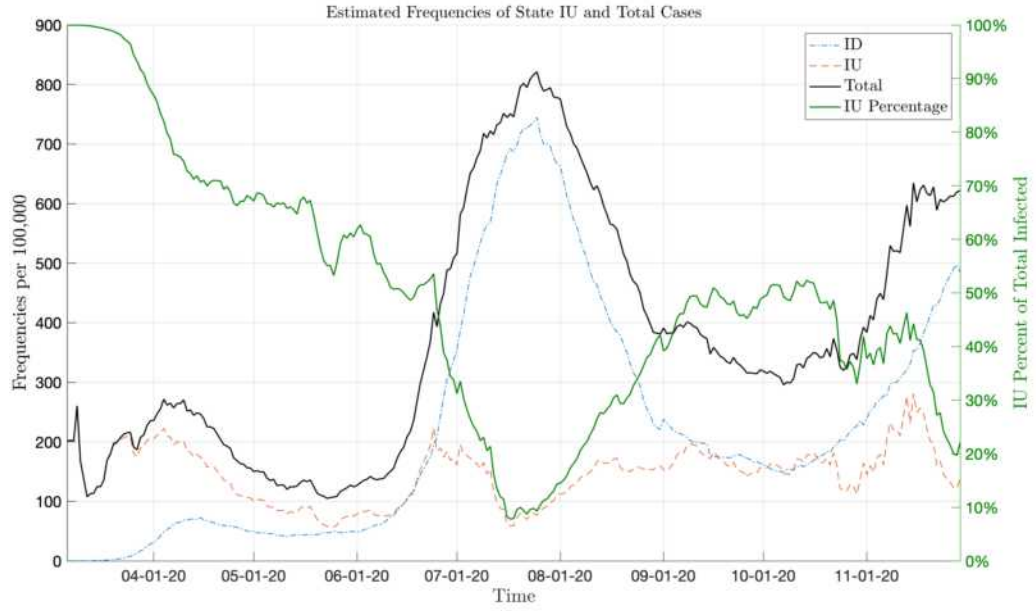


(a) **State Parameters**

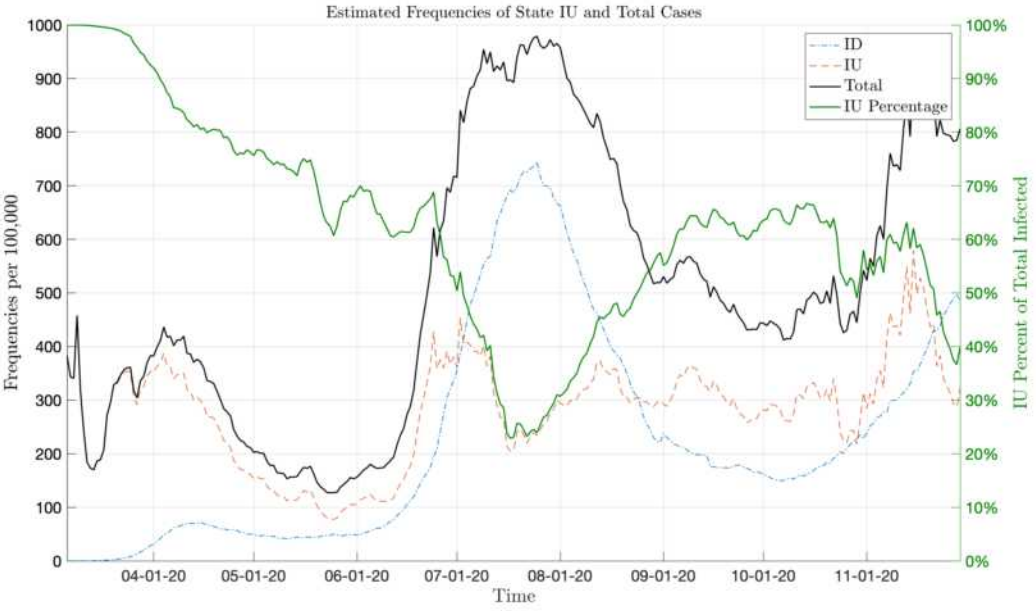


(b) **Average Parameters**

Figure 13: **Estimated Frequencies of State IU and Total Infections (CA)**
See Figure 4 for more details.

