Estimating Unreported COVID-19 Cases in the United States Based on the tvSIRu Model

Lingbo Liu

Wuhan University <https://orcid.org/0000-0002-9876-8506>

Tao Hu

Harvard University

Shuming Bao

China Data Institute

Hao Wu (✉ wh79@whu.edu.cn)

Wuhan University <https://orcid.org/0000-0001-7107-8081>

Zhenghong Peng

Wuhan University

Ru Wang

Wuhan University

Research Article

Keywords: SIR, Time-varying parameters, unreported infection rate, Infection fatality rate, COVID-19

Posted Date: October 7th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-85391/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Estimating unreported COVID-19 cases in the United States based on the tvSIRu model

Authors list

Lingbo Liu, School of Urban Design, Wuhan University, Wuhan, Hubei, China; lingbo.liu@whu.edu.cn;

Tao Hu, Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA; taohu@fas.harvard.edu

Shuming Bao, China Data Institute, Ann Arbor, MI, US; sbao@umich.edu

***Hao Wu**, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wh79@whu.edu.cn

Zhenghong Peng, School of Urban Design, Wuhan University, Wuhan, Hubei, China; pengzhenghong@whu.edu.cn;

Ru Wang, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wang_ru@whu.edu.cn;

*Corresponding author

***Hao Wu**, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wh79@whu.edu.cn

Summary

Background The potential unreported infection may impair and mislead policymaking for COVID-19 , and the contemporary spread of COVID-19 varies in different counties of the United States. It is necessary to estimate the cases that may be underestimated based on county-level data to take better countermeasures against COVID-19. We suggested taking time-varying SIR models with unreported infection rates (UIR) to estimate the factual COVID-19 cases in the United States.

Methods SIR integrated with unreported infection rates (SIRu) of fixed time effect and SIR with time-varying parameters (tvSIRu) were applied to estimate and compare the value of transmission rate (TR), UIR, and infection fatality rate (IFR) based on US county-level COVID-19 data.

Results Based on US county-level COVID-19 data from January 22 (T_1) to August 20 (T_{212}) in 2020, SIRu was first tested and verified by a general OLS regression. The further regression of SIRu at the country-level showed that the average values of TR, UIR, and IFR were 0.034, 19.5, 0.51% respectively. The range of TR, UIR, IFR of all states ranged were 0.007-0.157 (mean=0.048), 7.31-185.6 (mean=38.89), and 0.04%-2.22% (mean=0.22%). Among time-varying transmission rate equations, the power function showed better fitness, which indicated a decline in TR decreasing from 227.58 (T_1) to 0.022 (T_{212}). The general equation of tvSIRu showed that both the UIR and IFR were gradually increasing, wherein, the UIR has an estimate of 9.1 (95%CI = 5.7-14.0), and IFR was 0.70% (0.52%-0.95%) at T_{212} .

Interpretation Despite the decline in TR and IFR, the UIR of the United States is still on the rise, which had been supposed to decrease with sufficient tests or improved countersues. The US medical system may be largely affected by severe cases in the rapid spread of COVID-19.

Keywords

SIR, Time-varying parameters , unreported infection rate , Infection fatality rate, COVID-19

Conflict of interest statements

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Introduction

Although COVID-19 was reported several months ago¹, the coronavirus is still raging on a global scale, especially keeping surging in the United States, which is one of the most important engines of the global economic network. The epidemic in the United States will have an important impact on the global economy and politics. It is fundamental to make relatively accurate estimates of this pandemic for preventing and controlling COVID-19 in the United States^{2,3}, wherein the transmission rate (TR), infection fatality rate (IFR) are key indicators⁴.

The main obstacle to calculating such indicators is the unreported infection rate (UIR), which may be caused by insufficient testing, data depression of mild or asymptomatic patients, and time lag bias^{5,6}. Direct use of IFR values derived from official data might lead to larger error⁷. Similar research on SARs has pointed out that preferential ascertainment of severe cases and delayed reporting of deaths are the main two reasons for case fatality risk (CFR) error⁸. Beyond insufficient early testing, mild and asymptomatic patients may cause most unreported cases. In Brazil, only some moderate and severe infectives in hospitalizations have been recorded⁹. On the other hand, the time lag deviation could be explained by the incubation period of COVID-19, which floats in a wide range¹⁰ and still possess high transmittance¹¹. The incubation period is also correlated to the age of the infectives, which also directly affects IFR¹². It has been concluded that the unreported cases may lead to four kinds of uncertainty in IFR calibration, with the unclear denominator, unknown infection time, unknown incubation, and undiagnosed asymptomatic infections¹³.

Characterizing unreported cases has become a popular question in epidemic modeling on COVID-19. There is a surging body of literature calculating UIR or reported rate (RR) based on country-level data^{14,15 16}, wherein, single country-level data may lead to greater bias¹⁷. Moreover, the county-level data in the United States on recovered infectives has not been released, the calculation of IFR depends merely on the national aggregate data, which may further amplify the error. More and more studies use multinational data¹⁸, county-level data¹⁹, or country-county mixed regional data²⁰ for analysis, which greatly improves the accuracy of modeling by increasing the dimensionality and quantity of data.

