Project: The analysis of storm events

Zhiyu Chai (Level 6000)

1. Introduction

Storm is any kind of atmospheric disturbance that affects the environment or the surface of Earth. Usually, the definition of storm is the wind having a speed of more than 89 kilometers per hour. Storms usually brings bad weather, like strong winds (tropical cyclones), thunder and lightning (thunderstorms), heavy precipitation (snowstorms) or substances that move with wind in the atmosphere (sandstorms). Even though storms can bring some advantages to reduce the drought in some areas, storms can also cause loss of life and property. The common disasters include shipwrecks, damage of vehicles, buildings, bridges and other outside objects (agriculture), flooding and so on.

Because of these harms of storms, government needs to have their part of force to prevent and fix the damage from storms. The government needs to transfer people before storm comes and rescue people after storms. So, the analysis of storm data is necessary, thus government can place their force in certain area, prepare different kinds of equipment with different number to transfer or rescue people and fix the damage.

2. Data description

The storm datasets are collected from NCDC (National Climate Date Center). The data covers from 1950 through to the present and is updated monthly. In this project, two kind of dataset will be used. The first one describes the parameters of storms. These parameters include the time of storms happening, location of storms happening with the county name, latitude and longitude, the range of storms and the azimuth of the storm. Because this dataset includes the detailed location of storms, clustering will be used in this dataset. The second dataset describes the fatalities caused by storms. In this dataset, the time of fatality, fatality type (death or injure), the gender and age of the victim, place where the fatality happening. Based on these details, decision tree will be applied on this dataset.

3. Basic analysis

For the first dataset, the dataset with location, I use 20-years-data to do the basic analysis. In this analysis, I removed several kinds of data. The data includes the one which range is bigger than 20, which longitude is bigger than 0. The percent of these data is really small, less than 0.1%, but has a big impact on the plot.

First, I create the heatmap of numeric data, it shows that these variables are less relative to the others, so linear regression cannot be applied on this dataset.

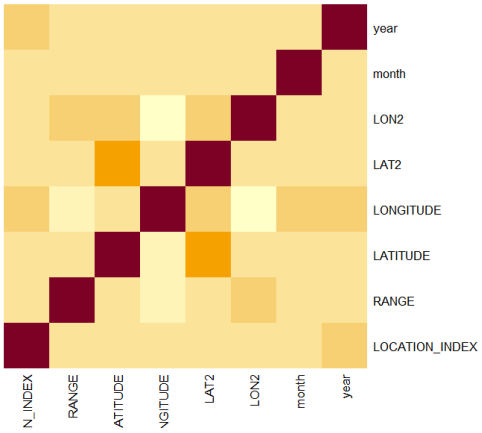

Fig.1 Hearmap of the location dataset

I choose four variables, range, latitude, longitude and month to do the basic analysis. The plots are shown below.

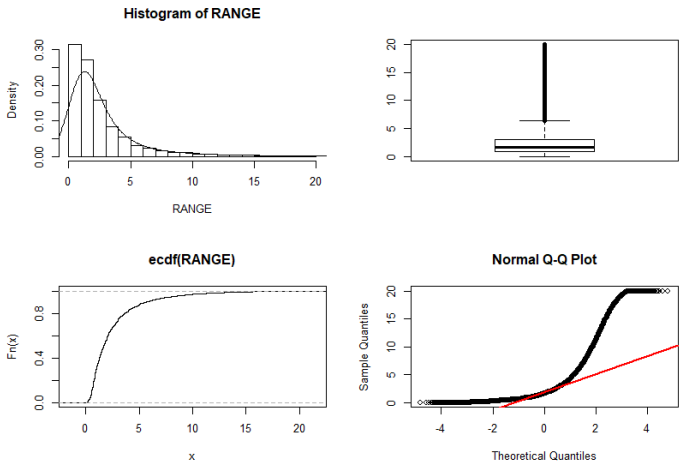

Fig.2 Four plot of the variable Range

These plots have shown that the range of most storms are smaller than 3. The max range is 20 and min is 0.1. The median of range is 2.571. The upper hinge is about 3, lower hinge is about 1. About 90% storm range is smaller than 5 and about less than 3% storm range is bigger than 10. The distribution of storm range is not a normal distribution.