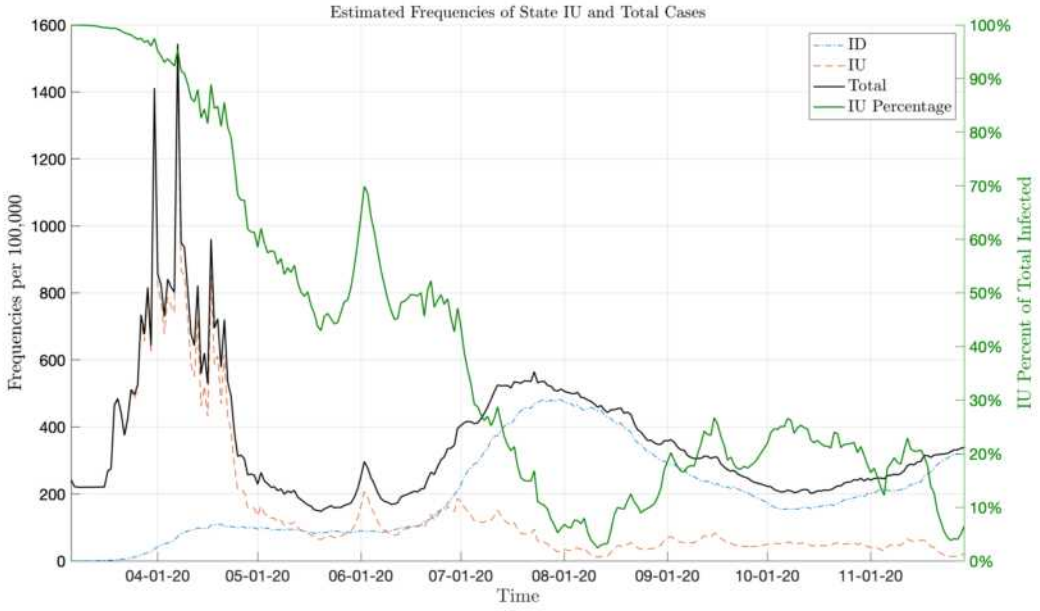


(a) **State Parameters**

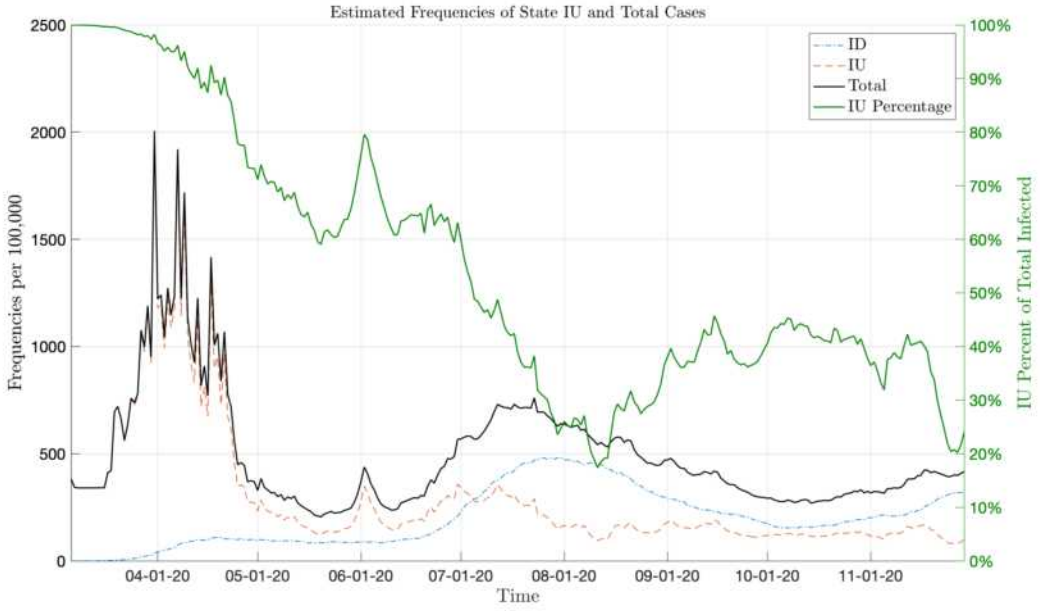


(b) **Average Parameters**

Figure 14: **Estimated Frequencies of State IU and Total Infections (FL)**
See Figure 4 for more details.

