Identifying Key Events During the Pandemic Using Covid-19 Time Series Data

Nathaniel Callens Jr, Hunter Lybbert, Jeremy Rawlings, Jake Snow

March 2022

Abstract

Time series data from Covid-19 cases, deaths and hospitalizations is detrended; and multiple seasoning methods are used to make the data covariance stationary. We also forecast potential alternative outcomes of case counts and deaths in the absence of new variants using an ARMA model. True death count attributed to Covid-19 for a given time period is estimated using the Kalman Filter and various state space models.

1. Introduction

Covid-19 has been a part of daily life since its outbreak at the beginning of 2020. Every day new cases are reported, and more people die as a result of this vicious infection. Modern technology has aided us in the fight against Covid-19, allowing for faster testing and reporting methods for people who think they are infected. Thanks to the efforts of many health organizations and volunteers, there is a sizable amount of daily data from the previous two years surrounding case counts, deaths, averages, and countries. Since each day gives the virus more time to spread, the corresponding data on new daily cases and deaths is a time series. Numerous insights can be gained by exploring the time series data generated by hundreds of millions of people worldwide during the pandemic. Our exploration focuses on two things: 1) getting a better estimate for the true number of deaths attributed to Covid-19 at a given time and 2) Forecasting Covid cases, deaths, and hospitalizations in various forms.

1.1 Motivation

Point number one addresses possible discrepancies in deaths attributed to Covid-19. The first few months of Covid saw relatively few deaths, and then spiked suddenly in April. We use a Kalman Filter together with a state space model to estimate the true number of deaths at the beginning of the pandemic. It is also known that there is a buffer period between infection and death. Using case counts from five weeks prior (the average amount of time from infection to death) we use a Kalman Filter to estimate the true number of deaths attributed to Covid.

Point number two is intended to use the data to forecast future and even previous cases, deaths, and hospitalizations using an ARMA model, the Kalman Filter, and seasonality and trend of the data. One question to explore in relation to this is how many cases and hospitalizations would have been expected in the absence of variants. To answer this, we forecast the expected cases and hospitalizations by training on data up to November 2021. The resulting predictions from November on are made in the absence of the omicron variant, giving an idea of what might have been possible had Covid not mutated.

Another predictive element is bred directly from the question: were deaths attributed to Covid-19 undercounted before we really knew what we were dealing with? With this question in mind we will predict backwards in time by beginning at some time already into the pandemic when

deaths appear low before spiking up.

2. Data

Covid-19 is perhaps the most well-documented disease in recent history. While there are discrepancies in reporting based on country, current governments and access to resources, most major countries have kept records as accurate as possible. The World Health Organization (WHO) compiles this data and uses it to establish guidelines and consult with policy-makers about preventative measures. The data includes features such as new daily cases, new daily deaths, country, day, and running 14 day average. For our purposes, we use the new daily cases, new daily deaths, countries, and date. The main dataset we used was downloaded from the World Health Organization's official website, as they have made it available for public use [1].

The WHO data contains the date, reporting country, new case counts for that day, and new deaths for that day, among other things.

Hospitalizations is another important metric of Covid-19 impact that we seek to forecast with. The hospitalization data that we have comes only from the United States and was found on The COVID Tracking Project website, a project sponsored by *The Atlantic*, a reputable multi-platform publisher [2].

3. Methods

3.1 Identifying Trend and Seasonality

Another classical way to decompose a time series Z_t is to break it down into the sum of the trend, seasonality, and remainder. This decomposition can be written as $Z_t = T_t + S_t + R_t$. We decompose the time series like this to find a covariance stationary time series (the R_t). Then we can fit the covariance stationary part (R_t) with an ARMA model, justified by the $Wold\ theorem$. For this model, we focused strictly on decomposing the daily new cases in the united states time series.

First, we calculate the trend via a moving average. This is done by creating a new sequence T_t where the case count on a given day in T_t is the average of the m days case counts around it. For more details on a moving

