**The association of the cost of property damage caused by weather events with its characteristics.**   
  
**Introduction.**

This is my final project in the online course [Data Analysis and Interpretation](https://www.coursera.org/specializations/data-analysis) organized by Wesleyan University (USA) through Coursera.org.  
The data contains various weather events, happened in the USA from 2013 to 2015. The Storm Event Database from [the National Center for Environmental Information](href=%22https:/www.ncdc.noaa.gov/) was the source of the data.

The purpose of the research is to identify what characteristics of various weather events are associated with the cost of property destroyed by this particular event.

The goal of the research to check the association of property damage with a climate region where it happened, month (relation to seasons), the scale of the event (did it happened on the county or zone level), the event duration and type of weather events.

Although I am a biologist and not climatologist, it is important for me to gain an experience with data which is not related to biology or medicine.  
The result of the research could be used for minimizing the damage caused by weather events. In process of city planning, house building other kinds of similar decision making it is important to know where in which month and what kind of weather event can be associated with a serious damage.

**Methods and data management**

**Sample**

The dataset contains N = 166068 weather events that took a place between January 2013 and October 2015 in the United States of America. This dataset is a part of the official publication of the National Oceanic and Atmospheric Administration (NOAA). The part of information could have been provided not by the National Weather Service (NWS), but by the media, law enforcement and/or other government agencies, private companies, individuals etc. Beyond ordinary weather events, rare or unusual phenomena and some other meteorological events like maximum or minimum temperature were written. After first attempts of univariate analysis, it became obvious that property damage is positively skewed and contains N= 101517 observations where no damage was registered and N=37033 where the damage took place. Thus, to perform a proper research, I divided it in two parts.

In the **part I**, the explanatory variable is categorical (whether the property damage took place or not) and the sample volume was N=138550. The sample included all observations except ones where property damage was not properly evaluated or it was unknown if property damage took place.

The goal of the **part II** was to find which features are mostly associated with the volume of property damage. The difference in the sample from the previous part is that also all observation with zero property damage were excluded, so the sample volume was N=37033.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 1. The division of the research in two parts.** | | | | |
| **Part** | **The goal** | **The response variable** | **The data management** | **The sample volume** |
| **I** | Find out which variables can be associated with the fact that property was damaged. | Categorical variable.  **1** – if the property was damaged  **0** – if the property was not damaged. | All observations where the amount of property damage was more or equal to zero. | N= 138550 |
| **II** | Find out with which variables the amount of property damage can be associated. | Quantitative variable.  The logarithm of the property damage with base 10  (more details in measures) | Only observations where the property damage was higher than zero. | N= 37033 |

**Measures**

The set of putative predictors was the same for both part of the research.

The cost of damage was entered as actual dollar amounts, but only in case if reasonably accurate estimate could be found. The estimation was provided by an insurance company or other individuals who were qualified enough to perform the evaluation. The observations with unknown or missing data about the property damage were removed. Because the property damage distribution was positively skewed and, the variable was modified. The logarithm with a base 10 became new quantitative response variable. As a consequence, skewness was significantly decreased.

The explanatory variables can be found in the table 2. The event designator(cz\_type) of weather event shows is the event happened in county(C), zone(Z) or Marine Zone(M). The county is administrative unit of state. Zone in this sample in this center means NWS Forecast zone which includes several counties. The designator shows which kind of events could happened as well as demonstrate the spread of the event. Marine zone was unintentionally ruled out of the sample after the data management.

A climate region originally was not presented in the data set. However, I have found the map on the [NOAA site](https://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php). Using states provided in the data set, I have created new variable. The District of Columbia was added to the Northeast climate region. Alaska was put to the separate category. Other events became a part of “Other” category. The last one category without Alaska and DC contains marine related places, mostly in equatorial climate.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2. The explanatory variables used in the analysis** | | | |
| **The variable name** | **The type of variable** | **Meaning** | **Methods where the variable was used** |
| “climate\_region” | Categorical | A climate region where the event took place | ANOVA, Chi-square, Decision Tree, Random Forest, Multiple regression |
| “cz\_type” | The designator showing if the event happened in county(C), zone(Z) or in the see(M). | ANOVA, Chi-square |
| “event\_type” | The type of the weather event (e.g. Flood, Marine Thunderstorm Wind etc.) | Decision Tree, Random Forest, Multiple regression |
| “month\_name” | A month when the event happened |
| “event\_duration” | Quantitative | The duration of the event in hours. |

**Analyses**

Univariate and bivariate tests were performed for the both parts of the research (with categorical and quantitative response variables). Distribution of every categorical variable was evaluated by frequency table. The bar charts with the number of events in each category were examined. Additionally, with quantitative variables histograms with distribution were examined.

