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 organized by Wesleyan University (USA) through Coursera.org.
The data contains various weather events, happened in the USA from January 2013 to October 2015. The Storm Event Database from the National Center for Environmental Information 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 presence and cost of property damage with a climate region where it happened, month (relation to seasons), the event type designator (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 possible 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 into two parts.

In **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. To perform machine learning analysis, all unknown or missing data were excluded and sample volume was N=90142.

The goal of **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. Because of missing or unknown data, the sample volume for the regression model was lower (N=20054).

Table 1. The division of the research in two parts.

Part	The goal	The response variable	Data management	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 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 a new quantitative response variable. As a consequence, skewness was significantly decreased.

The set of putative predictors was the same for both parts of the research. The explanatory variables can be found in table 2. The event designator (cz_type) of weather event shows is the event happened in a county(C), zone (Z) or Marine Zone (M). A county is an administrative unit of a state. Zone means NWS Forecast zone which includes several counties. The designator shows which kind of events has happened as well as demonstrate the spread of the event. Marine zone was unintentionally ruled out from the sample after the data management.

A climate region originally was not presented in the dataset. However, I have found the map on the NOAA site. Using states provided in the dataset, I have created a 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 an equatorial climate.

Table 2. The explanatory variables used in the analysis

Variable name	Variable type	Meaning	Approached methods
"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 both parts of the research (with categorical and quantitative response variables). Distribution of every categorical variable was evaluated by frequency tables. 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, month 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, posthoc test was done. For Chi-square test the Bonferroni adjustment was implemented and for ANOVA, I have used the Tukey test.

To perform a 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 into a 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. Because of unknown or missing data, the sample volume for regression model was N=20054.

Results

Descriptive statistics.

As mentioned above (table 1), for 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 into two parts as well.

All quantitative variables presented in table 3. Because in the part I the response variable is categorical, the only one quantitative variable is the event duration. In part II, the 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%