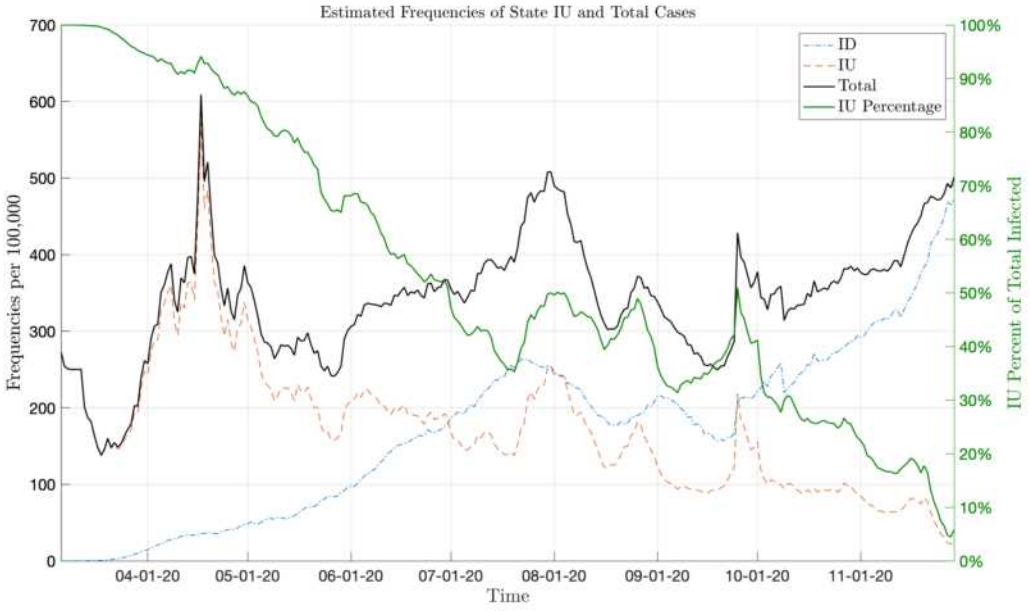


(a) **State Parameters**

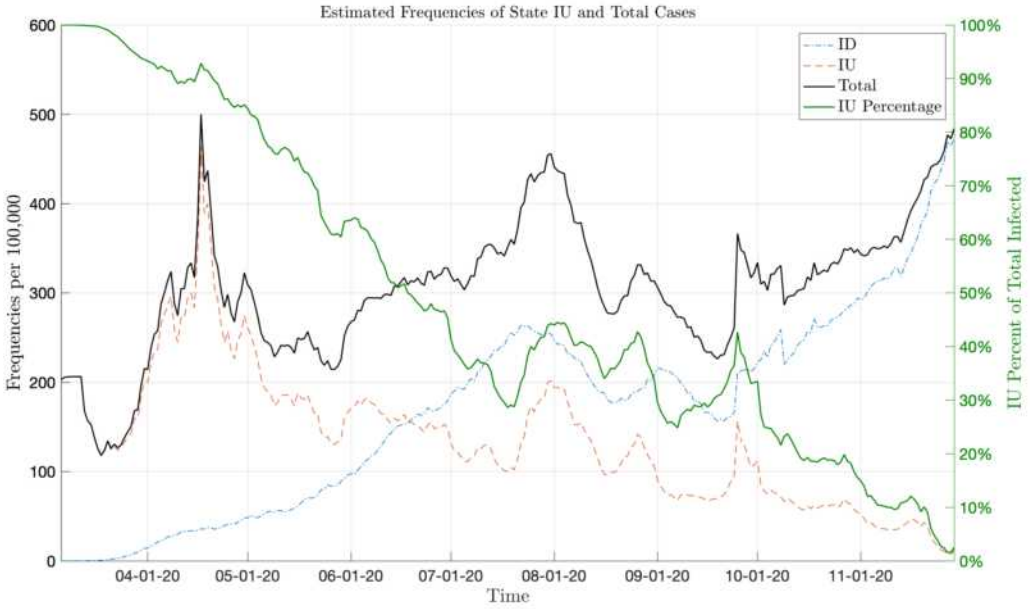


(b) **Average Parameters**

Figure 15: **Estimated Frequencies of State IU and Total Infections (GA)**
See Figure 4 for more details.

