Estimating Excess Deaths

Forecasting Oregon’s Deaths During Record Heat

# Final Report by Louie Ligon

**Problem Statement**

Climate change is now impacting the planet Earth and its inhabitants more than ever since humans began keeping record of weather. From temperature extremes to droughts and floods, from forest fires to monstrous hurricanes, we are experiencing an unprecedented frequency and magnitude of deadly natural disasters. In a continued effort to educate and motivate people, governments, and businesses, we must understand the consequences of climate change and our dire need to act.

**Introduction**

An extreme heat wave, estimated to be made 150 more times likely by climate change, affected much of Western North America from late June through mid-July 2021. In the United States, the heat wave affected Northern California, Idaho, Western Nevada, Oregon, and Washington. Often referenced informally as a heat dome, these record setting high temperatures were caused by a high-pressure system in which hot air was trapped over a single geographic area.

Oregon has made weekly deaths publicly available dating back to January 2020. This data includes three- and five-year averages going back to 2015. The scope of this project is to focus on Oregon death data in an effort to estimate the excess deaths caused by the heatwave. Given weekly averages for each multi-year timeframe, we can forecast expected deaths and compare the forecast to actual deaths. This will allow us to approximate the unexpected deaths related to unprecedented heat.

With the weekly deaths data provided by the Oregon Health Authority (OHA), in conjunction with daily temperature records, we can evaluate a time series of death trends. I explored the data to prepare for the forecast with consideration to an ARIMA model. I was prepared to use differencing to make the time series stationary, but this was not necessary after splitting the data into train and test. I proceeded to identify a model, estimate parameters, and check the model. The final output is a forecast used to compare to the actual deaths recorded by OHA. Evaluating the difference between forecast and actual resulted in my estimation of excess deaths during the period of record setting high temperatures.

**Data Wrangling**

I started by loading 96 weeks of death data for the state of Oregon, dating from January 2020 to October 2021. In addition to the death data, I found historical daily temperature readings for various weather stations throughout Oregon. The death data was for the entire state of Oregon, and knowing the heat wave affected all of Oregon, I selected Portland as the representative location for temperature data. Initial wrangling and exploration of the data resulted in dropping the first week of data as it represented only 4 days instead of 7 days.

After reviewing some of the features in greater detail, I elected to drop all temperature data except the date and temperature high and low. Throughout the project, I was only concerned with the high temperature readings. The death data had a number of formatting concerns that were addressed. While there were many columns not used in this project, like 3-year average deaths and excess deaths over the average, I kept the data in case of interest in future efforts. One calculated field was created by subtracting COVID deaths from the death data that would be used to build a model.

**Exploratory Data Analysis & Initial Findings**

EDA began with evaluating the temperature data with a plot. There was immediately a very obvious spike in temperature during the time of concern, the week of 6-27-2021. Another heat wave was evident later in the summer of 2021, but this was similar to high temperatures seen in the summer of 2020.

Chart, line chart

Description automatically generated

The vertical line in the plot above represents three days that had high temperatures above 103 degrees Fahrenheit. I did explore the deaths around the second spike in temperature and did not discover any evidence of excessive deaths.

Next, I moved into exploring the death data. I found the data to have excess deaths compared to 3 and 5 year averages but not the excess deaths given the trend for 2020-21 weekly deaths. I want to estimate the excess deaths for the week of record setting heat by forecasting the expected deaths using early 2020-21 historicals. The data appeared to show excess deaths beyond COVID-19 deaths, suspected to be indirectly related to COVID, most notably in the late months of 2021.

The deaths data is weekly instead of daily (like the temperature data). I considered calculating the average temperature maximum for each week, but it wasn’t necessary. My approach to use the temperature data just for reference and as a means to EDA was sufficient. The deaths data had columns representing 2020-21, plus average and excess deaths for 3 and 5 year ranges. With my approach to estimate excess deaths based on variance to a forecast, all I needed was the 2020-21 deaths and the COVID-19 deaths.

There is a clear spike in deaths during the heat wave. We can also see how the excess deaths accounts for an increase at this time. It's interesting to see a similar pattern between deaths per week and excess deaths, beyond the heatwave but with a greater magnitude than COVID deaths. This suggests that COVID might be causing excess deaths beyond the scope of what is considered a COVID death. Knowing how this disease works, it's fair to say that a number of complications related to COVID appear to be causing deaths that aren't considered COVID deaths.

Chart, line chart

Description automatically generated

If this is the case, we should see the 2020-21 deaths per week trend with a greater magnitude than the trends for averages. I removed COVID deaths from the 2020-21 deaths per week for a clearer picture. Before doing that, I took a look at the same trends as above but with averages instead of excess for the 3- and 5-year ranges.