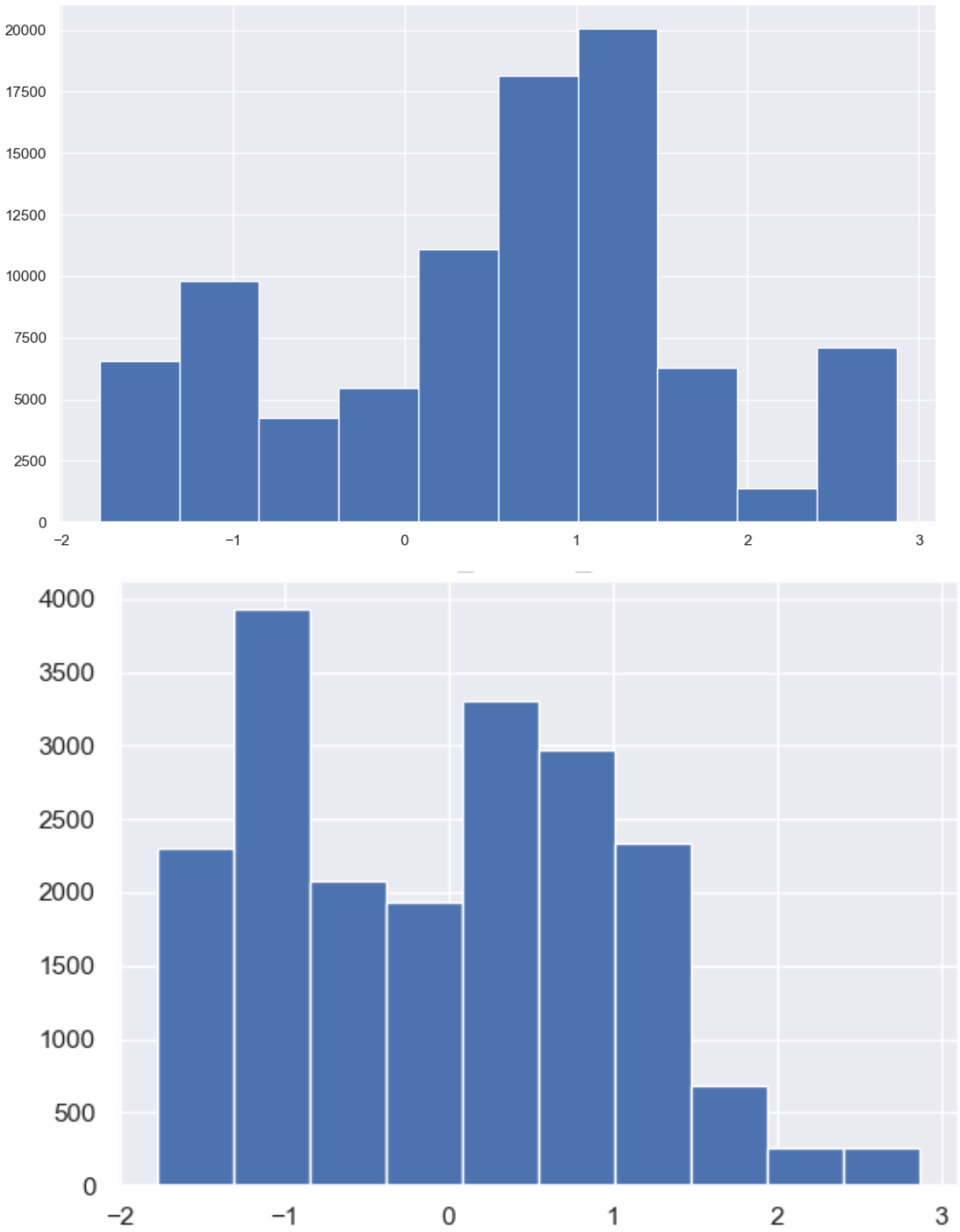


Figure 1. Histograms of the event duration(lg) in part I (top) and part II (bottom) of the research.

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 three modes. Anyway, logarithm has significantly decreased positive skewness. As for property damage, even from the table, it is easy

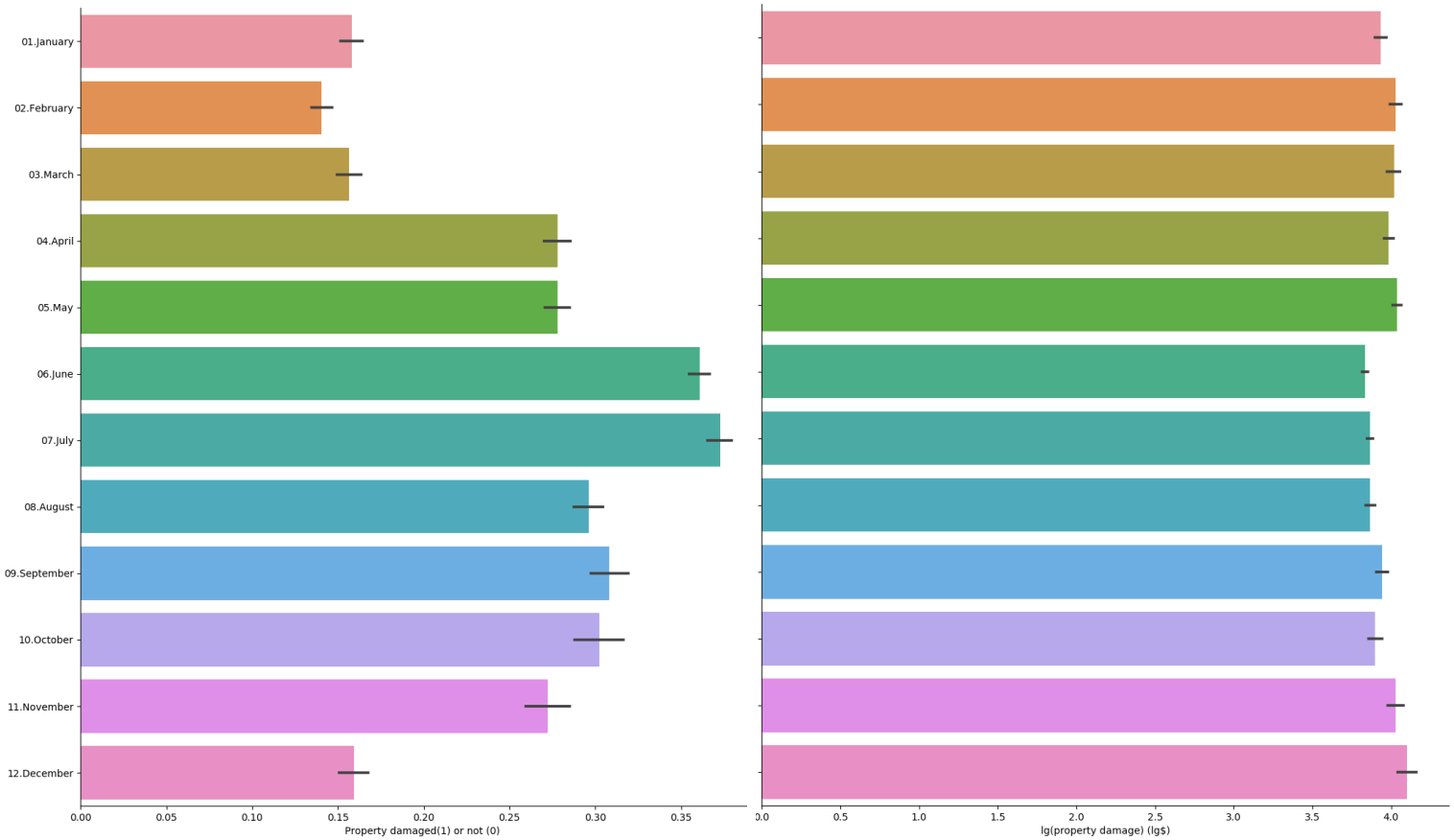
to see, that the distribution is still right skewed because most events still are associated with low property damage, even after using logarithm operation.

