Identifying regional COVID-19 presence early with time series analysis

Ryan Kruse and Suboh Alkhushayni

Department of Computer Information Science, Minnesota State University, Mankato, United States of America

E-mail: ryan.kruse@mnsu.edu

Keywords: COVID-19, ILINet, time series analysis, prediction intervals, residual hypothesis test, changepoint analysis

Abstract

The first confirmed case of COVID-19 in the United States was January 20, 2020 in Washington, while the first globally confirmed cases were in China in December 2019. The CDC’s Influenza-like Illness Surveillance Network is used to track the amount of people who seek medical attention for influenza-like illnesses, along with the illness cause. The metric rILI- is used to assess the amount of people who test negative for influenza or any other specific cause. To assess the evidence of COVID-19 presence in the US in late December 2019 or early January 2020, rILI- data from 2010 to mid-March 2020 was used to perform three types of analysis. First, we forecast prediction intervals using data until mid-November 2019 and compared the predictions with observed values for the subsequent 16 weeks. Second, we performed residual hypothesis testing by removing the trend and seasonality in order to compare residuals from before and after November 17, 2019. Third, we used changepoint analysis to identify major changes in trend and seasonality. This study provides strong evidence of COVID-19 presence in the US in late December 2019 or early January 2020. Combined with the knowledge that COVID-19 was spreading across other parts of the world, anomalous patterns in ILINet data should have been a warning sign that COVID-19 was already spreading in the US. The purpose of the study was not to identify specific states, but South Dakota has the strongest evidence of any US state, followed by California, Delaware, Maine, and New Mexico.

1. Introduction

In early 2020, confirmed cases of COVID-19 began to quickly spread throughout the United States, with the first coming in Washington state on January 20. By June 14, 2020, the US alone had over 2 million confirmed cases and over 115,000 confirmed deaths [1]. The first confirmed cases of the virus came in Wuhan, Hubei, China in December 2019, with the first confirmed fatality on January 11, 2020 [2]. Although the first US confirmed case was in January, many states did not have a confirmed case until March, with the latest coming in West Virginia on March 17 [3]. Social distancing measures were widely adopted in March to reduce spread and impact of the disease. The importance of early intervention in COVID-19 control is detailed by Pei et al [4]. Additionally, Chen et al describe the impact that widespread cooperation—combined with individual prevention efforts—have on early virus suppression when awareness of the disease is prevalent [5, 6].

The US Center for Disease Control’s (CDC) Influenza-like Illness Surveillance Network (ILINet) tracks the number of people seeking medical care for influenza-like illnesses, meaning they have a fever and at least one other flu-like symptom. ILINet also records if the patient tests positive for flu or any other known cause [7]. From this data, Reich et al derive rILI-, a metric indicating the rates and amount of people seeking medical care for influenza-like illnesses without a known cause [8]. Higher rILI- indicates greater evidence of some unknown virus, while lower rILI- indicates less evidence of some unknown virus.
2. Objectives

The objectives of this study are:

- To assess the evidence of COVID-19 presence in the United States in late December 2019 and early January 2020 using data from the CDC ILINet.
- To investigate methods for earlier detection and intervention of future disease outbreaks.

3. Materials and methods

Our null hypothesis is that COVID-19 did not impact any state’s rILI- values prior to the state’s first confirmed case. Our alternative hypothesis is that COVID-19 impacted rILI- values in some state(s) prior to the state’s first confirmed case. We conduct this analysis at the .05 significance level. See Multiple Comparisons below for multiple comparison considerations.

We used three unique methods of analysis to determine the strength of evidence against the null hypothesis: forecasting prediction intervals, hypothesis testing on residuals, and changepoint analysis.

3.1. Data

The data used in this analysis came from Reich et al[8] and is available on GitHub. It was originally pulled from the CDC’s ILINet [7]. The time series data of state-level rILI- used in this study began on Week 40 of 2010 and ended with Week 11 of 2020 (the week of March 13).

