The Association between the Number of Deaths and Pre-determinable Weather Event Factors

This analysis is to identify the best predictors of the number of fatalities from storm event characteristics including the month it happened, the designation of the event, the state it started in, and the type of event. From these we can produce a model that will hopefully have the best chance of producing accurate predictions.

As an insurance analyst, and due to the recent extreme weather partly caused by the El Niño last year (2016), I am interested in the potential impact on businesses and people. By identifying which characteristics are important to predicting negative factors, such as deaths, we can identify what information to gather from the insured and potentially warn them about the risks they are facing so they can react responsibly.

Through predicting risk reliably there will be increased accuracy in pricing the products and less information may be collected meaning it will be quicker, easier, and potentially cheaper to buy your insurance.

My sample includes N = 1,573 individual storm events of the total 166,048 in the original data set. The sample size is so proportionally low because I have only included entries where either deaths or injury occurred. If I had left them in it would have skewed my results with the bias towards there never being deaths, but a high accuracy rate for the model. The other filtering column I have used is the state. I have removed any values that are not one of the 50 American states, removing entries that are general areas such as Atlantic North.

The storm events are logged by their state and county, and occurred between January 2013 and October 2015. The main types of weather phenomenon that are recorded in the data are thunder storms, hail, winter weather, flash floods, winter storms, and drought.

The entries collected were from a variety of sources, some casual like individuals reporting an event they witnessed, and others that were scientific about their data collection such as the National Weather Service.

The National Oceanic and Atmospheric Administration published the data, it contains:

- Weather occurrences that cause loss of life, significant property damage, or disruption to business;
- Uncommon, media worthy, weather occurrences, e.g. snow flurries in South Florida;

- Other significant meteorological events, such as record temperatures.

My response variable is the number of deaths directly caused by the weather event. It is recorded as an integer that ranges from 0 to 43. The majority of storm events have a recorded value of 0.

My predictors are:

- The **state** in which the weather event occurred, I have only included records that fit with the full USA state name. For instance, records in the Atlantic North will be excluded as it is not the name of a state.
- **Country or zone categorisation**, which indicates where the event occurred, (C) county/parish or (Z) zone. Zone happens over a large area such as a drought or heavy fog, and county/parish relates to more localised events such as flash floods or hail.
- The **time** the event occurs, which I am using the month to measure.
- The **weather event label** that most accurately describes the meteorological event leading to fatalities, injuries, damage, etc. The top five occurring of these are noted in the sample description.

I used the month name the event happened in, which I have mapped from the month name to the month's number, for example, August becomes 8.

I investigated the distributions for the predictor variables, and of the response variable. By producing frequency tables for the USA state, the type of weather event, and the country / zone categorisation associated with the event. And by calculating the mean, standard deviation, maximum, and minimum values for the quantitative variables I got a better understanding of the spread and form of the variables.

For the numerical variables I have looked at scatter plots. I then used a Pearson Correlation test to examine my response variable against the quantitative variables to check for independence. I can then examine a scatter plot of the potential fit line overlaying the response vs explanatory variable to see how well a linear relationship fits.

For the categorical variables I will look at bar charts to see the relative counts of values within the variables. I will use an Analysis of Variance (ANOVA) test to check for relationships between each response and categorical variable. Once I have a potential association I will use Tukey's Honestly Significant Difference test (THSD) to perform the post hoc tests. THSD checks the elements pairwise against the response variable to make sure it is a true association. This will help reduce the chance of making a Type I error, which is rejecting a potential association incorrectly.

I will use a Random Forest to create my model. I think this is the best model to use as there are lots of categorical variables in my set of predictors so a Random Forest will be more likely to capture the relationships between variables. By randomly splitting my data into a test and a training set I can train my model and then test it to check its accuracy. I am going to use a 70% training, 30% test split of my data. Part of the Random Forest model is an indication of the variables importance to the model, I will use this to remove any low importance variables.

Data management

My data management started by applying the filters described above to the data set.

I then created several new columns for month, event type, state, and county / zone type that map the values to numbers so that they can be used in the Random Forest analysis.

I also categorised the number of deaths to 1 if there were any deaths and 0 otherwise. I felt that an indication of if there were deaths is more important than how many deaths occurred.

