# *Has the death rate in the United States changed with COVID-19?*

Mark T. Drummond - Fairfield University

## Executive Summary

The purpose of this analysis is twofold. First to see if the overall death rate in the United states has increased, decreased, or stayed the same during the Covid-19 pandemic compared to prior years. The second is to evaluate a number of different forecasting models using R.

Death rate is defined by total deaths reported and published on the Center for Disease Control and Prevention (CDC) website under the ICD-10 Codes.

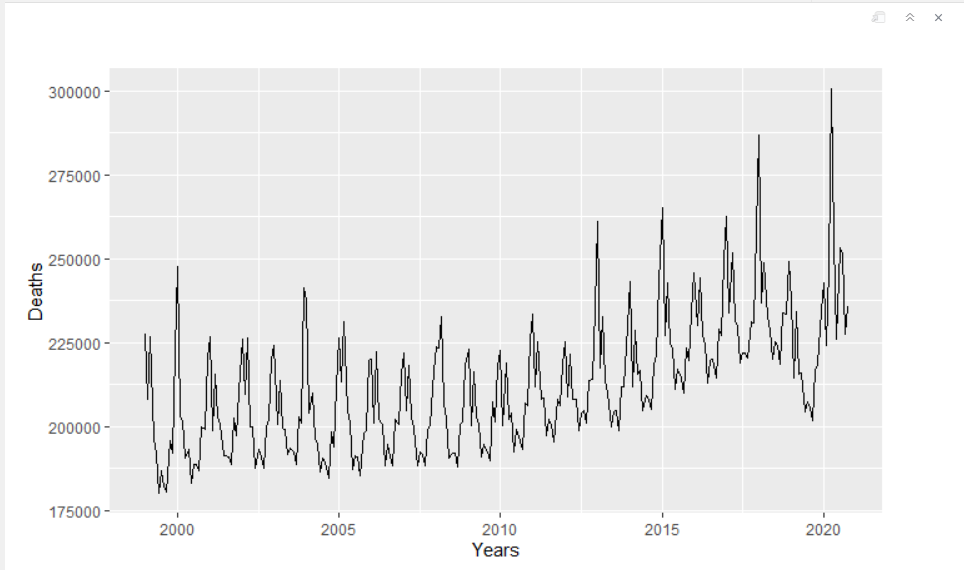
The data consist of two merged data sets one from 1999 to 2018, and one from 2019 – 2020.

1. [Underlying Cause of Death, 1999-2018](https://wonder.cdc.gov/controller/datarequest/D76) Counts of death by the month the deaths occurred from 1999-2018.
2. [Monthly provisional counts of deaths by age group, sex, and race ethnicity for select causes of death](https://data.cdc.gov/NCHS/Monthly-provisional-counts-of-deaths-by-age-group-/65mz-jvh5). Provisional counts of deaths by the month 2019-2020

## Death Trends

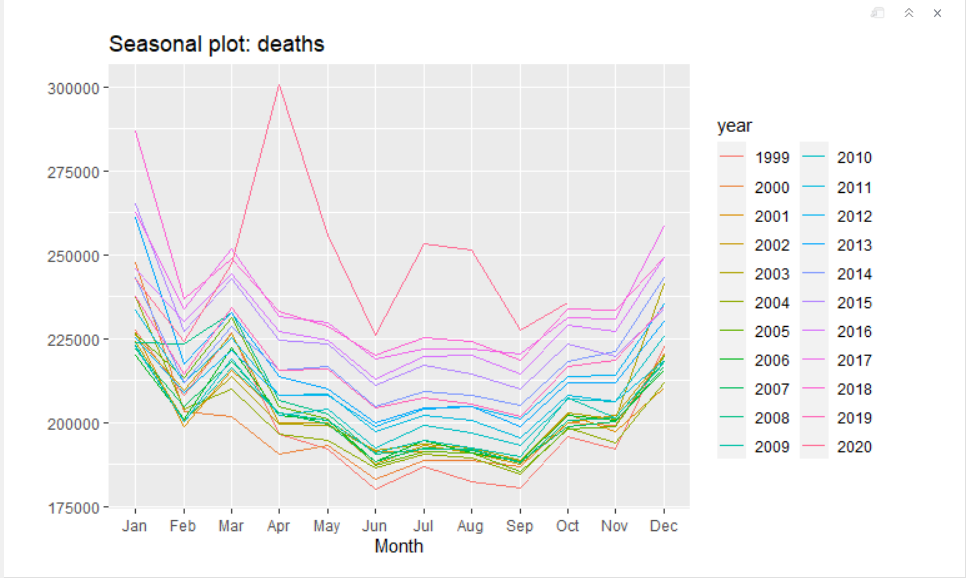
The overall death rate went up significantly due to COVID-19 as depicted in the following graph. The monthly average death for the months in the before Jan 2020 was 211,189 The average for Jan 2020-October 1, 2020 was 249,706 with a max of 300,662 reported for May 2020. This is an Average of 38,517 more deaths during COVID-19 so far. We have seen the number on the rise in November and going into December.