However, little study has investigated the time effect of UIR, which may affect the accuracy of all the indicators. A recent study suggests using the time-varying SIR model to capture the changing transmissive rate²¹. Moreover, the incubation period has been proved to change in different stages of transmission²². Some studies believe that the COVID-19 IFR of China has been overestimated, and the possible value would be 2.3%²³, while another study shows that the IFR of COVID-19 in Wuhan in the early stage might be as high as 20%²⁴. Such disputes may also imply a changing trend in IFR.

This study proposes a SIR regression model with unreported infection rate (SIRu) and SIRu with time-varying parameters (tvSIRu) to estimate the values of TR, UIR, and IFR, and to assess the impact of time effect as well. Data for this study are the county-level data in the United States released by Johan Hopkins University²⁵. This study provides the first insights into the time series values of TR, UIR, and IFR of COVID-19, contributing to a deeper understanding of the trend of COVID-19 in the United States.

Methods

Data

The COVID-19 data used in this article contains 3142 counties in the United States, which included the number of daily new infectives, cumulative infectives, and deaths, while the population of recovered infectives remained unreported.

The date of the data was ranging from Jan 22, 2020, to August 20, 2020, which contained 666104(3142 *212) records. As a time lag order($t, t+1$) was applied in data analysis, the number of whole records used for regression was 662,962 (3142 *211).

tvSIRu model with fixed UIR

In the classic SIR dynamic model, the number of daily infectives (I_d^{t+1}) at time t_1 can be expressed by

the function of infection rate β , the number of susceptible persons (S^{t_0}), infected persons (I^{t_0}), and total population (N) at time t_0 (Equation 1).

$$I_d^{t_1} = \frac{\beta S^{t_0} I^{t_0}}{N} \quad (1)$$

The SIR with unreported infection rate (SIRu) could be illustrated in Figure 1.

Figure 1 SIR model with unreported infection cases

As the population of recovered infectives was not released, two special parameters were added to the SIR model, φ for unreported infection rate (UIR), λ for recovery/death rate (RDR), which could be described as the following equation:

$$I_c^{t_0} = \varphi I_{cr}^{t_0}, \quad R_c^{t_0} = \lambda R_{dr}^{t_0} \quad (2)$$

Wherein, $I_c^{t_0}$ represented the total cumulative infectives at time t_0 , and $I_{cr}^{t_1}$ denoted the cumulative cases reported. Then, $R_c^{t_0}$ reflected the whole population of removals at time t_0 , $R_{dr}^{t_0}$ as the cumulative death reported at time t_0 . Wherein, RDR can also be transformed into an infection fatality rate (IFR).

$$\text{IFR} = 1 / (RDR + 1) \quad (3)$$

Then the two explanatory variables in Equation 1, S^{t_0} , I^{t_0} , could be calculated as

$$S^{t_0} = N - I_c^{t_0}, \quad I^{t_0} = I_c^{t_0} - R_c^{t_0} \quad (4)$$

While UIR was fixed without time effect, then the daily new infectives at time t_1 ($I_d^{t_1}$) could also be calculated by φ and the corresponding data reported ($I_{dr}^{t_1}$):

$$I_d^{t_1} = \varphi I_{dr}^{t_1} \quad (5)$$

The SIR model (Equation 1) could be developed into Equation 5 by substituting Equation 2-4, while the further simplification and operation of Equation 7 would be taken as the general tvSIRu model.

$$\varphi I_{dr}^{t_1} = \frac{\beta(N - \varphi I_{cr}^{t_0})(\varphi I_{cr}^{t_0} - \lambda R_{dr}^{t_0})}{N} \quad (6)$$

$$I_{dr}^{t_1} = \beta I_{cr}^{t_0} - \frac{\beta \lambda}{\varphi} R_{dr}^{t_0} - \beta \varphi \frac{I_{cr}^{t_0^2}}{N} + \beta \lambda \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N} \quad (A)$$

As the four variables can be calculated by the data released, Equation A.0 could be seen as a primary linear function of Equation A with coefficients, a, b, c, d:

$$I_{dr}^{t_1} = a I_{cr}^{t_0} + b R_{dr}^{t_0} + c \frac{I_{cr}^{t_0^2}}{N} + d \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N} \quad (A.1)$$

While considering the fixed time effect of all three parameters in Equation A, the corresponding average value

$(\beta, \lambda, \varphi)$ could be calculated in Equation B.

$$I_{dr}^{t_1} = \beta_0 I_{cr}^{t_0} - \frac{\beta_0 \lambda_0}{\varphi_0} R_{dr}^{t_0} - \beta_0 \varphi_0 \frac{I_{cr}^{t_0^2}}{N} + \beta_0 \lambda_0 \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N} \quad (A.2)$$

Wherein, the values of $\beta_0, \lambda_0, \varphi_0$ are constants.

tvSIRu model with time-varying UIR

If the UIR varies over time, the UIR of the cumulative case and daily new cases also be different, which was defined as φ and φ' respectively. Equation 7 would be rewritten as

$$\varphi' I_{dr}^{t_1} = \frac{\beta(N - \varphi I_{cr}^{t_0})(\varphi I_{cr}^{t_0} - \lambda R_{dr}^{t_0})}{N} \quad (7)$$