3.2. Forecasting prediction intervals

First, we selected a model to forecast future values of rILI-. The models considered included Holt Winters, ARIMA, and TBATS [9]. Holt Winters is an extension of simple exponential smoothing that accounts for trend and seasonality. ARIMA is a flexible model that focuses on autocorrelation. TBATS uses exponential smoothing and Fourier representations in a way that allows for forecasting multiple seasonal periods [10]. An introduction to each method is provided by Hyndman and Athanasopoulos [11]. We assumed that each state can be successfully modeled by one model type; instead of selecting a new model type for each state, we selected the model we felt would perform best overall, as explained below figure 1. Models were fit for a subset of the data ending Week 31 of 2019. The subsequent 16 weeks were used to assess the models based on MAPE for each state.

The MAPE results are summarized below in figure 2. The three models had comparable averages and medians. Ultimately, we chose to use a TBATS model because it had the lowest standard deviation.

Next, using the rILI- data up until November 17, 2019, we constructed TBATS models for each state in our dataset, creating prediction intervals at 99.9% confidence for the subsequent 16 weeks. This approach allowed us to see which states had observed rILI- values outside of the models’ prediction intervals.

![Figure 1. The first ten states of MAPE for each model.](image-url)
3.3. Hypothesis testing
Hypothesis testing was performed by first removing trend and seasonality with Loess using R’s stl package [12], then separating the residuals into before and after November 17, 2019. Population means of before and after groups were compared using a one-sided T test.

3.4. Changepoint analysis
Changepoint analysis was carried out using R’s changepoint package [13]. This analysis was used to detect if and where major changes in trend and seasonality occurred. The At Most One Change (AMOC) method was used, forcing the analysis to result in one or zero changepoints for each time series. Specifically, the AMOC approach identifies if and where a change in mean and variation occur in the time series. Only the changepoints occurring from Week 42 of 2019 to Week 6 of 2020 with p-value less than .001 were considered significant. All significant changepoints found occurred before the corresponding state’s first confirmed case.

3.5. Multiple comparisons
Tests were conducted for the 46 states with available data. To avoid false-positive errors caused by multiple comparisons, we used Bonferroni Correction, a conservative approach to account for multiple comparisons [14]. This required us to carry out each state’s test using a significance level equal to our desired significance level (.05) divided by the number of tests (46). Bonferroni Correction was deemed appropriate for several reasons. First, initial disease spread is likely to be sporadic, so the nature of the problem suggests not all tests must come back significant in order to reject the null hypothesis. Second, analysis was carried out in an attempt to establish results, as opposed to exclusively motivating further investigation. Third, we deemed a Type II error favorable to a Type I error. Thus, each state was analyzed using a .001 significance level. We rounded this down to .001, further reducing the risk of a Type I error, which consequentially increased the risk of a Type II error.

4. Results

4.1. Forecasting prediction intervals
Figure 3 below is the plot of South Dakota’s rILI- data. The gray area to the right shows the TBATS model’s 99.9% prediction intervals. The black and red lines indicate the observed rILI- values. We see significant seasonality in the time series, with peaks in the winter months and valleys in the summer. Note, the 99.9% prediction intervals extend to higher rILI- than seen in the previous six years’ observed data. Figure 4 below shows the same data focused to show late 2018 until the end of our prediction intervals. The observed data from November 17, 2019 to March 13, 2020 is again shown in red.

Clearly, the observed data falls outside the 99.9% prediction intervals. Because of our Bonferroni Correction, we have evidence at the .05 significance level that COVID-19 was in South Dakota in late December/early January, significantly before the first confirmed case on March 10, 2020.

In total, four states had observed rILI- values outside their 99.9% prediction intervals: Alaska, Delaware, Missouri, and Alaska. All states’ plots are available on GitHub [15].

4.2. Hypothesis testing
Below, figure 5 shows a plot of South Dakota’s data, decomposed into trend, seasonality, and residual (remainder) using Loess. The residuals followed an approximately normal distribution.

