Dating the First Case of COVID-19 Epidemic from a Probabilistic Perspective

Authors: Zhouwang Yang, Yunhe Hu, Zhiwei Ding, Tiande Guo, Tiande Guo

Date: 2021-09-22T00:00:00+00:00

Abstract

In the early stages of the coronavirus disease 2019 (COVID-19) epidemic, limited understanding of the pandemic, insufficient nucleic acid testing capacity, and delayed data reporting, among other factors, rendered the determination of the pandemic' s origin time challenging. Therefore, source tracing is crucial for infectious disease prevention and control. The objective of this paper is to infer the origin time of the COVID-19 pandemic based on a hybrid data- and model-driven method.

We model the testing positive rate to fit its actual trend and employ least squares estimation to obtain the optimal model parameters. Furthermore, kernel density estimation is applied to infer the pandemic' s origin time given a specific confidence probability.

By selecting 12 representative regions in the United States for analysis, the estimated dates of the first infected case with 50% confidence probability predominantly fall between August and October 2019, which is earlier than the officially announced date of the first confirmed case in the United States on January 20, 2020. The experimental results indicate that the COVID-19 pandemic in the United States likely began spreading around September 2019 with high confidence probability.

Additionally, existing confirmed cases from Wuhan City and Zhejiang Province in China are utilized to infer the origin time of COVID-19 and provide the associated confidence probability. The results indicate that the spread of the COVID-19 pandemic in China likely began in late December 2019.

Full Text

Preamble

Dating the First Case of COVID-19 Epidemic from a Probabilistic Perspective

Zhouwang Yang, Yunhe Hu, Zhiwei Ding, Tiande Guo*

University of Science and Technology of China
University of Chinese Academy of Sciences

*Corresponding author: tdguo@ucas.ac.cn (Tiande Guo)

ABSTRACT

In the early days of the coronavirus disease 2019 (COVID-19) epidemic, determining the origin time of the outbreak proved difficult due to limited understanding of the pandemic, inadequate nucleic acid testing capacity, and delays in data reporting. Source tracing is therefore crucial for infectious disease prevention and control. This paper aims to infer the origin time of the COVID-19 pandemic using a hybrid data- and model-driven approach.

We model the testing positive rate to fit its actual trend and employ least squares estimation to obtain optimal model parameters. Furthermore, kernel density estimation is applied to infer the pandemic' s origin time given a specific confidence probability.

By analyzing 12 representative regions in the United States, we find that the dates of the first infected case with 50% confidence probability mostly fall between August and October 2019—significantly earlier than the officially announced date of the first confirmed case in the United States on January 20, 2020. The experimental results indicate that the COVID-19 pandemic in the United States likely began spreading around September 2019 with high confidence probability.

Additionally, we utilize existing confirmed case data from Wuhan City and Zhejiang Province in China to infer the origin time of COVID-19 and provide corresponding confidence probabilities. The results suggest that the COVID-19 pandemic in China most likely began spreading in late December 2019.

KEYWORDS: Pandemic; COVID-19; Testing positive rate; Infectious disease dynamic model; Least squares optimization; Kernel density estimation.

INTRODUCTION

Throughout human history, identifying the origins of infectious diseases has often taken decades, and even now not all questions have been answered. Recently, investigating the origin, spread, and evolution of coronavirus disease 2019

(COVID-19) across more than 200 countries and regions has become a critical research subject for the global scientific community. Source tracing is essential for infectious disease prevention and control, yet the scientific demonstration process is complex, requiring extensive biological information and epidemiological evidence to converge into a mutually supportive chain of proof—a process that is both time-consuming and uncertain. Previous studies have demonstrated that the United States, Spain, France, Italy, Brazil, and other countries experienced coronavirus attacks before the outbreak in China.

The primary task of disease tracing is to identify the first case. Although the first known case in Wuhan, China, represents the first confirmed case reported, it does not necessarily represent the true index case sought by investigators. Throughout the history of humanity' s fight against infectious diseases, few successful precedents for origin tracing exist.

The main method for COVID-19 origin tracing is molecular traceability [1]. This approach first requires establishing a global coronavirus information database to integrate genomic, epidemiological, and clinical data. Subsequently, by analyzing molecular and epidemiological data, researchers can systematically investigate correlations and patterns between this coronavirus series and various exposure factors, providing important references for traceability.