To simply the computation, a new parameter β' was introduced:

$$\beta' = \beta / \varphi' \quad (8)$$

Then the time-varying model (tvSIRu) could be transformed into a similar form of equation A:

$$I_{dr}^{t_1} = \beta' I_{cr}^{t_0} - \frac{\beta' \lambda}{\varphi} R_{dr}^{t_0} - \beta' \varphi \frac{I_{cr}^{t_0^2}}{N} + \beta' \lambda \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N} \quad (B)$$

To verify the assumption of time-varying parameters, the coefficients in Equation A and B could be represented by initial values and time effect functions. Such functions were substituted into the two models gradually, result in several sub-equations with time effects.

$$\beta = \beta_0 g(t) \quad (9)$$

$$\beta = \beta_0 g(t) \quad , \quad \lambda = \lambda_0 h(t) \quad (10)$$

$$\beta' = \beta'_0 g'(t) \quad , \quad \lambda = \lambda_0 h(t) \quad , \quad \varphi = \varphi_0 f(t) \quad (11)$$

Substituting the Equation 9-11 into Equation A and B, then three complete functions could be generated:

$$I_{dr}^{t_1} = \beta_0 g(t) I_{cr}^{t_0} - \frac{\beta_0 \lambda g(t)}{\varphi} R_{dr}^{t_0} - \beta_0 g(t) \varphi \frac{I_{cr}^{t_0^2}}{N} + \beta_0 g(t) \lambda \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N} \quad (A.2)$$

$$I_{dr}^{t_1} = \beta_0 g(t) I_{cr}^{t_0} - \frac{\beta_0 g(t) \lambda_0 h(t)}{\varphi} R_{dr}^{t_0} - \beta_0 g(t) \varphi \frac{I_{cr}^{t_0^2}}{N} + \beta_0 g(t) \lambda_0 h(t) \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N} \quad (A.3)$$

$$I_{dr}^{t_1} = \beta'_0 g'(t) (\varphi_0 f(t) I_{cr}^{t_0} - \lambda_0 h(t) R_{dr}^{t_0} - \varphi_0^2 f(t)^2 \frac{I_{cr}^{t_0^2}}{N} + \lambda_0 h(t) \varphi_0 f(t) \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N}) \quad (B.0)$$

In terms of the specific functions reflecting time effect, the power, exponential, and periodic function had been tested and compared in this article:

$$\tau_1(t) = x^t \quad , \quad \tau_2(t) = t^x \quad , \quad \tau_3(t) = \frac{1}{2} \left(1 + \cos\left(\frac{t}{x}\pi\right) \right) \quad (12)$$

This paper tested the five equations, A.0, A.1, A2, A3, and B.0, wherein, A.0 is the OLS linear regression derived from SIRu model, Equation A.1 is SIRu with fixed time effect, A.2, A.3, and B.0 are tvSIRu with single time-varying β , time-varying β and λ , all time-varying parameters of β , λ and φ , respectively.

Results

OLS and SIRu Regressions

The linear regression derived from the SIRu model showed acceptable fitness, and the Adjusted R^2 was 0.4813 ($n=662,962$) (Table 1). The negative value of coefficients b and c were Consistent with the corresponding operation signs in Equation A.0. Such results may verify the assumption of the SIRu to a certain extend.

Table 1 Linear SIR Regression estimated by Equation A.0

	$I_{dr}^{t_1} = aI_{cr}^{t_0} + bR_{dr}^{t_0} + c\frac{I_{cr}^{t_0^2}}{N} + d\frac{I_{cr}^{t_0}R_{dr}^{t_0}}{N}$				
	Estimate	Std. Error	t value	P-value	Significance
Intercept	0.94454128	0.06176004	15.29	<0.001	***
a	0.02831670	0.00006719	421.44	<0.001	***
b	-0.18534723	0.00114767	-161.50	<0.001	***
c	-0.53924176	0.00236657	-227.86	<0.001	***
d	4.67187687	0.02415657	193.40	<0.001	***
Adjusted R²	0.4813			<0.001	***
AIC	7059288				

The SIRu with a fixed time effect in Equation A.1 further provided the estimated value of transmission rate β , unreported infection rate ϕ , and the recovery/death rate λ (Table 2). The result showed that the average β value from Jan 22 to August 20 was 0.0339 (95%CI 0.0338-0.0340), and the ϕ value was 19.5(95%CI 19.38-19.54), which implied that there might exist 19.5 cases undiagnosed while one infection reported in US counties averagely. Meanwhile, the λ value of 192.5 (95%CI 191.790-193.243) could be interpreted as an IFR value of 0.516%.