The residual values were then separated in into two groups: before and after November 17, 2019. The two groups’ population means were compared using a one-sided T test. The null hypothesis was that the population means were equal. The alternative hypothesis was that the after group had a greater mean than the before group.
According to our Bonferroni Correction, we again used a significance level of 0.001. The R output gave a p-value of 0.000 052 53, providing strong evidence to reject the null hypothesis. Thus, hypothesis testing also provided evidence of COVID-19 presence in South Dakota in late 2019.
In total, two states had p-values less than .001: Maryland and South Dakota. Interestingly, seven states had less than .01, and fourteen had less than .05. Each state’s p-value can be found in figure 7. Note, as time series residuals are often autocorrelated, time series hypothesis testing is prone to Type I errors.

4.3. Changepoint analysis

For our third method of analysis, we used changepoint analysis to detect major changes in the mean and variance of trend and seasonality. See figure 6.

The red vertical line represents the detected changepoint given by R’s changepoint package, which also gives an associated p-value. Changepoints were considered significant if they had a p-value less than .001 and occurred from Week 42 of 2019 to Week 6 of 2020, corresponding to the 470th and 486th weeks across our entire dataset. Four states were found to have significant changepoints: Arizona, California, Maine, and South Dakota.

5. Discussion

Using three methods of time series analysis of weekly, state-specific rILI- values, we found evidence at the .05 significance level that COVID-19 impacted some state rILI- values before the states’ first confirmed cases. In South Dakota, all three methods of analysis provided compelling evidence of COVID-19 presence, perhaps as early as December 2019—two and a half months before the state’s first confirmed case. The purpose of this analysis was to determine the strength of evidence COVID-19 was present in the United States earlier than previously thought. The purpose was not to identify specific states, but our approach indicates Arizona, California, Delaware, Maine, New Mexico, and South Dakota are the states this virus was present earlier than confirmed. A summary of each state’s analysis results can be seen in figure 7.

As described in section 3.5, multiple comparisons were accounted for by using the conservative Bonferroni Correction, so each state’s prediction intervals, hypothesis tests, and changepoint analysis required 99.9% confidence. Bonferroni Correction was justified for three reasons:

• the sporadic nature of disease spread suggests not all tests must come back significant in order to reject the null hypothesis
• we aimed to establish results, as opposed to exclusively motivating further research
• a Type II error was favorable to a Type I error.

Because of this conservative approach, this study was at higher risk of Type II errors, meaning there may have been more states with evidence of early COVID-19 spread than we found. Conversely, the conservative approach reduced the risk of Type I errors, perhaps strengthening the evidence that some states had COVID-19 several months before their first confirmed cases.

Figure 6. South Dakota’s detected changepoint indicated by the red vertical line.
6. Limitations

In early-March and perhaps before, multiple major news sources reported that some health care facilities were urging people not to seek care unless the situation was urgent [16, 17]. Similarly, news of COVID-19 around the world could have impacted each state’s rILI-data. Several states’ data were missing or incomplete, including Florida, Rhode Island, New Jersey, and New Hampshire. Additionally, data was not available for Washington DC or the five territories.

Our data had 52 weeks of entries for each year except 2014, which had 53 entries. To allow the seasonal period to be an integer, we removed Week 30 of 2014 from the data. This was assumed to have negligible impact on the analysis.

7. Conclusion

This study provides strong evidence of COVID-19 presence in the United States earlier than proven to date. Although the purpose of the study was not to identify which states may have had COVID-19 early, South Dakota seems like the state with the strongest evidence, followed by California, Delaware, Maine, and New Mexico. We conducted three types of time series analysis on CDC ILINet data. Most states did not have significant results, which was to be expected given the sporadic nature of new disease spread. Our approach accounted for this using Bonferroni Correction, which allowed us to draw conclusions from the individual states that had significant results. The study provides motivation for further research into when COVID-19 spread to each state.