There is a summary of descriptive values in **Table 1.** The number of rows in the table is 1,573 and there are no nulls in any of the columns. The maximum deaths directly caused by a weather event is much higher than the mean number of deaths, as shown below. There is a big spread of values over the different states and event types, the hidden relationships will hopefully be captured by the model.

Table 1.

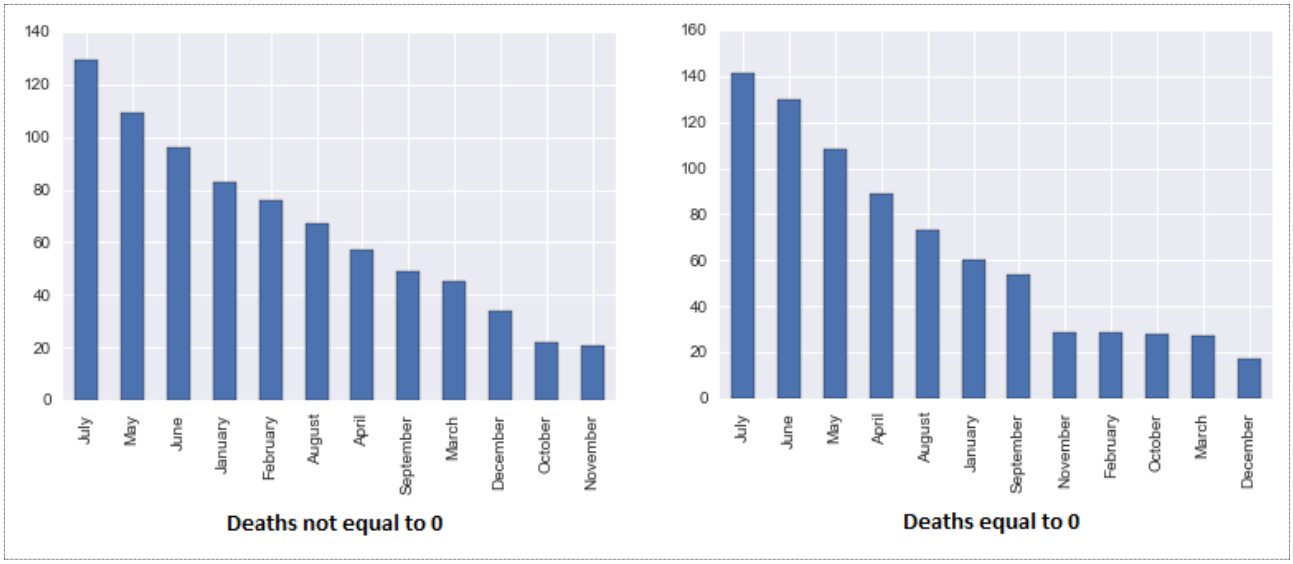
	count	mean	std	min	25%	50%	75%	max
DEATHS_DIRECT	1,573.00	0.72	1.76	0.00	0.00	1.00	1.00	43.00
DEATHS_DIRECT_CAT	1,573.00	0.50	0.50	0.00	0.00	1.00	1.00	1.00
MONTH_NUMBER	1,573.00	5.84	2.80	1.00	4.00	6.00	8.00	12.00
STATE_NUM	1,573.00	23.47	14.84	1.00	9.00	23.00	36.00	50.00
EVENT_TYPE_NUM	1,573.00	19.59	8.44	1.00	11.00	22.00	26.00	32.00
CZ_TYPE_NUM	1,573.00	1.41	0.49	1.00	1.00	1.00	2.00	2.00

For my categorical variables I looked at bar charts of the counts in each category when the number of deaths was not equal to 0 and when they were equal to 0, to see if there was any distinction

between the two death indicators. The important part of these graphs is the visual distribution across the whole subset.

