**COVID-19 Pandemic Excess Death Estimation and its Health Impact in the US**

Nuria Diaz-Tena, Jin Wang, Davit, Javier Cabrera, Michael Katehakis

**Abstract**

This paper is using statistical methods to estimate the impact of COVID-19 on the United States. Our paper provides a more granular view of which states and health conditions were more affected by the pandemic. It is vital to understand which policies performed the best to improve the health policies for the next pandemic. Only deaths with an autopsy are sure to be correct, the other deaths are classified by the most probable cause without any test to prove it. COVID-19 has brought further uncertainties given the additional cause of death; additional COVID-19 tests necessary to achieve the correct cause of death of the deassed Americans. The COVID-19 deaths data doesn’t catch everyone whose life was shortened by the pandemic and add other people whose primary reason for dying was not COVID-19. This paper shows how to select the best method to estimate excess deaths depending on the data being used and its granularity and how to improve the estimates once you know which model type provides the best estimate.

**Models used and performance.**

1. **Introduction**

An accurate measurement of the number of deaths due to the COVID-19 pandemic by state is necessary to understand which health policies were more effective to reduce the number of deaths. However, the reported COVID-19 deaths represent only a partial count of total death toll from the COVID-19 pandemic. Excess COVID-19 pandemic death, defined as the difference between the number of deaths during the pandemic and the number of expected deaths as if the pandemic would not have happened. The expected deaths without COVID-19 are forecasted using the historical data (before the pandemic). Gaps exist between reported and excess deaths related to the COVID-19 pandemic and we will observe how the gap decreases thru time when the new tests and policies to detect COVID-19 are in place.

Two different models were used to forecast the expected number of deaths without COVID-19: Exponential smoothing and sinusoidal models. And the CDC excess data provided using the Farrington algorithm is used as a comparison. The models have few parameters to estimate because we do not have lots of data for the model. The models were run on the training set (data from 2015 till February 2019) and evaluated on a hold-out period before COVID-19 (March 2019 to February 2020). Once the best model was selected, that model will be reran using data from 2015 to February 2020, including the previous hold-out period to include the latest data in the model before the forecast. The models are first fit at the state level to estimate the total excess death in the United States, and then fit by causes of deaths in each state. The results indicate which states over reported or under reported COVID-19 deaths.

1. **Methods** **and Results**
   1. **Data**

All deaths data by month and state were downloaded from the CDC1 from 20105 to June 2023. The data from the last month is not reliable due to the lag on reporting deaths and tit does not include the whole month (the last update was June 21st 20023) and the data was downloaded June 22nd 2023). There are 50 observations for each state to build the first models, from January 2015 to February 2019).

The population data by state and year was collected from the Census2 and the same population was used in 2023 as in 2022.

The computed excess death3 from the CDC using the Farrington algorithm were downloaded for comparison with our results.

* 1. **Excess death definition and metric to forecast.**

Excess death (1) is defined as the difference between the actual deaths and the forecasted deaths as if the pandemic would not have occurred. We prefer to forecast the daily crude rate (3) instead of the number of deaths because the daily crude rate is a more stable and smoother metric than the deaths or the crude rate (2). Plot 1 compares the actual monthly deaths in the US from 2015 to May 2023 to the monthly crude rate. The seasonality from both metrics is the same, but the crude rate does not increase as much as the number of deaths because is divided by the population and the population increases thru time. Plot 2 compares the monthly crude rate in the US from 2015 to May 2023 to the daily crude rate. The trend is the same with those two metrics, and the seasonality is smoother using the daily crude rate than the monthly crude rate.

1. Excess death = Actual Deaths – Forecasted death without COVID

is estimated by forecasting the daily crude rate (1) by month before COVID-19, estimating the expected death without the pandemic, and subtracting the observed deaths to the forecasted ones (2).

Plot 1. Comparison of the Actual Monthly death and the monthly crude rate in the US from 2015 to May 2023

A picture containing text, font, line, plot

Description automatically generated

1. https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-Jurisdiction-and-Age/y5bj-9g5w

2. <https://usafacts.org/data/topics/people-society/population-and-demographics/population-data/population/>

3. https://data.cdc.gov/NCHS/Excess-Deaths-Associated-with-COVID-19/xkkf-xrst

1. Crude Rate by month = 100,000\* (Monthly deaths)/Population
2. Daily Crude Rate = Crude Rate by month /Number of days in the month
3. Excess death = Observed deaths – Forecasted death without COVID = Observed death – Forecasted Daily Crude Rate \* Number of Days in a Month\* Population/100,000

Plot 2. Comparison of the Actual Monthly Crude Rate and the Daily Crude Rate in the US from 2015 to May 2023

A picture containing text, font, screenshot, line

Description automatically generated

The daily crude rate by month is used instead of the crude rate to have a smooth seasonality instead of a bumpy seasonality given that February has less days and less deaths and similarly with months with 30 days have less deaths than days with 31 days.