A picture containing text

Description automatically generated

It is very obvious that 2020-21 death rates are exceeding 3- and 5-year averages. Given the ongoing pandemic, this isn't a surprise, but the trend appears to be more than just from COVID deaths. Next, I removed COVID deaths from the 2020-21 weekly deaths and ploted the results. I kept the averages in my plot as well. I cleaned the data up a bit more by moving the dates of interest back a week. This was the week where only 4 days were used.

Chart

Description automatically generated

There it is again! The weekly deaths (excluding Covid!) dating back to January 2020 are noticeably greater than the average deaths over the last 5 years. This presents an interesting opportunity for future work with this project. Getting back to our original focus, let's take a look at a plot of just non-covid deaths.

Chart, line chart

Description automatically generated

The death spike during the week of the heat wave is really obvious. I recalled the second heat wave we saw earlier, and it looks like it might line up with the death spike that happened after the first. I pulled the dates for the second heat wave and plotted them with the deaths per week to see if they align.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

This plot made it evident that there were no excess deaths around the second heat wave. In fact, the deaths grow and maintain despite the begin of fall and cooler temperatures. Guess what thrives in cooler temperatures. COVID does.

**Modeling & Findings**

Having wrangled and explored the data, I could now begin to prepare the data for modeling. I used an ARIMA model to forecast weekly deaths in 2020-21. After doing so, I compared the forecast to actual and estimated excess deaths as the variance. First, I needed to check if our time series data was stationary. Non-stationary time series show the effect of seasonality, trends, or other measures dependent on the time index. This will make forecasting, with considerations to the mean and variance, rather difficult.

There were other time series forecast models to conside. In fact, we can break apart the ARIMA model into sub-models depending on the nature of our time series. An ARIMA model is a combination of the following components:

* AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
* I: Integrated. The use of differencing of raw observations (e.g., subtracting an observation from an observation at the previous time step) to make the time series stationary.
* MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The parameters of the ARIMA model are defined as follows:

* p: The number of lag observations included in the model, also called the lag order.
* d: The number of times that the raw observations are differenced, also called the degree of differencing.
* q: The size of the moving average window, also called the order of moving average.

Any of these parameters can be set to 0, effectively allowing for the ARIMA model to be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model. Given that I did not have to difference the data, d = 0, and I believe I technically used an ARMA model.

One of the first things I needed to determine was if our time series was stationary. Often you can determine this by simply looking at the plot. Think average daily temperature in the U.S. over a year. There will be clear seasonality making the time series non-stationary. If you consider world population over time, you'll have an example of a time series that isn't stationary due to trending. It was hard to say for certain if this data was stationary or not. There appears to be an inconsistent mean and variance but is it significant enough to call this non-stationary?

Chart, line chart

Description automatically generated

We needed to explore further to determine this. One method to determine if a time series is stationary is to plot the rolling mean and standard deviation for deaths per week. I'll be looking for flat lines, parallel to the x-axis. I'm going to use non-covid deaths only moving forward.

Chart, line chart

Description automatically generated

It was still too unclear to say from the plot. The window for rolling stats is 8 weeks. The standard deviation is mostly flat but with some jumps, and the mean trends up after the heat wave. Another method to determine stationaries is the Augmented Dickey-Fuller Test: The time series is considered stationary if the p-value is low (according to the null hypothesis) and the critical values at 1%, 5%, 10% confidence intervals are as close as possible to the ADF Statistics.

The results of the ADF Test, combined with the ambiguity of our plot, suggests that the time series is NOT stationary. Not surprising when you consider the deaths data includes the large spike from heat and indirect COVID deaths in late 2021. I went ahead and split the data for modeling. We can then check to see if our train data is stationary.

Train data was assigned to be all dates leading up to, but not including the heat wave (and subsequent increase in average deaths) while the test data was assigned the dates from the heat wave on. One benefit of doing this is that the ARIMA object's forecast() function will forecast the first step outside of training data. This will be the week of record heat. The training data passed stationarity testing and allowed me to start modeling.

The first thing I needed to do for the ARIMA model was to determine the model parameter values.

Here are the parameters we need to define:

* p: The number of lag observations included in the model, also called the lag order.
* d: The number of times that the raw observations are differenced, also called the degree of difference.
* q: The size of the moving average window, also called the order of moving average.