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

Other interesting differences can be observed in distribution between two parts of the research i.e. between the 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 do not consider observations without damage. Another example is that 7 event types do not cause any damage higher than zero, so they 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 parts of the research were examined together 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 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. The first group, if from December to March inclusively, when property damaged with relatively low frequency: from 14.04% to 15.02%. The second group contains June and July when the 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. Posthoc tests (not shown) confirm the difference between these three groups.

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.

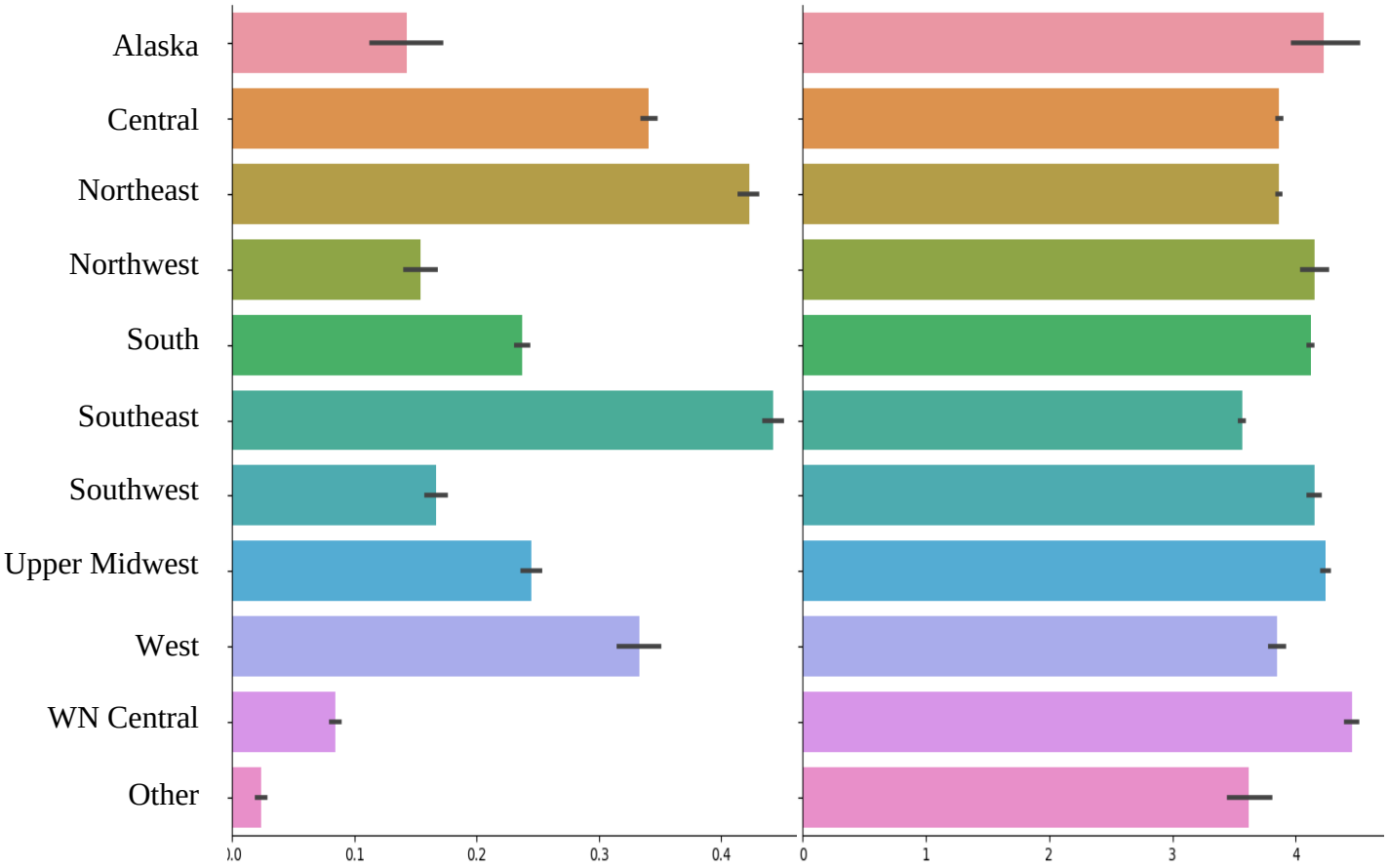
*	01.Jan	02.Feb	03.Mar	04.Apr	05.May	06.Jun	07.Jul	08.Aug	09.Sep	10.Oct	11.Nov	12.Dec
01.												
02.												
03												
04.												
05.												
06.												
07.												
08.												
09.												
10.												
11.												
12.												

