## Forecasting High dimensional tensor with relatively few observations to assess COVID-19 Pandemic Excess Death in the United States

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

**Abstract**

This paper is studying the COVID-19 excess deaths using a tensor time series of mortality data in the 50 states and the district of Columbia for the major 15 causes of death in 2020 and the rest of the deaths to be able to report the overall. COVID-19 effect in the United States. This data is collected monthly since 2015 (before COVID-19) until 2023, and the objective is to estimate the COVID-19 excess deaths during the pandemic. We propose the combination of a non-linear trend and seasonality model that explains the over-all structure of the data, followed by an autoregressive tensor model on the residuals of the initial model. We used different initial models for the trend and seasonality: (1) logarithm trend with monthly dummies and autoregressive terms, (2) exponential smoothing forecasting and (3) ARIMA models first, to use the autoregressive tensor of the rest of the causes of death to improve the initial model.

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 states performed the best to improve the health policies for the next pandemic. 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. We will use the model with the best performance – Exponential Smoothing forecast & autoregressive tensor to provide the COVID-19 excess death patterns by state and cause of death during the pandemic.

1. **Introduction**

An accurate measurement of the excess COVID-19 deaths 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, is 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.

Three different types of models were employed to forecast the expected number of deaths without COVID-19: (1) logarithmic trend with monthly dummies and autoregressive terms, (2) exponential smoothing, and (3) ARIMA models. The residuals from each of these models were utilized to fit another model using an autoregressive tensor for other causes of death. Additional information regarding changes in other causes of death could influence and enhance the forecast for a specific cause of death. For instance, if there is an increase in diabetes-related deaths in the last three months, this may affect heart disease deaths, potentially leading to a slight decrease from heart disease (competing risk between diseases), or slight increase from heart disease (due to the same reason diabetes death have increased). The models estimate few parameters because we do not have lots of data for the model. (50 months of data for the training set). 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.

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

All deaths data by month and state were downloaded from the CDC1-2 mortality data set from 2015 to September 2023. The data from the last six month is not reliable at the non-natural causes of death (unintentional accidents and Self-harm) because these causes of death have a lag of six month. There are 50 observations for each state (50 states, and DC) and 165causes of death to forecast (14 top causes of death except COVID-19, the combination of the not top 15 causes of death) to build the models, from January 2015 to February 2019.

The population data by state and year was collected from the Census3 and the same population was used in 2023 as in 2022.

* 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 should use the crude rate defined in (2) instead of the total excess death numbers to be able to compare different states with different population numbers. And we prefer to forecast the daily crude rate (3) instead of the crude rate because the daily crude rate is a more stable and smoother metric than the deaths or the crude rate. 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 because the daily crude rate considers the number of days in each month.

1. Excess death = Actual Deaths – Forecasted death without COVID
2. Crude Rate by month = 100,000\* (Monthly deaths)/Population
3. Daily Crude Rate = Crude Rate by month /Number of days in the month
4. <https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-Jurisdiction-and-Age/y5bj-9g5w>
5. <https://wonder.cdc.gov/controller/datarequest/D176;jsessionid=814C90129BD725B84779B60FE623>
6. <https://usafacts.org/data/topics/people-society/population-and-demographics/population-data/population/>
7. Excess death = Observed deaths – Forecasted death without COVID = Observed death – Forecasted Daily Crude Rate \* Number of Days in a Month\* Population/100,000

Excess death will be estimated using the dependency between excess death and daily crude rate and expressed in equation (4).

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

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

1. **Initial 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 simple models to estimate few parameters. These models will be improved by adding the cause autoregressive tensor models.

* 1. **Non-linear trend with seasonality. Explain mode structure NURIA to specify**
  2. **Exponential Smoothing. Typical Exponential Smoothing anyone to explain**
  3. **ARIMA model Typical Exponential Smoothing anyone to explain**
  4. **Adding Autoregressive Tensor to previous Models Explain autoregressive tensor**

**Specify any difference Nuria**

1. **Results** TO BE ALL UPDATED FROM HERE TO THE END

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 (without adjusting by age) before COVID are West Virginia, Maine, Mississippi, Arkansas, and Alabama. The states with the smallest crude rate before COVID are Utah, Alaska, DC, Colorado, Texas and California. 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) to know how COVID-19 affected different states because it considers the population by states and its prior crude rate.

