Advanced Data Analysis Project Report-Flood Data Analysis

Grop 11
12/11/2016

Grop members:
Huilong An, ha2399
Yiqun Nian, yn2289
Zida Lin, zl2446
Xiaohan Zhang, xz2436
Peijun Dai, pd2485
Yu Qin, yq2186
Yifei Hu, yh2781
Kexin Gu, kg2619

Content

Section I: Background and Obejectives 3
- Background 3
- What is flooding? What is the cause for flooding? 3
- Why are we interested in this topic? 3
- Why we finally chose New York City as our target? 3
- How this paper consists? ..... **Error! Bookmark not defined.**

Section II Analysis of Length of the Weather Events 4
- Material and Method 4
- Results 7
- Sensitivity analysis to validate the assumptions 8
- Conclusions 9

Section III: Flooding Prediction 10
- Background 10
- Material and Method 10
- Results 13
- Sensitive Test 14
- Conclusions 14

Section IV : Flooding in New York - Flooding Damage on Properties and Lives 15
- Background 15
- Material and Method 15
- Sensitive Test 34
- Conclusions 34

Section V: Summary 35

Reference 36

Appendix 37
- Appendix A 37
- Appendix B 42
- Appendix C 44

Section I: Background and Obejectives

• Background

We are motivated by keeping receiving flooding alerts after coming to New York City. After some researches, we found floods is a very common extreme weather and did cause lots of damage on both lives and properties. So, we want to have a deep insight into all extreme weathers in New York area.

• What is flooding? What is the cause for flooding?

Flooding is an overflowing of water onto land that is normally dry. Floods can happen during heavy rains, when ocean waves come on shore, when snow melts too fast, or when dams or levees break. Flooding may happen with only a few inches of water, or it may cover a house to the rooftop. They can occur quickly or over a long period and may last days, weeks, or longer. Floods are the most common and widespread of all weather-related natural disasters.

• Why are we interested in this topic?

Flooding can be deadly harmful to human beings' lives and properties. For example, flash floods are the most dangerous kind of floods, because they combine the destructive power of a flood with incredible speed and unpredictability. Flash floods occur when excessive water fills normally dry creeks or river beds along with currently flowing creeks and rivers, causing rapid rises of water in a short amount of time.
Floods are hard to predict. They can happen with little or no warning. So, we are wondering if we can use some statistical tools to predict this kind of flooding, and detect the factors that might be involved in a flooding event.

• Why we finally chose New York City as our target?

Flooding occurs in every U.S. state and territory, and is a threat experienced anywhere in the world that receives rain. In the U.S. floods kill more people each year than tornadoes, hurricanes or lightning.

Historically, some communities throughout New York City have been prone to flooding. Sections of Queens, Staten Island, the Bronx and Brooklyn, for instance, have periodically faced this problem. In recent years, however, flooding has occurred more frequently than in the past, affecting a broader range of communities than ever. Local topography, including lengthy river and ocean coastlines, dense urban development patterns, the

capacity of our aging sewer system and increasingly extreme weather are some of the biggest causes.

We are motivated by keeping receiving flooding alerts after coming to New York City. After some researches, we found floods is a very common extreme weather and did cause lots of damage on both lives and properties. So, we want to have a deep insight into all extreme weathers in New York area.

• Objective

Our project means to get a clear understanding about how flood is affecting our daily lives in NYC. And we did it in three sections:
1. At the very beginning, we focus on changing duration of extreme weather in the New York State because of the flood.
2. We then try to use the weather data to predict when the flood would occur in NYC.
3. Focus more on flooding damages. The dataset will be used in this part.

Section II Analysis of Length of the Weather Events

• Material and Method

– Data Source

The data set is collected from NOAA National Centers for Environmental Information from 1950 to 2015. It contains records of weather events from all over the states in America. According to the objective, we need to data in New York City. However, the data size will be too size, if filted by New York City. Thus, we extended our data set to include events happened in the whole New York State. Over 50 events are included in the such as heavy snow, blizzard, flash flood, etc. And we chose only16 typical types of events associated with our objective from the raw data. We have three types of flood, flash flood, flood, coastal flood. Different types of flood have different sources. The Flash flood usually shows rapidly during or after heavy rains by severe events like thunderstorm, hurricane meltwater, etc[1]. And coastal flood occurs when the lower-lying land is flooded by the seawater. Usually, it happens with strong winds that causes storm surges[2]. Thus, we retain the events that will cause heavy water fall or strong wind. The time of the data was also truncated. Due to the limitation of the recording technique, from 1950 to 1954, we only have Tornado. Only Tornado, Thunderstorm Wind and Hail were found from 1955 to 1996. To keep the completeness of the data, we included logs after 1996.

Variables in the dataset consist of general variables, like begin time, end time and longitude, and specific variables for different events, like Fuji Scale that describes the wind speed of tornado and tornado length. Our goal in this section is to give a general idea of how the duration of these events vary along with time. So, we used 3 variables: begin time, end time, event type.

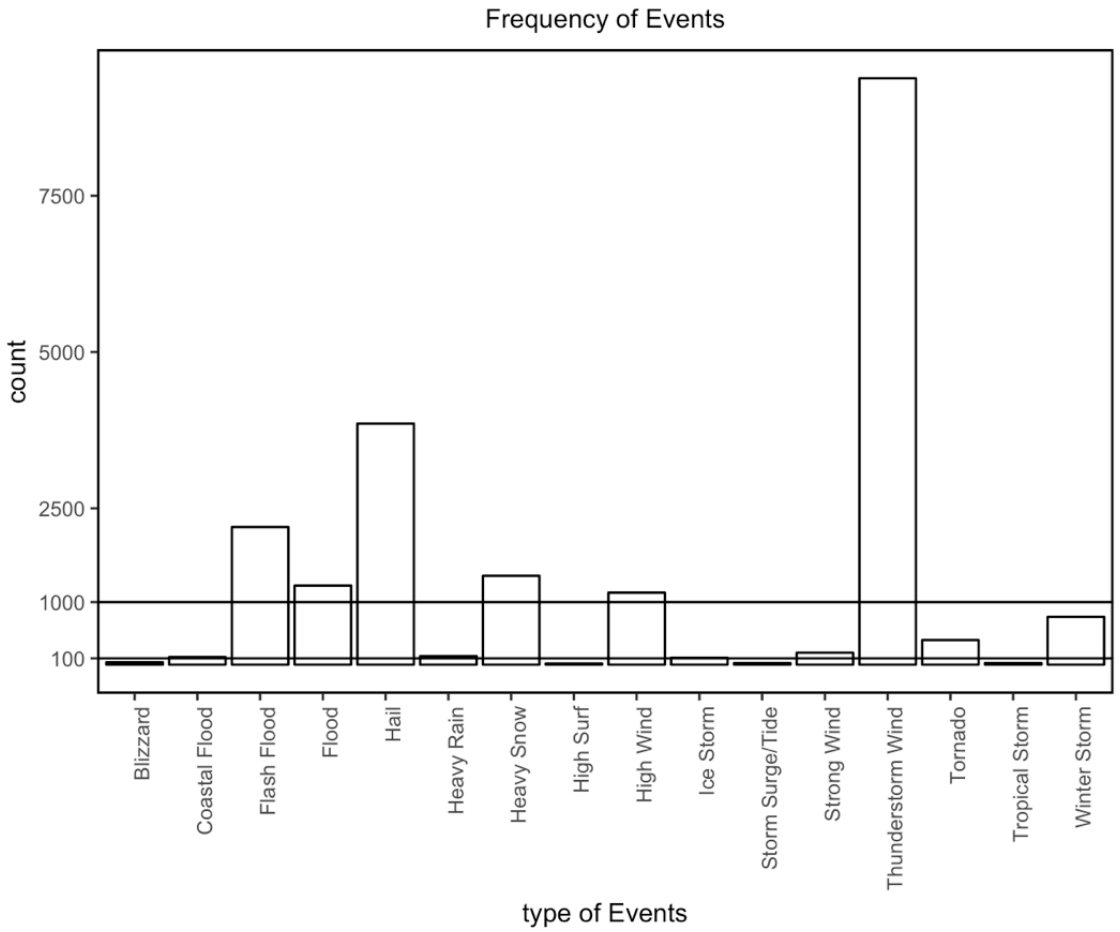


Figure 1. Frequency of the 16 events. The x-axis is the 16 events. And the y-axis is the count of appearing times of the 16 events.

If you count the frequencies of different types of events, a large variation will be found. Showed in figure 1. Only 25 tropical storms were recorded from 1996 to 2015, while thunderstorm wind happened 9379 times. The size of the data set usually determined the analytic methods thus we do only part. It is reasonable to differently treat the event type by the frequency. We divide the final data set into three groups, shown in table 1, large size data, middle size data and small size data. The large size data contains such weathers, like hail, high wind, thunderstorm wind, flash flood, heavy snow and flood, since they appeared more than 1000 times. The middle size data contains such weathers, like coastal flood and heavy rain, since their frequencies ranging from 100 to 1000. And the rest with less than 100 frequency are put into small data set. In this group, the max frequency is the one of blizzard, 37. That means there are less than 2 of the events happened in one year on average. To ensure accuracy, we decided not to analyze this group in the following sections.

Table 1 Group of Data Set

Group	Types
Large group	hail, high wind, thunderstorm wind, flash flood, heavy snow, flood
Medium Group	coastal flood, heavy rain, ice storm, strong wind, tornado, winter storm
Small group	blizzard, High surf, storm surge/ tide, tropical storm

– **Analytical plans**

Though out the analysis in this section, we assume the duration of each year is exponentially distributed, which is reasonable because in survival analysis exponential distribution is fundamental. And, we assume that the durations within one year is i.i.d distributed. They are also independent among the twenty years.

For the large size data set:

For this group, we set a model based on Bayes analysis[3]**.** For simplicity, we postulate the prior has a gamma distribution and so the posterior has a gamma distribution, by conjugacy as the exponential distribution.

To investigate the changes of the distribution of duration along the time, we set the null hypothesis as the mean of duration of one event in 20 years are the same. We try to do this in the following steps for each event:

1. Use Bootstrap and Bayes model to estimate the mean and variance of the prior distribution,
2. Use Monte Carlo Method to build the 95% confidence interval of the estimator, with 10000 times simulation,
3. Test if there exists a year that the interval does not overlap with any other intervals.

We planned to do the first step in this way. If our null hypothesis is right, it is reasonable to set the prior with the parameter of the whole sample. The parameters were solved by the following equation:

$$\frac{\alpha}{\beta} = \frac{1}{\bar{x}}$$

$$\frac{\alpha}{\beta^2} = \mathrm{var}\left(\frac{1}{\bar{x}}\right)$$

, where α and β is the parameters of the prior, and $\bar{x}$ is the sample mean, if Null is true. $\bar{\theta}$ is easy to get, while the variance of θ was not. Thus, we borrowed bootstrap. We repeated this procedure for 6 weathers in the data set.

For the middle size data set:

We consider of our task in this data set as a 20-group sample test. Permutation test[4][5] is implemented. We considered using permutation because, the amount of data is not large and the durations among the year were exchangeable.

We planned to test whether the exponential parameter is the same from 1996 to 2015, just like the one we had for large group of data. The trick here is that we set the null

hypothesis to be that the first moment is the same throughout the twenty years, because for exponential distribution the same parameter is equivalent to the same first moment.

The exponential assumption prohibits us from using ANOVA. So, we use Kruskal Wallis statistic as the kernel in the permutation. If the p-value is small, we rejected the null.

• Results

For large size data:
We select High Wind as an example of our result in this data set, the rest of them will be presented in Appendix A.
High Wind:

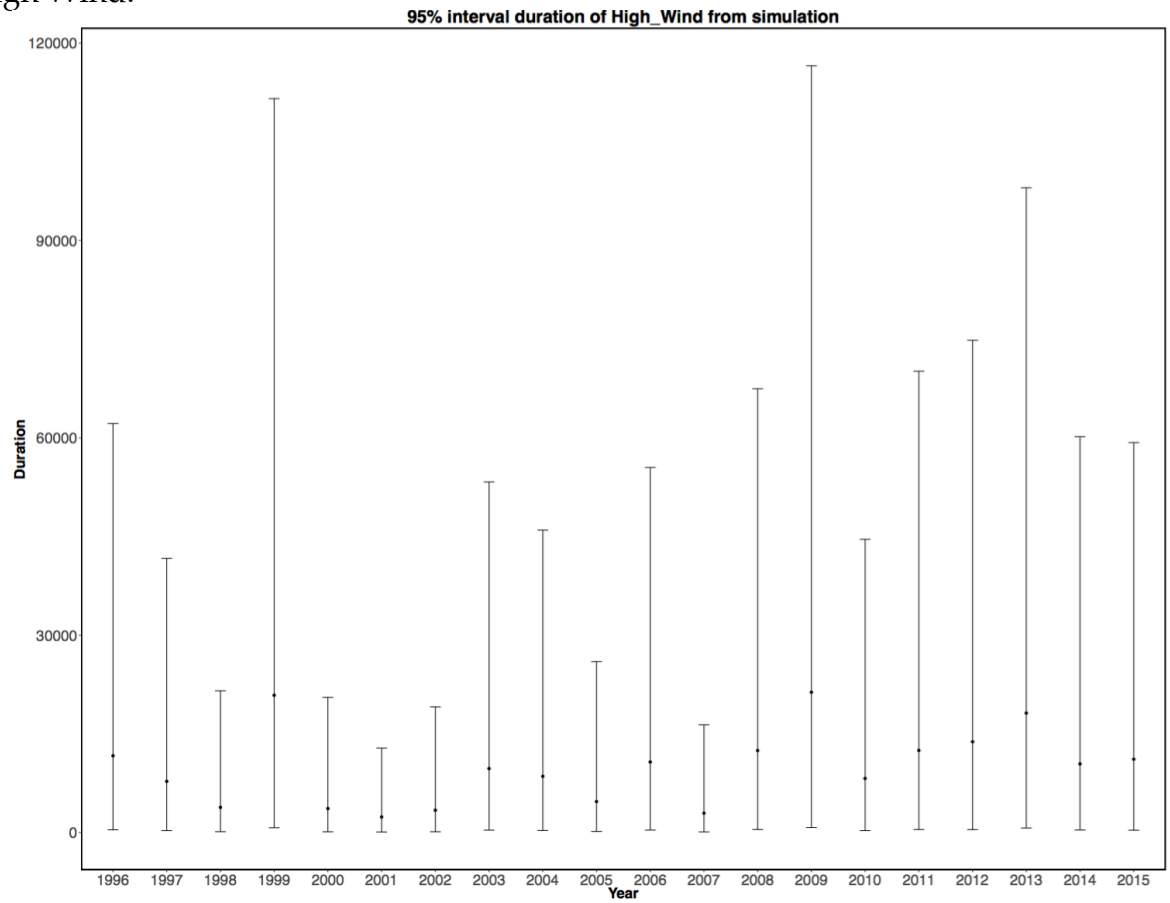


Figure 2. 95% interval of duration of high wind from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Figure 2 shows that overlapping happens for each pair of year. And the duration varies over time.

For middle size data set:
We select heavy rain as an example of our result in this data set. The result gave us a p-value of zero. This result is possible, if the true p-value is extremely small, 10000 times permutation is not enough to reach an extreme case. And if we perform KW test on the

raw sample, we will get a p-value around 2*10^-16. And for each of the six weathers in this data set, the p-value is 0. Those suggest us to reject the null hypothesis.

To give a more intuitive understanding, we make a plot, with the estimator of theta. And the Monte Carlo simulation of 95% confidence interval, assuming the estimator of θ is true for the sample for each year. If two of the interval do not overlap, it will be reasonable to believe that the sample is less likely collected from the null hypothesis.