[1] https://covid19.who.int/table

[2] https://covidtracking.com/data/

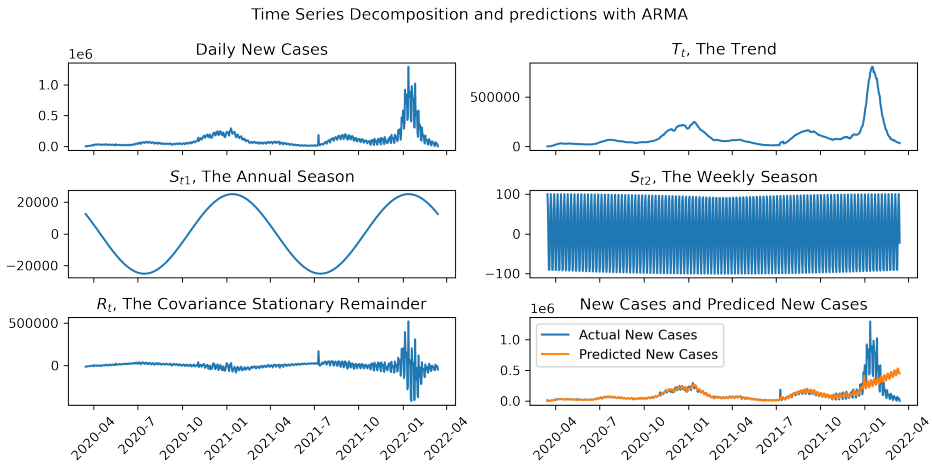

Figure 1: Decomposition of daily new cases in the United States time series into trend and seasonal components.

average see Volume 3 Chapter 14 Section 14.1 Definition 14.1.6. We chose m to be 7 so our trend component would represent a weekly average.

Next, we identify seasonality. In Covid-19 case counts there seem to be two different seasons at play: an annual season and a weekly season. Once we decided there were two seasonal components we chose periodic functions with specific amplitudes, and periods to fit each of the seasonal components.

After removing the trend and seasonal components from our data, we should be left with the remainder which we hope is covariance stationary, or white noise. We trained an ARMA model on the remainder time series. After getting the predictions from the this model we can generalize them to the actual time series by adding in both seasonal components and the trend.

3.1.1 Results of Trend and Seasonal Analysis

In order to compare our predictions accuracy to what actually happened, we trained our model on the data up to December 31st, 2021. We also had to fit a 4th degree polynomial to the trend to predict where the trend would go after December 31st, 2021. The predictions were relatively good in terms of predicting the increase in cases. However, the drastic jump in cases which is not predicted by ARMA would indicate there was something else at play. This can be explained by the fact that the Omicron variant was hitting the United States at about this time. It is important to acknowledge that our

R_t did not actually become fully covariance stationary. Overall, this method for predicting the future daily new case counts did not perform well.

3.2 SARIMA

The Seasonal Autoregressive Integrated Moving-Average (SARIMA) combines the ARIMA model with the ability to perform the same auto regression, differencing, and moving average modeling at the seasonal level.[3] SARIMA takes in two sets of parameters: order and seasonal order. For the respective AR and MA parameters, iterables are allowed. The model then automatically selects the best one. For a more in-depth understanding of the SARIMAX model, please see the documentation[4]. Since the Covid data from all of 2021 seemed very seasonal, we wanted to see how the case count would change if no additional variants ever mutated. Using a SARIMA model of iterables for the AR and MA parameters combined with an order of 1 for differencing we forecasted before the Omicron variant was reported.

Figure 2 demonstrates that with 95% confidence the Covid cases would remain typically mild compared to what actually happened with the Omicron variant. It's clear to see that predicting new variants and how they will affect populations using just historical data is very difficult.

3.2.1 ARMA Results

The first year of the Covid-19 pandemic was very unpredictable. In Figure 3 we can clearly see a 3-month seasonal trend in an increase in hospitalizations. In another SARIMA model with no differencing order and using iterables for the AR and MA parameters. We include a periodicity of 4 to represent different quarters of the year. We forecast from November 2020 to March 2021 and with 95% confidence that our prediction would increase followed by a slight decrease in hospitalizations, but the actual number of hospitalizations far surpassed our predictions. With further analysis we can note that around the holiday season in December and January we see the large increase in hospitalizations. Inferring that as people traveled home for the holidays and visited with relatives covid-19 hospitalized many more people. Future analysis with more current data could be able to verify this seasonal trend during the winter holidays.

[3] We used the SARIMAX model here, but without including any exogenous variables, thus it behaves exactly like the SARIMA model.