* 1. **Models**

There are many different forecasting models to use: from using deep learning models to Exponential smoothing forecast. The sophisticated methods tend to be very accurate if there are plenty of observations on the training set. However, given the scarcity of our data set, we choose the following models :

Definition of exponential smoothing forecast.

Definition of FARRINGTON FORECAST ALGORITHM because its forecast is provided by the CDC. JIN

Definition of sinusoidal models.

The exponential smoothing estimates 3 parameters by estate, while the FAARRINGTON. XXXXXX, and the sinusoidal models estimates 4 parameters by estate.

1. **Results** TO BE UPDATED with sinusoidal model

The training set for the exponential smoothing and sinusoidal models ranges from 2015 to February 2019 and the hold-out period ranges from March 2019 to February 2020. We compute the MAPE of the model (mean absolute percentage error) on the hold-out period and the smallest MAPE is the best fitted model and it should be use for the forecast. The Farrington algorithm is provided by the CDC and we could not chose the training set, neither the hold out period for comparison purposes. COMPARING MAPES BETWEEN EXP AN SINUSOIDAL

We observed that the crude rate by state is different before COVID-19. The crude rate is the number of Americans dying in each state by 100,000 of the population. The states with the largest crude rate before COVID are West Virginia, Tennessee, Kentucky, Alabama, Mississippi and Maine and the states with the smallest crude rate before COVID are: Alaska, Utah, Colorado, California, Texas and Wyoming. The crude rate depends on the proportion of old population, the proportions of minorities and the access of health care on those states. We will compute the Excess crude rate (5) that considers the population by states and its prior crude rate.

1. Excess crude rate = Actual crude rate – Expected crude rate without COVID-19

Table 1 provides the MAPE by state and type of model and the excess crude rate by state. The states better fit by the exponential smoothing models are Maine, Wisconsin, and Oregon with a MAPE smaller than 12%. The states with worst models using exponential smoothing with MAPES larger than 16% are Delaware and Oregon.

The overall excess crude rate using the exponential smoothing models and the Farrington models is e=very different. The exponential smoothing estimates an excess crude rate of 312 while the Farrington model estimates 72. There are 7 states with an excess crude rata larger than 500 using the exponential smoothing model: Oregon, West Virginia, Arizona, Arkansas, District of Columbia, Alaska and South Carolina. The Farrington model does not estimate and excess crude rate larger than 500 in state. The largest excess crude rate estimated by the Farrington model is 215 for West Virginia. The Farrington model estimates estates with an excess death less than 100 in 4 states, meaning that those states had less deaths due to the pandemic, those states are: Rhode Island, south Dakota, Massachusetts, and District of Columbia. Table1 provides all excess crude rate by state and the type of model and the MAPE is provided also for the Exponential smoothing and Sinusoidal models.