Table 2 SIR Regression estimated by Equation A.1

	$I_{dr}^{t_1} = \beta I_{cr}^{t_0} - \frac{\beta\lambda}{\phi} R_{dr}^{t_0} - \beta\phi \frac{I_{cr}^{t_0^2}}{N} + \beta\lambda \frac{I_{cr}^{t_0}R_{dr}^{t_0}}{N}$			
	Estimate	Std. Error	t value	P-value
β	0.03393158	0.00005614	604.4	<0.001
ϕ	19.46038052	0.04151802	468.7	<0.001
λ	192.51630090	0.37073915	519.3	<0.001
AIC	7080522			

SIRu at the State level

The study further utilized county-level data to compare state-level parameters based on fixed time effects. Figure 2 shows the fitness of Equation A.0 across the whole states, most of which were above 0.5 (Figure 2), and each state has different values of TR, UIR, and RDR in Equation A.1, which indicated an obvious spatial heterogeneity in the transmission of COVID-19(Figure 3). All the parameters and statistical descriptions are reported in Appendix A and B.

Figure 2 State-level fitness of Equation A.0 with county-level data

Most of the states had a TR between 0.018-0.053, Seven of the states with relatively high values are Illinois (0.146), Massachusetts (0.109), Connecticut (0.104), New Jersey (0.098), Nevada (0.080), Arizona (0.087), and Alaska (0.076) (*Figure 3a*).

In terms of UIR, most states are concentrated between 28-50(*Figure 3b*). Some states have relatively low values, such as New York (7.31) and Oregon (8.64), while the top five states were Main (122.84), Vermont (185.66), Alaska (85.69), West Virginia (80.90).

The fitting results on RDR in some cities are not significant, but most of the significant values are between 200-500, which is equivalent to the value of IFR ranging from 0.2% to 0.5% (*Figure 3c*). Wherein, eight cities were reported below 99 (IFR > 1%), including Ohio (44.29), Oklahoma (49.35), Florida (77.97), Alabama (66.40), Mississippi (98.32), Kentucky (74.40), Iowa (58.45), New Mexico (62.18), California (55.91).

The Pearson correlation between the three state-level indicators was also tested, showing a positive correlation between UIR and RDR. In other words, the lower the IFR, the higher the UIR (*Figure 3d*).

Figure 3 State-level parameters of Equation A.1 with county-level data:

(a) Transmissive rate; (b) Unreported infection Rate; (c) Recovery/Death Rate; (d) Correlation Test

tvSIRu regression at the country level

tvSIRu with time-varying TR was first tested by three sub-equations of Equation A.2, and the AIC of all equations was reduced comparing to the SIRu model of fixed time effect (*Table 3*). Meanwhile, all estimated IR displayed a declining trend (*Figure 4*). Wherein, the power function showed the best fitness with an initial extremely high value of 227.58 (95%CI 219.89-235.27) decreasing to 0.022 on August 20. Such a high value may reflect the high contagiousness of COVID-19 in the early stage. The corresponding UIR and RDR was 18.61 (95%CI 18.52-18.69) and 183.34 (95%CI 182.63-184.05), which were slightly higher than the values in Equation A.1.

Table 3 Time-varying TR estimated by Equation A.2

$$I_{dr}^{t_1} = \beta_0 g(t) I_{cr}^{t_0} - \frac{\beta_0 \lambda g(t)}{\phi} R_{dr}^{t_0} - \beta_0 g(t) \phi \frac{I_{cr}^{t_0^2}}{N} + \beta_0 g(t) \lambda \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N}$$

	$g(t) = m^t$ (A.2.1)		$g(t) = t^m$ (A.2.2)		$g(t) = \frac{1}{2} \left(1 + \cos \left(\frac{t}{m} \pi \right) \right)$ (A.2.3)	
	Estimate	P value	Estimate	P value	Estimate	P-value
β_0	0.24983960	<0.001	227.58625	<0.001	0.05253167	<0.001
φ	18.50696966	<0.001	18.610006	<0.001	19.7915931	<0.001
λ	181.9526655	<0.001	183.34371	<0.001	196.000582	<0.001
m	0.98836902	<0.001	-1.722910	<0.001	43.5300211	<0.001
AIC	6982233		6962783		7076624

Figure 4 Time-varying TR estimated by Equation A.2

When the time effect of RDR was further added in Equation A.3, the AIC of power function displayed a slight decrease in equation A.3 (Table 4). Wherein, the UIR was 19.02 (95%CI 18.93-19.12), which is similar to the value in Equation A. However, both equations showed decreasing trends in the changing RDR, implying an increase of IFR (Figure 5).

Table 4 Time-varying TR and RDR estimated by Equation A.3

$I_{dr}^{t_1} = \beta_0 g(t) I_{cr}^{t_0} - \frac{\beta_0 g(t) \lambda_0 h(t)}{\varphi} R_{dr}^{t_0} - \beta_0 g(t) \varphi \frac{I_{cr}^{t_0^2}}{N} + \beta_0 g(t) \lambda_0 h(t) \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N}$				
	$g(t) = m^t, h(t) = k^t$ (A.3.1)		$g(t) = t^m, h(t) = t^k$ (A.3.2)	
	Estimate	P-value	Estimate	P-value
β_0	0.25239807	<0.001	241.912633	<0.001
φ	18.66057736	<0.001	19.024919	<0.001
λ_0	196.63534702	<0.001	284.386081	<0.001
m^*	0.98828594	<0.001	-1.734960	<0.001
k	0.99949887	<0.001	-0.085439	<0.001
AIC	6980073		6959144

Figure 5 Time-varying RDR with 95%CI estimated by Equation A.3

The power function also showed better performance in tvSIRu with all three time-varying parameters estimated by Equation B.0, which indicated a gradual increase in both UIR and RDR (Table 5). This trend indicated that the initial UIR and RDR were relatively low (Figure 6). The value of UIR and RDR achieved 9.1 (95%CI = 5.7-14.0) and 141.706 (95%CI=103.3358-189.9486) at T_{212} on August 20 respectively. IFR could be calculated as 0.70% (95%CI =0.52%-0.95%). Based on the officially released data on August 20,2020, it can be concluded that about 30% of the whole population had been infected.