When actual vs the forecasted deaths are compared the forecast is off an average of 10%. This would indicate that a new factor is causing additional deaths (COVID-19) that was not there in past years.



**Total Deaths in the US 1999-2020**

Death plotted by month for each year we see a higher rate in the winter months (Jan, Mar and Dec). The Spike in April and then July, August 2020 during COVID-19



**Total Deaths by Month**

## Model Evaluation

The data was converted to a time series and then forecasted using the best fitting model in 4 different model categories.

Models from each category were evaluated using data from 1999 through 2019 and a “Best fit” was selected.

The original data and the forecasted data from the best model then used to create an excel file.

The forecasted data was then compared to the actual data

The models analyzed were the following:

* Simple moving Averages (SMA) & Exponential smoothing (SES)
  + The best model was the SES without order with a RSME of 13246.03
* Simple Linear Regression
  + The best fitting Simple Linear Regression model was with Trend and Season with a RMSE = 7431.877
* Holt, Holt Winters (HW), and Exponential smoothing state space model (ETS)
  + The best of the Holt, Holt Winters and ETS was the ETS method with a RSME of 5037.350
* Arima Models.
  + The best Arima Model was a RMSE = 4787.379

**The best overall model based on RMSE was the ARIMA model with RMSE = 4,787.379**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Group** | **Best Model in Group** | **RMSE** | **Rank Based on RMSE** |
| Arima Models | Manual Arima | 4787.379 | 1 |
| Holt, HW and ETS | ETS | 5037.350 | 2 |
| Simple Linear Regression | Trend + season | 7431.877 | 3 |
| SMA & SES | SES | 13246.03 | 4 |

## Forecasting Model Selection

1. Date observations and patterns

* The data is monthly from January 1999 to December,2019.
* The PACF plots using diff show that there are lags at 1 and 13
* The AFC plot there is a geometric decay geometric decay at each lag. This would indicate that a seasonal AR model would be used
* The Box Cox value is. 0.250792. This indicates that some transformation is needed
* The ndiff() value using the Box-Cox and lambda was 1 which indicates Non-seasonal differencing needed (stationarity).
* The nsdiff() value using the diff (Box-Cox and lambda) was 0 which indicated no Seasonality needed
* ggseasonplot() – The seasonal plot shows that yearly upward trend in January, March and then September through December. In 2020 there is an outlier for April and then July and August ostensibly due to the COVID-19 pandemic.
* The ggtsdisplay()
  + No real Trend in Plot
  + The PACF plots using diff show that there are lags at 1 and 13
  + The AFC plot there is a geometric decay geometric decay. This would indicate that a seasonal AR model would be used

