**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 first case 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 second part 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **The goal** | **The response variable** | **The data management** | **The sample volume** |
| 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 |
| 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.

For explanatory variables next ones were chosen:

* A **month** when the event happened;
* A **climate region** where the event took place;
* the **type** of the weather event (e.g. Flood, Marine Thunderstorm Wind etc.)
* The **duration** of the event in hours.
* So-called “**cz type**” which means if the event happened in county(C), zone(Z) or in the see(M).

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.

**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 month, a region, cz-type or type of weather event 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. To test if there is any association between property damage and event duration, Pearson correlation was implemented.

To perform a deeper analysis of 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.

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**

**Univariate analysis**

To check the distribution of every variable

**Bivariate analysis**

To perform simple analysis

**Machine learning approaches**

To analyze potential association with

**Regression model**

To

**Conclusion/Limitations**