Table 5 Time-varying UIR and RDR estimated by Equation B.0

$I_{dr}^{t_1} = \beta'_0 g'(t) I_{cr}^{t_0} - \frac{\beta'_0 g'(t) \lambda h(t)}{\varphi f(t)} R_{dr}^{t_0} - \beta'_0 g'(t) \varphi f(t) \frac{I_{cr}^{t_0^2}}{N} + \beta'_0 g'(t) \lambda h(t) \frac{I_{cr}^{t_0} R_{dr}^{t_0}}{N}$				
$g(t) = m^t, f(t) = n^t, h(t) = k^t$ (E.1)			$g(t) = t^m, f(t) = t^n, h(t) = t^k$ (E.2)	
	Estimate	P-value	Estimate	P-value
β'_0	0.25075898	<0.001	40.166018546	<0.001
φ_0	15.22876377	<0.001	0.010973001	<0.001
λ_0	143.41791888	<0.001	0.000087692	<0.001
m^*	0.98383494	<0.001	-2.689048329	<0.001
n	1.00187869	<0.001	1.255552967	<0.001
k	1.00135561	<0.001	2.668759146	<0.001
AIC	6969888		6813832

Figure 6 Time-varying UIR, RDR with 95%CI estimated by Equation B.0

Discussion

Few studies have analyzed the unreported infection rate (UIR) with time-variable effects of COVID-19, and it impacts on the estimation of the transmission rate (TR) and IFR. The initial purposes of this project were to estimate the value of UIR, TR, and IFR of time-fixed effect and time-varying effect with tvSIRu models based on county-level data.

In terms of fixed time effect, the results show that from January 22 to August 20, the average TR and UIR at the country level in the United States were 0.03 and 19.5 respectively, and the RDR was 19.5, which also mean the infection fatality rate was 0.516%. The IFR is slightly lower than the overall infection mortality rate of 0.66% estimated in China¹⁷, while the CDC recommends 0.65%²⁶.

In further analysis of the state level, the UIR of all states was ranging from 7.32-185.66 (mean=38), and the IFR range is 0.037%-2.20%(mean=0.21%). Comparing with related research on 20 US counties which have UIR of 4.32 - 776.68(mean=27.7) and IFR of 0.02% - 1.81%(mean=0.027%), the range of UIR estimated in this paper is more concentrated, and the IFRs has similar upper boundary²⁷. Four states' upper boundary of UIR had been reported in the previous study: Illinois(40.86), Massachusetts(38.28), New Jersey(29.22), and New York(35.17)¹⁹, of which the first three are similar to the value estimated by SIRu, which are 41.51, 39.22 and 31.83, only New York had a different value of 7.32. But interestingly, the study also pointed out that the UIR estimated by an antibody test in New York State in early May was around 7.6, which may indicate the stability of the proposed SIRu method.

Based on the time-varying effect at the country level, UIR and IFR are increasing following the power function rather than the exponent function which is the default used in many research²¹. Other than the average value of 0.03 in SIRu, the TR estimated by tvSIRu is decreasing from a large value of 227 to the value of 0.022 on August 20, which is much lower than the fixed value 0.05-0.06 reported in related research²¹. It may further explain the high contagiousness in the initial stage in COVID-19 transmission. The increasing UIR estimated by tvSIRu has a similar value of 9.1 (95%CI = 5.7-14.0) at T₂₁₂(August 20), which is very close to the value of 9 estimated in a former study on April²⁰, and the latest study in September. The UIR value is also close to the value reported in Brazil (Reported rate=9.2%, UIR=10.8)¹⁸. Such similarity in the estimated UIR in different periods may be caused by the fixed time effect in the former models, which only represent the average values of UIR as the SIRu model calculated. The increasing UIR means that the IFR is on a downward trend. The value of IFR on August 20 was 0.70% (95%CI =0.52%-0.95%), which is still close to the

value recommend by CDC²⁶.

Many studies have supposed that the UIR will decrease with the improvement of COVID-19 testing and increased hygiene awareness, but our research shows that UIR in the United States is increasing, which may have a great impact on policy-making for COVID-19 prevention. On the other hand, empirical IF is often used in contemporary COVID-19 modeling, but the tvSIRu model indicates that the COVID-19 infection rate changes dramatically. The initial value of TR is 246, reflecting that this pandemic is extremely contagious in the early transmission stage of the United States. Previous SIR modeling has seldom characterized such a feature, which may lead to large estimation errors. The reducing TR, IFR, and increasing UIR indicated by the model show that the epidemic is spreading rapidly in the United States with a large number of self-healing populations, however, it is noteworthy the potential increasing cases of severe illnesses would greatly affect the medical system, and relevant departments still need to provide more protection to high-risk groups.