[4] `https://www.statsmodels.org/devel/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html`

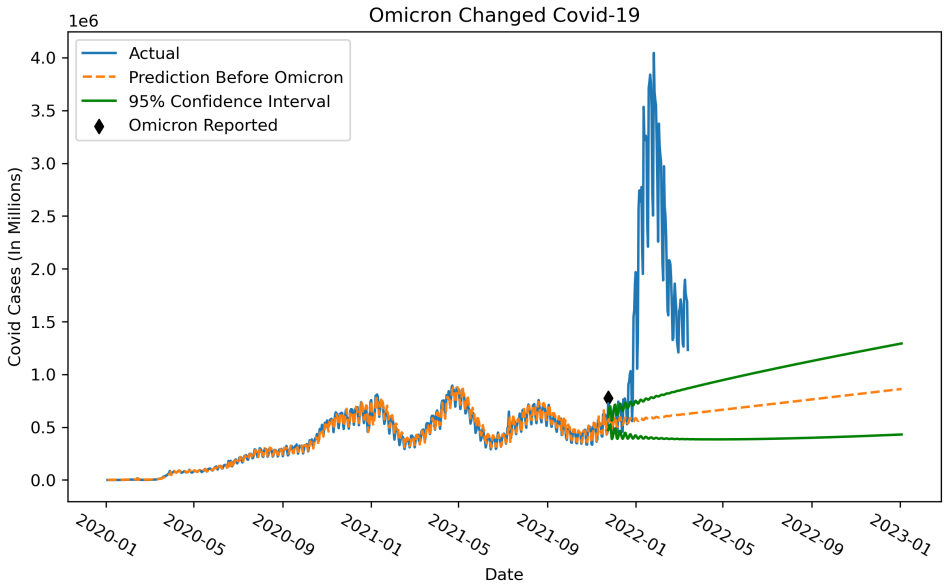

Figure 2: World wide covid-19 cases prediction using the SARIMAX algorithm before the Omicron variant was reported.

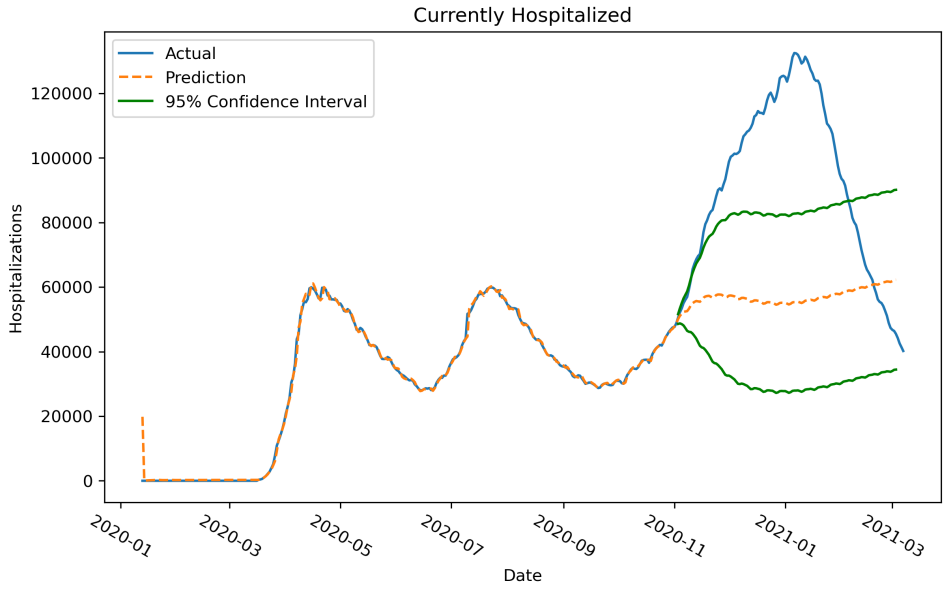

Figure 3: USA covid-19 hospitalization prediction using the SARIMAX algorithm the first year of the pandemic.

3.3 Kalman Filter

The first deaths attributed to Covid-19 were recorded in January 2020, but it wasn't until April of that year that deaths spiked to more than 7500 per day. It was around this same time that the general population became publicly affected by the disease, as businesses and schools began to shut down and sporting leagues suspended play. Covid-19 was certainly an outbreak, but is there a chance that Covid was responsible for more deaths in the months before we knew much about it than we think? The sudden spike in deaths does not correspond to a sudden spike in cases during the same time period. Our goal is to try to model the first 100 days of Covid deaths and determine if it's possible that the sickness was causing deaths it was not receiving credit for.