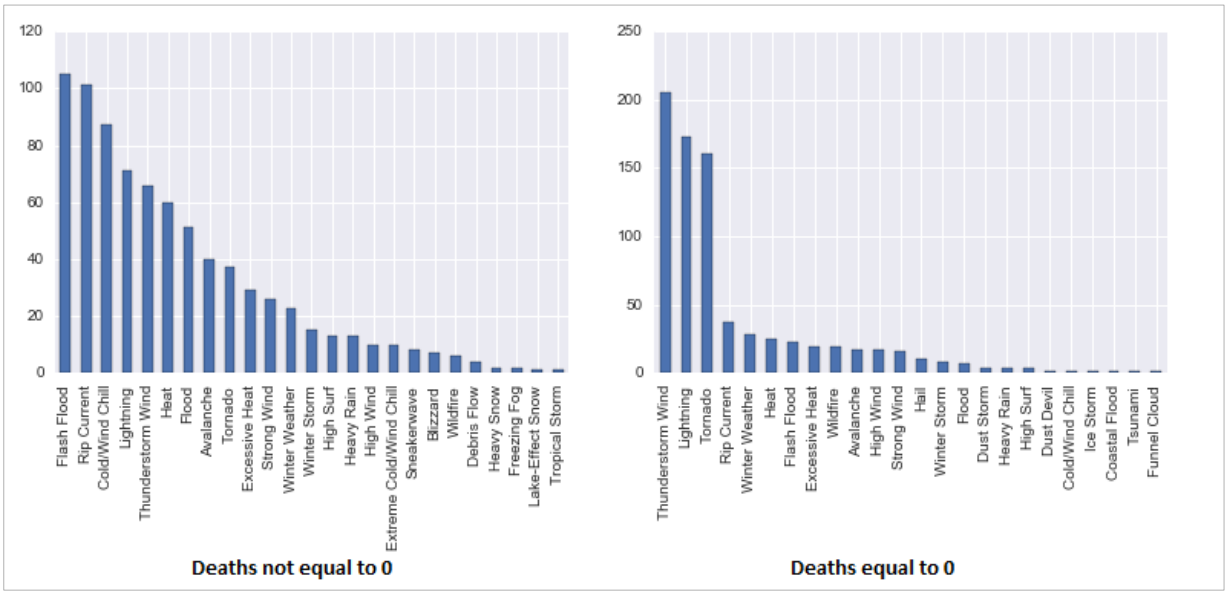
In **Figure 1.** I look at the month the event occurred in, and we can clearly see that there are no big differences, which would indicate that there was no impact on the response variable from the predictor.

Figure 1.



In **Figure 2.** I look at the event type of the weather event, and it is clear that there is a very different distribution of counts for when there are deaths and when there aren't. It is likely that this predictor will have an effect on the model.

Figure 2.



County / zone type, as well as state to a lesser degree, also shows strong indication that there is a difference between when there is deaths and when there isn't. Indicating that the variables will have an impact on the model. All the observations from the bar charts are supported by the frequency tables.

For the quantitative variable I looked at the frequency tables. The deaths indicator has an even distribution between it's two values, 0 and 1.

Bivariate analysis

I used an ANOVA test on each categorical variable to check for independence.

For state there was a p-value less than 0.05 for five States: Nevada, Arizona, Oklahoma, Utah, Washington. The THSD test indicated that only Nevada was a true positive, making it the only significant association.

For month, only March has a p-value less than 0.05 from the ANOVA and the THSD indicated that this was a false positive, further indicating that month is not significantly associated with deaths.

For event type, three categories had a significant association with the deaths indicator, and they were all supported by the THSD test. They are debris flow, lightning, and thunderstorm wind.

Finally, for county / zone type, the association is very significant p-value less than 0.000 and the THSD supports this.

For my bivariate analysis for my qualitative variables I looked at scatter plots of the values against the number of deaths. I also plotted a linear line of best fit, as produced by the Pearsons analysis. These linear lines of fit visually represent the Pearson coefficients.

The deaths indicator has a significant but slightly negative association coefficient (-0.0669) with the month number (p-value 0.008).