Conclusion

This article indicates that there may be an increasing number of unrecorded COVID-19 cases in official U.S. data, wherein, the tvSIRu model provides a simple, convenient and relatively accurate calculation of the unreported parameters of COVID-19 with time effect based on official released data. Moreover, this method can be easily transplanted to analyze or epidemic modeling of other countries.

It must be admitted that if single geography unites of data is used, the independent variables may display strong collinearity, leading to overfitting, it is necessary to use proper sub-geographical level data to fit the national-level or state-level data. Furthermore, the non-linear model regression was based on the Gauss-Newton iteration, which could be further optimized with machine learning models.

Appendix A

Parameters estimated by Equation A on State level

State	R ²	β (TR)	ϕ (UIR)	β (RDR)	IFR
Alabama	0.6800	0.0384	28.1118	66.4050	0.0148
Alaska	0.7043	0.0758	85.6993	2677.0740	0.0004
Arizona	0.8344	0.0873	31.8510	459.3959	0.0022
Arkansas	0.4323	0.0202	14.4224	-103.5099*	
California	0.7199	0.0337	31.9797	55.9131	0.0176

Colorado	0.6327	0.0324	55.2171	395.2257	0.0025
Connecticut	0.4209	0.1039	55.4896	389.6164	0.0026
Delaware	0.3171	0.0875	38.4571	708.5320	0.0014
District of Columbia	0.7049	0.1575	50.9941	780.3851	0.0013
Florida	0.7076	0.0391	13.4101	79.9765	0.0123
Georgia	0.7657	0.0402	28.1156	195.6581	0.0051
Hawaii	0.8935	0.0928	70.3956	2208.2030	0.0005
Idaho	0.7048	0.0588	33.2108	466.1497	0.0021
Illinois	0.7909	0.1457	41.5079	693.7160	0.0014
Indiana	0.6182	0.0301	29.8719	255.2491	0.0039
Iowa	0.4786	0.0190	15.7639	58.4500	0.0168
Kansas	0.5360	0.0230	16.0095	-81.6163*	
Kentucky	0.5781	0.0259	25.9078	74.4067	0.0133
Louisiana	0.4316	0.0658	25.3544	284.9328	0.0035
Maine	0.5311	0.0223	122.8401	-543.0694*	
Maryland	0.6333	0.0406	35.1383	325.5810	0.0031
Massachusetts	0.6904	0.1089	39.2230	448.5265	0.0022
Michigan	0.4950	0.0533	56.0760	339.5087	0.0029
Minnesota	0.7441	0.0068	30.1427	-1098.4030*	
Mississippi	0.5179	0.0349	21.9778	98.3266	0.0101
Missouri	0.7096	0.0412	31.3774	188.5405	0.0053
Montana	0.5230	0.0229	11.8879	-139.4672*	

Nebraska	0.5644	0.0250	13.2247	400.8106	0.0025
Nevada	0.9250	0.0806	35.9869	514.9621	0.0019
New Hampshire	0.4695	0.0446	40.4871	486.1162	0.0021
New Jersey	0.5094	0.0986	31.8312	323.4239	0.0031
New Mexico	0.5075	0.0197	13.4675	62.1788	0.0158
New York	0.4734	-0.0221*	7.3187	244.7202	0.0041
North Carolina	0.7307	0.0355	39.0629	182.6843	0.0054
North Dakota	0.6131	0.0511	49.7167	715.7587	0.0014
Ohio	0.6007	0.0213	28.9136	46.3505	0.0211
Oklahoma	0.7452	0.0362	33.4416	44.2904	0.0221
Oregon	0.6663	0.0166	8.6465	-24.6175*	
Pennsylvania	0.5268	0.0547	47.0019	407.8571	0.0024
Rhode Island	0.2214	0.0373	40.5705	-56191.3200*	
South Carolina	0.6950	0.0563	31.2758	433.8097	0.0023
South Dakota	0.4069	0.0500	39.8930	1084.9280	0.0009
Tennessee	0.6854	0.0200	17.9252	-388.7768*	
Texas	0.5605	0.0424	28.6138	250.5731	0.0040
Utah	0.9043	0.0361	42.7415	119.1223	0.0083
Vermont	0.1490	0.0322	185.6625	1314.2170	0.0008
Virginia	0.5568	0.0299	48.3013	259.4586	0.0038
Washington	0.5531	0.0213	15.1840	115.7273	0.0086
West Virginia	0.3163	0.0346	80.9098	497.3075	0.0020

Wisconsin	0.7368	0.0206	28.0754	-231.9747*	
Wyoming	0.2622	0.0185	34.7665	404.7622	0.0025

* P value>0.05

Appendix B

Statistical Prescription of State level Parameters on COVID-19

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
TR	0.006824	0.0235	0.036772	0.047826	0.055902	0.157463
UIR	7.318659	25.63112	31.9797	38.89121	42.12469	185.6625
RDR	44.29042	135.0128	332.5448	456.1626	481.1246	2677.074

Declarations

Ethical Approval and Consent to participate

Not applicable

Consent for publication

We confirmed that all authors have approved the manuscript for submission

Availability of supporting data

The data used in this paper is available in the Github repository by John Hopkins

University

<https://github.com/CSSEGISandData/COVID-19>

Competing interests

The authors declare that they have no known competing financial interests or personal

relationships that could have appeared to influence the work reported in this paper.

