**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 |
| “month\_name” | A month when the event happened |
| “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 |
| “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 and month were variables 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. For random forest 25 trees were observed.

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 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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% |

|  |
| --- |
|  |
|  |
| 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 unknown. After ruling out the observations with zero damage we can see decreasing of the mean. The distribution related to the part I (Figure 1, top) is symmetrical, but not unimodal. A different picture can be observed the one from the part II (Figure 1, bottom) is positively skewed and probably has two modes. Anyway, logarithm has significantly decreased positive skewness. As for property damage, even from the table, it is easy to see, that the distribution is still right skewed, because most events still are associated with low property damage, even after using logarithm operation.

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.

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**

Because all explanatory variables chosen for bivariate analysis are categorical (Table 1), only bar charts were built. All the bar chats from both part of the research was examined parallelly in order to see the whole picture.

|  |  |
| --- | --- |
|  | |
| *Chi-square value: 5113.28; p-value: 0.0* | *F-statistic: 31.64; p-value: 1.36e-67* |
| Figure 2. The association of the categorical(left) and quantitative(right) property damage variables with a month when the event took place. For every category mean is calculated and confidence intervals are shown. | |

On the figure 2 we can see the bar charts showing the association between property damage and month when it happened. In both cases p-values are very low, so we can say that there is some difference between the groups. From this graph we can see three different groups of months. First group, if from December to March inclusively, when property damaged with relatively low frequency: from 14,04% to 15.02%. The second group contain June and July when frequency of property damage is significantly higher than in the other months (36.07% and 37.27% respectively). The last group includes April, May and months from August to November. In these months’ frequency varies from 27.21% to 30.79% So, we can see seasonal changes in the frequency of damaged property. Post-hoc tests (not shown) strongly confirm this idea.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5. The comparison table of post hoc Tukey test.  All pairs which are significantly different mark from green (bigger mean difference)  to yellow (less mean difference). Red means no significant difference. | | | | | | | | | | | | |
| \* | 01.Jan | 02.Feb | 03.Mar | 04.Apr | 05.May | 06.Jun | 07.Jul | 08.Aug | 09.Sep | 10.Oct | 11.Nov | 12.Dec |
| 01. |  |  |  |  |  |  |  |  |  |  |  |  |
| 02. |  |  |  |  |  |  |  |  |  |  |  |  |
| 03 |  |  |  |  |  |  |  |  |  |  |  |  |
| 04. |  |  |  |  |  |  |  |  |  |  |  |  |
| 05. |  |  |  |  |  |  |  |  |  |  |  |  |
| 06. |  |  |  |  |  |  |  |  |  |  |  |  |
| 07. |  |  |  |  |  |  |  |  |  |  |  |  |
| 08. |  |  |  |  |  |  |  |  |  |  |  |  |
| 09. |  |  |  |  |  |  |  |  |  |  |  |  |
| 10. |  |  |  |  |  |  |  |  |  |  |  |  |
| 11. |  |  |  |  |  |  |  |  |  |  |  |  |
| 12. |  |  |  |  |  |  |  |  |  |  |  |  |

The variance in the amount of property damage among months is lower than it was with the categorical response variable. The values of the property damage logarithm were from 3.833316(in June) to 4.100127 (in December) which means from 6812$ to 12592$. There is some association of the amount with group of months because p-value is pretty low, which means that not all groups are equal and we have to reject the null hypothesis. The table 5 shows that some of the months are significantly different from others. From the graph we can see that average damage property drops in June. Then we do not see any significant difference between neighbor months until October, which is confidently lower than November. From October to February all months are different with the highest value in December. The months from February to May are not different with each other in terms of the property damage amount. So, damage amount has two stable periods from February to May and from June to October, then we can see some changes.

Thus, there are strong and clear association between the fact of damaged property with the season when it happened. About the amount of damage, there are two periods without significant change in the amount of property damage and a winter period with significant difference but without clear pattern.