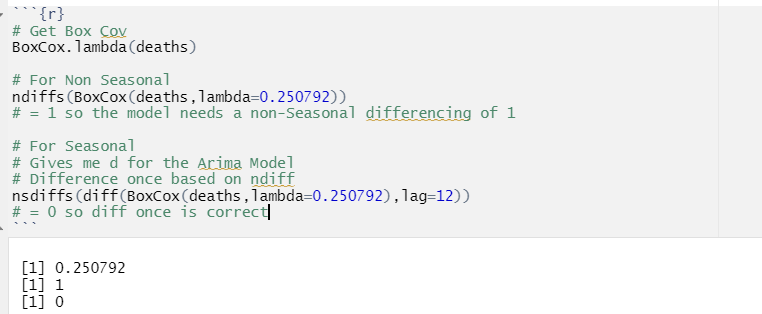


Figure 1 - Diffs and Box Cox

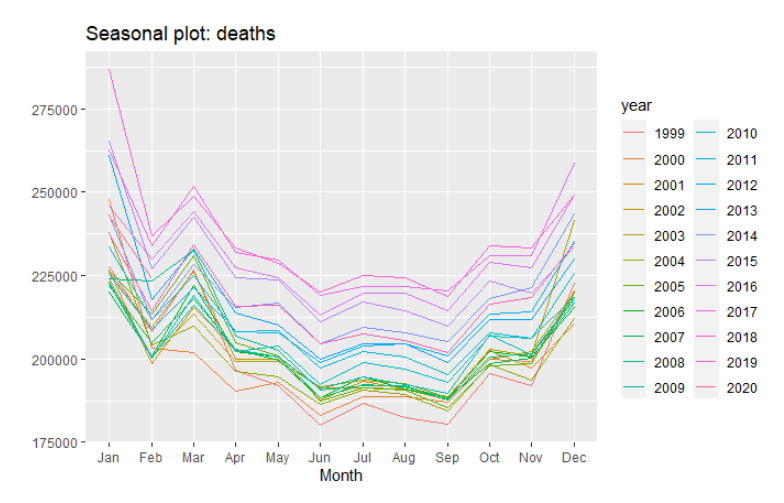


Figure 2 - ggtsdisplay()

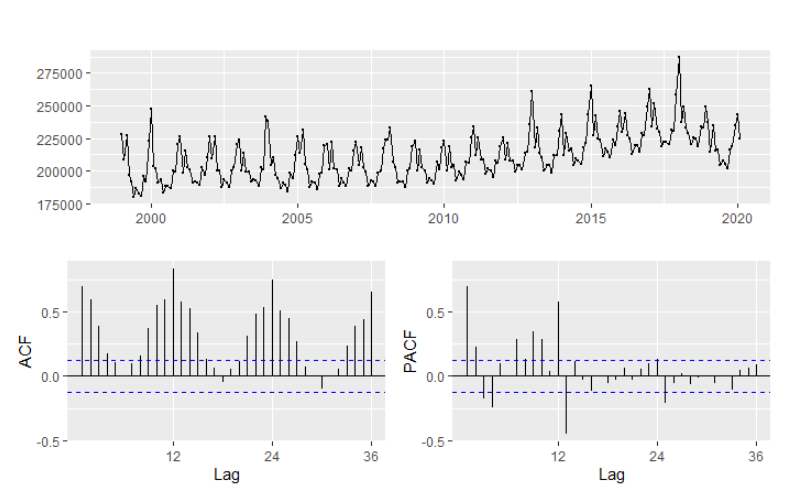


Figure 3 - ggtsdisplay ()

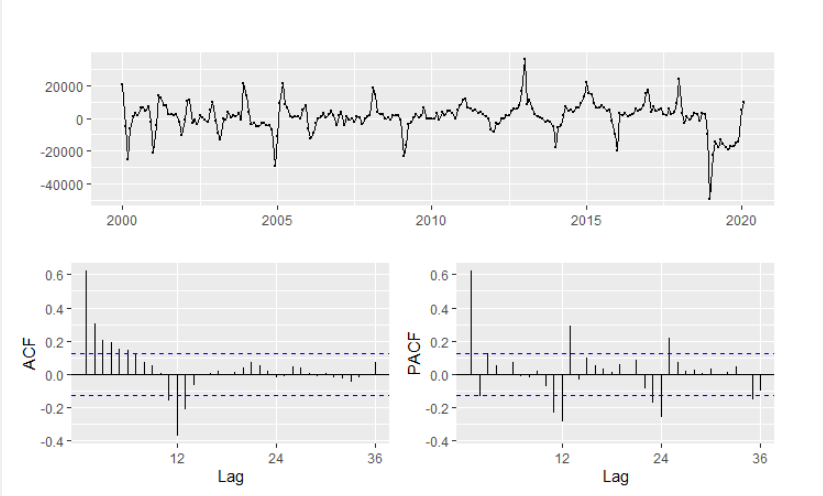


Figure 4 - ggtsdisplay () with diff and lag 12

## Model Selection and analysis

* Simple Moving Averages and Exponential smoothing forecasts
  + Two SMA models were evaluated.
    1. One with Lambda of 0.250792
    2. One with an order of 12 and Lambda of 0.250792
    3. Models with order 12 had a better RMSE of 13252.08 vs 14255.52