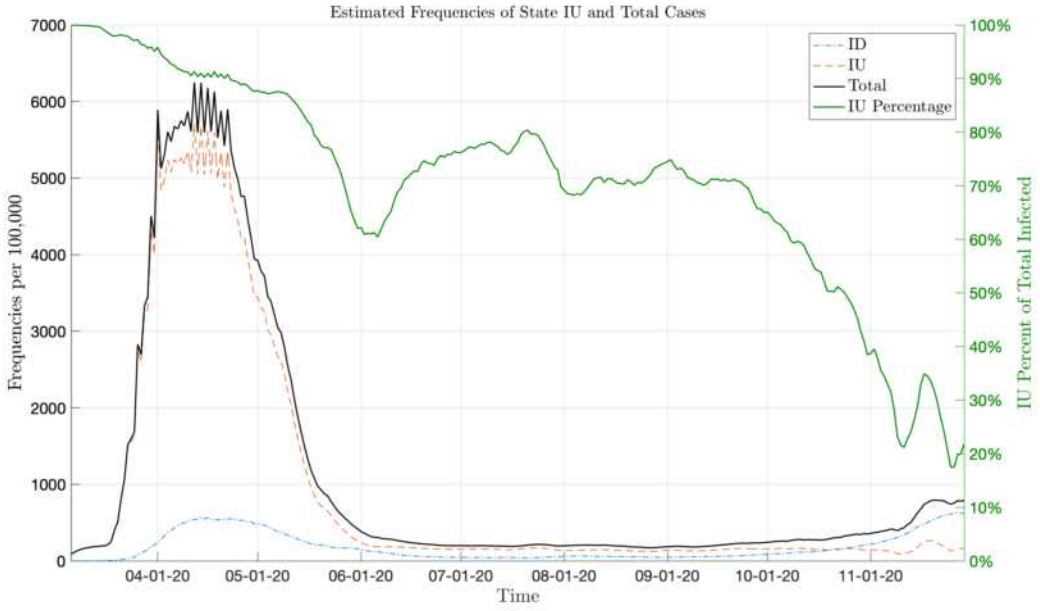


(a) **State Parameters**

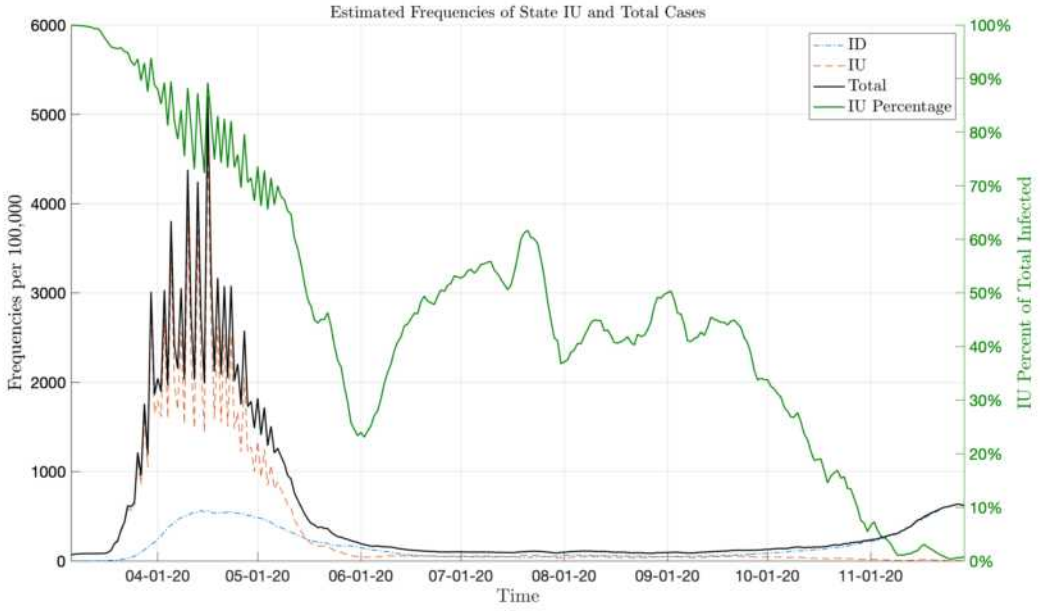


(b) **Average Parameters**

Figure 16: **Estimated Frequencies of State IU and Total Infections (NC)**
See Figure 4 for more details.