Table1. MAPE during the hold-out period (March 2019 to February 2020) for the exponential smoothing and sinusoidal models, excess death crude rate during from March 2020 to May 2023 for each model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Exponential Smoothing** | | |  | **Sinusoidal** | | **Farrington Algorithm** |
| **State** | **MAPE** | **Significant Excess crude rate** | **Percent COVID REported** |  | **MAPE** | **Excess crude rate** | **Excess crude rate** |
| **Overall** | 13.40% | **311.76** |  |  |  |  | **72.34** |
| Oregon | 11.95% | 571 |  |  |  |  | 157 |
| West Virginia | 12.85% | 571 |  |  |  |  | 215 |
| Arizona | 12.01% | 567 |  |  |  |  | 174 |
| Arkansas | 12.87% | 567 |  |  |  |  | 122 |
| District of Columbia | 12.65% | 521 |  |  |  |  | (101) |
| Alaska | 13.83% | 507 |  |  |  |  | 131 |
| South Carolina | 13.50% | 506 |  |  |  |  | 124 |
| Montana | 15.21% | 491 |  |  |  |  | 78 |
| Nevada | 12.41% | 450 |  |  |  |  | 6 |
| New Mexico | 14.27% | 450 |  |  |  |  | 146 |
| North Carolina | 12.74% | 407 |  |  |  |  | 47 |
| Tennessee | 13.27% | 407 |  |  |  |  | 75 |
| Kentucky | 14.31% | 407 |  |  |  |  | (9) |
| Georgia | 12.29% | 399 |  |  |  |  | 151 |
| Louisiana | 13.16% | 399 |  |  |  |  | 172 |
| Texas | 13.05% | 392 |  |  |  |  | 107 |
| Oklahoma | 14.70% | 391 |  |  |  |  | 75 |
| Florida | 12.38% | 384 |  |  |  |  | 44 |
| South Dakota | 12.84% | 377 |  |  |  |  | (129) |
| Colorado | 12.38% | 362 |  |  |  |  | 48 |
| Kansas | 12.98% | 359 |  |  |  |  | 30 |
| California | 13.23% | 356 |  |  |  |  | 85 |
| North Dakota | 13.09% | 352 |  |  |  |  | (62) |
| Alabama | 13.91% | 342 |  |  |  |  | 166 |
| Michigan | 12.54% | 332 |  |  |  |  | 41 |
| Virginia | 12.77% | 326 |  |  |  |  | 88 |
| Maine | 11.52% | 325 |  |  |  |  | 10 |
| Washington | 12.35% | 325 |  |  |  |  | 74 |
| Wyoming | 15.17% | 318 |  |  |  |  | 170 |
| Indiana | 12.82% | 301 |  |  |  |  | (9) |
| New York | 13.92% | 292 |  |  |  |  | 5 |
| Vermont | 14.77% | 271 |  |  |  |  | 191 |
| Missouri | 13.09% | 260 |  |  |  |  | 28 |
| Nebraska | 12.57% | 256 |  |  |  |  | (14) |
| Illinois | 12.99% | 237 |  |  |  |  | 60 |
| Iowa | 12.52% | 233 |  |  |  |  | (44) |
| Pennsylvania | 13.46% | 223 |  |  |  |  | (65) |
| Rhode Island | 15.14% | 212 |  |  |  |  | (141) |
| Minnesota | 12.54% | 211 |  |  |  |  | (44) |
| Ohio | 14.62% | 208 |  |  |  |  | (53) |
| Connecticut | 13.97% | 206 |  |  |  |  | (52) |
| Wisconsin | 11.86% | 198 |  |  |  |  | 4 |
| Utah | 12.72% | 195 |  |  |  |  | 20 |
| Idaho | 15.19% | 188 |  |  |  |  | 83 |
| Delaware | 16.92% | 176 |  |  |  |  | 26 |
| Hawaii | 12.60% | 151 |  |  |  |  | 20 |
| Puerto Rico | 13.12% | 131 |  |  |  |  | 214 |
| Massachusetts | 13.58% | 128 |  |  |  |  | (116) |
| Maryland | 13.94% | 80 |  |  |  |  | (91) |
| New Jersey | 15.12% | 67 |  |  |  |  | (13) |
| New Hampshire | 16.24% | 59 |  |  |  |  | (72) |

COMPARIOSN BETWEEN EXPONEENTIA SMOOTHING AND SINUSOIDAL EXCESSD EATHS. The overall results will differ depending on which algorithm is used.

We can provide the United States excess death using the best models at the state level. The best model will be rerun at the state level using the data till February 2020 for the exponential smoothing and sinusoidal models. The Farrington model estimates 85% less excess death than hte exponential smoothing model using the initial model. The total excess deaths are provided in Table 2.

Table 2 US Excess death using only the training set in the model and using data till February 2020 once the model was evaluated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Excess Deaths** | **Exponential Smoothing** | **Sinusoidal** | **Arrington Algorithm** | **Best Model** |
| **Using model till February 2019** | **1,093,495** |  | **166,593** |  |
| **Using model till February 2020** |  |  |  |  |
| **Pct difference** |  |  |  |  |

**Case study with 3 different states: New York (first state hit by COVID-19), West Virginia (worst crude rate before the pandemic) and Alaska (best crude rate before the Pandemic)**