On the figure 3 we can see the bar charts showing the association of the property damage variables with climate region. In both cases (the fact of property damage and the amount of it) p-value so low, that it was considered like zero. Like with the months when we divide observations to climate region we can see bigger variance of the frequency of events caused property damage than variance in the amount of it.

|  |  |  |
| --- | --- | --- |
|  |  | |
| Alaska |
| Central |
| Northeast |
| Northwest |
| South |
| Southeast |
| Southwest |
| Upper Midwest |
| West |
| WN Central |
| Other |
|  |
|  | *Chi-square value: 5113.28; p-value: 0.0* | *F-statistic:* *321.4; p-value: 0.0* |
| Figure 3. The association of the categorical(left) and quantitative(right) property damage variables with a climate region where the event took place. For every category mean is calculated and confidence intervals are shown. | | |

Southeast climate region demonstrates the highest frequency (44.27%) of events caused property damage. Interestingly, that the average damage amount is one of the lowest in the same climate region. The lowest frequency of events with damaged property (2.34%) observed in the Other category (which is related to marine zones and islands). Additionally, “Other” climate region demonstrates the lowest mean value of property damage logarithm (3.62 or 4168$). The highest amount of damage (4.45 or 28183$) was found in the West North Central region, however, this group of observation significantly different from all regions except Alaska.

As for event type designator, the picture is different for categorical and quantitative property damage variables. As we can see after causing-damage weather events related to C(county) designator happens significantly more often 38,66% of this-kind of events against 9.58% of events with Z (zone) designator. The same time, the mean amount of property damage caused by C-type events is higher than by Z-type ones (4.07 (11748$) and 3.9 (7943$) respectively). Because p-values related to both graphs are very low, both the differences are significantly different, moreover confidence intervals are so low, that it is difficult to see them on the graphs.

|  |  |
| --- | --- |
|  | |
| *Chi-square value:* *14463.76; p-value:0.0* | *F-statistic:188.8; p-value:* *7.37e-43* |
| Figure 4. The association of the categorical(left) and quantitative(right) property damage variables with an event type designators Z(blue) and C(orange). For every category mean is calculated and confidence intervals are shown. | |

To summarize, we have observed the association of the fact of property damage and the amount of it with three different probable predictors. The frequency of events where property damage was different from zero is highly associated with months, climate regions and event type designator. The intergroup variance in property damage amount is lower between the same groups, however there are definitely different (p-value is pretty low) and some significant differences were observed in every pair.

**Machine learning approaches**

To perform multivariate analysis of the categorical property damage variable (whether property damage was bigger than zero (1) or not (0)), the Random Forest approach was implemented. As explanatory variables logarithm of the event duration, type of event, climate region and month were chosen. Because all categorical explanatory variables contained more than two categories, One Hot Encoding was performed. As a consequence, every event type, every month and every climate region were turn to binary variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 6. The result of the Random Forest model. | | | | |
|  | Precision | Recall | f1-score | Support |
| No damage (0) | 0.91 | 0.93 | 0.92 | 21048 |
| Damage > 0 (1) | 0.74 | 0.68 | 00.71 | 5995 |
|  |  |  |  |  |
| Micro avg | 0.88 | 0.88 | 0.88 | 27043 |
| Macro avg | 0.83 | 0.81 | 0.82 | 27043 |
| Weighted avg | 0.87 | 0.88 | 0.87 | 23043 |

The Random Forest confusion matrix has shown 19651 successfully predicted observations without damage and 4055 with it. The false positive (with damage) and false negative (without it) number of observations were 1940 and 1397, respectively. The accuracy score of the model was 0.8776. More detailed report can be found in table 6. The table shows that the model better predicts observations without property damage than with it. In the table 7 there are explanatory variables with feature importance higher than 1%. The highest score related to the event duration (38.84%). All other variables were related to either event type or climate region. Months also were in the full feature importance table, however the highest feature importance related to month were with May and June and was only 0.7%. Among the event type the Thunderstorm Wind holds the largest feature importance (10.46%). Another event types like Strong Wind, Flash Flood, Tornado etc. have relatively high feature importance too. As for climate regions, the Southeast climate region has the highest feature importance (2.24%) in its category, but not so high like some event types.