The variability 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 between the damage cost with a 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 from 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 is a strong and clear association between the facts 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.

On 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 a bigger variability of the frequency of events caused property damage.

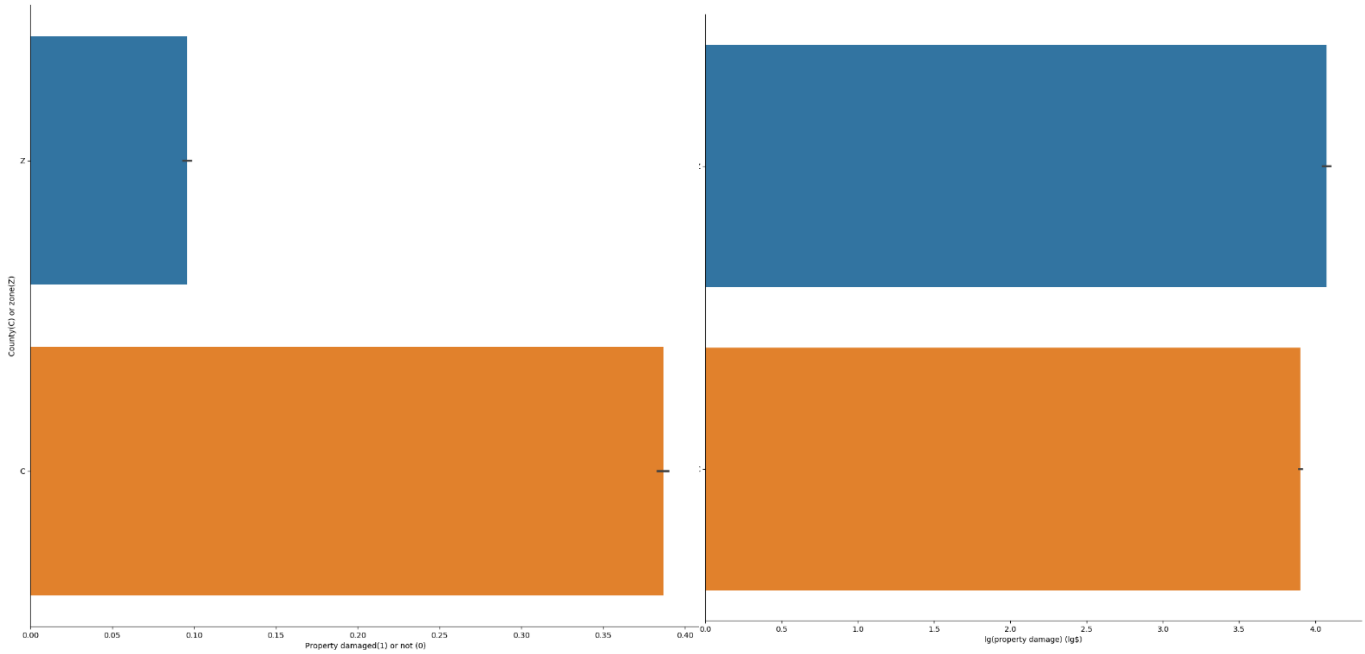


Chi-square value: 5113.28; p-value: F-statistic: 321.4; p-value: 0.0 0.0

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 cost 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 cost 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 variability in property damage amount is lower between the same groups, however, there are definitely different (p-values are pretty low) and some significant differences were observed in every pair.

Machine learning approaches

To perform a 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 turned to a binary variable.

The Random Forest confusion matrix has shown 19651 successfully predicted observations without damage and 4055 with it. The false positive and false negative number of observations were 1940 and 1397, respectively. The accuracy score of the model was 0.8776. The more detailed report can be found in table 6. According to the values of precision, recall and f1-score the model better predicts observations without property damage than with it. In 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 was 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 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

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 the y-axis and the number of trees on x-axis was built (Figure 5).