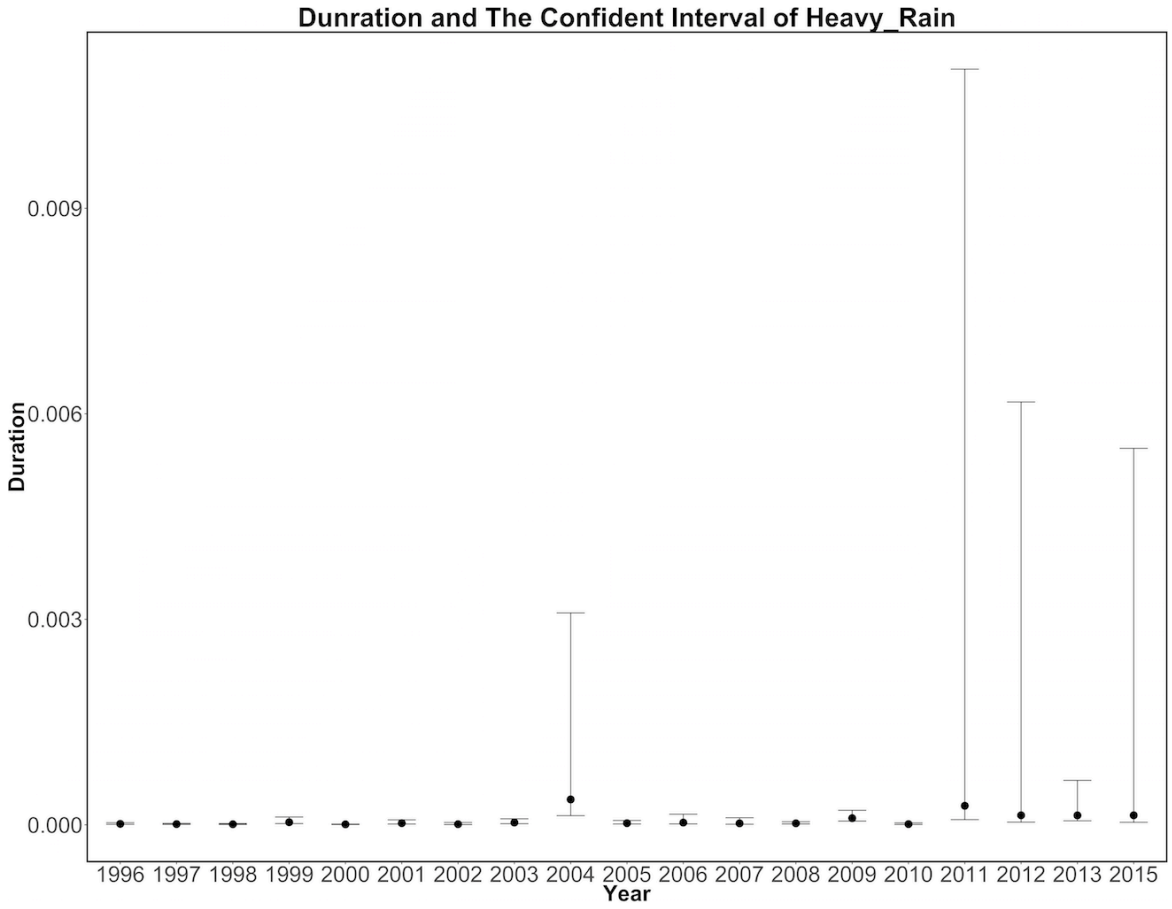


Figure 3. 95% interval of duration of heavy rain from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Figure 3 shows that for 1998 and 2004, the intervals do not overlap. It validates the permutation results. The graphs for other plots are in the Appendix A.

• Sensitivity analysis to validate the assumptions

For large size data, we assume that the duration of each event obeys exponential distribution.

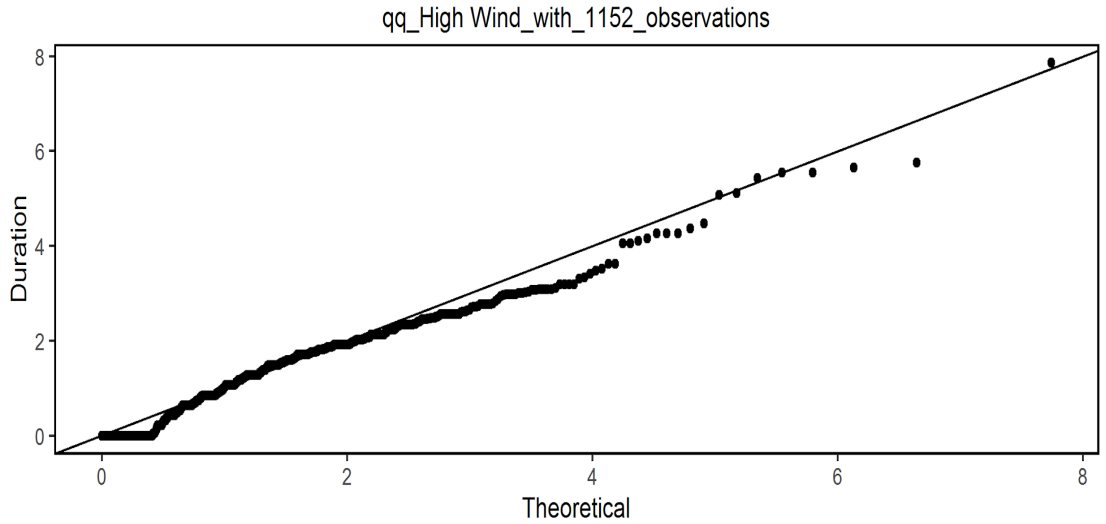


Figure 4. qqplot of testing event the distribution of high wind is exponential distribution. The x-axis is the theoretical value of qqplot. And the y-axis is the duration of the event.

We see from the figure 4, the assumption is valid. The graphs for other plots are in the appendix.

Since for middle size data, we use permutation test, the part doesn't have assumptions.

• Conclusions

– **Summary of major findings**

For large size data, we can see that for each of the six events, the duration in each year from 1996 to 2015 really varies, but we can't figure out a trend (We don't know whether the duration will go up or down.)
For middle size data, we should reject the null hypothesis and conclude that for each one of the six weathers, the mean of the duration of each year is not the same. However, we still can't figure out a trend.

– **Limitation of studies**

For the Bayes model, we can't test whether the null hypothesis that for each one of the weathers, the mean of the duration of each year is the same. And since we use conjugacy, the prior we used may not be the real distribution of that.

And the nature of the exponential limits the accuracy of our estimation. Fisher information[6] can show this. If we treat the mean of sample as the parameter, then

$$I\left(\theta = \frac{1}{\lambda}\right) = \frac{n}{\theta^2}$$

For large θ, which is true in this case, we get low Fisher information and incredibly volatility. And this will bring large interval. So, we should try to use λ as the parameter during the analysis in the future.

The number of years are relatively small for trend analysis like time series. One possible solution is do more detail analyze the trend weather condition not only extreme weather events.

Section III: Flooding Prediction

• Background

Because we want to predict the occurrence of flood, we need to get weather data in New York as predictors. We planned to use some machine learning methods containing stochastic process, such as Hidden Markov Model, but due to the limit of the occurrence times of flood, it was very hard to use those methods.

So, we turned to use some other supervised machine learning algorithms. We labeled every day (excluding some days do not have weather data) with 0(means no flood) or 1(means flood occurrence), and treated this label as target variable.

• Material and Method

– **Data Source**

The weather data of New York City [7] was obtained in website Weather Underground. The data set we got was from 2000 from 2016. We have variables as below:

1. Max Temperature, 2. Mean Temperature, 3. Min Temperature, 4. Max Dew Point, 5. 6. Min Dew Point, 7. Max Humidity, 8. Mean Humidity, 9. Min Humidity, 10. Max Sea Level, 11. Mean Sea Level, 12. Min Sea Level, 13. Max Visibility, 14. Mean Visibility, 15. Min Visibility, 16. Max Wind Speed, 17. Mean Wind Speed, 18. Max Gust Speed, 19. Precipitation, 20. Cloud Cover, 21. Wind Direction Degrees.

The source data from Weather Underground is provided as csv format. We imported it into R and did some raw data processing. Because most of the NA values were from Precipitation, we just directly set the NA values to 0.

Other part of data came from NOAA (National Oceanic and Atmospheric Administration) website [8], it contained all extreme weather happened in New York City from 2000-2016. The extreme weathers include Tornado, Thunderstorm Wind, Hail, Blizzard, Coastal Flood, Flash Flood, Flood, Heavy Rain, Heavy Show, High Surf, High Wind, Ice Storm, Storm Surge/Tide, Strong Wind, Thunderstorm Wind, Tornado, Tropical Storm, Winter Storm.

We only used the Coastal Flood, Flash Flood, Flood data in this part. In order to predict weather, it will occur flood, we combine the weather type data with lots of weather conditions, which is the weather data in website Weather Underground. Then we delete all NA lines and get relatively clean 5496 lines of data, each line represents one day's data.

For each day in data set, we have an unique label 0 or 1 to show if there is flood in this day. And it has 546 days with flood in total at this time period.

- **Methodology**

Method I:

At first, we used Neural Network to predict the occurrence of flood. We used Multilayer Perceptron to do the predictions. [9]

Data Processing:

1. Different variables in this data set have different scales. For example, most value of Precipitation is ranging from 0 to 1 and most value of Humidity is ranging from 30 to 100. So, we need to scale the data set first to set the mean of every variable to be 0 and the standard deviation to be 1.

2. Also, because the raw data set is consecutive, we shuffled the data set before training the model to eliminate time factors.

3. In order to avoid overfitting, as what we usually do in Deep Learning, we divided the data set into three parts, training set, validation set and test set. The whole data set contains 5496 observations. After setting, the training set contained 3496 observations, the validation contained 1000 observations and the test data set contained 1000 observations.

4. Because of the characteristic of neural network, we did not need to do feature selections. We can put all features into our models.

Tools and language:

In order to train the model more efficiently, we used Theano, which is very popular in deep learning. All of the code is written in python (excluding some raw data processing part). We ran the code in Amazon Web Service with GPU.

Neural Network Structure:

We tried different numbers of layers and hidden units in each layer. Finally, we found 5 layers and 500 hidden units in each layer resulted in the best result. As for the activation function, we used tanh function. Finally, we put the logistic regression layer to get the classification.

Cost function:

Because only 10% days from 2000 to 2016 have flood, if we classify all observations to be label 0, we can also get the accuracy about 90%, but actually we do very poorly in finding days labeled with flood.

In order to do better in finding the label 1, we need to make the cost of wrongly labeling 1 to 0 higher than the cost of wrongly labeling 0 to 1. In this way, we can do better in classifying label 1, but will sacrifice some accuracy in classifying label 0.

In my opinion, we can set the cost ratio according to different requirements. For example, if we need to pay more attention to the occurrence in order to alleviate the bad influence of flood, we can set cost ratio to be 3, 4 or even higher. On the other hand, if we do not need to care that much about the occurrence of flood, we can set the cost ratio to be a little bit lower.

For the cost that we used to train the model, we used negative log likelihood, which was similar to cross-entropy, of course, we combined it with the cost ratio. In our code, it was shown as below:

```
# y.shape[0] is (symbolically) the number of rows in y, i.e.,
# number of examples (call it n) in the minibatch
# T.arange(y.shape[0]) is a symbolic vector which will contain
# [0,1,2,... n-1] T.log(self.p_y_given_x) is a matrix of
# Log-Probabilities (call it LP) with one row per example and
# one column per class LP[T.arange(y.shape[0]),y] is a vector
# v containing [LP[0,y[0]], LP[1,y[1]], LP[2,y[2]], ...,
# LP[n-1,y[n-1]]] and T.mean(LP[T.arange(y.shape[0]),y]) is
# the mean (across minibatch examples) of the elements in v,
# i.e., the mean log-likelihood across the minibatch.
return -T.mean(T.log(self.p_y_given_x)[T.arange(y.shape[0]), y] * y_cost)
```

For the accuracy in validation and testing part, we just used the error rates, and they were also combined with error rates. The code is shown as below:

```
def errors_2(self, y, y_cost):
    """Return a float representing the number of errors in the minibatch
    over the total number of examples of the minibatch ; zero one
    loss over the size of the minibatch

    :type y: theano.tensor.TensorType
    :param y: corresponds to a vector that gives for each example the
              correct label
    """

    # check if y has same dimension of y_pred
    if y.ndim != self.y_pred.ndim:
        raise TypeError(
            'y should have the same shape as self.y_pred',
            ('y', y.type, 'y_pred', self.y_pred.type)
        )
    # check if y is of the correct datatype
    if y.dtype.startswith('int'):
        return T.mean(T.neq(self.y_pred, y) * y_cost)
    else:
        raise NotImplementedError()
```

Optimization and Regularization:

For optimization, we did not use complex method, we just used normal gradient descent to try to minimize the cost function.

For regularization, we used L1 and L2 regularization term in cost function. Also, we used early stopping to do the regularization, which was also very popular in deep learning. We set the patience to be 10000, the patience increase to be, and the improvement threshold to be 0.995.

Parameter setting:

After some experiments, we got the best result using the following parameter settings: (1) learning rate was 0.05, (2) L1_reg was 0.000, (3) L2_reg was 0.0000005, (4) batch size was 1000.

Method II:

We also tried several different supervised methods in model building, such as gradient boosting machine, random forest, adaboost and logistic regression[10].

But each of them performs not good enough at beginning. We can see these bad performances from ROC curves (see Appendix B). Then we dug more at the data set, and we found only around 10% data have label 1, others are all with label 0. In this situation, if we predicted all data as label 0, we would a predict accuracy more than 90%.

So the imbalanced data distribution makes us increase the cost of predicting 0, one of the efficient ways to do this is to adjust the initial weights of observations. After tuning with train data, we finally used 3:1 weights in training data with label 1 and label 0.

Then, all of the four methods above improved their performances in predicting 1 with a little fall back in predicting 0. In order to take advantage of all these four methods, we combined these four methods together and gave our final prediction based on the predict results above.

Here is our algorithm: if one of the methods above gave prediction of one test data is 1, then we predict this test data as label 1. In other words, this method keeps all 1 we predicted in former four methods.

For both above methods, in order to make the model more robust and avoid overfitting, we used cross validation to train the models.

• Results

For Method I, with the cost ratio to be 3 and training the model for 2000 epochs, we got the final result as below:

```
print ("Whole accuracy is: ", (num_0_right + num_1_right) * 1.0 / (num_0 + num_1))
```

```
('test accuracy of label 0 is: ', 0.8092105263157895)
('test accuracy of label 1 is: ', 0.5)
('Whole accuracy is: ', 0.782)
```

In this experiment, in all observations labeled with 1, we got the accuracy of 50%. In all observations labeled with 0, we got the accuracy of 80.9%. In the whole dataset, we got the accuracy of 78.2%.

And for Method II, we have 87% predict accuracy as our final result. Actually, as we mentioned before, the predict accuracy of predicting 0 for all test data is above 90%. But this biased prediction definitely cannot work in real circumstances. So we pay more attention to the TPR (true positive rate), and we have around 53% TPR. Although we can do more adaptive steps to make TPR even more better, the FPR (false positive rate) may drop significantly, which is also not the ideal case. All in all, we have this relatively balanced result for predicting 1 and predicting 0.

Table 2 Result Table for Model II

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
TPR	0.6	0.63963964	0.571428571	0.582089552	0.5
Accauracy	0.862602366	0.866242038	0.868971793	0.872611465	0.868061874

• Sensitive Test

For this method, we randomly picked 80% data as train data set and 20% of data as test data set. We tried 100 times and used their average as our final prediction. We performed quantile analysis for these results, the results were pretty converged and the variance for these results was also at a small scale. So as a result, this model performed well in this sensitive test. Even we change a part of data in ± 1%, the results was stable and also concentrated.

Table 3 Result Table of Sensitive Test

	-1%	-0.50%	0	0.50%	1%
TPR	0.582524272	0.533898305	0.471698113	0.494845361	0.605504587
Accauracy	0.847133758	0.857142857	0.84622384	0.866242038	0.856232939

• Conclusions

– Summary of major findings

On average, we can successfully classify about 50% of the floods. We can also adjust the cost ratio to get different results.

Due to the limitation of the numbers of occurrence of floods and the limitation of size of the whole data set, this result is not bad.

– **Limitation of studies and Improvement:**

Since the data set is not big enough, and flood situations are rare among all observations, the predictions which we made performed not good enough. Although we adjusted our methods based on the data base, it still has a tendency to predict "No flood" for every situation. Also, we treated data of everyday as independent observations, so the correlations and connections between close dates were not used so far.

Based on these drawbacks, the first thing we think is to find more abundant data set with more details. And also, some important factors related to flood may be ignored. In a word, we need to enlarge our data set.