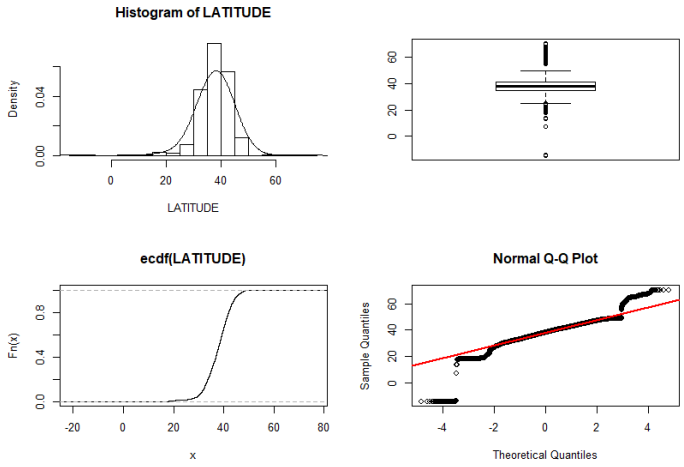

Fig.3 Four plot of the variable Latitude

These plots have shown that the latitude of most storms is in the range of 25 to 45. The max latitude is 70.5 and min is -14.56. The median of range is 38.11. The upper hinge is about 42, lower hinge is about 37. About 60% storm latitude is about 41 and about less than 5% storm latitude is bigger than 50. The distribution of storm latitude is not a normal distribution.

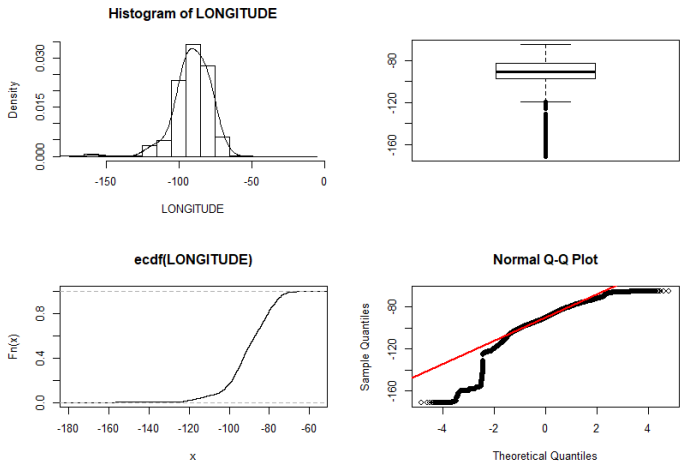

Fig.4 Four plot of the variable Longitude

These plots have shown that the longitude of most storms is in the range of -105 to -75. The max latitude is -64.57 and min is -14.56. The median of range is -90.84. The upper hinge is about -79, lower hinge is about -95. Less than 18% storm longitude is smaller than -100 and about less than 25% storm latitude is bigger than -80. The distribution of storm longitude is not a normal distribution.

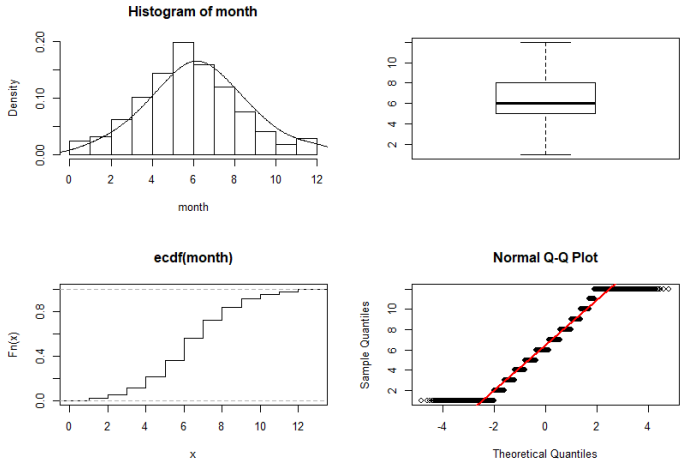

Fig.5 Four plot of the variable month

These plots have shown that most storm happens in June. Also, about 60% of storm happens in the first half of year. The distribution of storm month is not a normal distribution.