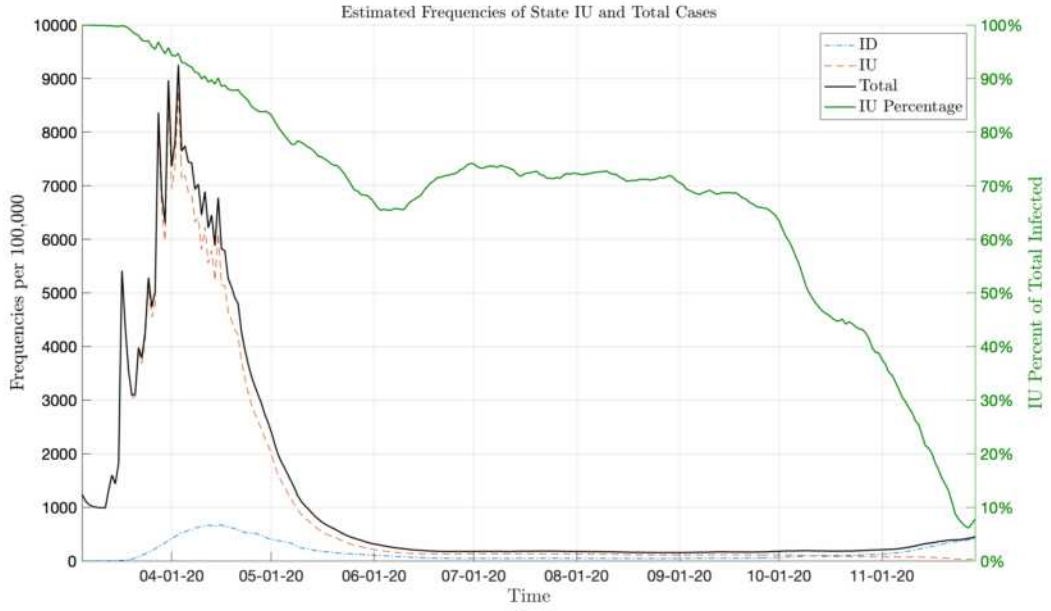


(a) **State Parameters**

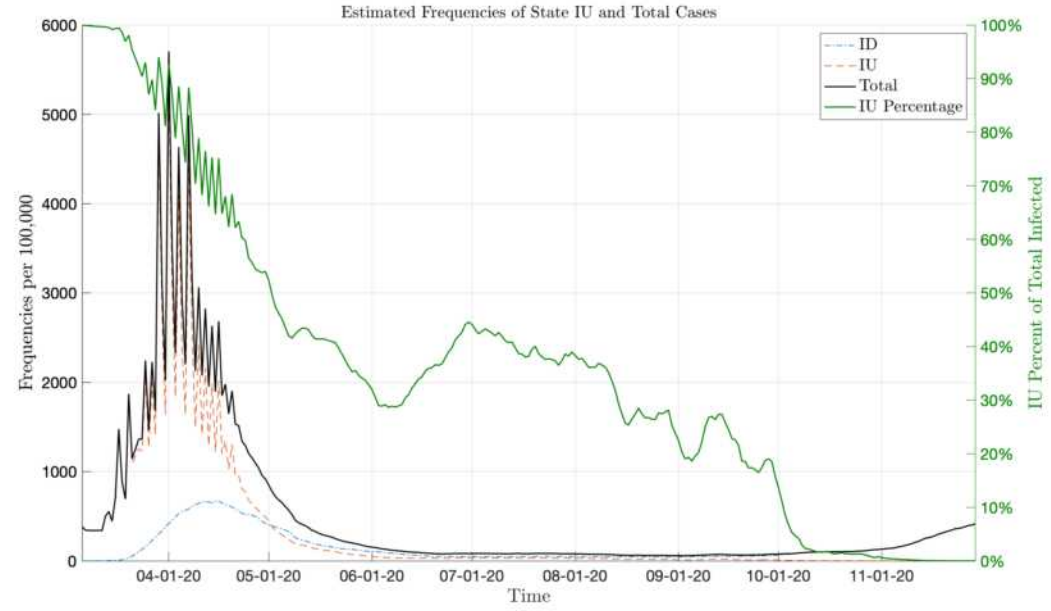


(b) **Average Parameters**

Figure 17: **Estimated Frequencies of State IU and Total Infections (NJ)**
See Figure 4 for more details.