Epidemic spread is a complex process involving numerous factors [2], some of which remain difficult to ascertain. However, epidemic data implicitly reflect the comprehensive influence of these factors. Theoretically, analyzing these big data can also reveal the laws of epidemic spread. Therefore, another method for COVID-19 origin tracing relies on big data analysis. Combining mathematical models with artificial intelligence technology enables qualitative and quantitative analysis of infectious diseases, revealing epidemic patterns and detecting disease origins and development trends. While many studies have used epidemic models and data for forward prediction [3-7], few have employed mathematical modeling and big data analysis for backward tracing [8,9].

In the early days of COVID-19, most countries, including the United States and China, lacked fundamental knowledge of the epidemic situation, and nucleic acid testing was not widely implemented. Additional problems such as data scarcity, reporting lags, and distortion further complicated determination of the epidemic' s origin time. To address this, we collected daily epidemic data for the U.S., including the number of newly confirmed cases, new deaths, Nucleic Acid Amplification Tests (NAATs), and test positive rates. After analyzing the characteristics of these data, we selected the number of nucleic acid tests and the testing positive rate for each U.S. state as our modeling data. Based on classical infectious disease models and statistical methods, we established an optimization model and obtained model parameters using least squares estimation. We then inferred the origin time of the epidemic in selected U.S. states and used kernel density estimation to determine the dates of the first infection, 50 infections, and 100 infections in these states at probability levels of 0.5, 0.6, 0.7, and 0.8.

RESULTS

Data Description

The primary data used for modeling in this study is the daily testing positive rate, defined as the daily proportion of positive nucleic acid tests relative to the total number of COVID-19 nucleic acid tests. Data on the total number of tests and positive tests for each U.S. state were obtained from the official website of the United States Department of Health and Human Services [10]. By examining the early testing positive rate curves of more than 50 U.S. states, we identified that 13 states and the District of Columbia (primarily in the Northeast), where excess mortality peaked earlier in 2020 [11], share the same pattern: the testing positive rate rises rapidly to a peak after a brief fluctuation.

Table 1 shows the cumulative number of tests, population, and test ratio (percentage of total tests in the population) for these 13 states and the District of Columbia at the peak of the testing positive rate. When the testing positive rate peaked in New Jersey and Vermont, the cumulative number of tests was less than 1,000, yielding too small a test ratio for reliable analysis; consequently, these two states were excluded from subsequent analysis.

Using Maryland as an example (Figure 1 [Figure 1: see original paper]), the black dots represent the calculated daily testing positive rate, while the red line represents values after 15-day smoothing (7 days before and after each day). Smoothing reduces the impact of data fluctuations, and unless otherwise specified, the testing positive rate refers to this smoothed value. In Maryland, the positive rate began increasing from 8% on March 18, reached a peak of nearly 30% on April 15, and subsequently declined.

The remaining 10 states and the District of Columbia exhibit the same characteristics as Maryland, as shown in Figure 2 [Figure 2: see original paper]. All states began opening commercial nucleic acid testing around March 15; before this date, limited testing capacity and insufficient test numbers caused the positive rate to fluctuate and not reflect the actual situation. Therefore, we selected only the steadily increasing sequence from the first valley to the first peak for modeling.

The 11 states and District of Columbia share the characteristic that the testing positive rate rises to a peak shortly after the beginning, reflecting approximately natural propagation during the early U.S. epidemic stage. If our infectious disease model can accurately fit this initial rising phase of the positive rate for each state, it indicates that the model effectively captures the temporal spread of the epidemic, enabling us to trace its origin by looking backward in time.

Date Tracing Process

The process of tracing the pandemic' s origin time consists of three main steps.

Step 1. Perform 15-day smoothing on the daily positive rate of the target dis-

trict to reduce random noise impact. If abnormal fluctuations occur in the early period, discard that data portion and select only the sequence corresponding to the first stably rising period. Denote this time interval as T, with length τ, where the endpoints correspond to the trough time and crest time, respectively.