|  |  |
| --- | --- |
| Table 7. The feature importance with the value higher than 0.01 | |
| lg (event duration) | 0.3884240 |
| event type: Thunderstorm Wind | 0.1046360 |
| event type: Strong Wind | 0.0733599 |
| event type: Flash Flood | 0.0475370 |
| event type: Tornado | 0.0313422 |
| climate region: Southeast | 0.0224069 |
| event type: Flood | 0.0216158 |
| event type: Winter Weather | 0.0183642 |
| climate region: Northeast | 0.0171283 |
| event type: High Wind | 0.0161539 |
| climate region: West North Central | 0.0158418 |
| event type: Drought | 0.0152262 |
| event type: Hail | 0.0152017 |
| climate region: Central | 0.0132139 |
| climate region: South | 0.0125632 |
| climate region: Upper Midwest | 0.0111124 |

To estimate if 25 trees were enough to create an accurate model, the model was rerun with numbers of trees from 1 to 25 and the plot showing the accuracy score on y-axis and the number of trees on x-axis was built (Figure 5).

|  |
| --- |
|  |
| Figure 5. The accuracy score depending of number of trees in the random forest model. |

All value in the plot (Figure 5) are between 0.8666 and 0.8753. After 5th tree the accuracy changes even less. So, there is no need to rerun the model with a higher number of trees. Moreover, because the accuracy score is pretty high even for one tree, it can be a good idea to visualize the data, running the decision tree with the same set of explanatory variables.

The accuracy score of the decision tree was 0.8395. According to the confusion matrix there were 20334 successful predictions of lack of property damage and 2371 with the presence of it. The number false positive (with damage) predictions was 726 and false negative 3612. Like it was with random forest, the table report (not shown) demonstrated better precision for observations without damage. To summarize, the accuracy score is lower with decision tree than with random forest, but it is still high and all statistics are more or less the same.

On the figure 6 the decision tree is visualized. The Gini impurity value before the first split is equals 0.346. The split could happen only of it leads to improve Gini impurity index. The most successful parts are about events which belongs to Strong Wind type (Gini=0.007, 1084 samples). The best result related to the lack of property damage was with Gini index 0.108, and it is related to the events whose type is not thunderstorm/strong wind, flash/regular flood, tornado, ice storm or lightning and that did not happen in Northeast climate region. Also, thunderstorm wind is the first point of split, and if the event was not of this kind the Gini impurity was improved from 0.347 to 0.288.

Thus, the visualized model shows that the event type (especially thunderstorm and strong winds) associated with the presence of property damage. The climate region (especially Southeast) also have an association with the target, but weaker than the event type. The event duration is not shown

|  |  |  |
| --- | --- | --- |
|  | | |
|  | Figure 6. The decision tree predicting whether the property damage doesn’t take a place in the observation (first value in every square) or it does (second value).  The orange shows the nodes, where more observations with damage.  The blue shows the nodes with a higher number of observations with property damage.  The intensity of blue/orange shows how big the part of Damage/No damage observations.  Gini impurity index is shown in every node. |  |

To summarize, the machine learning technics helped to show the most important explanatory variables associated with the property damage presence. The random forest model’s accuracy score was 87% and separate decision tree’s one was 84%. The biggest contribution belongs to the event duration, the event types and the climate regions. The “best” event types are thunderstorm and strong winds, flash flood, tornado and flood. The climate region which is mostly associated with property damage is Southeast. Other climate regions and weather events had lower than 2% feature importance score to the model.

**Regression model**

To analyze the property damage amount and how is it associated with event duration, event type, climate region and month when it happened, multiple regression analysis was implemented.

The table 8 shows that

**Conclusion/Limitations**

In process