To make these predictions, we use a one-dimensional Kalman Filter where the state is the number of deaths on a given day. The observations are how many deaths are reported on a given day. We start with the state-space model denoted by:

$$x_t = Fx_{t-1} + Gu + w$$

$$z_t = Hx_t + v$$

where x_t is the true state, and z_t is the observation. We choose F to be a scalar representing the proportion of deaths from one day to the next. Since at the start of Covid no preventative measures were being taken, we set $F = 1.036$, meaning that the next day has 3.6% more deaths than the previous day. This estimate arises from the average increase in deaths found by using 100 straight days of data from another point in the pandemic (to be discussed later). Since we have no control, $u = 0$. We let H describe the relationship between the true state and what we observe. Using information obtained from various news sources about Covid reporting[5], we set $H = 1.08$. This is because hospitals had a tendency to over-count deaths during the pandemic. In this case, H represents the observed value being 8% larger than the true value. Finally, w and v are random variables distributed normally as $\mathcal{N}(0, 900)$ and $\mathcal{N}(0, 300)$, respectively. The variance here is impossible to measure quantitatively and so these numbers were set based on personal observation and intuition.

The system then becomes:

$$x_t = 1.036x_{t-1} + w$$

[5] https://www.msn.com/en-us/health/medical/are-we-overcounting-covid-19-numbers/ar-AASMh22

$$z_t = 1.08x_t + v$$

The Kalman Filter predicts backward in time by starting with an initial guess for the state at a given time. If we are trying to predict T time steps, then let x_T be our initial guess at time T. x_T is then used to find x_{T-1} by the relation

$$x_{T-1} = \frac{x_T - u}{F}$$

With $u = 0$, the update will be $x_{T-1} = \frac{x_T}{F}$ in this case. Repeating for T time steps (so the interval is $[0, T]$) and keeping track of each successive x yields a list of predictions from time 0 to T.

3.3.1 Kalman Results

We chose to predict the first 100 days of Covid deaths worldwide because a significant spike occurred between days 80 and 100. We want to model the potential deaths caused by Covid in this window where there are very little deaths before going up rapidly. Running a predictive Kalman Filter using the system described above on the Covid-19 deaths gives the graph shown in 1. Notice the natural continuation of the trend going backwards towards the true beginning of the pandemic. This predicted line is in stark contrast to the numbers reflected in the reported deaths, leading us to believe that, at least at the beginning of the pandemic, actual deaths were a lot higher than reported.

The Kalman Filter can also be used to estimate the true state measurements over a period of time. We use this fact to estimate the true number of deaths attributed to Covid in the U.S. from October 2020 to January 2021. The observations are the case counts from five weeks prior, because that is the average amount of time between infection and death. The hidden states are the true number of deaths five weeks after cases we observe. In the 100 day period from Oct. 2020 to Jan. 2021, the average increase in deaths from day to day was 3.6% (this was how we got 3.6% for the model described above). This is also F for the new state space model. We also use a University of Oxford estimate that in the U.S. the rate of death among infected individuals is about 2% [6]. This means that given a number of deaths, multiplying that by 50 yields a rough estimate of how many cases were reported exactly 5 weeks earlier (according to our rough estimates). The following model combines all this information.

$$x_t = 1.036x_{t-1} + w$$

[6] https://ourworldindata.org/mortality-risk-covid

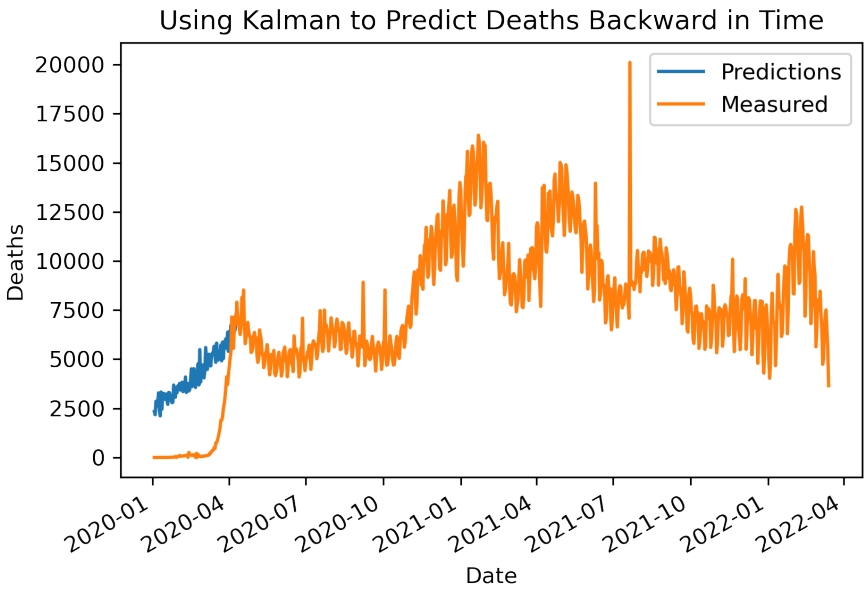

Figure 4: Using the Kalman Filter to predict worldwide Covid deaths backward through time for days 1-100 (Jan. 1, 2020 to April 10, 2020). The orange is the actual data, the blue is the Filter's prediction. Noise for the prediction was added in.