And the time series model can be considered for next step. Due to the inaccuracy of our former move in predicting flood, the model focusing on correlation between close dates may help.

Section IV : Flooding in New York - Flooding Damage on Properties and Lives

• Background

For all the extreme weathers, flooding is the most common one in New York area, including 5 boroughs – Manhattan, Queens, Kings (Brooklyn), Richmond, Bronx. As our final goal is to have a deep insight into the extreme weathers' influence on humans' properties and lives. We use latest 10 years' flooding events data in New York area to do some explorations on this influence.

• Material and Method

– **Dataset Description**

The dataset used here is all the flooding events from 2006 to 2016, and it has 37 variables, they are:
event_id,cz_name_str,begin_location,begin_date,begin_time,event_type,magnitude,tor_f_scale,deaths_direct,injuries_direct,damage_property_num,damage_crops_num,state_abbr,cz_timezone,magnitude_type,episode_id,cz_type,cz_fips,wfo,injuries_indirect,deaths_indirect,source,flood_cause,tor_length,tor_width,begin_range,begin_azimuth,end_range,end_azimuth,end_location,begin_lat,begin_lon,end_lat,end_lon,absolute_rownumber.

They are all very clear to understand by names. Like the begin_location, which means the location where the event begins. However, there are still some variables not so intuitive. Here I made some explains:

About event-id and episode-id: an episode can be several event, and event is a specific storm episode, and only event-id is primary key in this database.
co. and zone: co means county and zone means zone.

- **Methodology overview**

We first will use some visualization plots to show the information contained in this dataset, and have a quick view of the flooding events situations in New York area. By these plots, we can actually have a quick familiarity to the flooding events in different areas in New York, and can get some characteristics of different areas.

And then we will implement some tests to detect the relationships among those variables, like whether there is some connection between flooding events frequency and area. By these exploratory tests, we can find some important information, like whether there are some trends or rules about the flooding event in New York 5 boroughs.

Then we will consider about building a predictive model. But our goal is not to do prediction, because lots of information contained in the dataset cannot be known in advance. Like the begin location and end location, we can only know them after flooding event happened. So this kind of predictive model is more like to help us to understand the flooding events. It will act as a kind of aggregated detector.

4.1 Overview of flooding events in New York Area

Symbols Description:
The red point stands for where the event began, and the blue point stands for where it ended. Pink line is the path of how event moved. Damage-cause or death-cause are labels will be placed nearby the point, where there causes corps or property damage, or indirect or direct death or injury. The size of the points stands for the range of that event, if the event range is larger, then the point size will be bigger.

Bronx:

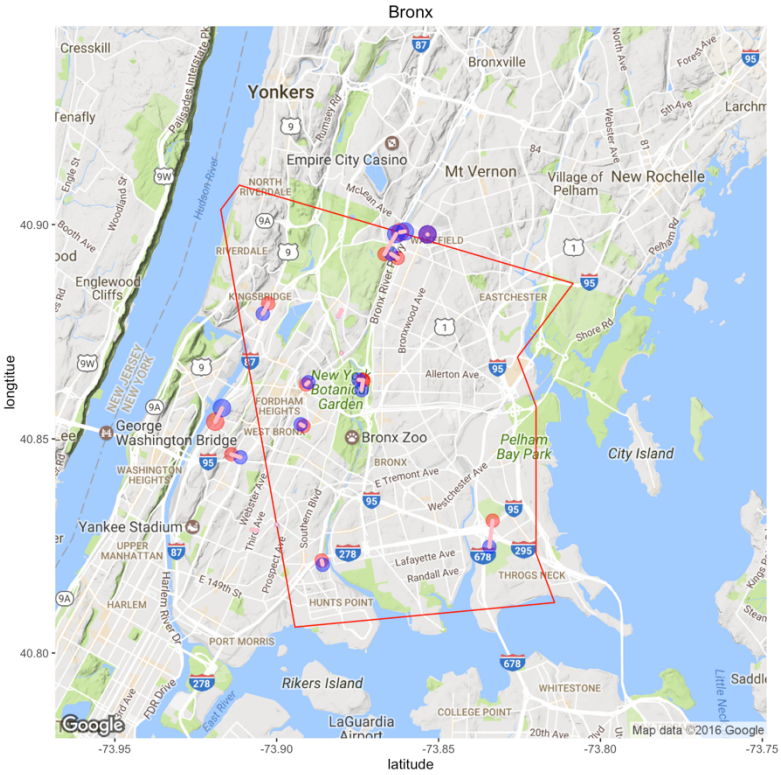


Figure 5 Flood in Bronx

There is no large-scale event, I mean, all the path length is small. Because only the event in county have the records of its initial place and end place. So we can only say that events in Bronx county, there is no event involved with damage or death or injury.

Kings:

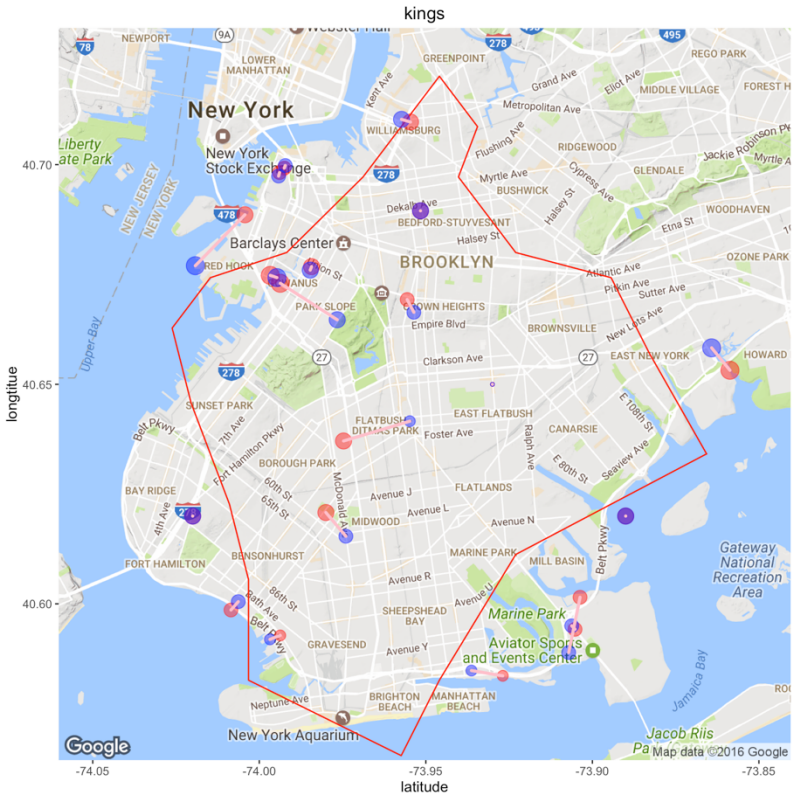


Figure 6 Flood in Kings

We can find an obvious characteristic of its events. They're concentrated in the marginal area. And the event scale is a little larger than Bronx.

Manhattan:
Some large-scale events, and all events have small range.

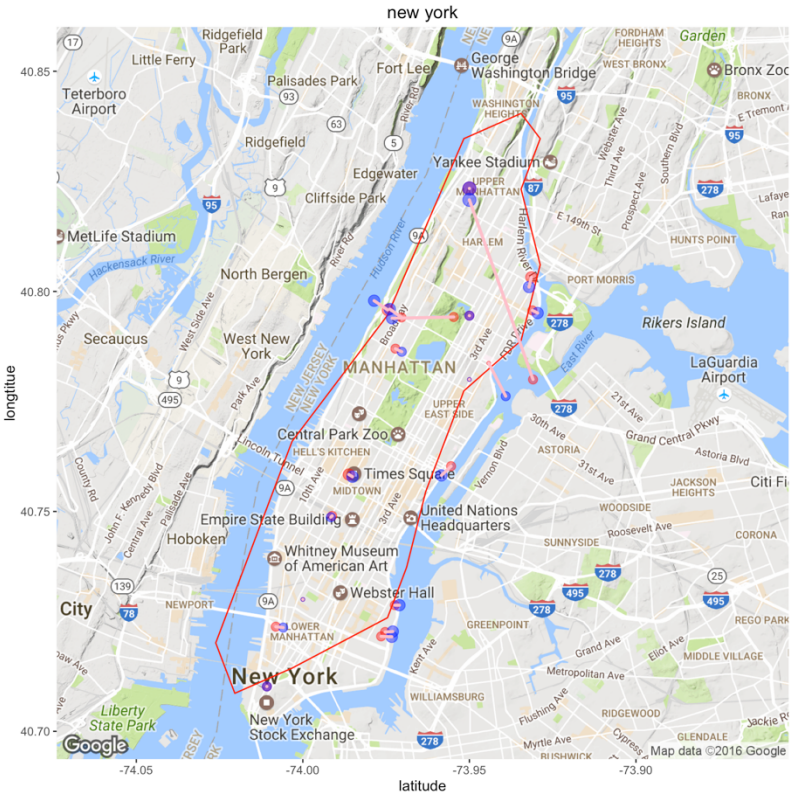


Figure 7 Flood in Manhattan

Queens:
Many large-scale events, and events are concentrated in the central area, and range of events varies much. And there is a damage-cause label, means there is an event accompanied with corps or property damage.

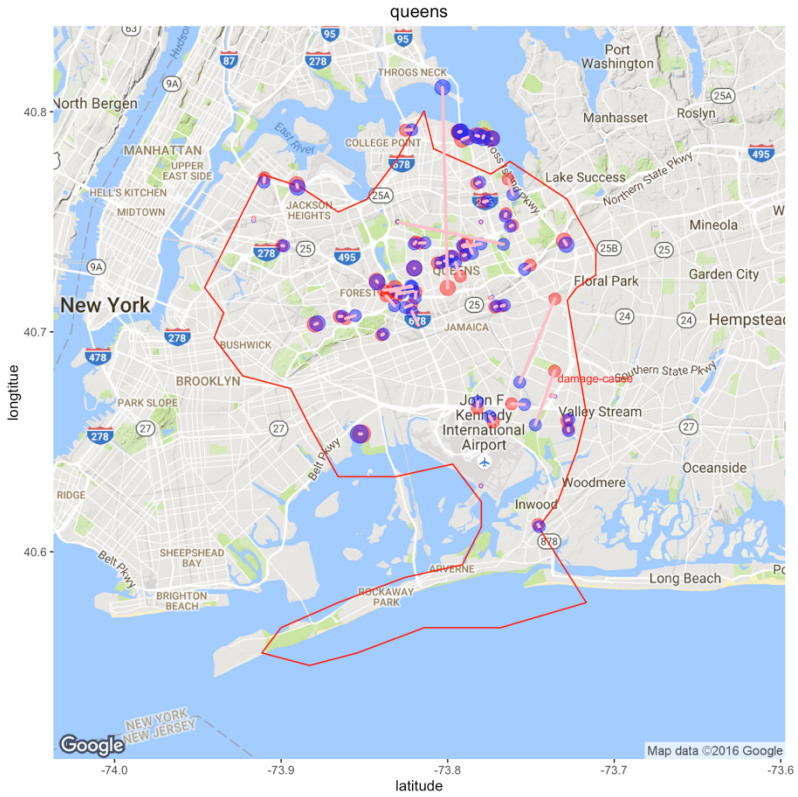


Figure 8 Flood in Queens

Richmond:
Just a few events, and median scale, and not so large range.

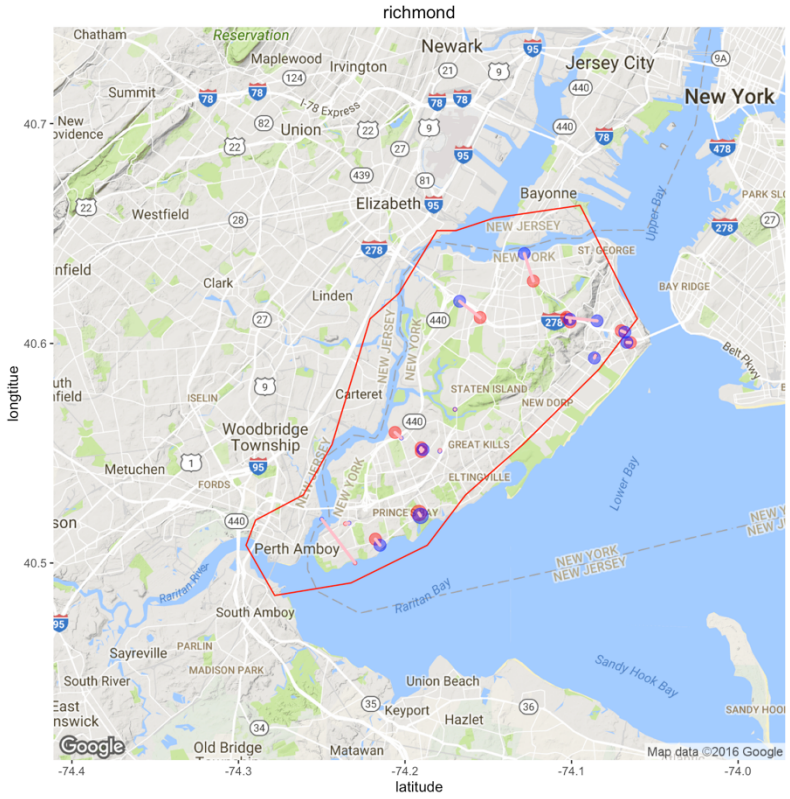


Figure 9 Flood in Richmond

After the individual plots, we can have a bigger picture to show all these event in New York area.

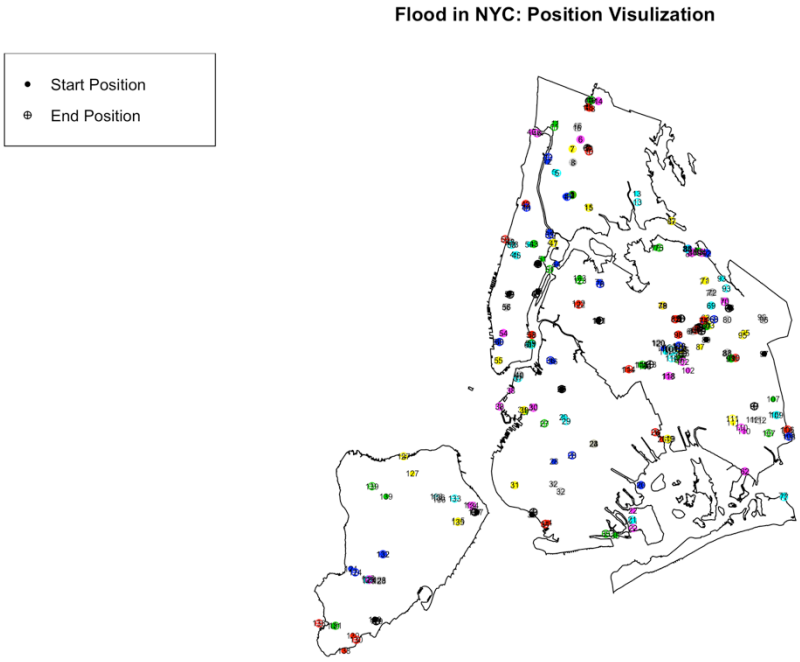


Figure 10 Flood in Whole New York City

Then we will use bar chart, and pie chart and also scatterplot to have a better understanding of the data. we should say that, all the data in the graphics are without NAs.