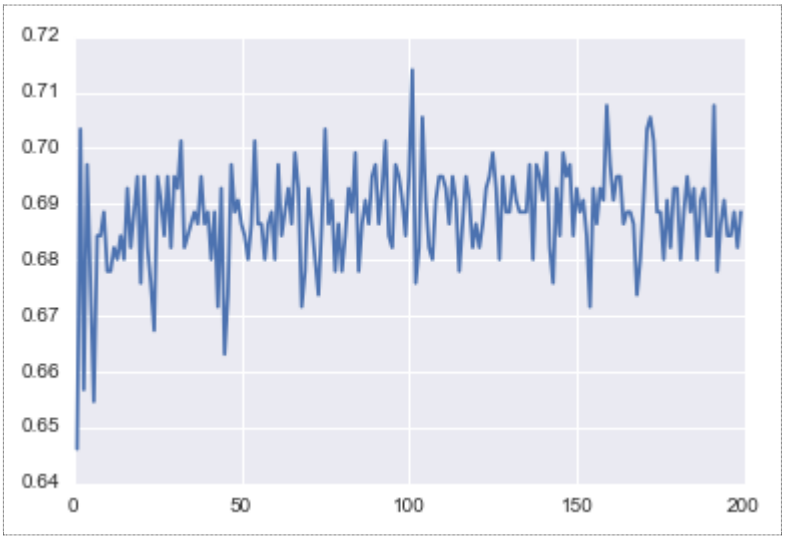
Multivariate analysis

I used a Random Forest to analyse my variables. I included the state, event type, and month number which I established had an association with my response variable deaths happened indicator.

I split my data set into 30% test set, 70% training set as I wanted a good amount of data to test my model on. I used a specific random state to split the data by so that my results were reproducible. The outputted test and training sets had the proportions training 59.73%, test 40.27%.

I identified the number of estimators that had the most stability for the model, which was exceptionally high at 125, which I was excepting because of the high number of categories over all the variables. The Random Forest model had an accuracy of 70%. The accuracy of the model for different estimator values is represented graphically in **Figure 3.** Over the different estimator numbers the model seems to stabilise, fluctuating mainly between 67% and 71% accuracy. The mode accuracy value for the different estimator numbers tested was 69%, which is slightly lower than our final model that uses 125 estimators and has an accuracy score of 71.6%.

Figure 3.



The accuracy of the model is represented by the confusion matrix, displayed in **Table 2.** The accuracy doesn't look amazing, but it is reasonably accurate.

Table 2.

prediction value	0	1	Total
true value			
0	185	53	238
1	90	144	234
Total	275	197	472

The importance of the variables passed to the model is 34% for the state, 50% for the event type, and 16% for the month the event occurred in. This indicates that all the included variables have a

significant impact on the model, but that the state and event type are substantially more significant than the month number.

Conclusions

The sample set of 1,573 entries was used to create a model that could predict if deaths were going to occur from a weather event. The sample includes all events where a death or injury occurred as a direct result of that weather event. The deaths indicator took values either 0 or 1 indicating if any deaths occurred or not, the proportion of 0's and 1's is 50-50, an even split. However the actual number of deaths that were directly caused by a weather event ranged from 0 to 43. This indicates a lot of variability but also that there was a need to be filled, the prevention of deaths from weather events.

Much as we would expect, the state and the event type are the most significant factors in the model for predicting if deaths occurred or not due to the weather event. That the model makes these predictions with an accuracy of 71.6% indicates that these two factors are likely the most important of any, including ones that may be included in the future, in predicting if deaths will occur from a weather event.

By inputting potential weather events that are either expected to occur because they happen seasonally, or are predicted to occur by predictive weather algorithms, we can identify to 71.6% accuracy whether there will be deaths or not. From knowing if there is potential for deaths emergency services can be forewarned, the public can be warned, and local authorities can issue guidance on how to avoid becoming a victim of the weather.

Limitations and future plans

There is not much data. The sample data set is limited to the weather events where either death or injury occurs. By limiting it to this subset we are more likely to build a working model that predicts accurately, however, as the model is trained on the subset it may not extrapolate nicely to the whole population.

The model is not very accessible in python, I cannot display the actual tree that is selected by the sklearn algorithm to use as the predictive model.

For future analysis it would be good to merge this data set with others so we gain information about spending on weather defences by local authorities, and frequencies of similar events happening in those locations. I believe that both will act as confounding factors. If an area already

has high spending on weather defence and education on weather then it is likely that there will be fewer deaths from severe weather as people are prepared for it.