Funding

National Key Research and Development Project (2019YFB2101803)

National Natural Science Foundation of China (52078390)

Wuhan University Experiment Technology Project Funding

Authors' contributions

Lingbo Liu contributed to the conception of the study;

Lingbo Liu, Hao Wu performed the experiment;

Lingbo Liu, Tao Hu contributed significantly to analysis and manuscript preparation;

Lingbo Liu, Ru Wang performed the data analyses and wrote the manuscript;

Shuming Bao, Zhenghong Peng helped perform the analysis with constructive discussions.

Acknowledgments

The authors would like to acknowledge Xun Shi and other experts for their suggestions on the presentation of Data Statistical Analysis and Spatio-temporal Prediction Models of COVID-19 based on Workflow in the COVID-19 Data Analysis Webinar.

Authors' information

Lingbo Liu, School of Urban Design, Wuhan University, Wuhan, Hubei, China; lingbo.liu@whu.edu.cn;

Tao Hu, Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA; taohu@fas.harvard.edu

Shuming Bao, China Data Institute, Ann Arbor, MI, US; sbao@umich.edu

*Hao Wu, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wh79@whu.edu.cn

Zhenghong Peng, School of Urban Design, Wuhan University, Wuhan, Hubei, China; pengzhenghong@whu.edu.cn;

Ru Wang, School of Urban Design, Wuhan University, Wuhan, Hubei, China; wang_ru@whu.edu.cn;

*Corresponding author

Reference

1. Peng Z, Wang R, Liu L, Wu H. Exploring Urban Spatial Features of COVID-19 Transmission in Wuhan Based on Social Media Data. *ISPRS International Journal of Geo-Information* 2020; **9**(6): 402.
2. Hu T, Guan WW, Zhu X, et al. Building an Open Resources Repository for COVID-19 Research. *Data and Information Management* 2020; **4**(3): 130.
3. Yang C, Sha D, Liu Q, et al. Taking the pulse of COVID-19: a spatiotemporal perspective. *International Journal of Digital Earth* 2020: 1-26.
4. Leon DA, Shkolnikov VM, Smeeth L, Magnus P, Pechholdová M, Jarvis CI. COVID-19: a need for real-time monitoring of weekly excess deaths. *The Lancet* 2020; **395**(10234): e81.
5. Jung S-m, Akhmetzhanov AR, Hayashi K, et al. Real-Time Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases. *Journal of Clinical Medicine* 2020; **9**(2): 523.
6. Spychalski P, Błazyńska-Spychalska A, Kobiela J. Estimating case fatality rates of COVID-19. *Lancet Infect Dis* 2020; **20**(7): 774-5.
7. Abdollahi E, Champredon D, Langley JM, Galvani AP, Moghadas SM. Temporal estimates of case-fatality rate for COVID-19 outbreaks in Canada and the United States. *Canadian Medical Association Journal* 2020; **192**(25): E666-E70.
8. Lipsitch M, Donnelly CA, Fraser C, et al. Potential Biases in Estimating Absolute and Relative Case-Fatality Risks during Outbreaks. *Plos Neglect Trop D* 2015; **9**(7): e0003846-e.
9. Sousa GJB, Garces TS, Cestari VRF, Florêncio RS, Moreira TMM, Pereira MLD. Mortality and survival of COVID-19. *Epidemiol Infect* 2020; **148**: e123.

10. Guan W-j, Ni Z-y, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine* 2020.
11. Wang X, Zhou Q, He Y, et al. Nosocomial outbreak of COVID-19 pneumonia in Wuhan, China. *European Respiratory Journal* 2020; **55**(6): 2000544.
12. Donnelly CA, Ghani AC, Leung GM, et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *The Lancet* 2003; **361**(9371): 1761-6.
13. Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet* 2020; **395**(10228): 931-4.
14. Lau H, Khosrawipour V, Kocbach P, et al. Internationally lost COVID-19 cases. *Journal of Microbiology, Immunology and Infection* 2020; **53**(3): 454-8.
15. Liu Z, Magal P, Seydi O, Webb G. A COVID-19 epidemic model with latency period. *Infectious Disease Modelling* 2020; **5**: 323-37.
16. Cakmakli C, Simsek Y. Bridging the COVID-19 Data and the Epidemiological Model using Time Varying Parameter SIRD Model 2020.
<https://ui.adsabs.harvard.edu/abs/2020arXiv200702726C> (accessed July 01, 2020).
17. Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020; **20**(6): 669-77.
18. Prado MFd, Antunes BBdP, Bastos LDSL, et al. Analysis of COVID-19 under-reporting in Brazil. *Rev Bras Ter Intensiva* 2020; **32**(2): 224-8.
19. Srivastava A, Prasanna V. Data-driven Identification of Number of Unreported Cases for

COVID-19: Bounds and Limitations; 2020.