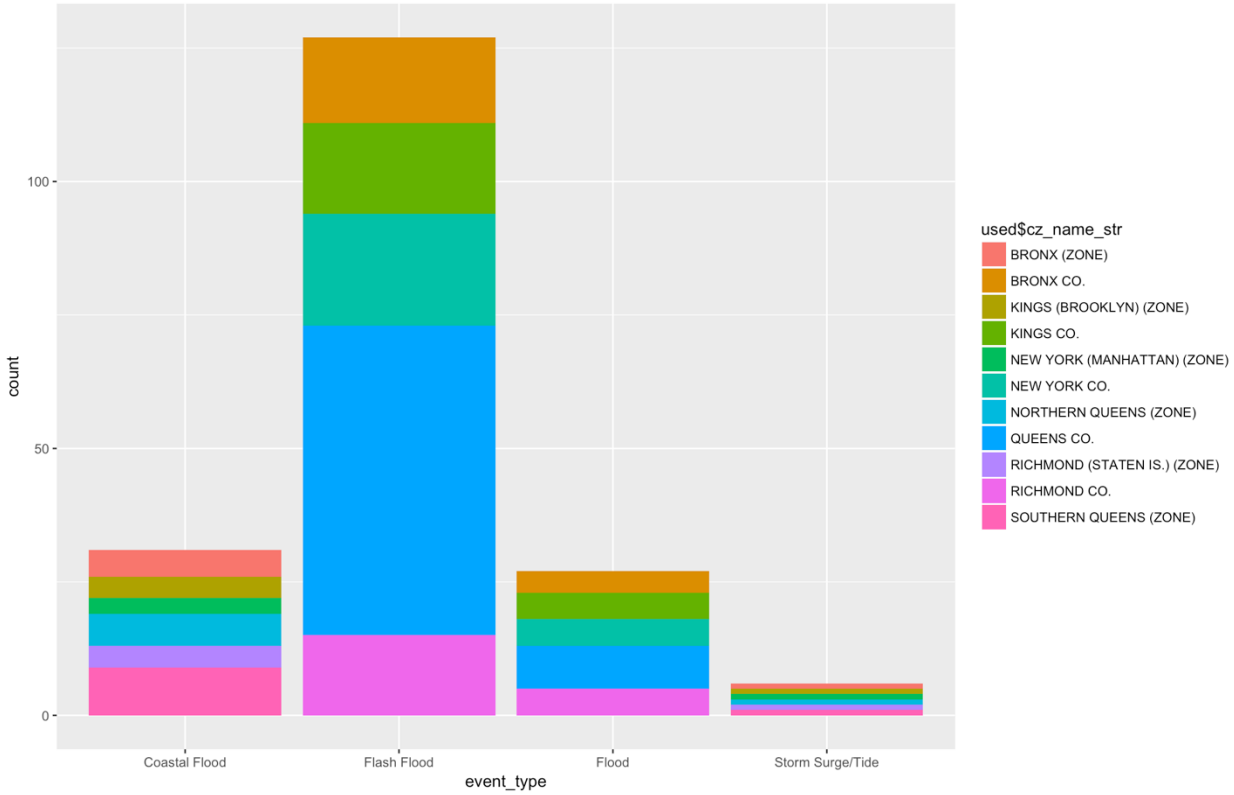


Figure 11 Aggregated Bar Chart

In the bar chart, we can see how the event are consisted of. Like flash flood, Queens county has the most flash flood events. And coastal flood, its southern queens.

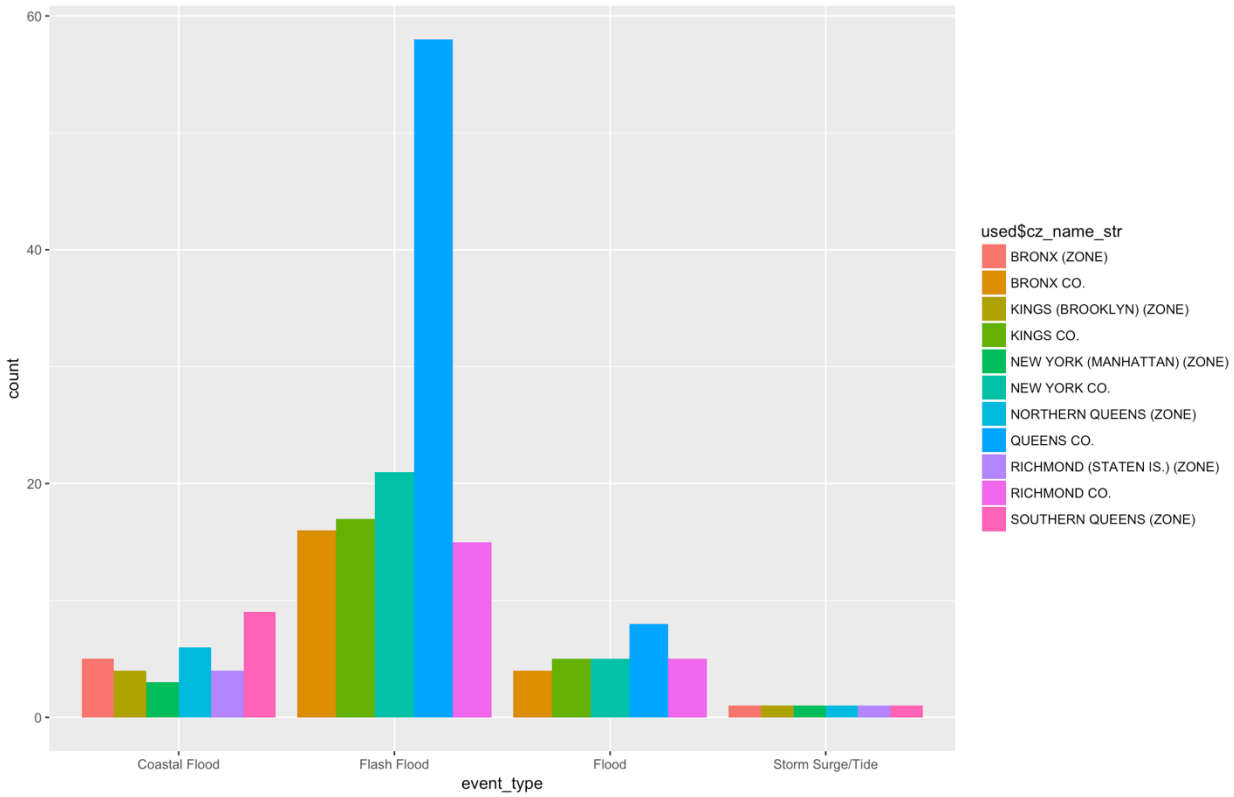


Figure 12 Seperated Bar Chart

In the separate bar chart, we can see how event are consisted of in each area. Like, in Queens county, the event types are flash flood and flood. This chart is good to do compare. For different areas, when compared New York county with Richmond county, New York seems have more flash flood events, and comparable number of flood events like Richmond.

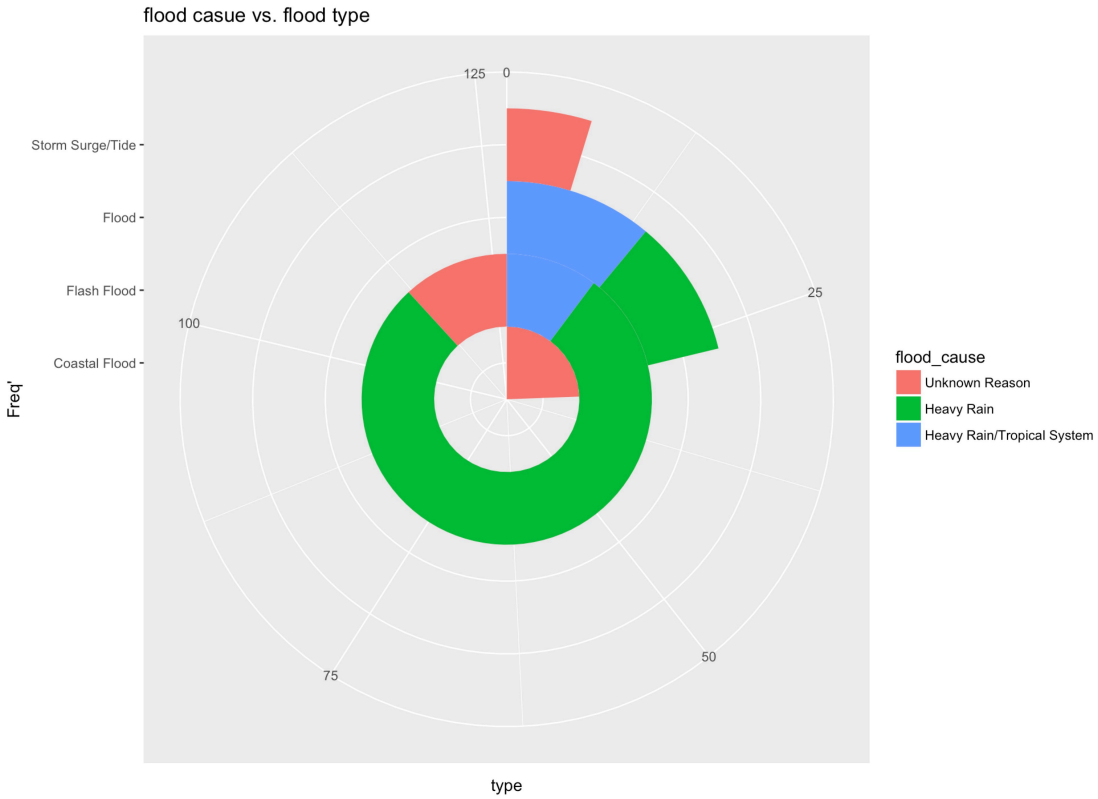


Figure 13 Pie Chart

In the pie chart, it's easy to see how reasons are consisted of in each event type. Like the flash flood, the second circle, its main cause is heavy rain.

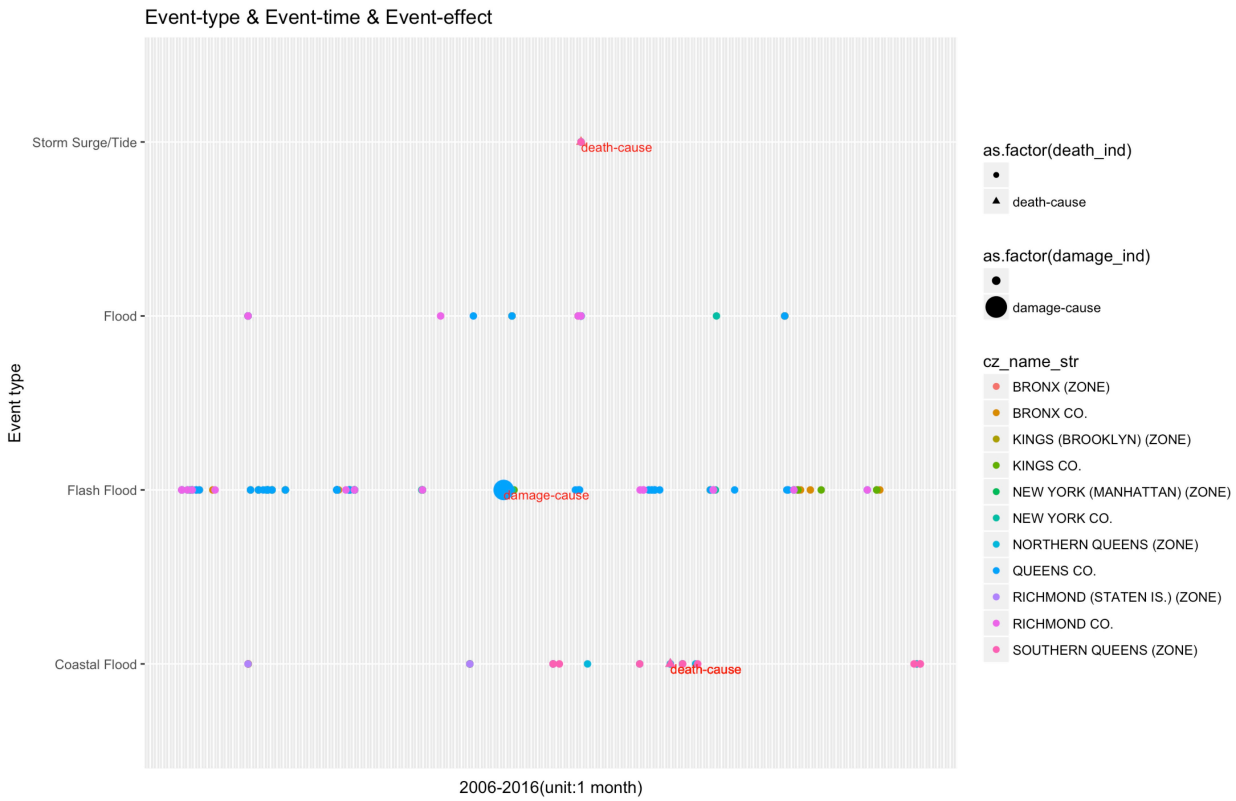


Figure 14 Scatter Plot of Event Time

In scatterplot, we can see some trends. Like what is the time in a year that an event is more likely to happen. We can see it in a large version. In Queens county, the flood events are concentrated among July to August.

4.2 Correlation Detection

Then we want to show some basic data analysis tools and how to use it in this dataset, here we mainly used Pearson test, because there are lots of categorical variables in this dataset.
For the theoretical part, Pearson test is a chi-square test, and the residual is the observation minus expectation divide its variance. Observation here means the observed number of that event, and expectation means if event-type and area are independent, how many events should be observed. So, the residuals can give us some clues about their correlation.

First, we want to see the relationship between event-type and area. This plot is called mosaic plot.

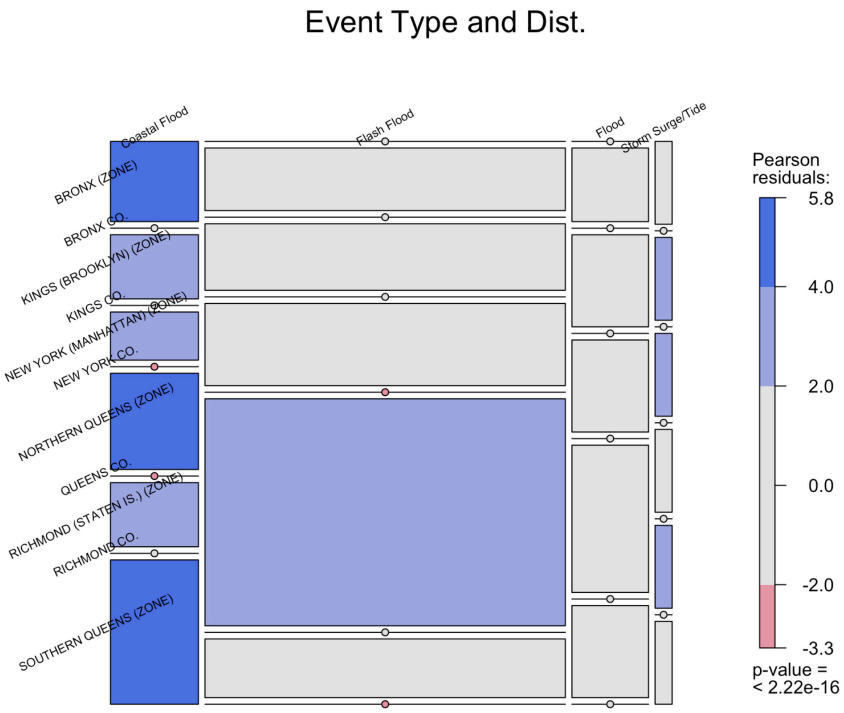


Figure 15 Correlation between Event Type and Event District

We can see from the residual plot that p-value is extremely small, which means there is overwhelming evidence show that event type and area are not independent.
Interpretation of Pearson residuals in the mosaic plot: Coastal Flood in Bronx, Queens (North and south) happened more than expectation, Flash flood in Queen happened more than expectation, and Storm Surge/Tide happened more than expectation in Kings, New York and Richmond.

In other words, Bronx, Queens tend to have more coastal flood, and Queens tends to have more flash flood, and Kings, New York and Richmond tends to have more Storm Surge/Tide.

Then comes the event type and reason.