Step 2. Take 14 consecutive days from interval T as fitting data $y_i, i = 1, \dots, 14$, use the two-parameter exponential epidemic model of the testing positive rate to obtain the fitting function $\widehat{y}$, and record the fitting accuracy index MAPE (Mean Absolute Percentage Error) $|\widehat{y}_i - y_i|$. Denote the number of people engaged in NAATs in the target district as M. Extend the positive rate fitting function $\widehat{y}$ backward historically and solve for the time $\bar{t}_1$ when $\widehat{y}(t)M = 1$, representing the occurrence time of the first case in the target district. Similarly, solve for time $\bar{t}_{50}$ when $\widehat{y}(t)M = 50$ and time $\bar{t}_{100}$ when $\widehat{y}(t)M = 100$, representing the occurrence times of 50 and 100 cases, respectively. Since $\widehat{y}(t) \to 0$ as $t \to -\infty$, these equations must have solutions.

Step 3. Use 14 days as the fitting window size and 1 day as the step size to perform sliding sampling on interval T. Repeat Step 2 for each window to obtain $\tau - 13$ retrospective dates and MAPE values. Apply kernel density estimation to obtain the probability distribution of the origin time and calculate the average MAPE as the evaluation index of overall fitting accuracy.

Origin Time of 11 States and District of Columbia in the U.S.

Using Maryland as an example to trace the epidemic' s origin, the rising period of the testing positive rate spans from March 18, 2020, to April 15, 2020. The observation data, transmission model data, and retrospective data are shown in Figure 3 [Figure 3: see original paper]. The interval between the blue dotted lines corresponds to March 18, 2020, to April 15, 2020, while the interval between the gray dotted lines represents one of the sliding data windows used for fitting.

Through sliding the fitting window, we obtain several inferred dates for the first case in Maryland, with the corresponding probability density shown in Figure 4A [Figure 4: see original paper]. Similarly, dates for 50 and 100 cases are inferred, with their probability densities shown in Figures 4B and 4C. The two red lines in each figure represent the mean line (left) and the density peak line (right).

Table 2 presents the results of inferring COVID-19 pandemic origins for 11 states and the District of Columbia, including dates of the first infection, 50 infections, and 100 infections at probability levels of 50%, 60%, 70%, and 80%. For Maryland, the probabilities that the first infection occurred before September 22, 2019, October 6, 2019, October 19, 2019, and November 2, 2019, are 50%, 60%, 70%, and 80%, respectively.

The average MAPE for modeling each state's testing positive rate is less than 5%, indicating high model accuracy. In several states with short positive rate rising periods, we appropriately reduced the fitting window length to obtain more

origin inference results, ensuring the accuracy of kernel density estimation.

Our analysis used the cumulative number of positive tests to trace COVID-19 origins. However, due to testing limitations, the actual number of infected individuals far exceeds the number of positive tests. Therefore, these inferences represent relatively conservative estimates, with inferred origin dates being relatively late. Authoritative studies have shown that actual COVID-19 infections in the United States are 3 to 20 times the number of confirmed cases [12], indicating that early epidemic detection in the U.S. was severely insufficient, resulting in significant underestimation of infected cases. To address this, we expanded the cumulative number of people involved in nucleic acid testing up to the positive rate peak by factors of 3, 5, 10, 15, and 20, respectively, to infer earlier epidemic onset dates. Maryland' s testing rate at the positive rate peak is at a medium level among selected regions; as a representative example, we expanded Maryland' s cumulative test numbers before the positive rate peak by these multiples, with corresponding origin dates and probabilities shown in Table 3 .

Origin Time of Wuhan City and Zhejiang Province in China

The number of "existing confirmed cases" is defined as the "cumulative number of confirmed cases" minus the "sum of cumulative recovered cases and cumulative deaths." Since China adopts a mass testing strategy for epidemic prevention and control [13], the testing positive rate remains very low most of the time, making it unsuitable for modeling. However, this strategy ensures that the number of existing confirmed cases in China more closely approximates the actual number of infections. Therefore, we directly use the number of existing confirmed cases instead of the testing positive rate, selecting Wuhan City and Zhejiang Province as two representative regions to trace COVID-19' s origin.

Changes in existing confirmed cases in Wuhan City and Zhejiang Province [14,15] are shown in Figure 5 [Figure 5: see original paper]. For Wuhan (Figure 5A), existing confirmed cases increased sharply on February 12, 2020, primarily due to the revision of confirmed case definitions—specifically, including clinically diagnosed cases in the confirmed case count. The number of existing confirmed cases peaked on February 18, 2020. For Zhejiang (Figure 5B), existing confirmed cases peaked on February 7, 2020.