For a bivariate test of the association of a climate region and the event type designator with the damage level, the Chi-square test and the Analysis of Variances (ANOVA) were used in case of the categorical and quantitative response variables, respectively. Because a climate region was variable with more than two categories, post-hoc test was done. For Chi-square test the Bonferroni adjustment was implemented and for ANOVA, I have used the Tukey test.

To perform deeper analysis (regression and machine learning approaches) more explanatory variables were used (Table 2).

With categorical response variable (whether the property damage happened or not), the decision tree and the random forest methods were implemented. All observations with unknown or missing data related to putative predictors were excluded from the sample. As a result, the sample volume for machine learning approaches was N=90142. Then, the dataset was divided into a training set (70%) N = 63099 and test sets (30%) N = 27043. Because all categorical variables chosen for machine learning contained more than 2 categories, One Hot Encoding was performed. As a result, every category in the variables was transformed to separate binary variable.

To analyze predictors for the quantitative response variable (the amount of property damage), the multiple regression model was used. The quantitative explanatory variable “event duration” were centered by subtracting the mean.

**Results**

**Descriptive statistics.**

As mentioned above (table 1), for the both parts of the research the set of explanatory categories is the same, but the response variable and sample are different. So descriptive statistics should be divided in two parts as well.

All quantitative variables presented in the table 3. Because in the **part I** the response variable is categorical, the only one quantitative variable is the event duration. In the **part II**, response variable is quantitative, which make 2 continuous variables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 3. The quantitative variables used in the research.** | | | | | | |
| **Variable** | **Research part** | **N** | **Mean** | **Std Dev** | **Min** | **Max** |
| Event duration(lg) | I | 90142 | 0.566537 | 1.193012 | -1.778151 | 2.871563 |
| II | 20054 | -0.057898 | 1.075208 | -1.778151 | 2.871563 |
| Property damage(lg) | II | 37033 | 3.925944 | 0.848408 | 1.000000 | 9.301030 |

|  |
| --- |
|  |
|  |
| **Figure 1. Histograms of the event duration(lg) in part I (top) and part II (bottom) of the research.** |

As we can see the numbers of value in the event duration are different from the sample volumes, because some of them are missing or unknow. After ruling out the observations with zero damage we can see decreasing of the mean Additionally, histograms (Figure 1) of the event duration logarithm are different between two parts of the research. Even from the table with mean, standard deviation and minimum and maximum value, it is easy to sea, that the distribution is still right skewed, because most events still are associated with low property damage, even after using logarithm operation. The distribution related to the part I (Figure 1, top) is symmetrical, but not unimodal. The one from the part II (Figure 1, bottom) is positively skewed and probably has two modes. Anyway, logarithm has significantly decreased positive skewness.

Categorical variables are presented in the table 4. All observations not equally distributed among the categories, that can be concluded even from the frequency of top (most frequent) categories. Especially it can be seen with the event type in the part II, where thunderstorm winds are in 53% of the all observations among 44 categories.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 4. The categorical variables used in the research.** | | | | | |
| **Variable** | **Part** | **N** | **Unique categories** | **Top category** | **Frequency**  **(Top)** |
| Property damaged | I | 138550 | 2 | 0 (no damage) | 101517 / 73.27102% |
| Month | 138550 | 12 | June | 22800 / 16.456153% |
| Climate region | 138550 | 11 | South | 26774 / 19.324432% |
| Designator | 138550 | 2 | C | 81709 / 58.974377% |
| Event type | 138550 | 51 | Thunderstorm Wind | 34138 / 24.63948% |
| Month | II | 37033 | 12 | June | 8223 / 22.204520% |
| Climate region | 37033 | 11 | Central | 8351 / 22.550158% |
| Designator | 37033 | 2 | C | 31585 / 85.288796% |
| Event type | 37033 | 44 | Thunderstorm Wind | 19659 / 53.085086% |

Other interesting differences can be observed in distribution between two parts of the research i.e. between sample with and without observations with zero damage, respectively. For example, the frequency of thunderstorm wind is significantly higher (from 24.6% to 53.1%) if we throw row without damage away. Another example it is that 7 event types are not associated with a property damage at all.

**Bivariate analysis**

To perform simple analysis

**Machine learning approaches**

To analyze potential association with

**Regression model**

To

**Conclusion/Limitations**