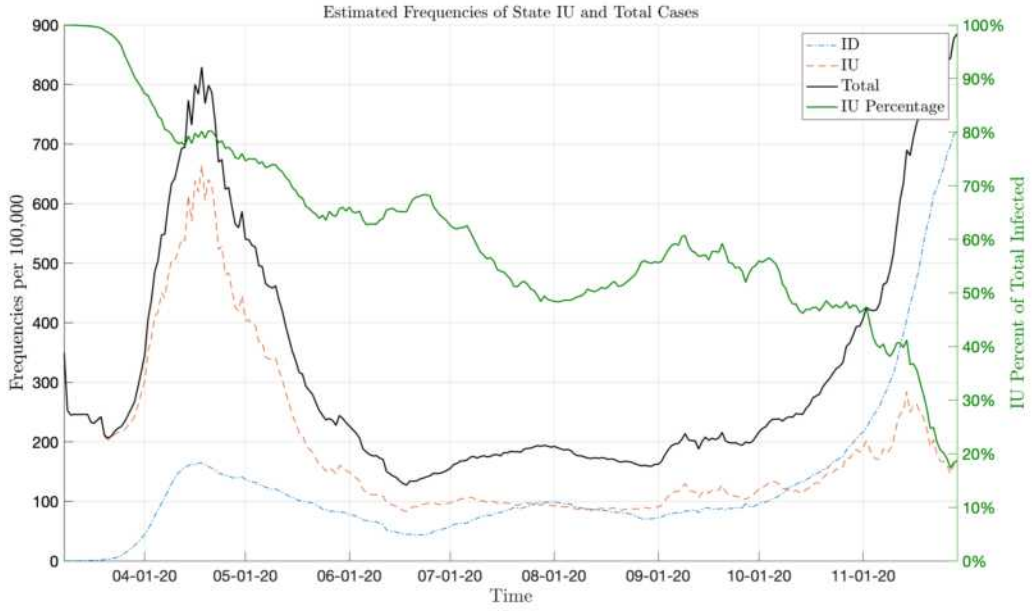


(a) **State Parameters**

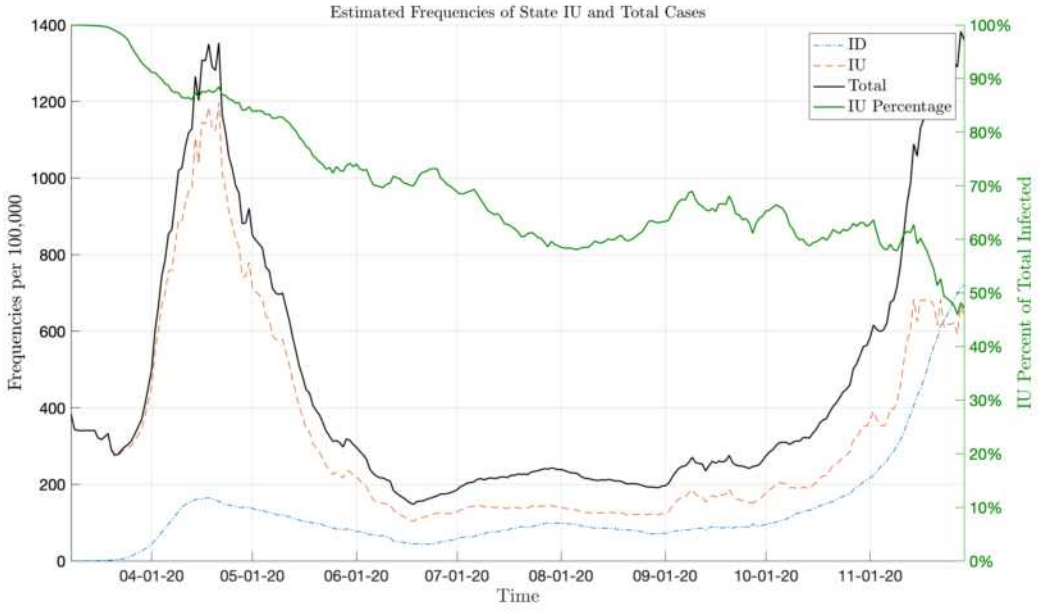


(b) **Average Parameters**

Figure 18: **Estimated Frequencies of State IU and Total Infections (NY)**
See Figure 4 for more details.