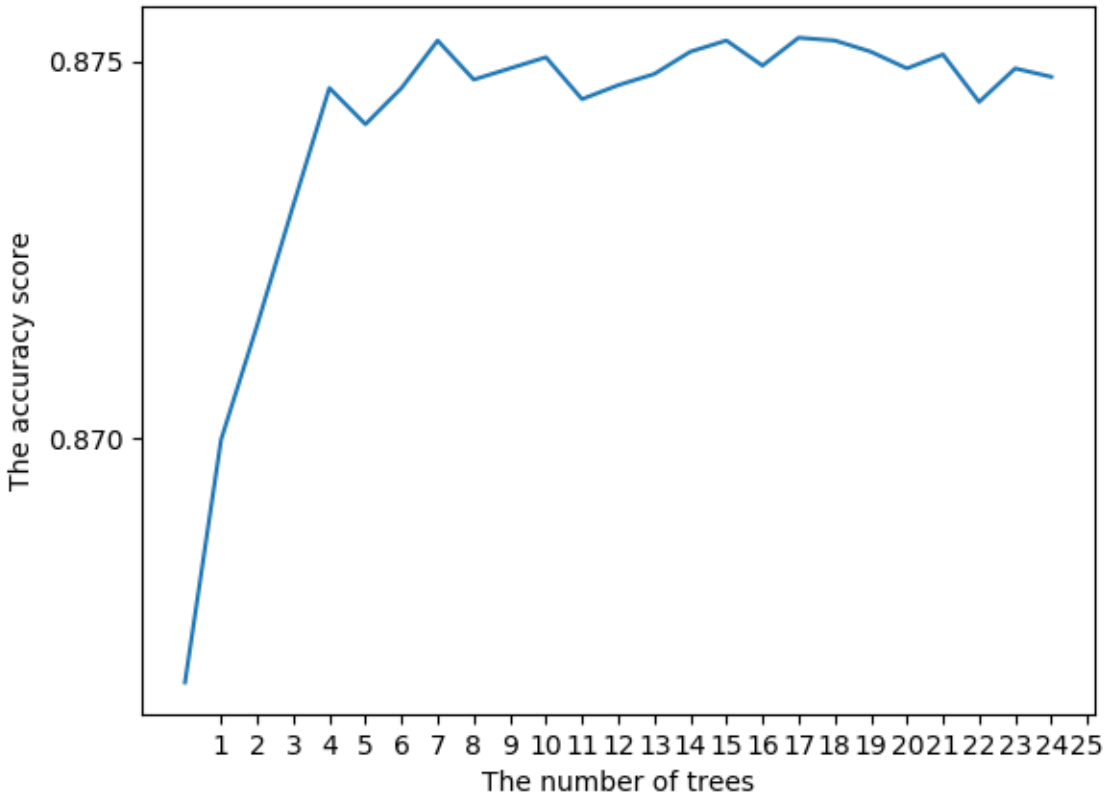


Figure 5. The accuracy score depending of number of trees in the random forest model.

All value in the plot (Figure 5) is 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 the 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 the decision tree than with random forest, but it is still high and all statistics are more or less the same.

On figure 6 the decision tree is visualized. The Gini impurity value before the first split equals 0.346. The split could happen only if it leads to improving Gini impurity index. The most successful parts are about events which belong 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 the 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.