To improve result reliability, we varied the time interval and sliding window size for multiple numerical experiments, selecting the model with the smallest MAPE as the final model. For Wuhan, we used January 27, 2020, to February 11, 2020, and January 27, 2020, to February 18, 2020, as fitting time intervals while varying window size. For Zhejiang, we used January 22, 2020, to February 7, 2020; January 23, 2020, to February 7, 2020; January 24, 2020, to February 7, 2020; and January 25, 2020, to February 7, 2020, as fitting time intervals. Model results for both regions are listed in Table 5 , with the selected model for origin time inference marked with a star in the MAPE column.

Through sliding the fitting window, we obtain multiple inferred dates for the first case, 50 cases, and 100 cases in Wuhan, with corresponding probability density shown in Figure 6 [Figure 6: see original paper]. Table 5 shows that the probabilities of the first infection occurring in Wuhan before December 20, 2019, December 22, 2019, December 24, 2019, and December 26, 2019, are 50%, 60%, 70%, and 80%, respectively. The probabilities of the first infection occurring in Zhejiang before December 23, 2019, December 31, 2019, January 6, 2020, and January 14, 2020, are 50%, 60%, 70%, and 80%, respectively.

CONCLUSION

Based on an infectious disease transmission model and big data analysis methods, this paper establishes an optimization model. Using daily data from 12 representative U.S. regions, we obtain model parameters for each region and infer the dates of the first case, 50 cases, and 100 cases of COVID-19 infection with corresponding probabilities. For these 12 representative regions, the dates of first infection at 50% probability mostly fall between August and October 2019, with the earliest being April 26, 2019, for Rhode Island and the latest being November 2019 for Delaware—all earlier than January 20, 2020, the officially announced date of the first confirmed U.S. case. The results indicate that the COVID-19 epidemic in the United States most likely began spreading around September 2019 with high probability.

Applying this model to daily existing confirmed case numbers in Wuhan City and Zhejiang Province, China, we obtain model parameters and infer infection times with corresponding probabilities. For Wuhan, the date of the first COVID-19 case at 50% probability is inferred as December 20, 2019, while for Zhejiang, the first case date is inferred as December 23, 2019. These results show that the COVID-19 epidemic in China most likely began spreading in late December 2019.

If early-stage detection data from other countries or regions are relatively accurate, this method can be used to infer epidemic origin times and provide dates for the first case or specific case numbers at given probability levels.

METHODS

Epidemic Model

The classic infectious disease dynamic model assumes that the number of infected persons increases exponentially in the early epidemic stage under non-intervention conditions with approximately natural transmission, intuitively presenting a J-shaped curve. This assumption aligns with the early U.S. epidemic situation. Literature [16] proposed the following infectious disease transmission model:

$$N(t) = N(t_0)\exp\{a_t(t - t_0)\}, \tag{1}$$

where $N(t)$ is the number of existing infections at time t, $N(t_0)$ and t_0 are constants, and a_t varies with time t. Due to the short initial spread period before viral mutation, we can assume a_t remains constant over time, denoted as a. After simplifying model (1), we obtain a two-parameter exponential model for the number of existing infections:

$$N(t) = e^{b+at}. \tag{2}$$

Let S denote the set of the cumulative tested population up to the first peak of the testing positive rate in the target district. The daily testing population can be regarded as random sampling from set S, with the test positive rate representing the infection rate of set S. Multiplying the positive rate by M (the number of elements in S) yields the number of existing infections in the test population, which is much lower than the actual number of existing infections in the target district during the same period. The testing positive rate curve synchronizes with the existing infection curve, differing only by a constant multiple; thus, model (2) also applies to the testing positive rate. Denoting the testing positive rate in the target district as y, we obtain from the two-parameter exponential model (2):

$$y(t) = e^{c_0+c_{1t}}, \tag{3}$$

where c_0, c_1 are model parameters to be determined and can be estimated from observational testing positive rate data.