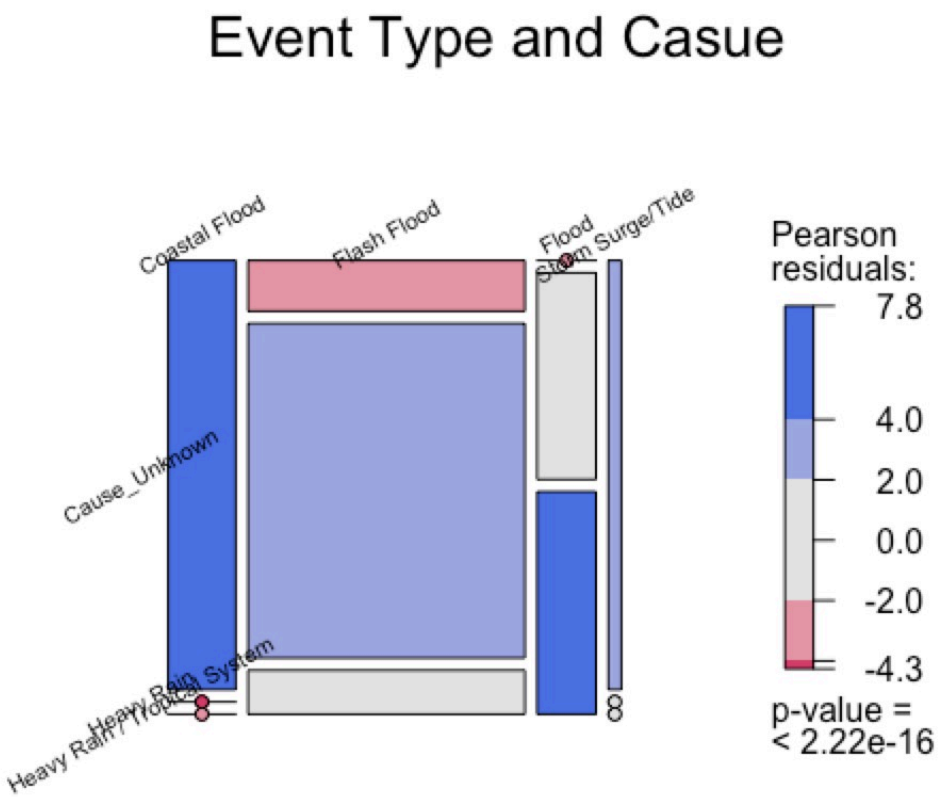


Figure 16 Correlation of Event Type and Event Cause

Same as above, but less useful because of the "Not Specified". Overwhelming evidence shows dependency between flood cause and event type.

Heavy Rain is moderately probably to cause flash flood, which means it is not a good indicator for the event type. But unspecified reason will have some influence on our result. So without unspecified reason, we can get a more useful idea.

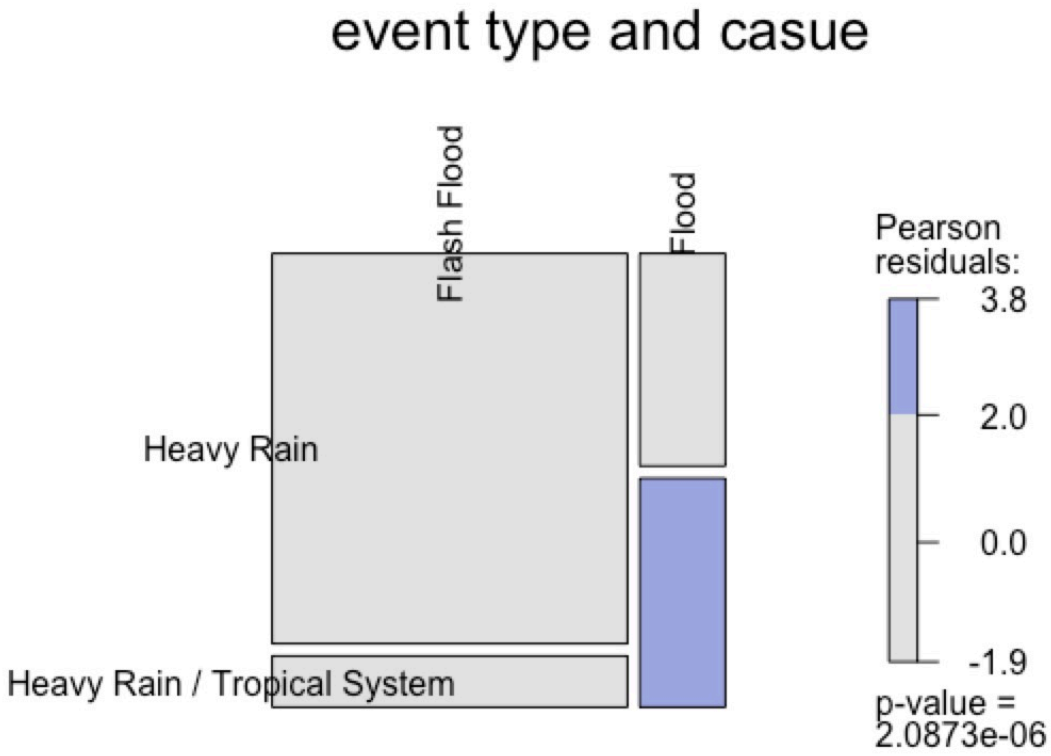


Figure 17 Correlation of Event Type and Event Cause

Here we get the idea that Heavy Rain/Tropical System is more probably to cause flood.

Then comes the event type, area and movement.

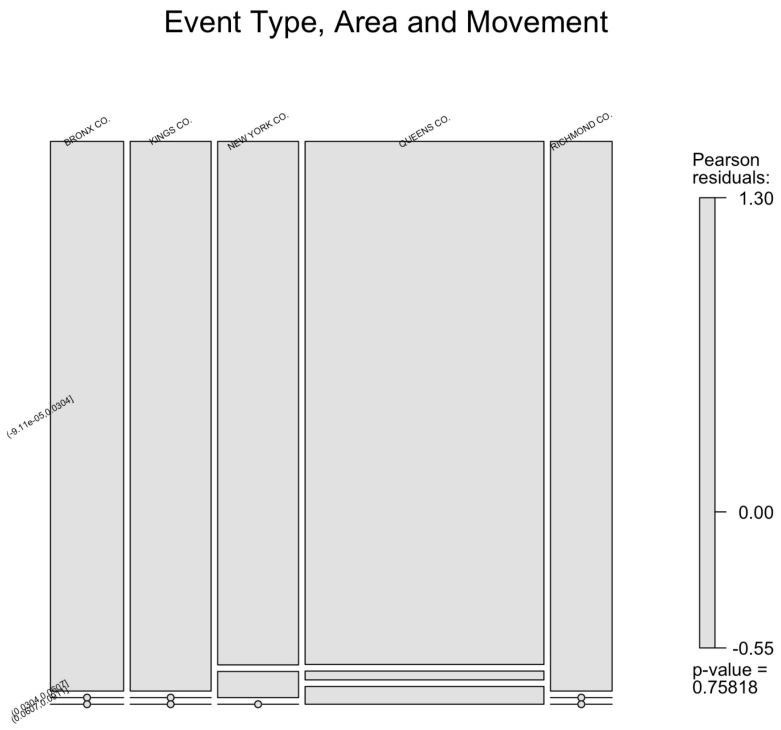


Figure 18 Correlation of Event Type Event Area and Movement

We can find that event range is somewhat random, independent of area and type.

Then comes the area and event range.

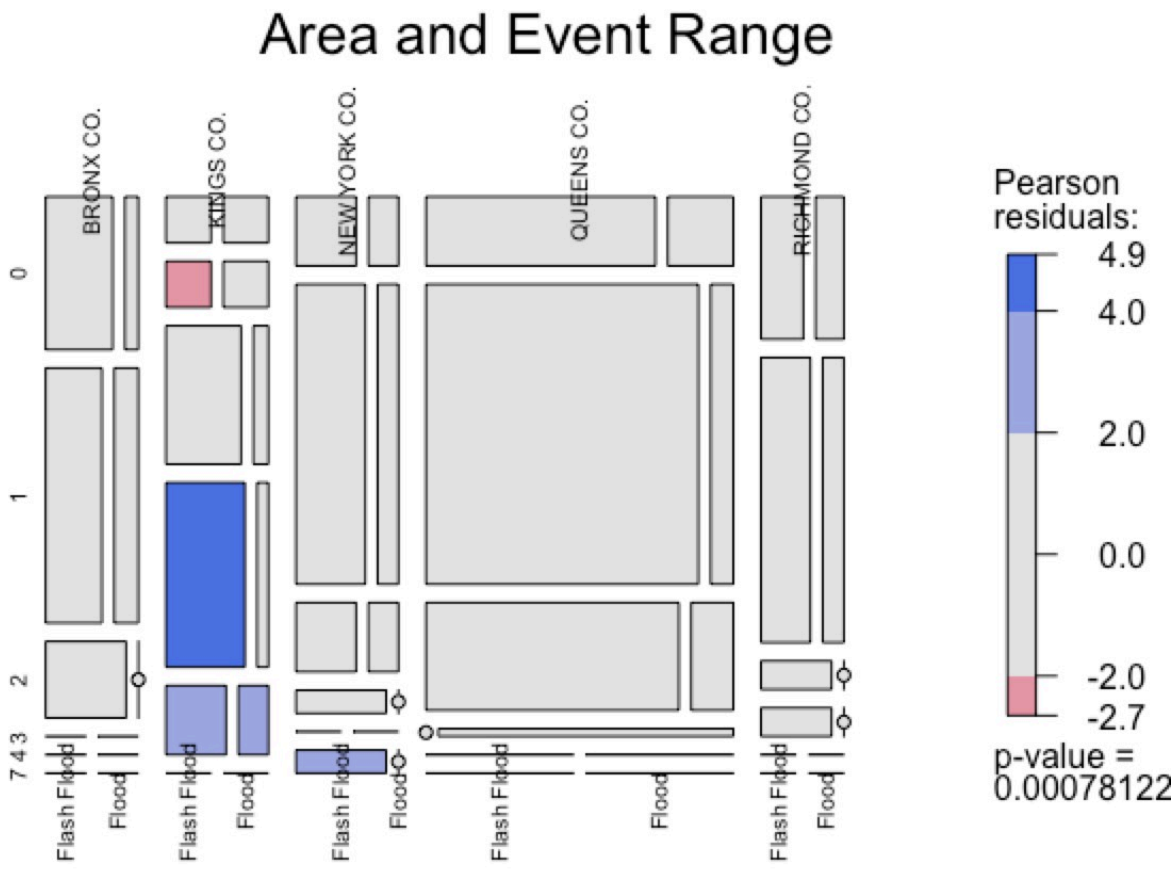


Figure 19 Correlation of Event Area and Event Range

.

It seems in Kings area, the range will tend to much higher than other areas. (Note that begin_range and end_range are highly correlated, the correlation coefficient is 0.913).

Then comes event area and event effect, event effect which means cause any damage, death or injury.

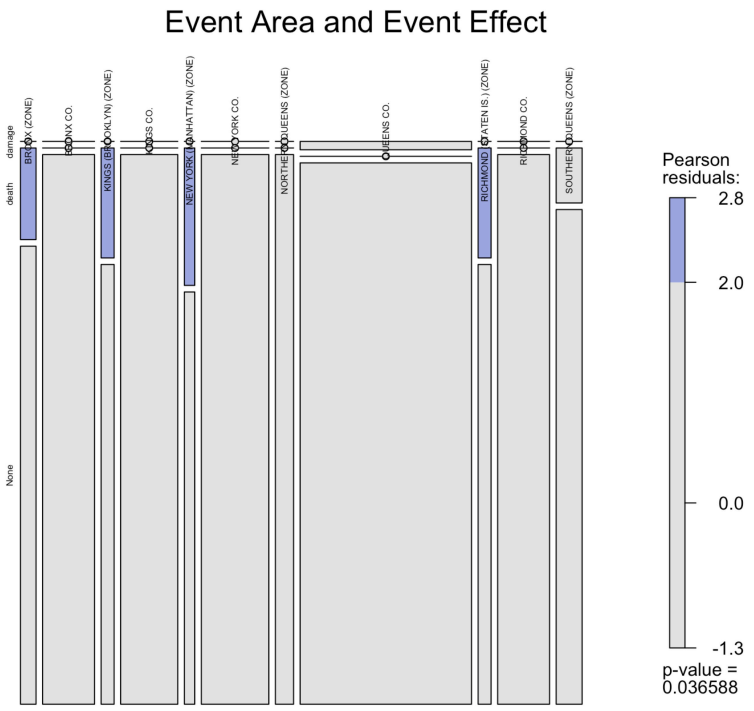


Figure 20 Correlation of Event Area and Event Effect

It seems in the zone of Bronx, Kings, New York, and Richmond, there tend to have bad effect caused by the event.

4.3 Build Predictive Model

Here, we will consider two frameworks to build a predictive model. One by machine learning methods, and the other is by Bayesian model.

Framework 1: Supervised learning to build predictive model

Model to choose: logistic regression, linear discriminant analysis, support vector machine, NN etc.

Step1:data preparation (deal with NAs, variable type transformation, variable aggregation)
Step2: choose input and output. Here is a plausible solution:
For outputs:
Take variables "deaths_direct" , "injuries_direct" , "injuries_indirect" , "deaths_indirect" as a factor and set “casue-death-or-injury” as 1 and “none-death-or-injury” as 0; take variables "damage_property_num", "damage_crops_num" as another factor, and set “cause-loss” as 1 and “none-loss” as 0.
For inputs:
Take variable "cz_name_str" as a factor, and “begin_range”, ”end_range”(kind of like an indicator for the power of the event, notice that they might highly correlated), and

absolute movement distance (Euclidean distance calculated by begin_lon,_lat and end_lon,_lat, another descriptive variable for the power of the event).
Step3: build model and use CV or simulation method (mainly bootstrap) to assess the model (prediction error).
But the result cannot be precise, because of huge problem with data! There are only 191 observations, and only 6 of them caused death or damage on properties. Output is not evenly distributed, which will cause calculation problem for parameter estimations.

So, we decide to use Bayesian model.

Framework 2 : Bayesian model

Bayesians focus on probabilities, so we believe data will not matter so much! The only difficulty is
toughness with how to choose prior and likelihood model.

Step1: variable selection
Output: effect (whether the event cause death, injury, crops damage or property damage)
Input: region(coded as 1:11 for different region); type(coded as 1:4 for different event type);
brange(begin range); erange(end range); move(Euclidean distance of the event movement distance)
Step2: prior and likelihood model
Here I choose to use logit link function (called logistic regression in GLM in frequentist field).

The model (likelihood):
effect~ Bernoulli(logit^{-1}($\beta_1 + \beta_2 * region + \beta_3 * type + \beta_4 * brange + \beta_5 * erange + \beta_6 * move$))

prior:
$(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) \propto 1$(informative prior: uniform)

The result is:

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
beta[1]	8254.25	5726.41	9778.99	123.95	1566.82	4048.96	9661.55	33004.03
beta[2]	0.41	0.18	0.99	-1.23	-0.23	0.31	0.92	3.04
beta[3]	-4132.72	2863.72	4889.52	-16506.94	-4837.75	-2029.66	-788.54	-65.77
beta[4]	1.65	0.47	2.82	-3.49	-0.21	1.65	3.37	8.01
beta[5]	-4.87	0.55	3.62	-13.53	-6.93	-4.61	-2.47	1.06
beta[6]	137.50	9.92	64.30	26.61	87.10	137.93	180.67	271.20

We can see from the above table 95% posterior interval of the parameters (2.5% column and 97.5% column). We can only get the result that length of movement of the event has a positive effect on the probability of causing negative effect. Actually, we can read more information in the odds simulation result.

The trace plots is shown below.

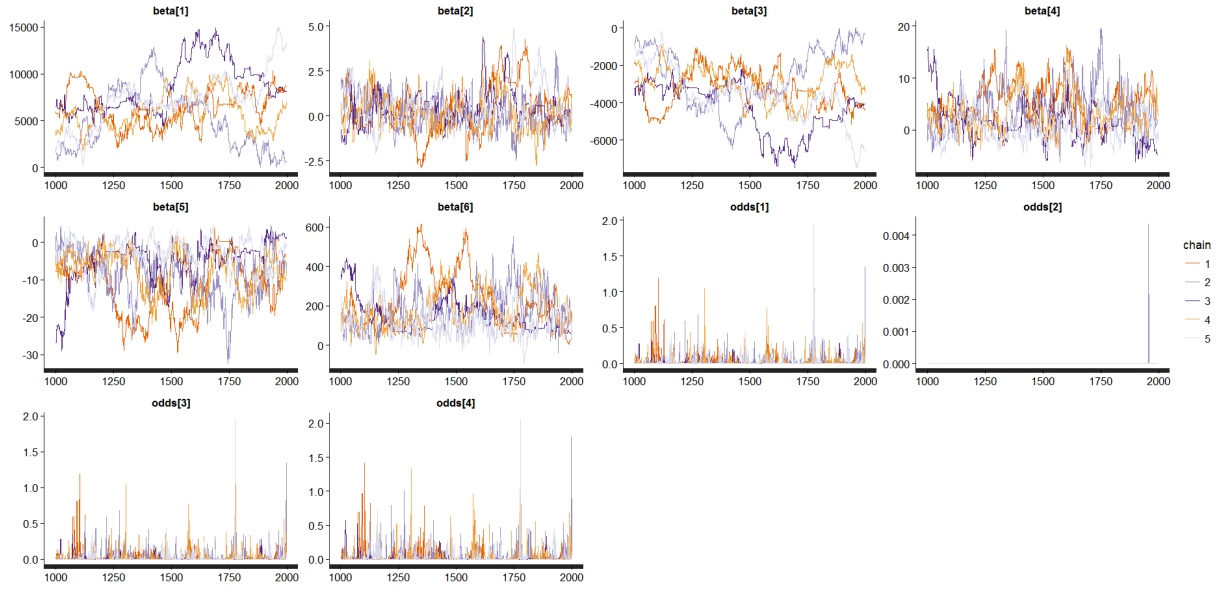


Figure 21 Sampling Trace Plot

We can see sampling procedure converged, which means this model is fairly good enough.