So, based on these analyses, the clustering model will be applied on the data that its month is June, latitude is in the range of 25 and 45, longitude is in the range of -105 to -75 and its range is under 20. If the model can be applied successfully, then the government can deploy their equipment in certain place where the rescue team can depart to rescue everyone in the area.

For the second dataset, I also use 20-years-data to do the basic analysis. In this analysis, I removed some data without fatality_age or 0 value of fatality_age and some data without fatality_sex. These data will have a big influence in the application of decision tree model.

After cleaning, some basic details have shown. The percent of injured people in the past 20 years is about 17.18% and the percent of people who died from storm is 82.82%. About 34.20% victims are female and 65.80% victims are male. Initially, it has shown that if there is a fatality, the victim is more likely to loss his or her life. Also, males have their life danger more than females.

First, I create the heatmap of numeric data, it shows that these variables are less relative to the others. Only fatality_time and fatality_day seems to be a little relative. So, linear regression cannot be applied on this dataset.

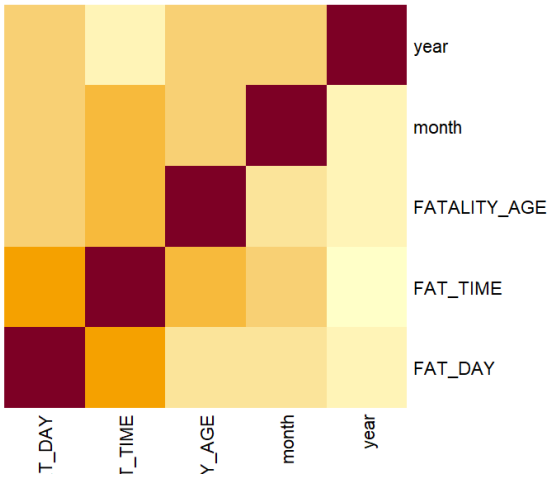

Fig.6 Hearmap of the fatality dataset

I choose two fatality_age and month to do the basic analysis. The plots are shown below.

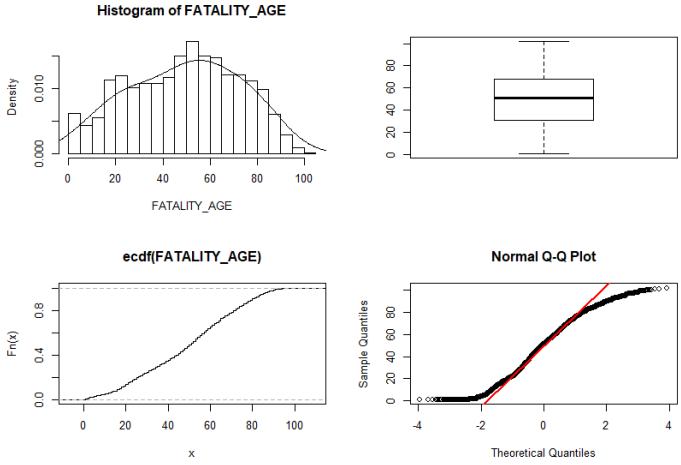

Fig.7 Four plot of the variable fatality_age

These plots have shown that the fatality_age distribution is evenly distributed. The oldest victim is 102 and the youngest is 1. The median of range is 51. The upper hinge is about 70, lower hinge is about 35. Less than 40% victims are younger than 40 and about less than 15% victims are older than 80. The distribution of storm longitude looks like a normal distribution, but still it is not a normal distribution.

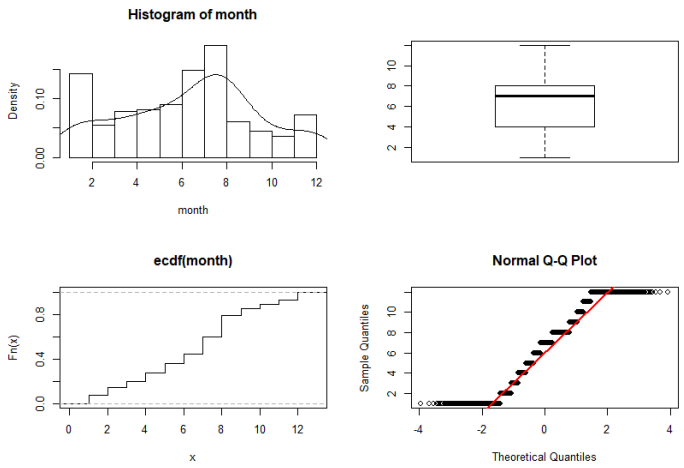

Fig.8 Four plot of the variable fatality_month

These plots have shown that most cases happen in July. Also, about 45% of storm happens in the first half of year. The distribution of storm month is not a normal distribution.