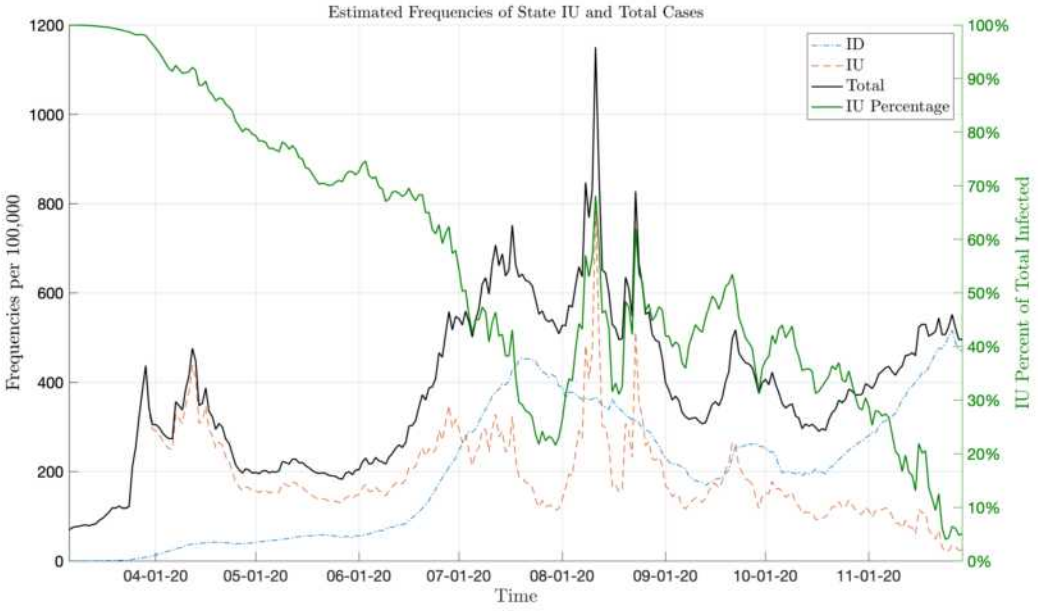


(a) **State Parameters**

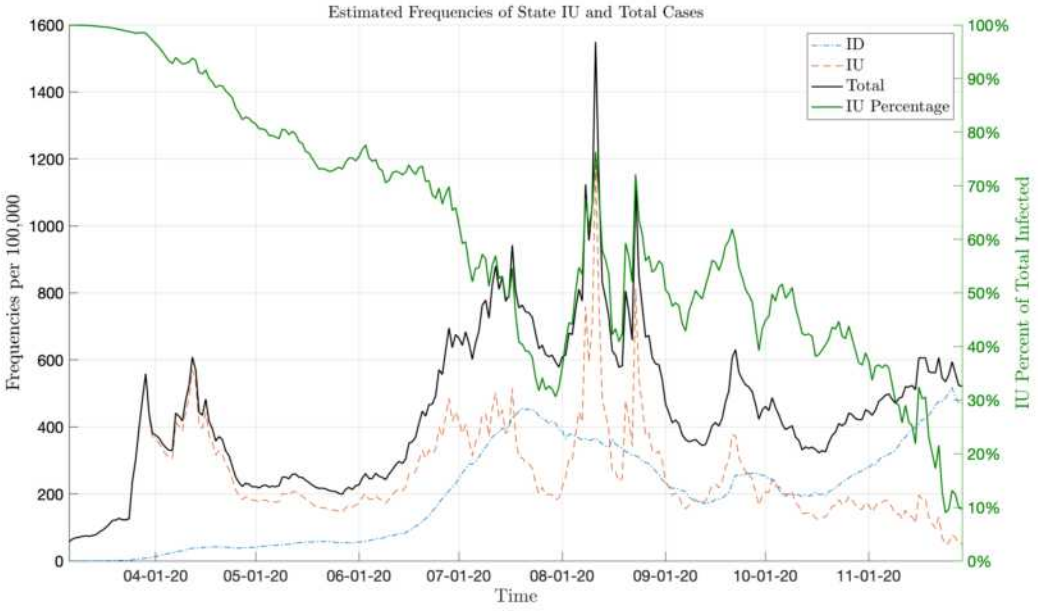


(b) **Average Parameters**

Figure 19: **Estimated Frequencies of State IU and Total Infections (PA)**
See Figure 4 for more details.



(a) **State Parameters**



(b) **Average Parameters**

Figure 20: **Estimated Frequencies of State IU and Total Infections (TX)**
See Figure 4 for more details.