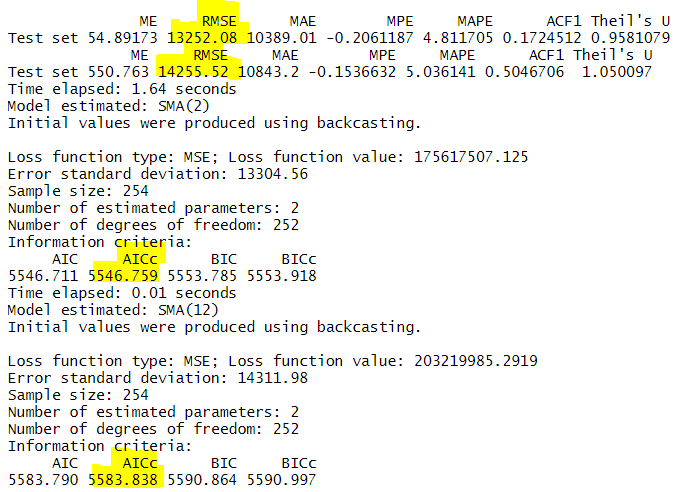
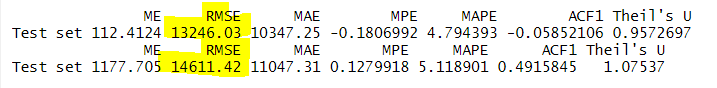


Figure 5 - SMA results

* + Two SES models were evaluated
    1. One without an Alpha and Lambda of 0.250792
    2. One with an Alpha of .1 and Lambda of 0.250792
    3. Without Alpha had a better of RMSE of RMSE 13978.23 vs 15382.81.



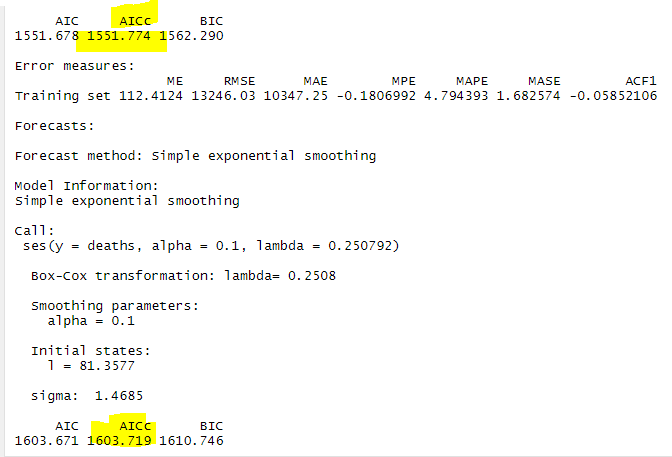


Figure 6 - SES Results

* + The SMA models are the best choice of the 4 models with a RMSE of 6.05978.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **AICc** | **Note** |
| SMA | 13252.08 | 5546.759 |  |
| SMA order 12 | 14255.52 | 5583.838 |  |
| SES | 13246.03 | 1551.774 | Best Model |
| SES with alpha =.1 | 14611.42 | 1603.719 |  |

Figure 7 – Model results

* Simple Linear Regression
  + Two different linear models with time series components were tested
    1. Trend had a RMSE = 14866.79
    2. TREND + SEASON had a RMSE = 7431.877
  + The best fitting model was just TREND+ season, R-square: 0.8271 and RMSE = 7431.877 and no T-values below 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **T-value** | **RMSE** | **R-squared** | **Adj R-squared** | **Note** |
| Trend | 10.59 | 14866.79 | 0.308 | 0.3052 |  |
| Trend + season | All > 2 | 7431.877 | 0.8271 | 0.8185 | Best Model |