20. Chow C, California J, Gerkin R, Vattikuti S. Global prediction of unreported SARS-CoV2 infection from observed COVID-19 cases; 2020.

21. Zhou YaW, Lili and Zhang, Leyao and Shi, Lan and Yang, Kangping and He, Jie and Zhao, Bangyao and Overton, William and Purkayastha, Soumik and Song, Peter. A Spatiotemporal Epidemiological Prediction Model to Inform County-Level COVID-19 Risk in the United States. *Harvard Data Science Review* 2020.

22. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England Journal of Medicine* 2020.

23. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* 2020; **323**(13): 1239-42.

24. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. *The Lancet Infectious Diseases* 2020; **20**(7): 773.

25. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 2020; **20**(5): 533-4.

26. Meyerowitz-Katz G, Merone L. A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. *medRxiv* 2020: 2020.05.03.20089854.

27. Hortaçsu AL, Jiarui ; Schweg, Timothy. Estimating the Fraction of Unreported Infections in Epidemics with a Known Epicenter: An Application to COVID-19. *SSRN* 2020; **University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2020-37**: No. 2020-37.

Figure legend

Figure 1 SIR model with unreported infection cases

SIRu model introduced a parameter of unreported infection in the traditional SIR model, which is supposed to have the same parameters of transmission rate, infection fatality rate as the reported infection.

Figure 2 State-level fitness of Equation A.0 with county-level data

The scaled density curve of adjusted R^2 shows that Equation A was generally applicable, and its mapping indicated that the potential spatial heterogeneity of the states will affect the results of SIRu modeling. Among them, the states in the southeastern, the west coast, and the Great Lakes Region show higher adaptability.

Figure 3 State-level parameters of Equation A.1 with county-level data

(a) Transmission rate. three obvious clusters could be identified, Nevada-Arizona, Illinois, and Massachusetts-New Jersey, wherein the coefficient of New York could not be applied due to the non-significant p-value; (b) Unreported infection Rate. The UIR in the northeast is relatively high, but there were also two central states with high values; (c) Recovery/Death Rate. Blank blocks indicate that the RDR in the area is not applicable due to the insignificant p-value, wherein, the RDR of northeastern cities is relatively higher, while the west coast states have both high TR and RDR. (d) Correlation Test. The Pearson correlation test of all states' parameters with significant p-values shows an obvious connection between RDR and TR, UIR.

Figure 4 Time-varying TR estimated by Equation A.2

Although the initial values of the power function are much higher than the exponential function, in the medium term, the two values tend to be the same, while the periodic function shows that it was in the third wave.

Figure 5 Time-varying RDR with 95%CI estimated by Equation A.3

If the time effect of UIR is not considered, the fitting results show that RDR exhibits a decreasing effect over time, which means that IFR may be slowly increasing.

Figure 6 Time-varying UIR, RDR with 95%CI estimated by Equation B.0

Equation B only provides the estimated values of UIR and RDR. Both the power function and the exponential function imply an increasing effect, wherein, the power function is much smaller than the exponential function in terms of UIR estimation.

Figures

Figure 1

SIR model with unreported infection cases SIRu model introduced a parameter of unreported infection in the traditional SIR model, which is supposed to have the same parameters of transmission rate, infection fatality rate as the reported infection

Figure 2

State-level fitness of Equation A.0 with county-level data The scaled density curve of adjusted R2 shows that Equation A was generally applicable, and its mapping indicated that the potential spatial heterogeneity of the states will affect the results of SIRu modeling. Among them, the states in the southeastern, the west coast, and the Great Lakes Region show higher adaptability.

Figure 3

State-level parameters of Equation A.1 with county-level data (a) Transmission rate. three obvious clusters could be identified, Nevada-Arizona, Illinois, and Massachusetts-New Jersey, wherein the coefficient of New York could not be applied due to the non-significant p-value;(b) Unreported infection Rate. The UIR in the northeast is relatively high, but there were also two central states with high values; (c) Recovery/Death Rate. Blank blocks indicate that the RDR in the area is not applicable due to the insignificant p-value, wherein, the RDR of northeastern cities is relatively higher, while the west coast states have both high TR and RDR. (d)Correlation Test. The Pearson correlation test of all states' parameters with significant p-values shows an obvious connection between RDR and TR, UIR.

Figure 4

Time-varying TR estimated by Equation A.2 Although the initial values of the power function are much higher than the exponential function, in the medium term, the two values tend to be the same, while the periodic function shows that it was in the third wave.

Figure 5

Time-varying RDR with 95%CI estimated by Equation A.3 If the time effect of UIR is not considered, the fitting results show that RDR exhibits a decreasing effect over time, which means that IFR may be slowly increasing.

Figure 6

Time-varying UIR, RDR with 95%CI estimated by Equation B.0 Equation B only provides the estimated values of UIR and RDR. Both the power function and the exponential function imply an increasing effect,

wherein, the power function is much smaller than the exponential function in terms of UIR estimation.