We can look at plots for the autocorrelation function (ACF) and partial autocorrelation function (PACF) to determine the parameters p and q. I'll set the value of d to 0 since we did not have to difference our raw observations. Differencing is done to transform data to stationary if needed. Timeline

Description automatically generated

The shaded blue horizontals represent the significance thresholds. The vertical lines represent the ACF and PACF values at a point in time. Only the vertical lines that exceed the horizontal lines are considered significant. The ACF can be used to figure out the best order of the MA model. Counting only the values that exceed the significance thresholds, I used only use the preceding 6 weeks.

We can use PACF to figure out the best order of the AR model. The lag count determines the order of the model. With the first 2 values above threshold, I used the preceding 2 weeks in the autoregression equation. I proceeded with defining my model instance with the ARIMA parameter values (p=2, d=0, q=6).

We get a nice summary from our model results. The key areas to focus on are the coefficients and p-values. If the p-value is less than or equal to the significance level (0.05 by default), we can conclude that the coefficient is statistically significant. If the p-value is greater than the significance level, we cannot conclude that the coefficient is statistically significant. I may want to refit the model without the term. I attempted to refit the model without the statistically insignificant coefficients and found no meaningful improvements to the model.

The model prediction looks pretty good! The most obvious errors are around May 1 and during what appears to be COVID related excess deaths during the Thanksgiving and Christmas holidays.

Chart, line chart

Description automatically generated

The forecast appears to decay quickly beyond the train data. Previous noise is no longer evident, and we appear to be working with a flat forecast

Chart, line chart

Description automatically generated

I was now able to use the ARIMA model's forecast() function to forecast a single step beyond the training data. Recall that our training data runs right up until the week of the heat wave, so I was able to compare this forecast value to the actual value for our estimate of excess deaths.

Using the ARIMA model from above, I estimate 161 excess deaths during the week of June 27, 2021. The range of excess deaths based on the 95% confidence interval is 109 to 213 deaths. A final interesting plot to view includes the 95% confidence interval.

Graphical user interface, chart

Description automatically generated

We can see the same information as previous plots. In addition, we have a shaded region around the future forecast. This region represents the 95% confidence interval for our forecast. While the forecast itself nears static, we can see the full interval captures most of the historical death values and provides a reasonable estimate for deaths. I believe this also provides further evidence of COVID related deaths that have not been classified as COVID deaths. We can see the excess deaths at the end of 2020, where we know COVID deaths were high in Oregon, and again after the heat wave. It just so happens that late August to early October, 2021 had the highest number of COVID deaths according the the Oregon Heatlh Authority. Recall the death data here is non-COVID deaths only.

According to the AP News, "Oregon blamed 116 deaths on the heat... More people died from the heat in the greater Portland area this June than in the entire state over the past 20 years, authorities said Tuesday."\* My excess death estimate is 161 compared to Oregon's estimate of 116. If we assume the model was good, I believe the significant variance can be explained.

I went back to the original death data sourced by OHA in the first notebook of this project. The data includes a reference to excess deaths over 3- and 5-year averages for death rates. After removing the recorded COVID deaths, I calculated non-COVID excess deaths and noticed a trend. The beginning of 2020 starts with a below average trend that slowly picks up starting in March/April of 2020, when COVID started spreading. It remains higher than average, going from single digit percentage points to sometimes more than 20% above average starting at the end of June 2021.

This implies a recent excess death count above what is considered a COVID death. I believe the forecast is not predicting this excess as it was not seen in the training data. The forecast plus actual heat related deaths is less than actual because the actual includes excess deaths beyond COVID and heat. I believe these deaths are COVID related as they coincide with COVID deaths.

**Future Work**

The plotting of 2021 deaths against 3- and 5-year averages presented an interesting opportunity for future work on this project. The original intent was to estimate excess deaths due to the June '21 heat dome. We can now also consider estimates for pandemic related extra deaths that weren't directly due to Covid. I think to do so, I'll have to use the averages for the model forecast rather than 2020-21. When using the 2020-21 deaths to forecast, I now know that excess deaths are involved. This wasn’t a problem for the forecast considering the excess deaths are most pronounced at the heat wave and beyond. The problem arises when you compare forecast to actual as the model was trained without reference to the late increase in excess deaths.

Sticking with the original model and intent, I would like to spend more time evaluating the data. I have learned a great deal about time series forecasting and have some questions around my approach. I want to know more about the assumptions made when using statistical tests to check for stationarity. Does this mean I’m assuming a normal distribution of the death data? I would like to evaluate for such in addition to experimenting further. Some experiments would include the parameter values for the ARIMA model. I would like to conduct hyperparameter tuning to select the best values. Ultimately, I’d like to continue work on this topic to further educate people about climate change. Perhaps one day we can use forecasts like this to motivate behavior, if not for long-term change like carbon dioxide emissions, then at least for things like evacuation when a hurricane or forest fire is eminent.