Figure 8 – Model results

* Holt, Holt Winters and ETS models
  + Holt's method – RMSE = 13247.62
  + Holt-Winters' additive method (HW) – RMSR = 5186.121
  + Holt-Winters' multiplicative method (HW) – RMSR = 5535.074
  + Exponential smoothing state space model (ETS) – RMSE = 5037.350
  + The best fitting in this group is the Exponential smoothing with a RSME of 5037.350

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **AICc** | **Note** |
| Holt's method | 13247.62 | 1555.969 |  |
| Holt-Winters' additive method | 5186.121 | 5788.377 |  |
| Holt-Winters' multiplicative method | 5535.074 | 5793.842 |  |
| Exponential smoothing | 5037.350 | 5743.942 | Best Model |

Figure 9 - Model Results

* Arima Models
  + Looking at the ndiff the data needed to be transformed
  + Used diff(BoxCox(deaths,lambda=0.250792),lag = 12)
* The PACF plots show that there are significant lags at 1 then minor ones at 12,13,24,25.
* The AFC plot there is a geometric decay at lag 12. This would indicate that a seasonal AR model should also be tried.

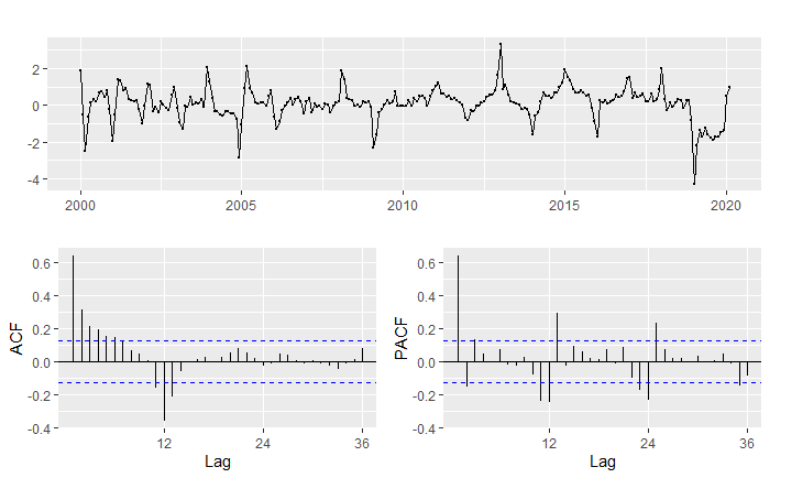


Figure 10 - ggtsdisplay(diff(BoxCox(deaths,lambda=0.250792),lag = 12))

* Manual ARIMA Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **AICc** | **Note** |
| c(2,0,0),seasonal=c(2,1,1),lambda=0.250792) | 5117.492 | 386.65 |  |
| c(1,0,1),seasonal=c(2,1,1),lambda=0.250792) | 5104.949 | 385.70 |  |
| c(2,0,2),seasonal=c(2,1,0),lambda=0.250792) | 5088.977 | 382.31 |  |
| c(2,0,1),seasonal=c(1,1,1),lambda=0.250792) | 4813.429 | 373.47 |  |
| c(2,0,2),seasonal=c(2,1,1),lambda=0.250792) | 4914.212 | 369.15 |  |
| c(1,0,2),seasonal=c(2,1,2),lambda=0.250792) | 4705.119 | 369.15 |  |
| c(2,0,2),seasonal=c(1,1,1),lambda=0.250792) | 4787.379 | 368.60 | Best Model |

Figure 11 - Model Results

* Auto Arima – RMSE = 4919.212 , AICc = 369.15

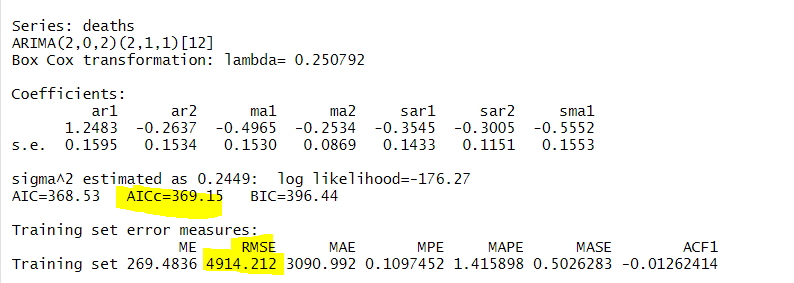


Figure 12 - Model Results

* Auto Vs Manual Arima Model
  + The Manual Arima Model had a better AICc then the Auto Arima did. 368.60 vs 369.15
  + Looking at the residuals
    1. There is a deep spike in the early months
    2. In the ACF graph all the values are within the blue lines except at 36.
    3. The distribution look normal