Least Squares Optimization

Assuming a total of n days of observational data $(t_i, y_i), i = 1, \dots, n$ are collected, we establish the following least squares optimization model to estimate parameters c_0 and c_1:

$$\min_{c_0,c_1} \sum (y_i - e^{c_0+c_{1t}i})^2. \tag{4}$$

To simplify calculations, we first take the natural logarithm of both sides of model (3) to obtain $\log y = c_0 + c_{1t}$, then reformulate the least squares optimization model for the transformed data $(t_i, \log y_i), i = 1, \dots, n$:

$$\min_{c_0,c_1} \sum (\log y_i - c_0 - c_{1t}i)^2. \tag{5}$$

Solving model (5) yields the optimal solution $(\hat{c}_0, \hat{c}_1)$, thus obtaining the epidemic spread model $\hat{y}(t) = e^{\hat{c}_0+\hat{c}_{1t}}$.

Kernel Density Estimation

Kernel density estimation, as a non-parametric method, can estimate unknown probability distributions without prior knowledge. The principle states that if a certain value appears in observations, its probability density is relatively large; values close to it also have relatively large probability density, while distant values have relatively small probability density. Therefore, a function satisfying these conditions can approximate the probability density for each observed value, and summing all such functions yields the probability density function after normalization.

Assuming several possible origin times for the target district are calculated as $x_i, i = 1, \ldots, m$ based on different data fitting intervals, the probability density of the origin date at x is:

$$h(x) = \sum K\left(\frac{x - x_i}{h}\right),$$

where the kernel function K satisfies $\int K(x)dx = 1$, and the smoothing parameter h is called bandwidth. The kernel function is generally a symmetric, unimodal probability density function. Here, we select the commonly used Gaussian kernel:

$$K(x) = \exp(-x^2/2)/\sqrt{2\pi},$$

and the bandwidth choice follows Silverman' s rule of thumb [17].

FUNDING

This work is supported by the Anhui Center for Applied Mathematics and the NSF of China (No. 11871447).

REFERENCES

[1] Shan K J, Wei C, Wang Y, Huan Q, Qian W. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process[J]. The Innovation, 2021, doi: https://doi.org/10.1016/j.xinn.2021.100159.

[2] Wu R, Ai S, J Cai, et al. Predictive Model and Risk Factors for Case Fatality of COVID-19: A Cohort of 21,392 Cases in Hubei, China[J]. The Innovation, 2020, 1(2): 100022.

[3] Shen, M., Peng, Z., Xiao, Y. and Zhang, L. Modeling the Epidemic Trend of the 2019 Novel Coronavirus Outbreak in China. The Innovation, 2020, 1(3): 100048.

[4] Sun H, Qiu Y, Yan H, Huang Y, Zhu Y and Chen S. Tracking and Predicting COVID-19 Epidemic in China Mainland. medRxiv, 2020.

[5] Chen Y, Lu P, Chang C and Liu T. A Time-Dependent SIR Model for COVID-19 With Undetectable Infected Persons. IEEE Transactions on Network Science and Engineering, 2020, 7(4).

[6] Giordano G, Blanchini F and Bruno R. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine, 2020.

[7] Zhang Y, You C, Cai Z, et al. Prediction of the COVID-19 outbreak in China based on a new stochastic dynamic model. Scientific Reports, 2020.

[8] Roberts DL, Rossman JS, Jarić I. Dating first cases of COVID-19. PLoS Pathog, 2021, 17(6): e1009620.

[9] Zhang C, Wang M. Origin time and epidemic dynamics of the 2019 novel coronavirus. bioRxiv,

[10] https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-

[11] https://www.nytimes.com/interactive/2020/us/covid-death-toll-us.html

[12] Wu S L, Mertens A N, Crider Y S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States[J]. Nature Communications, 2020, 11(1): 4507.

[13] Shen M, Xiao Y, Zhuang G, Li Y and Zhang L. Mass testing—An underexplored strategy for COVID-19 control. The Innovation, 2021.

[14] https://github.com/CSSEGISandData/COVID-19

[15] http://www.nhc.gov.cn/xcs/yqtb/list_{gzbd}.shtml

[16] Huang N E, Qiao F. A data driven time-dependent transmission rate for tracking an epidemic: a case study of 2019-nCoV[J]. Science Bulletin, 2020, 65(6): 425-427.

[17] Silverman B W. Density Estimation for Statistics and Data Analysis[M]. Chapman and Hall,

Note: Figure translations are in progress. See original paper for figures.