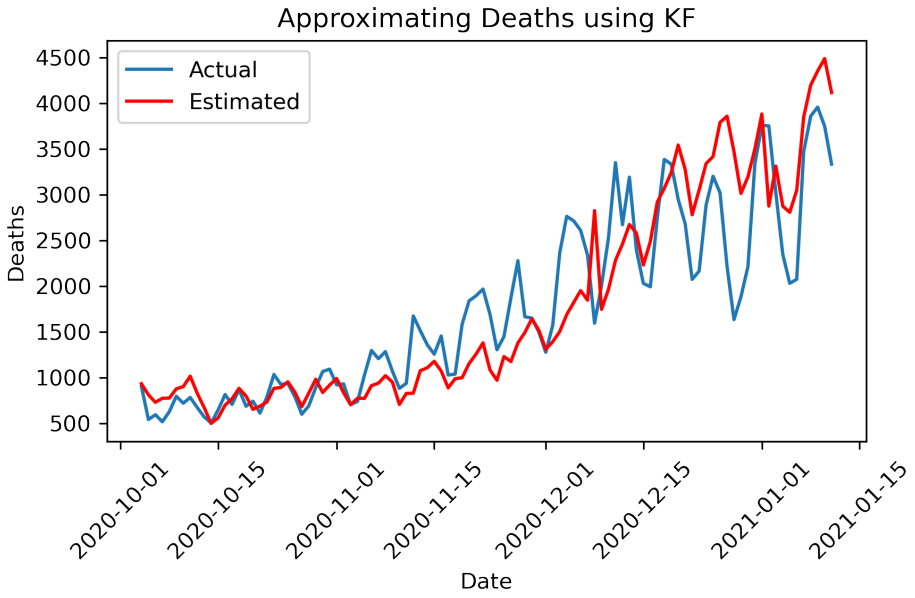

Figure 5: The Kalman Filter predicts the true number of new deaths per day between Oct. 2, 2020 and January 3, 2021. The blue line are the actual deaths measured while the red line are the Kalman Filter's prediction for true number of deaths based on cases from five weeks prior.

$$z_t = 50x_t + v$$

Here, x_t are the deaths five weeks after reported cases, where reported cases is represented as z_t. Once again, w and v are noise. Running the Kalman Filter for the 100 days between the months previously mentioned yields an estimate displayed in **??**. As seen in the figure, the filter's estimates for true number of deaths follows the actual deaths quite closely. At some points between November 1 and December 1 the model estimates fewer deaths than there actually were. This might indicate overcounting of deaths, and it could also be attributed to the choice of five weeks as a lag time. The model's estimates overcount the true number of deaths in the middle of December, possibly attributable again to a set lag time (not all deaths occur five weeks after infection).

4. Ethics

When working with any type of medical data, it is always important to keep in mind that we are talking about people and not numbers. Every number in the sick, recovered, or deceased categories represents a person afflicted or otherwise affected by COVID-19. It follows that preserving anonymity and ensuring data security is of the utmost importance with these kinds of problems. Since the only feature we use is the sick/death count for each country listed, we don't have to worry too much about a lack of confidentiality.

We should, on the other hand, concern ourselves with how we choose to use the information learned from the outcomes of our analysis. The trends identified in this paper are interesting, but it is important to realize that, when talking about hidden Markov models, the dimension of the state space is usually unknown. What this means in the context of the ethical ramifications of our analysis is that we don't know what exactly was going on behind the scenes that might have affected the trends in the data. For example, we don't know if a decline in the trend was due to the effectiveness of long-term mask-wearing or a result of increased herd immunity. Therefore, we should avoid making policy based on this type of analysis since a change in policy could fundamentally change the state-space of the system.

5. Conclusion

By analyzing data gathered during the COVID-19 pandemic, we were able to identify a number of interesting trends hidden throughout the data. We identified fundamental changes to the system such as the introduction of Omicron by looking at data outside of a 95% confidence interval. We also analyzed the accuracy of COVID-19 reported data by predicting deaths as well as reversing the Kalman filter to make predictions for the beginning of the pandemic.

Not only are these trends interesting and fun to analyze, they also lend themselves to insights for future pandemics. We hope that this analysis can help future analysts and epidemiologists know the best course of action when they find themselves in similar situations.