**New York**

Let’s use the state of New York as the case study and example of the previous computations. This will help on understanding he differences between the Exponential smoothing and the Farrington results. Plot 3 shows the daily crude rate for New York from 2015 to May 2023. The peak on the plot shows the increase in deaths in April 2020 due to the COVID-19 pandemic in New York city. The exponential smoothing model uses the prior data, and its forecast is much lower because it does not take COVID-19 in the forecast. Instead, the Farrington model seems to increase the expected death in April 2020 while it is not supposed to do so because the forecast should not consider the COVID deaths. The Farrington model forecast is too high, and it does not discount the COVID-19 deaths, that provides a bias excess crude rate estimation. Focus on the April peak to verify the prior statement by plotting the April daily deaths in Plot 4 an comparing the actual daily crude rate and the forecasted using Farrington and Exponential smoothing algorithms. The actual April daily deaths are 2 in 2015 and 2016, it increases to 2.5 in 2017 and 2019, the forecast using exponential smoothing is 2.5 from 2019 to 2022 (following the trend from 2015 to 2018. Instead, the Farrington algorithm forecast 5 daily crude rate in 2020 without following the previous data without pandemic deaths. The 95% prediction interval ranges from 1.9 to 3.1 while the actual daily crude rate is 6.1 in 2020, outside the prediction interval as shown in Plot 5. We can report the excess death considering the difference between the actual and the forecast and we can report the significant excess death only considering any quantity that lies outside the prediction interval. Table 3 shows the excess daily crude rate and the significant excess crude rate in New York. In 2020 the significant daily crude rate is larger than the excess daily crude rate because it does not consider the insignificant negative excess

Plot3. Comparing the Actual daily crude rate in New York estate and its daily crude rate forecast using Farrington and exponential models’ algorithms.

A picture containing text, font, line, screenshot

Description automatically generated

hold-out period

Plot4. Comparing the Actual APRIL daily crude rate in New York estate and its daily crude rate forecast using Farrington and exponential models’ algorithms.

A picture containing text, line, receipt, plot

Description automatically generated

daily crude rate from June to October. The excess daily crude rate and the significant excess daily crude decrease over time and it is not positive in 2023 (the end of the pandemic). The cumulative significant Excess crude rate is more smoother than the not significant one as seen in Plot 6 and it is flat from 2022 till now, meaning that the pandemic is not adding new deaths in America.

Plot5. Actual APRIL daily crude rate in New York estate and its daily crude rate forecast using exponential models’ algorithms and its prediction interval.

A picture containing line, plot, screenshot

Description automatically generated

Table 3. Actual APRIL daily crude rate in New York estate and its daily crude rate forecast using exponential models’ algorithms and its prediction interval.

|  |  |  |  |
| --- | --- | --- | --- |
| **New York** | **Excess daily crude rate** | **Significant Excess daily crude rate** | **Pct COVID-19 deaths Reported** |
| 2020 | 6.45 | 7.09 |  |
| 2021 | 1.80 | 0.60 |  |
| 2022 | 1.18 | 1.32 |  |
| 2023 | (0.08) | - |  |

Plot6. Cumulative crude rate and significant crude rate in New York estate

A picture containing text, plot, line, screenshot

Description automatically generated

1. Sinusoidal models are used to build a model on the training set, the model is evaluated in the hold-out period and the MAPE is computed.
2. The model with the smallest MAPE will be chosen to provide the initial excess crude rate rate and excess death.
3. The best model will be rerun till February 2020 and the final

The Farrington algorithm forecast is very close to the actual deaths before and after COVID. Even though we took the estimated excluding COVID, it does not seem to ignore the COVID

Plot1. New York’s total actual daily crude rate New York’s forecasted daily crude rate using Farrington and Exponential smoothing forecasts.

Plot explanations

Table3. Final Excess crude rate using data till February 2020 for the first year the second year and third year of the pandemic.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Excess crude rate** | **Model Type** | **Excess crude rate 1st year** | **Excess crude rate 2nd Year** | **Excess crude rate 3rd Year** |
|  |  |  |  |  |
| Alaska |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Point out difference between years and the previous forecast.

MAP with data on pervious table

We want to know if COVID-19 was overreported or under-reported by state. We will compare the excess death to the reported COVID cases and provide the percent of overreported or underreported COVID deaths by state. If COVID-19 was over reported or underreported means that other causes of death were affected by the pandemic also. We will use the same methodology described above with data at the state and cause of death level. We will know the excess death by state and disease and that will let us know which causes of death were more affected by the COVID-19 pandemic.

Table4 percent over or under-reported covid by state. And map

Table 4.with percent change excess death by disease and bar plot

Map with the percent change of excess death by state for major diseases or diseases with the most change.

Task:

1. Meeting with Davit to divide and conquer results.
2. Collect data till now and add it to the data set
3. Collapse the data a t the state level
4. Run models on daily crude rate with 4 year of forecast
   1. ETS
   2. Sinusoidal
   3. Econometric
5. Compare diagnostics by state and choose best Table 1
6. Re-ran models till February 2020 and forecast 3 years. Table 2 and Table 3 and Map, Tabl4 with map and bar plot
7. Models at state-disease and us plot and map of different diseases excess crude rate by state.