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

The training set for all models ranges from 2015 to February 2019 and the hold-out period ranges from March 2019 to February 2020. We compute the MAE of the model (mean absolute Error) on the hold-out period and the smallest MAE is the best fitted model, and it should be use for the forecast. Table 1 and Plot 3 provide the MAE by cause of death and type of model. TO BE INSERTED

for the hold out period (March 2019 to February 2020) where we are testing the model. The Exponential Smoothing Forecast is the one performing the best with this data as seen in Table 1. COMMENTS ON CAUSES

Table1. MAPE during the hold-out period (March 2019 to February 2020) for the three initial models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Exponential Smoothing** | | **Non-Linear Trend and Seasonality** | | **ARIMA** | |
| **State** | **MAE initial Model** | **Average Daily crude rate** | **MAE initial Model** | **Average Daily crude rate** | **MAE initial Model** | **Average Daily crude rate** |
| **United States** | **13%** | **131** |  |  | **395** |  |
| Diabetes | 13% | 118 |  |  | 649 |  |
| Neoplasms | 12% | 100 |  |  | 535 |  |
| Influenza | 13% | 85 |  |  | 217 |  |

Plot 3. MAE by cause as Table 1

Table2 and Plot 4 provides the MAE of the final models besides the improvement of the cause tensor on the initial models. We can observe that the Cause Tensor improves the initial models by xxx% being the best model the exponential smoothing with cause tensor.

Table2. MAE during the hold-out period (March 2019 to February 2020) using the cause tensor on the residual of the previous models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Exponential Smoothing & Cause Tensor** | | **Non-Linear Trend and Seasonality & Causse Tensor** | | **ARIMA & Cause Tensor** | |
| **State** | **MAE** | **Pct improvement from Initial Model** | **MAE** | **Pct improvement from Initial Model** | **MAE** | **Pct improvement from Initial Model** |
| **United States** | **13%** | **131** |  |  | **395** |  |
| Diabetes | 13% | 118 |  |  | 649 |  |
| Neoplasms | 12% | 100 |  |  | 535 |  |
| Influenza | 13% | 85 |  |  | 217 |  |

Plot 4. Overall MAE comparing the 6 models by disease

Some results on the cause tensor models showing which cause of death influenced other causes of deaths.

Once the best model is identified, the models can be rerun on the entire data set before COVID-19 to add 12 months of data to the training set. The forecast will be done till May of 2022 (end of the Pandemic). These kinds of models cannot be forecasted too long because other excess death can pick up instead as: change in policies, climate disasters and others.

Table 3 US Excess death differences among different models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Excess Deaths** | **Reported** | **Exponential Smoothing & Cause Tensor** | **Non-Linear forecast and Cause Tensor** | **ARIMA & Cause Tensor** |
| **COVID-19 Deaths** | **1,134,722** | **866,175** | **832,909** | **vvv** |
| **Percent Reported** | **100%** | **76%** | **73%** | **0%** |

We can provide the Forecasted Daily Crude Rate, the reported Crude Rate and the excess crude rate by disease in Plot 3. Observations of the plot. The same plot at the state level will provide the states most affected by COVID-19 in Plot 4. Observations of the plot. And the plot can be provided at the month level to see our recovery from COVID-19 in Plot 5. Observations of the plot.

Plot 3. Forecasted Daily Crude Rate, the reported Crude Rate and the excess crude rate by disease

COMMENTS ON CAUSES INSTEAD OF STATES. states that better fit by the exponential smoothing models are Maine, Wisconsin, and Oregon with the smallest MAE. The states with worst MAE are Delaware and Oregon. Add possible causes for bad models like poor data , etc once you know the states performing the worst.

The exponential smoothing estimates an excess crude rate of 312 while the ARIMA 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.

Plot 4. Forecasted Daily Crude Rate, the reported Crude Rate and the excess crude rate by state

Plot 5 Forecasted Daily Crude Rate, the reported Crude Rate and the excess crude rate by month

Breaking down the pandemic on variants and compare the effect of those variant son the population.

**Conclusion** on the importance of finding the best model to most accurate find our forecast and the ultimate results. How the causse tensor has helped on improving the model

I do not think we need 3 case studies

**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 rates 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

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, screenshot, line

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, plot, screenshot

Description automatically generated

than the excess daily crude rate because it does not consider the insignificant negative excess

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

A picture containing text, line, font, screenshot

Description automatically generated

Plot6. Actual daily crude rate in New York estate and its daily crude rate forecast using Farrington algorithms and its prediction interval.

A picture containing line, text, screenshot, font

Description automatically generated

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Exponential Smoothing** | | **Farrington** | |
| **New York** | **Excess daily crude rate** | **Significant Excess daily crude rate** | **Excess daily crude rate** | **Significant Excess daily crude rate** |
| 2020 | 6.45 | 7.09 | 7.42 | 7.17 |
| 2021 | 1.80 | 0.60 | 3.59 | 3.49 |
| 2022 | 1.18 | 1.32 | 2.28 | 2.05 |
| 2023 | (0.08) | - | 0.04 | 0.16 |

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

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. Switch to muilti-dimensional data with few observations
2. Send programs to Jin
3. QC results
4. Write…