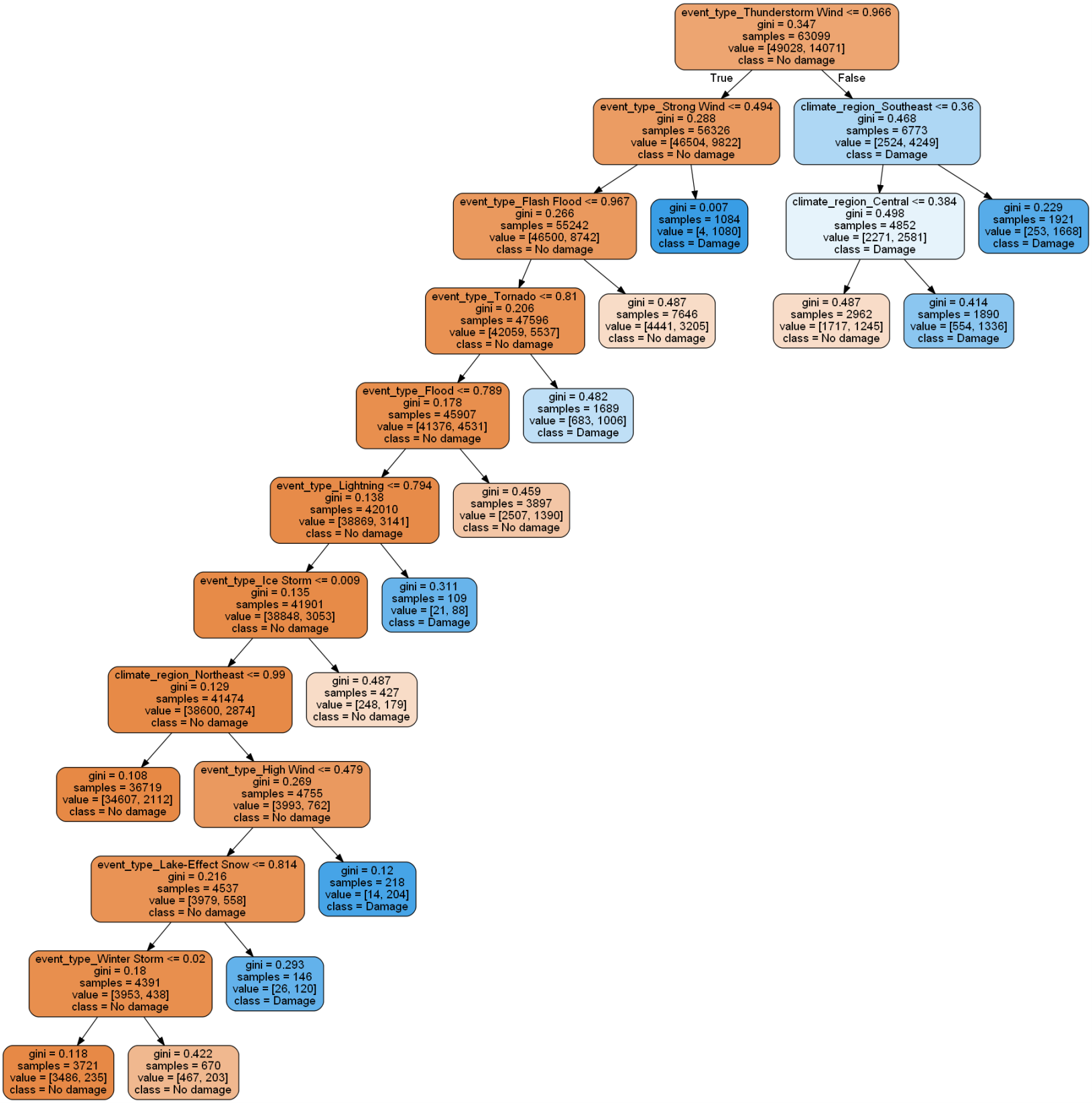


Figure 6. The decision tree predicting whether the property damage doesn't take a place in the observation (first value in every square) or it does (second value).
The orange shows the nodes, where more observations with damage.
The blue shows the nodes with a higher number of observations with property 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 in the 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.

From every categorical model, references with the least regression coefficient were chosen. The hypothetical reference event is drought happened in January in the "other" climate region (not related to US mainland or Alaska) which took 0.8751 hours (or 52 min 31 sec). Intercept equals 1.8189, which means that this event causes property damage in 65.90$. The results are summarized in Table 8. The majority of categories, including the only one quantitative variable, hold p-value less than 0.05. The p-value for the whole model is close to zero and R-square was 0.26, so it explains 26% of the variability.

Among the months February and August did not show significant difference with the property damage amount in January. The highest regression coefficient is in November (0.2356), so in the model, the month category contributes up to 0.2356 to the logarithm of property damage amount.

As for climate regions, Alaska, Southeast and West are not confidently different from the reference "other" climate region. The West North Central region demonstrates the biggest regression coefficient among regions (0.6778). Looking at other values, we can conclude that according to the model climate regions can be more related to the damage property cost than a month when it happened.

The event types regression coefficients are higher than ones related to other categorical variables, most of them are bigger than 1. The property damage caused by an avalanche, lakeshore flood, marine lighting or marine strong wind did not demonstrate a significant difference with it caused by drought. The highest regression coefficient is 3.3344 and related to storm surge or tide. Some other types revealed in the machine learning analysis like a most probable to cause a property damage also has high regression coefficients in the model: 2.0961 for thunderstorm find, 1.7673 for flash flood and 2.9551 for a tornado.

Write a general formula describing the model is a bit of complication because of many categories. For example, if we try to predict the property damage in the West North Central region in November caused by thunderstorm wind, the formula will be:

$$\mathrm{pd} = 10^{4.2686+\lg(t)}$$

where pd – property damage amount and t – the event duration in hours. So, if the thunderstorm wind takes one hour, then the property damage will be approximately 18560.94$. If we change

thunderstorm wind to strong wind, it changes the formula and the prediction of damage is 1229.42$.

Table 8. The result of multiple regression. The lines where p-value higher than 0.05 marked filled grey. C – categorical variable.