• Sensitive Test

There is no potential outliers, which has been reflected by our visualization plots.

• Conclusions

– **Summary of major findings**

Floods in New York area do have some difference in different boroughs. And for different flood type, it will show some different characteristics. For example, floods in Queens are more likely to be flash flood, and very likely to cause damage on lives and properties. What is more, there are some trends in the happening period of flood events.

By Bayesian model, we actually can find that for all boroughs, just like our intuition, the range and scale of the event will have positive effect on the likelihood to cause damage. And the effect of begin range and end range (e.g. the scale of events at beginning and end) will not be very clear.

– **Limitation of studies**

However, there are some problems with this method:

1. It cannot be used to do prediction, because it is based on the information being already collected. But we still can learn something from it (which I have stated in the beginning).

2. Some of the parameters have not converged so much in some sense. Maybe we have not included enough variables into our model. At this point, we should explore more about the related variables.
3. Because of limited knowledge, we have not set up the procedure to validate the prior and likelihood model. That might be a potential problem with our model.

Section V: Summary

We first test the relation between year and duration of events. There is significant variation but no obvious trend. Then, we try to predict whether floods will occur under certain events with test accuracy of 0.85 and true positive rate of 50%. At last, we do inference on a predictive bayesian model to find that flood range at the beginning and the end has little effect on losses.

Reference

[1] https://en.wikipedia.org/wiki/Flash_flood
[2] https://en.wikipedia.org/wiki/Coastal_flood
[3] Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian data analysis. Vol. 2. Boca Raton, FL, USA: Chapman & Hall/CRC, 2014.
[4] Higgins, James J. "Introduction to modern nonparametric statistics." (2003).
[5] Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.
[6] DeGroot, Morris H. Morris H. Probability and statistics. No. 04; QA273, D4 1986. 1986.
[7]https://www.wunderground.com/history/airport/KNYC/2016/12/10/DailyHistory.html?HideSpecis=1
[8] fttp://ftp.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/
[9] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[10] James G, Witten D, Hastie T, et al. An introduction to statistical learning[M]. New York: springer, 2013.

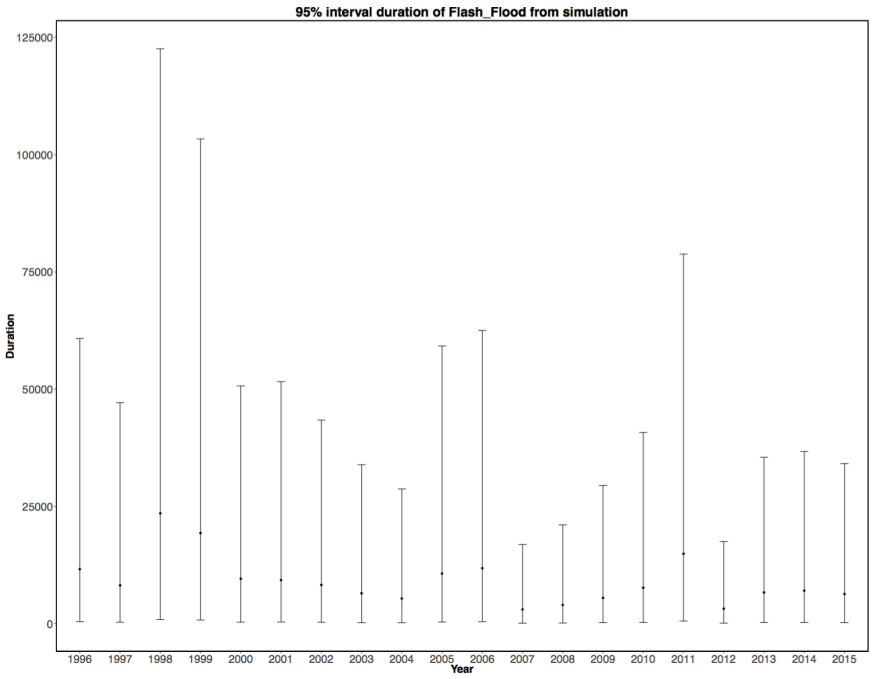
Appendix

• Appendix A

For large size data:

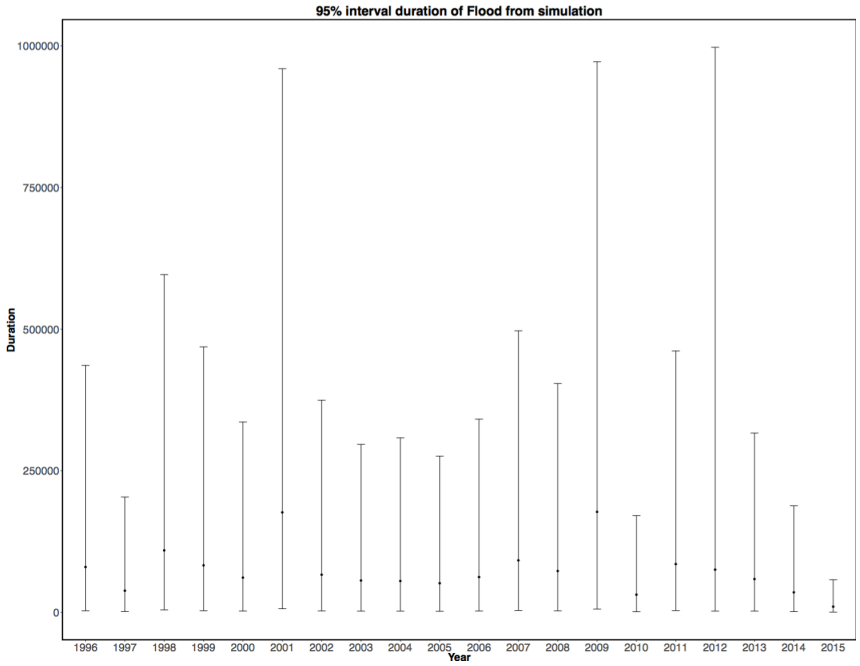
Simulation from Bayes model:

Flash flood:



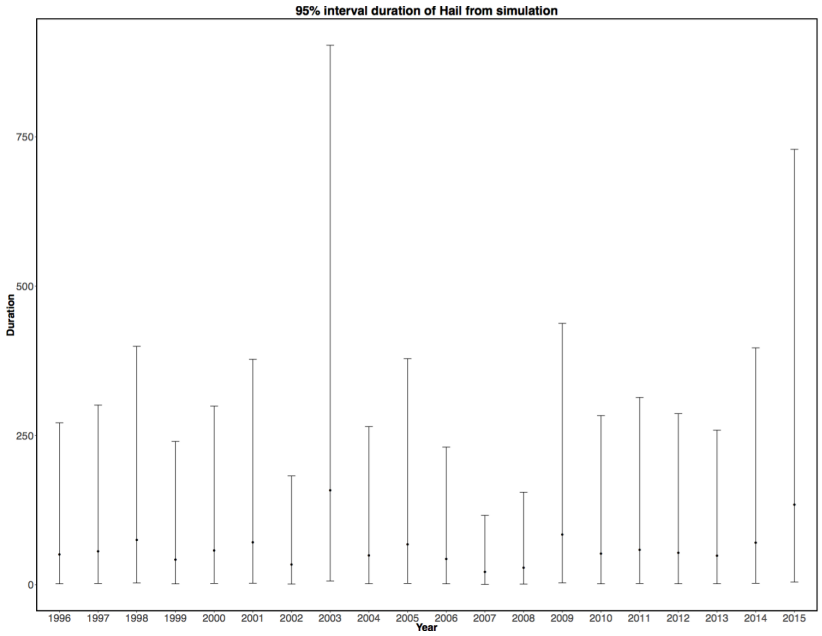
95% interval of duration of flash flood from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Flood:



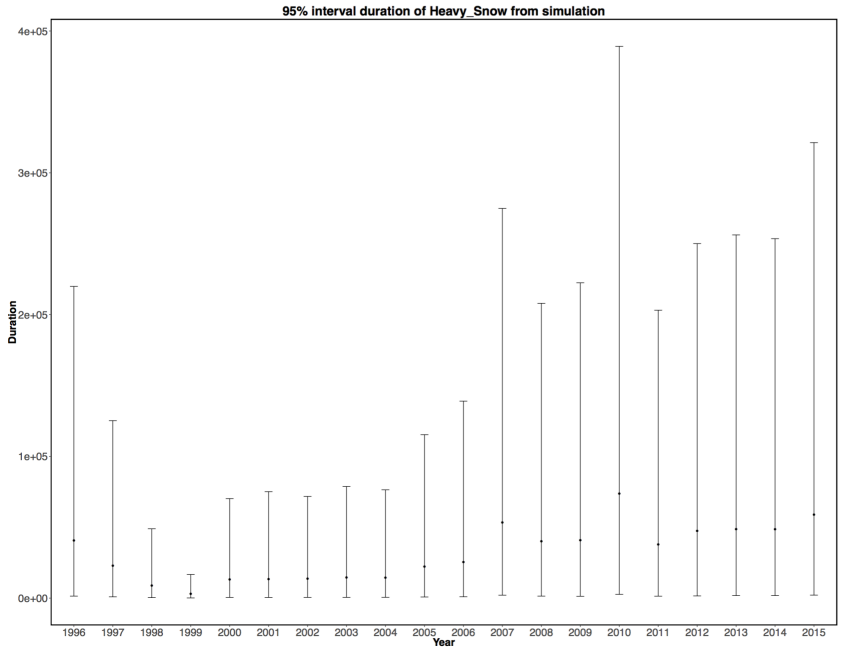
95% interval of duration of flood from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Hail:



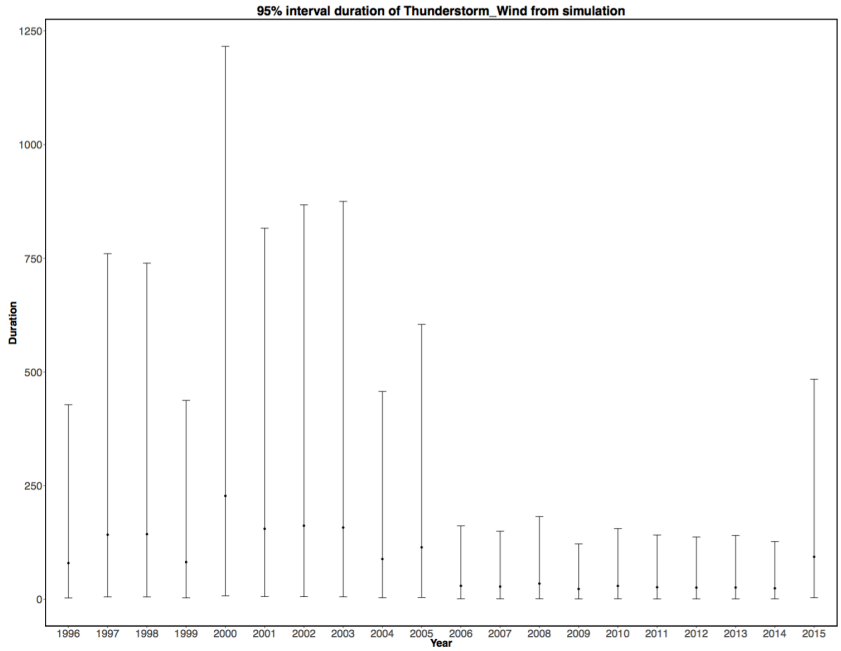
95% interval of duration of hail from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Heavy snow:



95% interval of duration of heavy snow from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

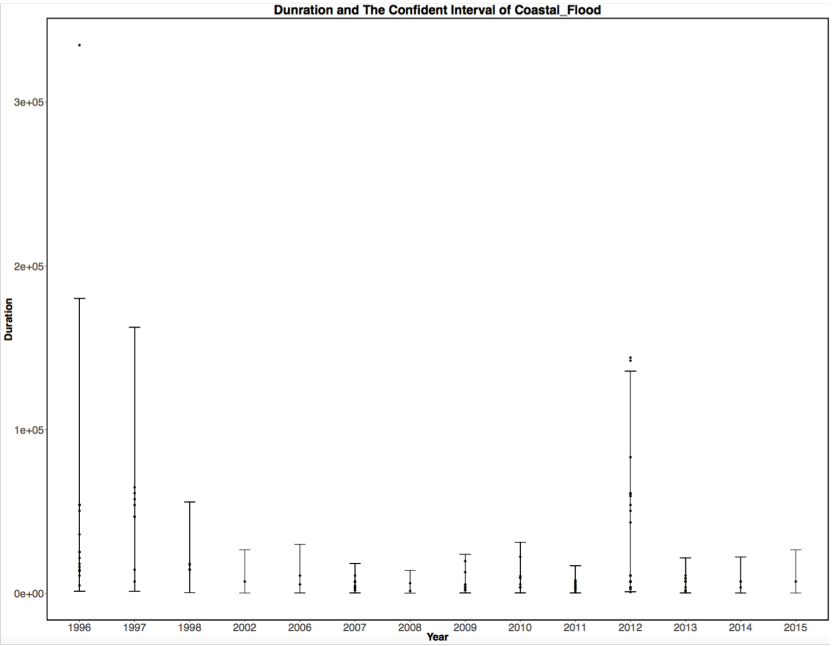
Thunderstorm wind:



95% interval of duration of thunderstorm wind from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

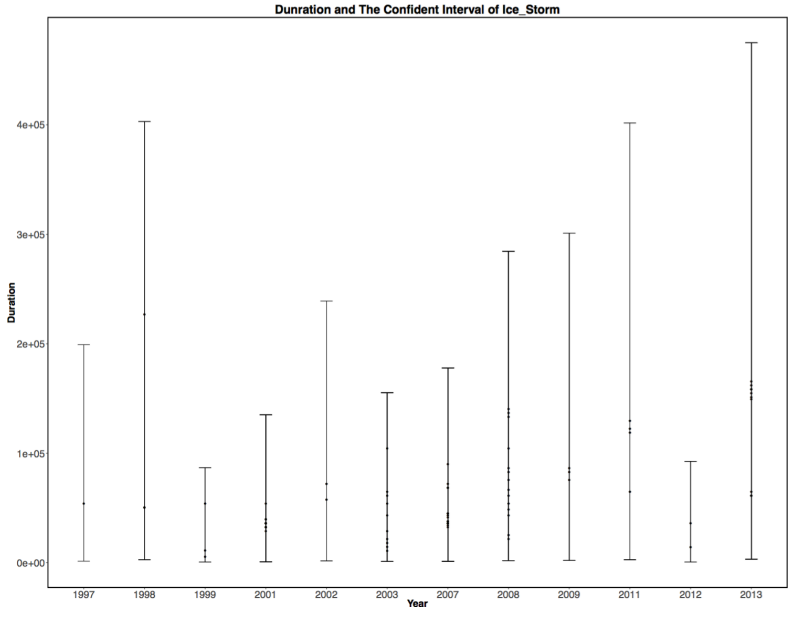
For middle size data:

Coastal flood:



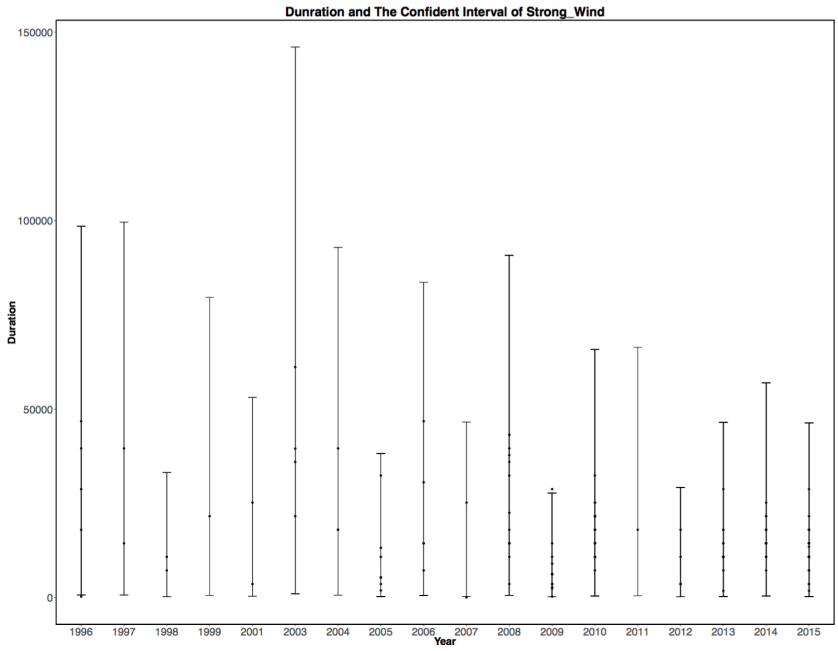
95% interval of duration of coastal flood from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Ice storm:



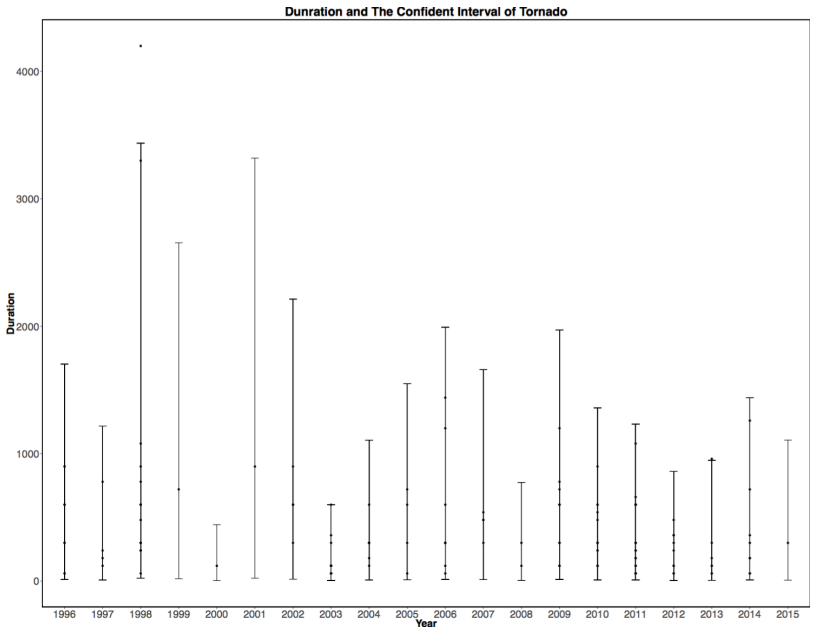
95% interval of duration of ice storm from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Strong wind:



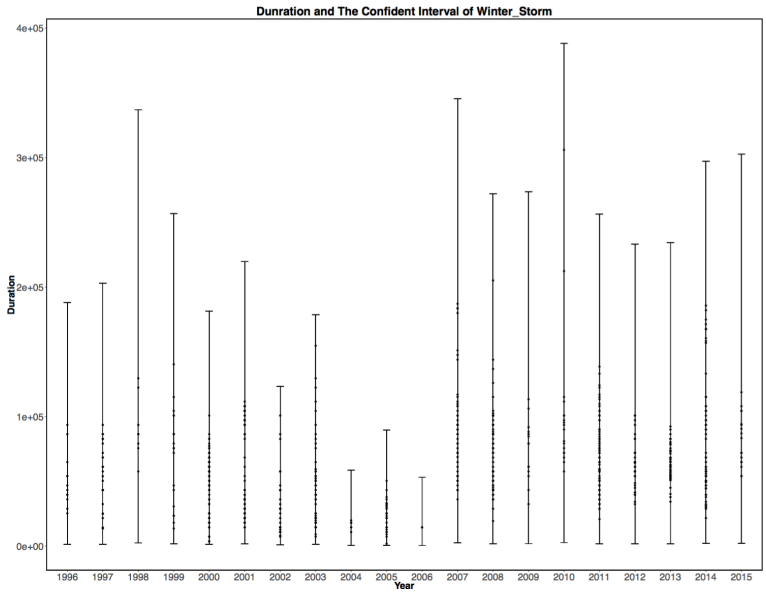
95% interval of duration of strong wind from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

Tornado:



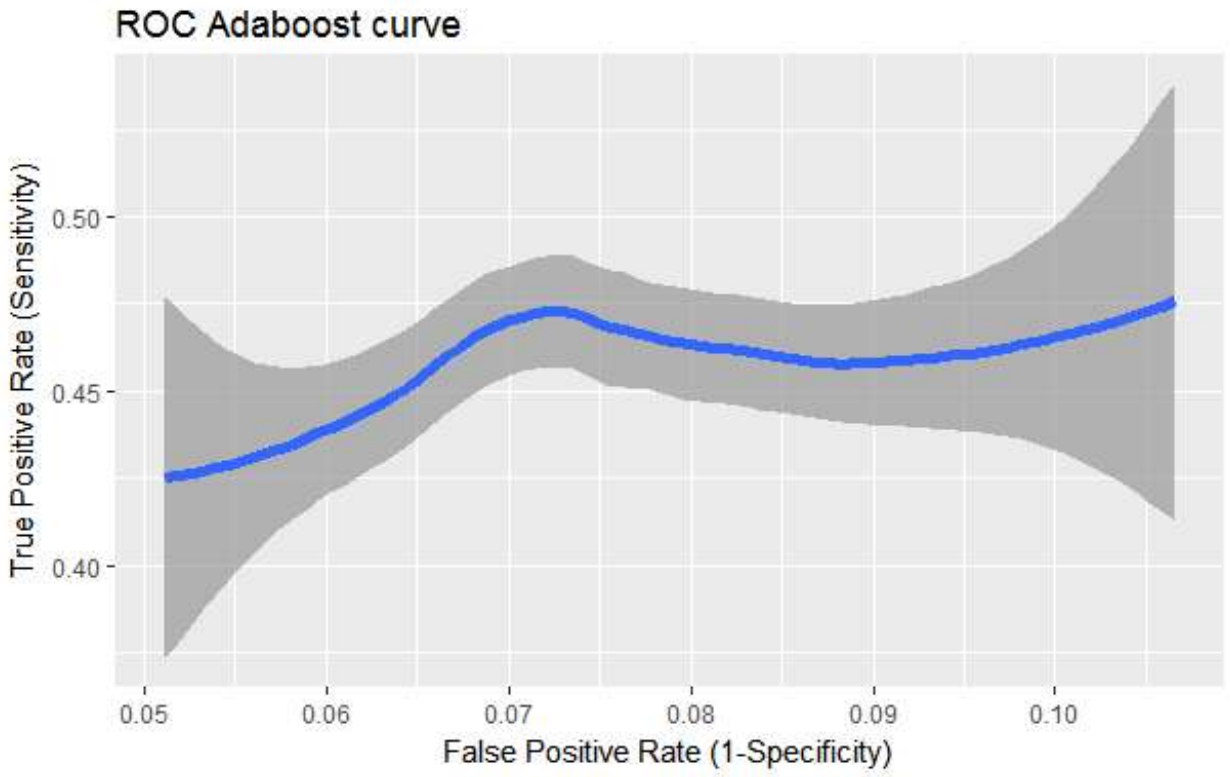
95% interval of duration of tornado from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

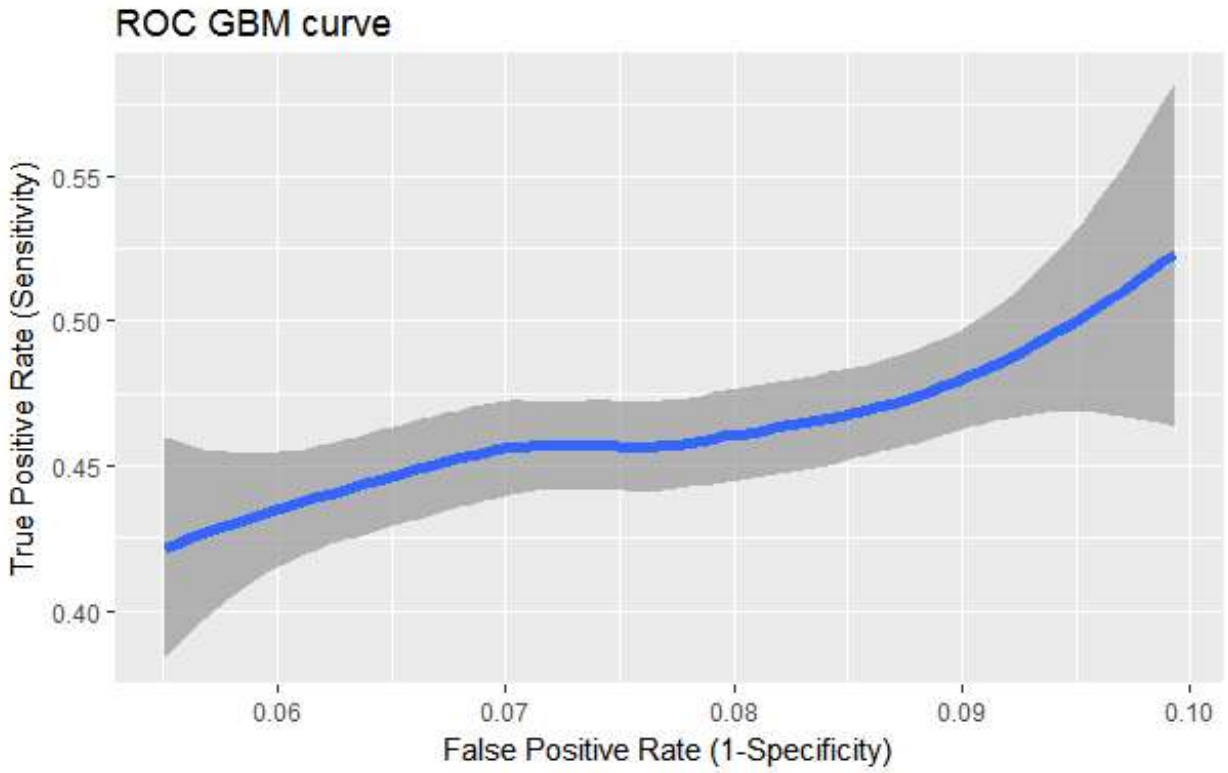
Winter storm:

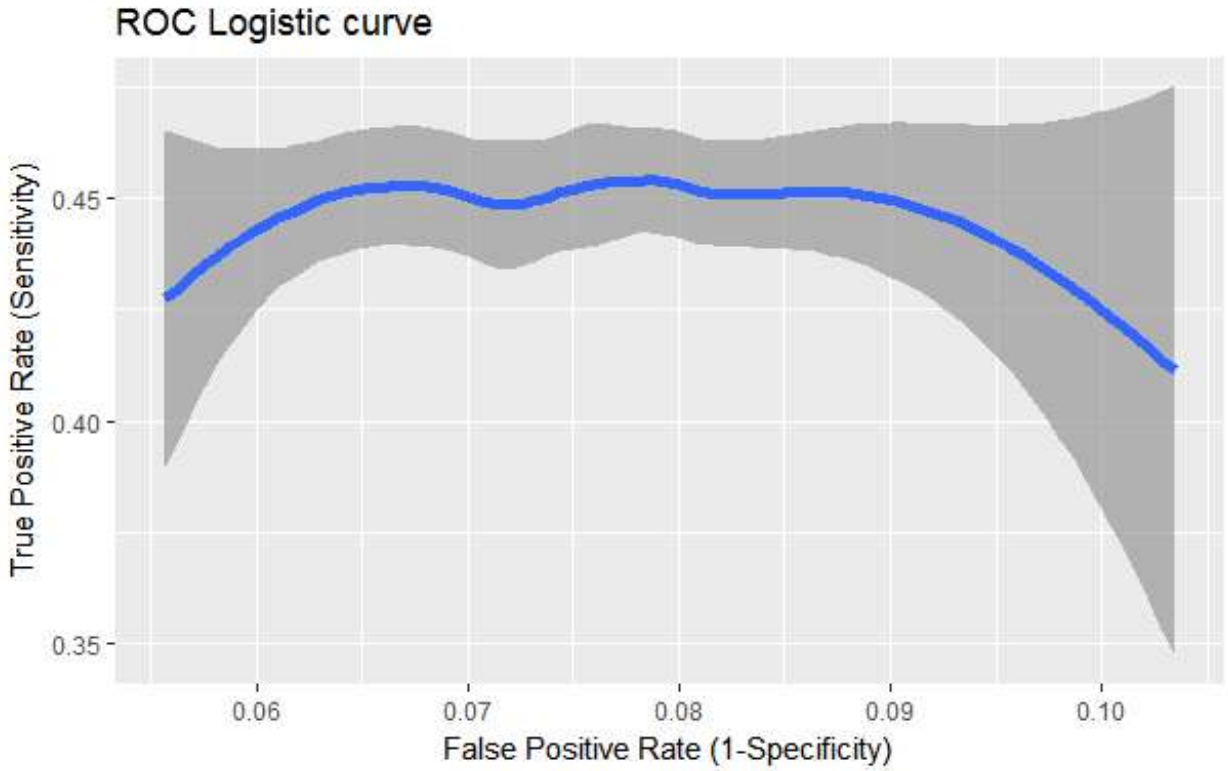


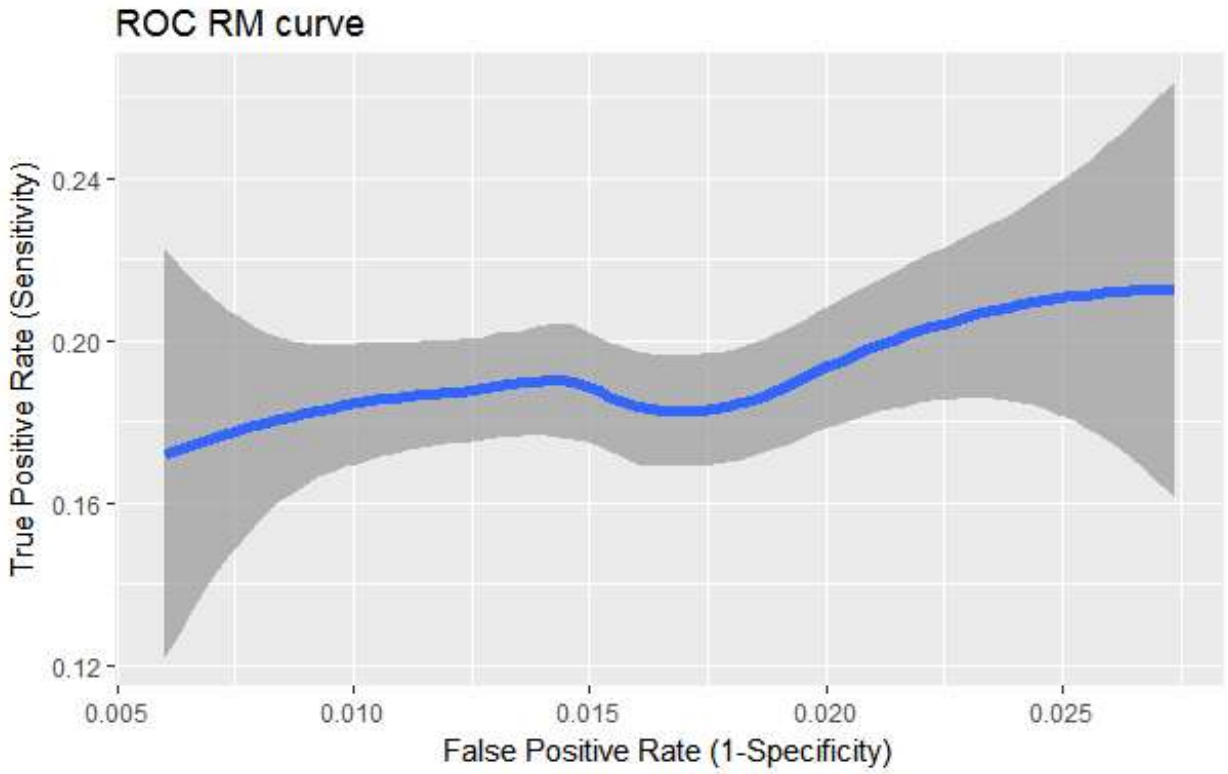
95% interval of duration of winter storm from 1996 to 2015 from simulation. The x-axis is the year. And the y-axis is the duration of the event.