While this analysis is likely too late to be helpful for the COVID-19 pandemic, in the future this type of analysis could be used to provide early warning signs of new viruses. When a novel virus is spreading, testing capacity is often limited early, making the virus difficult to track and contain. The findings of this analysis could guide a targeted approach to early containment and testing efforts, perhaps limiting the spread of future viruses.
References

[1] Centers for Disease Control and Prevention 2020 Cases of coronavirus disease (COVID-19) in the U.S. Accessed on: June 14, 2020. Available: https://covid.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

[2] Wang C, Horby P W, Hayden F G and Gao G F 2020 A novel coronavirus outbreak of global health concern Lancet 395 470–3 PMID: 31986257. Accessed on: April 4, 2020. Available: https://thelancet.com/journals/lancet/article/PIIS0140-6736(20)30185-9/fulltext

[3] New York Times Watch How the Coronavirus Spread Across the United States. Accessed on: April 4, 2020. Available: https://nytimes.com/interactive/2020/03/21/us/coronavirus-us-cases-spread.html

[4] Pei S, Kandula S and Shaman J May 29, 2020 Differential Effects of Intervention Timing on COVID-19 Spread in the United States. Available: https://medrxiv.org/content/10.1101/2020.05.15.20103655v2

[5] Chen X, Liu Q, Wang R, Li Q and Wang W 2020 Self-awareness-based resource allocation strategy for containment of epidemic spreading Complexity 2020 1–12

[6] Chen X, Zhou T, Feng L, Liang J, Liljeros F, Havlin S and Hu Y 2019 Nontrivial resource requirement in the early stage for containment of epidemics Phys. Rev. 100 032310 doi:10.1103/PhysRevE.100.032310

[7] Centers for Disease Control U.S. Influenza surveillance system: purpose and methods. Accessed on: April 4, 2020. Available: https://cdc.gov/flu/weekly/overview.htm

[8] Reich N G, Ray E L, Gibson G C, Cramer E and Rivers C M 2020 Looking for evidence of a high burden of COVID-19 in the united states from influenza-like illness data Accessed on: April 2, 2020. Available: https://github.com/reichlab/ncov/blob/master/analyses/il-latest-report.pdf

[9] RDocumentation Forecasting functions for time series and linear models. Accessed on: May 10, 2020. Available: https://rd documentation.org/packages/forecast

[10] De Livera A M, Hyndman R J and Snyder R D 2011 Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing J. Am. Stat. Assoc. 106 1513–27

[11] Hyndman R J and Athanasopoulos G 2018 Forecasting: principles and practice 2nd edn (Melbourne, Australia: OTexts) Available: https://otexts.com/fpp/holt-winters.html

[12] RDocumentation Seasonal decomposition of time series using loess. Accessed on: May 10, 2020. Available: https://rd documentation.org/packages/stats/topics/stl

[13] RDocumentation Methods for changepoint detection. Accessed on: May 10, 2020. Available: https://rdocumentation.org/packages/changepoint versions/2.2.2

[14] Armstrong R A 2014 When to use Bonferroni Correction Ophthalmic Physiol Opt 34 502–508 Wiley Online Library. Accessed on: July 7, 2020.

[15] Kruse R 2020 Accessed on: April 6, 2020. Available: https://github.com/kruser1/COVID19/blob/master/COVID19-early-signs/state-forecasts-99-9-pi.pdf

[16] The Washington Post Fearful doctors and nurses at walk-in clinics have a message for patients: stay away. Accessed on: May 10, 2020. Available: https://washingtonpost.com/local/fearful-doctors-and-nurses-at-walk-in-clinics-have-a-message-for-patients-stay-away/2020/03/20/710c194c-6a1c-11ea-9923-57073adce27c_story.html

[17] Wired Worried about Covid-19? hospitals have a request: stay home. Accessed on: May 10, 2020. Available: https://www.wired.com/story/worried-about-covid-19-hospitals-have-a-request-stay-home/