	Coef.	Std err	p-value
Intercept	1.8189	0.153	0
C (month_name, Treatment (reference='01. January')) [02. February]	0.0457	0.031	0.135
C (month_name, Treatment (reference='01. January')) [03. March]	0.0797	0.032	0.013
C (month_name, Treatment (reference='01. January')) [04. April]	0.1658	0.03	0
C (month_name, Treatment (reference='01. January')) [05. May]	0.2132	0.03	0
C (month_name, Treatment (reference='01. January')) [06. June]	0.0889	0.028	0.002
C (month_name, Treatment (reference='01. January')) [07. July]	0.1057	0.029	0
C (month_name, Treatment (reference='01. January')) [08. August]	0.0566	0.032	0.076
C (month_name, Treatment (reference='01. January')) [09. September]	0.0763	0.035	0.032
C (month_name, Treatment (reference='01. January')) [10. October]	0.1922	0.042	0
C (month_name, Treatment (reference='01. January')) [11. November]	0.2356	0.034	0
C (month_name, Treatment (reference='01. January')) [12. December]	0.1659	0.038	0
C (climate_region, Treatment(reference='Other')) [Alaska]	0.1972	0.137	0.151
C (climate_region, Treatment(reference='Other')) [Central]	0.2037	0.102	0.045
C (climate_region, Treatment(reference='Other')) [Northeast]	0.3936	0.102	0
C (climate_region, Treatment(reference='Other')) [Northwest]	0.3114	0.109	0.004
C (climate_region, Treatment(reference='Other')) [South]	0.49	0.102	0
C (climate_region, Treatment(reference='Other')) [Southeast]	0.1285	0.102	0.206
C (climate_region, Treatment(reference='Other')) [Southwest]	0.4106	0.104	0
C (climate_region, Treatment(reference='Other')) [Upper Midwest]	0.6523	0.103	0
C (climate_region, Treatment(reference='Other')) [West]	0.1353	0.105	0.197
C (climate_region, Treatment(reference='Other')) [West North Central]	0.6778	0.105	0
C (event_type, Treatment(reference='Drought')) [Avalanche]	1.0213	0.824	0.215
C (event_type, Treatment(reference='Drought')) [Blizzard]	1.7356	0.131	0
C (event_type, Treatment(reference='Drought')) [Coastal Flood]	1.9202	0.15	0
C (event_type, Treatment(reference='Drought')) [Told/Wind Chill]	1.9722	0.258	0
C (event_type, Treatment(reference='Drought')) [Debris Flow]	1.892	0.174	0
C (event_type, Treatment(reference='Drought')) [Tense Fog]	2.3728	0.236	0
C (event_type, Treatment(reference='Drought')) [Dense Smoke]	1.7143	0.819	0.036
C (event_type, Treatment(reference='Drought')) [Dust Devil]	1.6531	0.241	0
C (event_type, Treatment(reference='Drought')) [Dust Storm]	2.1926	0.205	0
C (event_type, Treatment(reference='Drought')) [Excessive Heat]	1.3829	0.584	0.018
C (event_type, Treatment(reference='Drought')) [Extreme Cold/Wind Chill]	1.8331	0.135	0
C (event_type, Treatment(reference='Drought')) [Flash Flood]	1.7673	0.11	0
C (event_type, Treatment(reference='Drought')) [Flood]	1.3081	0.108	0
C (event_type, Treatment(reference='Drought')) [Freezing Fog]	1.3222	0.379	0
C (event_type, Treatment(reference='Drought')) [Frost/Freeze]	2.5091	0.175	0
C (event_type, Treatment(reference='Drought')) [Hail]	2.3884	0.119	0
C (event_type, Treatment(reference='Drought')) [Heat]	2.3987	0.481	0
C (event_type, Treatment(reference='Drought')) [Heavy Rain]	1.6855	0.135	0