- **Appendix B**




ROC GBM curve
True Positive Rate (Sensitivity)
0.55
0.50
0.45
0.40
0.06
0.07
0.08
0.09
0.10
False Positive Rate (1-Specificity)


ROC Logistic curve
True Positive Rate (Sensitivity)
0.45
0.40
0.35
0.06
0.07
0.08
0.09
0.10
False Positive Rate (1-Specificity)



- **Appendix C**

Model Information:
Inference for Stan model: bf9075d3fd55f9865f16b47c8720201a.
5 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=5000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
beta[1]	8254.25	5726.41	9778.99	123.95	1566.82	4048.96	9661.55	33004.03
beta[2]	0.41	0.18	0.99	-1.23	-0.23	0.31	0.92	3.04
beta[3]	-4132.72	2863.72	4889.52	-16506.94	-4837.75	-2029.66	-788.54	-65.77
beta[4]	1.65	0.47	2.82	-3.49	-0.21	1.65	3.37	8.01
beta[5]	-4.87	0.55	3.62	-13.53	-6.93	-4.61	-2.47	1.06
beta[6]	137.50	9.92	64.30	26.61	87.10	137.93	180.67	271.20
odds[1]	0.02	0.00	0.07	0.00	0.00	0.00	0.01	0.16
odds[2]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[3]	0.02	0.00	0.07	0.00	0.00	0.00	0.01	0.16
odds[4]	0.02	0.00	0.09	0.00	0.00	0.00	0.01	0.21
odds[5]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[6]	0.02	0.00	0.07	0.00	0.00	0.00	0.01	0.16
odds[7]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[8]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[9]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

odds[10]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[11]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[12]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[13]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02
odds[14]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[15]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[16]	0.02	0.00	0.09	0.00	0.00	0.00	0.01	0.21
odds[17]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
odds[18]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[19]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[20]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[21]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[22]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[23]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[24]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.05
odds[25]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[26]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[27]	0.01	0.00	0.08	0.00	0.00	0.00	0.00	0.06
odds[28]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
odds[29]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[30]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[31]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[32]	0.01	0.00	0.05	0.00	0.00	0.00	0.00	0.04
odds[33]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[34]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[35]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[36]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[37]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[38]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[39]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[40]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[41]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03
odds[42]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[43]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[44]	0.01	0.00	0.10	0.00	0.00	0.00	0.00	0.05
odds[45]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[46]	0.01	0.00	0.30	0.00	0.00	0.00	0.00	0.03
odds[47]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[48]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[49]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[50]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[51]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[52]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[53]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[54]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[55]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[56]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[57]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[58]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

odds[59]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[60]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[61]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[62]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[63]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[64]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[65]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[66]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[67]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[68]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[69]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.05
odds[70]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[71]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[72]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[73]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[74]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[75]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[76]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[77]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[78]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.05
odds[79]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.05
odds[80]	0.70	0.37	4.45	0.00	0.00	0.03	0.21	4.48
odds[81]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[82]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[83]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[84]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[85]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[86]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[87]	6.83	1.91	101.69	0.00	0.03	0.25	1.24	29.13
odds[88]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.05
odds[89]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[90]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[91]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[92]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[93]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[94]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[95]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[96]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[97]	0.02	0.00	0.05	0.00	0.00	0.01	0.03	0.15
odds[98]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[99]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[100]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[101]	0.01	0.00	0.12	0.00	0.00	0.00	0.00	0.11
odds[102]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
odds[103]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[104]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[105]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[106]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[107]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03

odds[108]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[109]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[110]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[111]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[112]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[113]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[114]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[115]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[116]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[117]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[118]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[119]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[120]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[121]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[122]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.05
odds[123]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
odds[124]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
odds[125]	0.35	0.08	3.06	0.00	0.00	0.01	0.13	2.19
odds[126]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
odds[127]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.07
odds[128]	0.02	0.00	0.16	0.00	0.00	0.00	0.01	0.18
odds[129]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
odds[130]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.04
odds[131]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[132]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[133]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[134]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
odds[135]	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.03
odds[136]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
odds[137]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
odds[138]	0.65	0.13	4.06	0.00	0.01	0.07	0.34	4.13
odds[139]	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.07
prob[1]	0.01	0.00	0.05	0.00	0.00	0.00	0.01	0.14
prob[2]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[3]	0.01	0.00	0.05	0.00	0.00	0.00	0.01	0.14
prob[4]	0.02	0.00	0.05	0.00	0.00	0.00	0.01	0.17
prob[5]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[6]	0.01	0.00	0.05	0.00	0.00	0.00	0.01	0.14
prob[7]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[8]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[9]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[10]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[11]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[12]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[13]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02
prob[14]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[15]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[16]	0.02	0.00	0.05	0.00	0.00	0.00	0.01	0.17
prob[17]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00

prob[18]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[19]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[20]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[21]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[22]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[23]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[24]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.05
prob[25]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[26]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[27]	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.05
prob[28]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
prob[29]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[30]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[31]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[32]	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.04
prob[33]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[34]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[35]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[36]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[37]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[38]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[39]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[40]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[41]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03
prob[42]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[43]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[44]	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.05
prob[45]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[46]	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.03
prob[47]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[48]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[49]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[50]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[51]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[52]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[53]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[54]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[55]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[56]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[57]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[58]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[59]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[60]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[61]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[62]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[63]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[64]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[65]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[66]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

prob[67]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[68]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[69]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.04
prob[70]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[71]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[72]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[73]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[74]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[75]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[76]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[77]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[78]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.04
prob[79]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.04
prob[80]	0.14	0.03	0.23	0.00	0.00	0.03	0.17	0.82
prob[81]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[82]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[83]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[84]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[85]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[86]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[87]	0.31	0.02	0.31	0.00	0.03	0.20	0.55	0.97
prob[88]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.05
prob[89]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[90]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[91]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[92]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[93]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[94]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[95]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[96]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[97]	0.02	0.00	0.04	0.00	0.00	0.01	0.03	0.13
prob[98]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[99]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[100]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[101]	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.10
prob[102]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
prob[103]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[104]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[105]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[106]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[107]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03
prob[108]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[109]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[110]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[111]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[112]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[113]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[114]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[115]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

prob[116]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[117]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[118]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[119]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[120]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[121]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[122]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.05
prob[123]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
prob[124]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
prob[125]	0.10	0.02	0.18	0.00	0.00	0.01	0.12	0.69
prob[126]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03
prob[127]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.06
prob[128]	0.02	0.00	0.05	0.00	0.00	0.00	0.01	0.15
prob[129]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
prob[130]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
prob[131]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[132]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[133]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[134]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prob[135]	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03
prob[136]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03
prob[137]	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03
prob[138]	0.17	0.03	0.23	0.00	0.01	0.06	0.25	0.81
prob[139]	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.06
lp__	-6.69	0.20	1.73	-10.92	-7.68	-6.37	-5.35	-4.30

	n_eff	Rhat
beta[1]	3	3.88
beta[2]	30	1.19
beta[3]	3	3.88
beta[4]	36	1.13
beta[5]	44	1.10
beta[6]	42	1.14
odds[1]	429	1.02
odds[2]	5000	1.00
odds[3]	429	1.02
odds[4]	451	1.01
odds[5]	471	1.01
odds[6]	429	1.02
odds[7]	487	1.01
odds[8]	486	1.01
odds[9]	5000	1.00
odds[10]	5000	1.00
odds[11]	5000	1.00
odds[12]	726	1.00
odds[13]	369	1.01
odds[14]	733	1.00
odds[15]	486	1.01
odds[16]	452	1.01
odds[17]	350	1.01

odds[18] 475 1.01
odds[19] 729 1.00
odds[20] 527 1.01
odds[21] 544 1.01
odds[22] 561 1.01
odds[23] 5000 1.00
odds[24] 655 1.01
odds[25] 527 1.01
odds[26] 5000 1.00
odds[27] 1434 1.00
odds[28] 5000 1.00
odds[29] 5000 1.00
odds[30] 527 1.01
odds[31] 527 1.01
odds[32] 1263 1.00
odds[33] 543 1.01
odds[34] 494 1.01
odds[35] 5000 1.00
odds[36] 531 1.01
odds[37] 544 1.01
odds[38] 1210 1.00
odds[39] 614 1.00
odds[40] 544 1.01
odds[41] 648 1.01
odds[42] 387 1.01
odds[43] 415 1.01
odds[44] 1652 1.00
odds[45] 415 1.01
odds[46] 4537 1.00
odds[47] 917 1.01
odds[48] 836 1.01
odds[49] 850 1.01
odds[50] 5000 1.00
odds[51] 613 1.01
odds[52] 346 1.01
odds[53] 5000 1.00
odds[54] 5000 1.00
odds[55] 415 1.01
odds[56] 415 1.01
odds[57] 348 1.01
odds[58] 928 1.01
odds[59] 895 1.01
odds[60] 5000 1.00
odds[61] 5000 1.00
odds[62] 5000 1.00
odds[63] 156 1.03
odds[64] 157 1.03
odds[65] 158 1.03
odds[66] 158 1.03

odds[67] 158 1.03
odds[68] 159 1.03
odds[69] 720 1.01
odds[70] 158 1.03
odds[71] 157 1.03
odds[72] 157 1.03
odds[73] 158 1.03
odds[74] 158 1.03
odds[75] 156 1.03
odds[76] 271 1.02
odds[77] 158 1.03
odds[78] 720 1.01
odds[79] 720 1.01
odds[80] 141 1.03
odds[81] 157 1.03
odds[82] 5000 1.00
odds[83] 271 1.02
odds[84] 5000 1.00
odds[85] 5000 1.00
odds[86] 268 1.02
odds[87] 2823 1.00
odds[88] 681 1.01
odds[89] 156 1.03
odds[90] 157 1.03
odds[91] 158 1.03
odds[92] 273 1.02
odds[93] 156 1.03
odds[94] 271 1.02
odds[95] 156 1.03
odds[96] 270 1.02
odds[97] 1027 1.01
odds[98] 273 1.02
odds[99] 156 1.03
odds[100] 161 1.03
odds[101] 1455 1.00
odds[102] 581 1.01
odds[103] 158 1.03
odds[104] 157 1.03
odds[105] 157 1.03
odds[106] 158 1.03
odds[107] 335 1.02
odds[108] 158 1.03
odds[109] 5000 1.00
odds[110] 157 1.03
odds[111] 5000 1.00
odds[112] 157 1.03
odds[113] 156 1.03
odds[114] 271 1.02
odds[115] 158 1.03

odds[116] 5000 1.00
odds[117] 156 1.03
odds[118] 158 1.03
odds[119] 5000 1.00
odds[120] 272 1.02
odds[121] 158 1.03
odds[122] 679 1.01
odds[123] 157 1.03
odds[124] 5000 1.00
odds[125] 1659 1.00
odds[126] 284 1.02
odds[127] 312 1.01
odds[128] 1075 1.00
odds[129] 496 1.01
odds[130] 286 1.02
odds[131] 5000 1.00
odds[132] 5000 1.00
odds[133] 5000 1.00
odds[134] 5000 1.00
odds[135] 951 1.01
odds[136] 284 1.02
odds[137] 285 1.02
odds[138] 912 1.01
odds[139] 314 1.01
prob[1] 257 1.02
prob[2] 5000 1.00
prob[3] 257 1.02
prob[4] 233 1.02
prob[5] 444 1.01
prob[6] 257 1.02
prob[7] 462 1.01
prob[8] 461 1.01
prob[9] 5000 1.00
prob[10] 5000 1.00
prob[11] 5000 1.00
prob[12] 695 1.00
prob[13] 337 1.01
prob[14] 710 1.00
prob[15] 461 1.01
prob[16] 233 1.02
prob[17] 329 1.01
prob[18] 449 1.01
prob[19] 704 1.00
prob[20] 521 1.01
prob[21] 536 1.01
prob[22] 548 1.01
prob[23] 5000 1.00
prob[24] 576 1.01
prob[25] 521 1.01

prob[26]	5000	1.00
prob[27]	839	1.01
prob[28]	5000	1.00
prob[29]	5000	1.00
prob[30]	521	1.01
prob[31]	521	1.01
prob[32]	656	1.01
prob[33]	535	1.01
prob[34]	483	1.01
prob[35]	5000	1.00
prob[36]	524	1.01
prob[37]	536	1.01
prob[38]	1069	1.00
prob[39]	604	1.00
prob[40]	536	1.01
prob[41]	560	1.01
prob[42]	382	1.01
prob[43]	407	1.02
prob[44]	704	1.00
prob[45]	406	1.02
prob[46]	1547	1.01
prob[47]	867	1.01
prob[48]	791	1.01
prob[49]	804	1.01
prob[50]	5000	1.00
prob[51]	522	1.01
prob[52]	342	1.01
prob[53]	5000	1.00
prob[54]	5000	1.00
prob[55]	407	1.02
prob[56]	406	1.02
prob[57]	344	1.01
prob[58]	878	1.01
prob[59]	847	1.01
prob[60]	5000	1.00
prob[61]	5000	1.00
prob[62]	5000	1.00
prob[63]	153	1.03
prob[64]	155	1.03
prob[65]	156	1.03
prob[66]	156	1.03
prob[67]	156	1.03
prob[68]	155	1.03
prob[69]	413	1.02
prob[70]	156	1.03
prob[71]	155	1.03
prob[72]	155	1.03
prob[73]	155	1.03
prob[74]	155	1.03

prob[75] 154 1.03
prob[76] 265 1.02
prob[77] 156 1.03
prob[78] 413 1.02
prob[79] 413 1.02
prob[80] 53 1.11
prob[81] 154 1.03
prob[82] 5000 1.00
prob[83] 265 1.02
prob[84] 5000 1.00
prob[85] 5000 1.00
prob[86] 262 1.02
prob[87] 328 1.01
prob[88] 402 1.02
prob[89] 153 1.03
prob[90] 155 1.03
prob[91] 156 1.03
prob[92] 268 1.02
prob[93] 154 1.03
prob[94] 266 1.02
prob[95] 154 1.03
prob[96] 263 1.02
prob[97] 797 1.01
prob[98] 268 1.02
prob[99] 154 1.03
prob[100] 158 1.03
prob[101] 348 1.01
prob[102] 689 1.01
prob[103] 155 1.03
prob[104] 154 1.03
prob[105] 154 1.03
prob[106] 155 1.03
prob[107] 322 1.02
prob[108] 156 1.03
prob[109] 5000 1.00
prob[110] 154 1.03
prob[111] 5000 1.00
prob[112] 154 1.03
prob[113] 153 1.03
prob[114] 265 1.02
prob[115] 155 1.03
prob[116] 5000 1.00
prob[117] 153 1.03
prob[118] 156 1.03
prob[119] 5000 1.00
prob[120] 266 1.02
prob[121] 156 1.03
prob[122] 402 1.02
prob[123] 155 1.03

prob[124] 5000 1.00
prob[125]   108 1.05
prob[126]   239 1.02
prob[127]   242 1.02
prob[128]   295 1.02
prob[129]   442 1.01
prob[130]   238 1.02
prob[131] 5000 1.00
prob[132] 5000 1.00
prob[133] 5000 1.00
prob[134] 5000 1.00
prob[135]   669 1.01
prob[136]   239 1.02
prob[137]   239 1.02
prob[138]   55 1.05
prob[139]   242 1.02
lp__   78 1.06

Samples were drawn using NUTS(diag_e) at Sun Aug 21 19:10:31 2016. For each parameter, n_eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat=1).