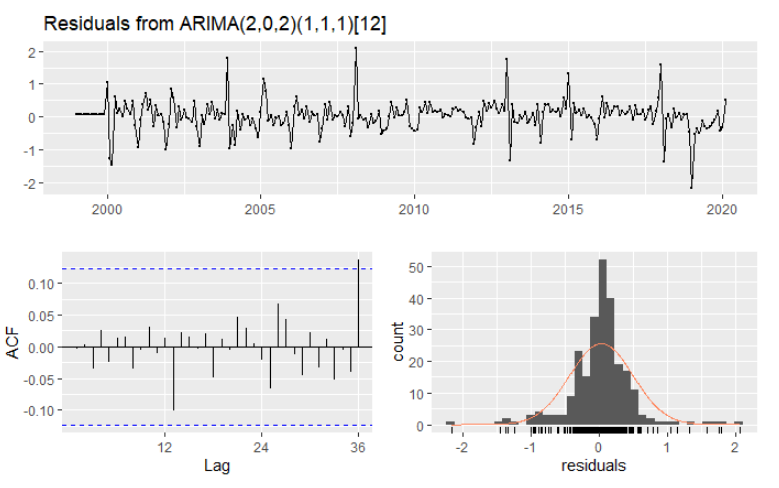


Figure 13 - ARIMA(2,0,2)(1,1,1)[12]

## Forecast vs actuals

* The Arima model was used to forecast what the death rate would be. The fitted values were then compared to the actuals. The different you would be a good estimate of what was caused by the pandemic (COVID-19)
* Below are the results. The model was off more that 10% half of the time.
* This would indicate that there is a new factor (COVID-19) that is affecting the death rates in the United States

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Month** | **Point Forecast** | **Hi.95** | **Actual** | **Diff** | **Percent Diff** |
| Mar-20 | 233,472 | 241,802 | 247,518 | (5,716) | -2% |
| Apr-20 | 214,578 | 223,332 | 300,662 | (77,330) | -26% |
| May-20 | 214,221 | 223,795 | 255,873 | (32,078) | -13% |
| Jun-20 | 203,025 | 213,070 | 225,958 | (12,888) | -6% |
| Jul-20 | 207,836 | 218,655 | 253,269 | (34,614) | -14% |
| Aug-20 | 205,762 | 217,175 | 251,288 | (34,113) | -14% |
| Sep-20 | 203,270 | 215,226 | 227,512 | (12,286) | -5% |
| Oct-20 | 216,494 | 229,288 | 235,569 | (6,281) | -3% |

Figure 11 - Model Results

## **References**

The data used is from the CDC website and consisted of two merged data sets. The two data sets are the following:

1. [Monthly provisional counts of deaths by age group, sex, and race ethnicity for select causes of death](https://data.cdc.gov/NCHS/Monthly-provisional-counts-of-deaths-by-age-group-/65mz-jvh5). Provisional counts of deaths by the month. Data is from January 1999 to October 2020.
2. [Underlying Cause of Death, 1999-2018](https://wonder.cdc.gov/controller/datarequest/D76) Counts of death by the month the deaths occurred from January 1999 through December 2018.

The selection criteria used for the second data set was all the defaults except the group by. The data was grouped by Census Region, Year, and Month. The defaults included All States, All 2013 Urbanization Categories, All Ages, All Gender, All Races, All Hispanic Origin, All Dates (1999-2018), All Weekdays, Autopsy values and Place of Death, All ICD-10 Codes Cases of Death. Without showing Totals

1. The Raw data files and aggregation are:
   1. Monthly\_provisional\_counts\_of\_deaths\_by\_age\_group\_\_sex\_\_and\_race\_ethnicity\_for\_select\_causes\_of\_death V2.xlsm
   2. ALL Deaths 1999 2018.xlsx