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Exponential Smoothing** | | | **Non-Linear Trend and Seasonality** | | | **ARIMA** | |  |
| **State** | **MAE initial Model** | **MAE with cause Tensor** | **Percent Reported COVID-19 cases** | **MAE initial Model** | **MAE with cause Tensor** | **Percent Reported COVID-19 cases** | **MAE initial Model** | **MAE with cause Tensor** | **Percent Reported COVID-19 cases** |
| **United States** | **13%** | **258** | **131%** |  |  |  | **395** | **86%** |  |
| Mississippi | 13% | 424 | 118% |  |  |  | 649 | 77% |  |
| Arizona | 12% | 414 | 100% |  |  |  | 535 | 77% |  |
| District of Columbia | 13% | 392 | 85% |  |  |  | 217 | 153% |  |
| West Virginia | 13% | 391 | 118% |  |  |  | 518 | 89% |  |
| South Carolina | 13% | 373 | 107% |  |  |  | 507 | 78% |  |
| New Mexico | 14% | 369 | 114% |  |  |  | 488 | 86% |  |
| Alabama | 14% | 351 | 120% |  |  |  | 552 | 76% |  |
| Alaska | 14% | 347 | 57% |  |  |  | 237 | 84% |  |
| Oklahoma | 15% | 346 | 133% |  |  |  | 489 | 94% |  |
| Oregon | 12% | 344 | 62% |  |  |  | 340 | 63% |  |
| Louisiana | 13% | 343 | 110% |  |  |  | 489 | 78% |  |
| Tennessee | 13% | 332 | 130% |  |  |  | 468 | 92% |  |
| Montana | 15% | 328 | 109% |  |  |  | 384 | 93% |  |
| Arkansas | 13% | 314 | 132% |  |  |  | 485 | 85% |  |
| Georgia | 12% | 305 | 110% |  |  |  | 474 | 71% |  |
| Kentucky | 14% | 299 | 149% |  |  |  | 451 | 99% |  |
| Ohio | 15% | 297 | 141% |  |  |  | 389 | 108% |  |
| New York | 14% | 297 | 134% |  |  |  | 391 | 102% |  |
| Texas | 13% | 295 | 119% |  |  |  | 439 | 80% |  |
| Nevada | 12% | 292 | 133% |  |  |  | 380 | 102% |  |
| South Dakota | 13% | 292 | 133% |  |  |  | 262 | 148% |  |
| Florida | 12% | 292 | 126% |  |  |  | 424 | 87% |  |
| Michigan | 13% | 283 | 129% |  |  |  | 392 | 93% |  |
| New Jersey | 15% | 277 | 138% |  |  |  | 340 | 112% |  |
| Kansas | 13% | 270 | 127% |  |  |  | 314 | 109% |  |
| Wyoming | 15% | 268 | 118% |  |  |  | 308 | 103% |  |
| North Carolina | 13% | 266 | 121% |  |  |  | 376 | 85% |  |
| Indiana | 13% | 239 | 164% |  |  |  | 383 | 102% |  |
| Connecticut | 14% | 236 | 146% |  |  |  | 261 | 133% |  |
| Pennsylvania | 13% | 233 | 174% |  |  |  | 357 | 113% |  |
| Colorado | 12% | 233 | 113% |  |  |  | 316 | 83% |  |
| Missouri | 13% | 231 | 163% |  |  |  | 382 | 99% |  |
| Idaho | 15% | 230 | 128% |  |  |  | 333 | 88% |  |
| Illinois | 13% | 227 | 133% |  |  |  | 363 | 83% |  |
| California | 13% | 225 | 122% |  |  |  | 355 | 78% |  |
| Virginia | 13% | 219 | 121% |  |  |  | 320 | 83% |  |
| North Dakota | 13% | 216 | 183% |  |  |  | 256 | 155% |  |
| Iowa | 13% | 213 | 161% |  |  |  | 274 | 125% |  |
| Massachusetts | 14% | 204 | 147% |  |  |  | 223 | 135% |  |
| Wisconsin | 12% | 198 | 145% |  |  |  | 274 | 105% |  |
| Nebraska | 13% | 180 | 163% |  |  |  | 232 | 127% |  |
| Maine | 12% | 178 | 130% |  |  |  | 223 | 103% |  |
| Rhode Island | 15% | 170 | 217% |  |  |  | 195 | 189% |  |
| Delaware | 17% | 164 | 206% |  |  |  | 275 | 123% |  |
| Washington | 12% | 157 | 121% |  |  |  | 239 | 80% |  |
| Maryland | 14% | 156 | 186% |  |  |  | 222 | 131% |  |
| Utah | 13% | 138 | 127% |  |  |  | 188 | 93% |  |
| Vermont | 15% | 134 | 109% |  |  |  | 226 | 64% |  |
| Minnesota | 13% | 123 | 215% |  |  |  | 200 | 132% |  |
| Puerto Rico | 13% | 113 | 164% |  |  |  | 272 | 68% |  |
| New Hampshire | 16% | 107 | 206% |  |  |  | 127 | 173% |  |
| Hawaii | 13% | 89 | 144% |  |  |  | 73 | 176% |  |