C (event_type, Treatment(reference='Drought')) [Heavy Snow]	1.3836	0.12	0
C (event_type, Treatment(reference='Drought')) [High Surf]	2.6129	0.381	0
C (event_type, Treatment(reference='Drought')) [High Wind]	1.6508	0.113	0
C (event_type, Treatment(reference='Drought')) [Hurricane]	2.9828	0.274	0
C (event_type, Treatment(reference='Drought')) [Ice Storm]	2.055	0.12	0
C (event_type, Treatment(reference='Drought')) [Lake-Effect Snow]	1.51	0.124	0
C (event_type, Treatment(reference='Drought')) [Lakeshore Flood]	0.6318	0.379	0.095
C (event_type, Treatment(reference='Drought')) [Lightning]	2.1073	0.137	0
C (event_type, Treatment(reference='Drought')) [Marine Dense Fog]	1.7794	0.592	0.003
C (event_type, Treatment(reference='Drought')) [Marine High Wind]	3.3154	0.493	0
C (event_type, Treatment(reference='Drought')) [Marine Lightning]	<0.001	<0.001	0.79
C (event_type, Treatment(reference='Drought')) [Marine Strong Wind]	1.2202	0.826	0.139
C (event_type, Treatment(reference='Drought')) [Marine Thunderstorm Wind]	1.6894	0.595	0.004
C (event_type, Treatment(reference='Drought')) [Seiche]	1.5027	0.481	0.002
C (event_type, Treatment(reference='Drought')) [Sleet]	1.6822	0.258	0
C (event_type, Treatment(reference='Drought')) [Storm Surge/Tide]	3.3344	0.583	0
C (event_type, Treatment(reference='Drought')) [Strong Wind]	0.9172	0.112	0
C (event_type, Treatment(reference='Drought')) [Thunderstorm Wind]	2.0961	0.116	0
C (event_type, Treatment(reference='Drought')) [Tornado]	2.9551	0.118	0
C (event_type, Treatment(reference='Drought')) [Tropical Depression]	1.5487	0.584	0.008
C (event_type, Treatment(reference='Drought')) [Tropical Storm]	1.7003	0.163	0
C (event_type, Treatment(reference='Drought')) [Waterspout]	1.9337	0.826	0.019
C (event_type, Treatment(reference='Drought')) [Wildfire]	1.7743	0.123	0
C (event_type, Treatment(reference='Drought')) [Winter Storm]	1.7623	0.111	0
C (event_type, Treatment(reference='Drought')) [Winter Weather]	1.4533	0.118	0
lg (event duration)	0.4544	0.012	0

Thus, all variables (but not every category) have an association with the amount of property damage. The strongest association related to the event type. The event duration is positively associated with the response variable. The categories which are plays important role in the damage amount are not always the same with ones contributed to the presence of property damage.

Conclusion/Limitations

The N=90142 weather events happened in the USA between January 2013 and October 2015 were analyzed by the random forest analysis. The regression model worked with a sample of N=20054 observations from the same period. Other approaches (descriptive and bivariate analyses) were performed with up to N = 138550 (depending on the approach) weather events. The random forest was used to identify months, event types, climate regions which can predict the presence of damage property as well as evaluate how strong the association of the event duration in the process. Multiple regression was implemented to predict the amount of property damage using the same set of explanatory variables.

The model predicting the presence of property damage has 87% accuracy, but better predicts negative results than positive ones. The model predicting the damage property amount explains

only 26% of the variability, however, adding more explanatory variables from the dataset could lead to bigger confounding.

Both response variables have demonstrated the association with the explanatory variable, which significantly depends on the particular category. The important moment is that not always the category which associated with the presence of property damage also is associated with the highest cost of it. Event duration is associated both with the fact of property damage as well as with its volume (highest feature importance in Random Forest; high regression coefficient). The event types also demonstrated higher feature importance scores and higher regression coefficients than other categorical variables. Especially, it is clear with thunderstorm winds, flash floods, and tornado. Although the regression coefficients related to this event were high, the strongest association with the property damage amount was demonstrated by storm surge (or tide) and marine high wind (which have shown low association with the fact of property damage). Strong wind has shown relatively strong association with the presence of property damage, but the correlation with the property damage amount was one the lowest among the other weather types. The climate regions have a weaker association with the response variables than event types. The Southeast and Northeast climate regions were more associated with the fact of property damage than other regions, but West North Central and Upper Midwest were "champions" in the correlation with the damage cost. Months had the weakest association in both models, however, the contribution can be still significant (which can be also supported by the clear seasonal deference demonstrated in the bivariate analysis with bigger sample). June and July were the months who are best associated with the fact of the property damage, and November and May were better associated with the damage amount than other months.

To summarize, both models demonstrate that property damage caused by a weather event is associated with different features of the event like its duration, type, place and season. For example, according to the models, the thunderstorm wind happening in the Southeast region in June has the highest probability to cause the property damage. The predicted highest damage amount takes a place in South climate region in November and caused by storm surge (of course there is a low probability of storm surge in West North Central or Upper Midwest region, so South region was taken). As higher the event duration, as a bigger probability of consequent property damage as well as a higher damage cost can be. Although most associations related to both models are clear, there were some problems during the research. The first problem it is very positively skewed distribution of the property damage amount and the event duration. Although logarithm partly has significantly decreased the skewness, there still was a too high value of damage cost which is related to the most serious disasters happened in the period (like Moore tornado in 2013). Another problem that actual event type cannot be independent of a month or a climate region, so confounding could take a place too. One more problem is mostly related to a model related to property damage amount prediction. The model does not predict the exact amount, but can more serve to compare which events where and when can cause more or less damage.

To build more successful prediction models, it can be good to separate the research by different climate zones. The reason for that is because every climate zone has its own set of the event types (for example on the coastal regions, there are more marine-related events).