So, based on these analyses, the decision tree model will be applied on this dataset. The variable fatality_location will be used to be independent value and the rest variables will be the dependent value. This decision tree model can help the rescue team to identify the location of the victim and help them to get the right equipment before they depart, if the model can be applied successfully.

4. Model: Clustering

For the location dataset, the variables latitude and longitude can be gotten. In this section, because the data size is too big, it's difficult for the computer to run the model. What's more, the data of the early years has lost some values, so only the data in the past ten years will be used and they will be applied separately. According to the basic analysis, only the data with month of June, latitude in the range of 25 and 45, longitude in the range of -105 to -75 and range under 20 will be applied to the model.

In this model, function clusGap() is used to determined the number of clusters, and then function kmeans() is used to apply the clustering model. Finally, function silhouette() is used to validate the model. The pictures and table below have shown the result. Each picture includes four plots, the plot of latitude and longitude, the optimal number of clusters, cluster plot and silhouette width.

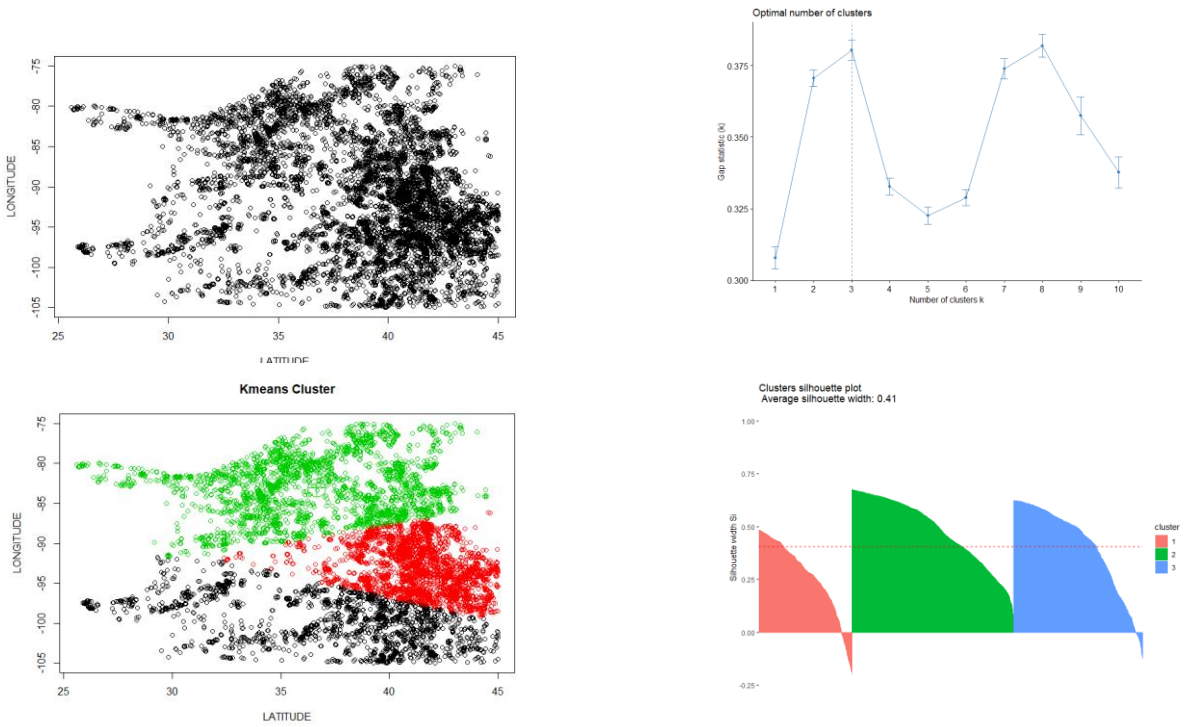

Fig.9 Four plot of the location data of 2010

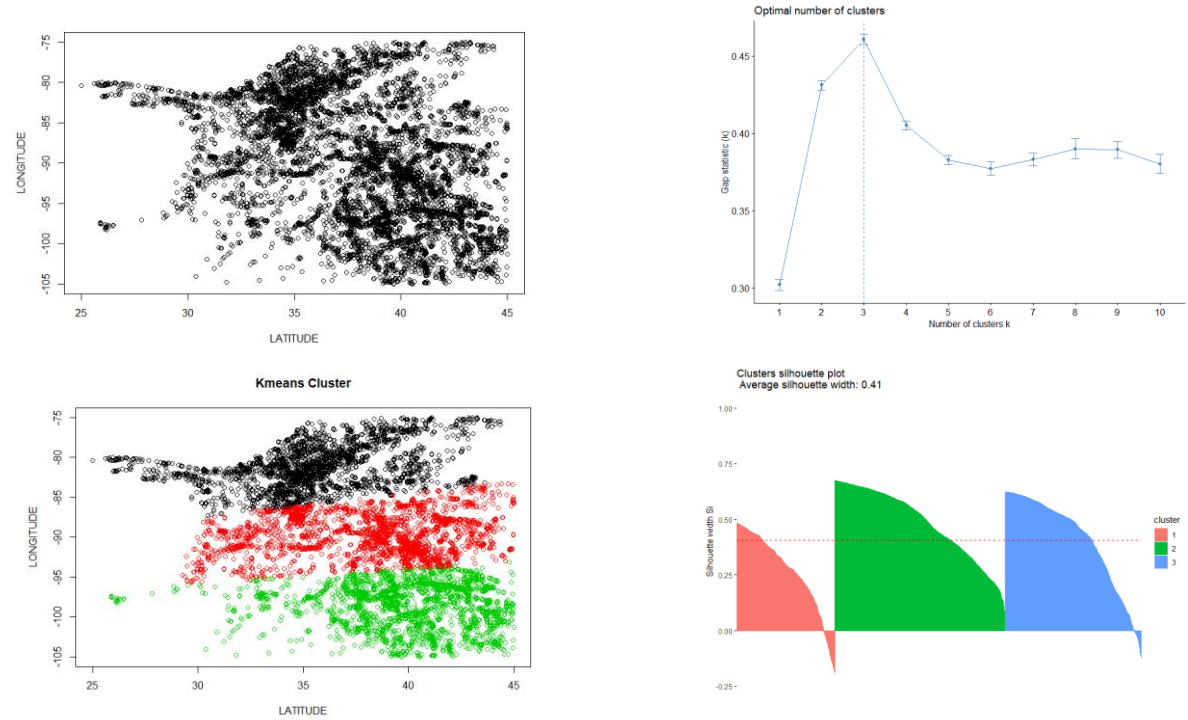

Fig.10 Four plot of the location data of 2011

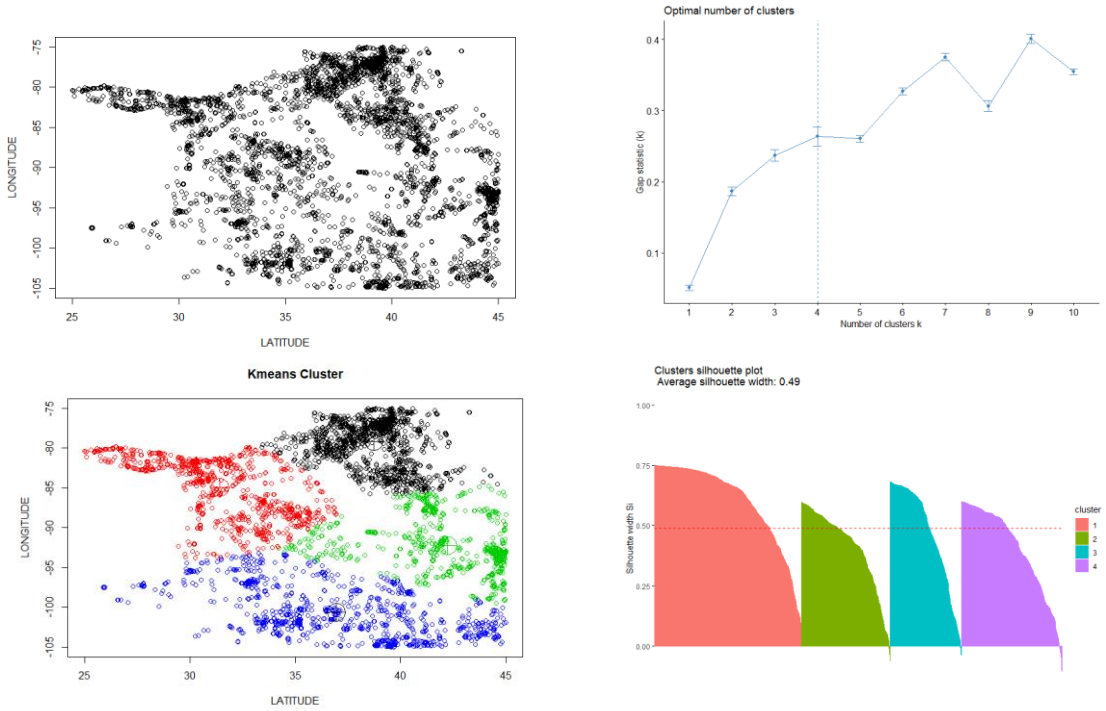

Fig.11 Four plot of the location data of 2012

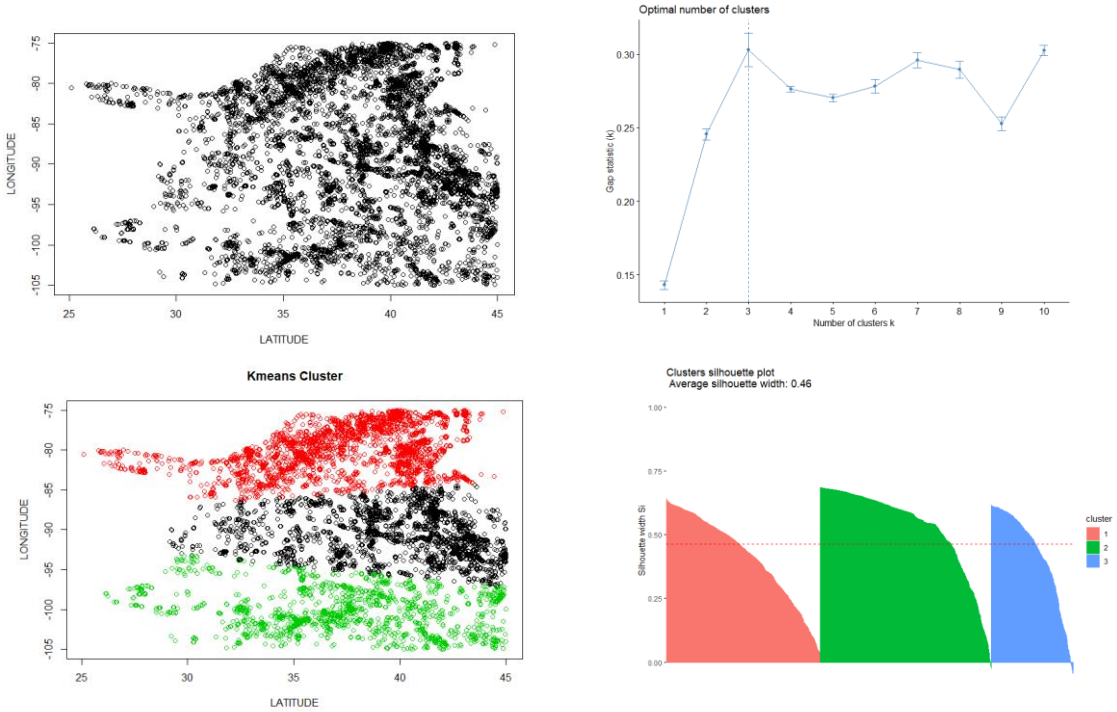

Fig.12 Four plot of the location data of 2013

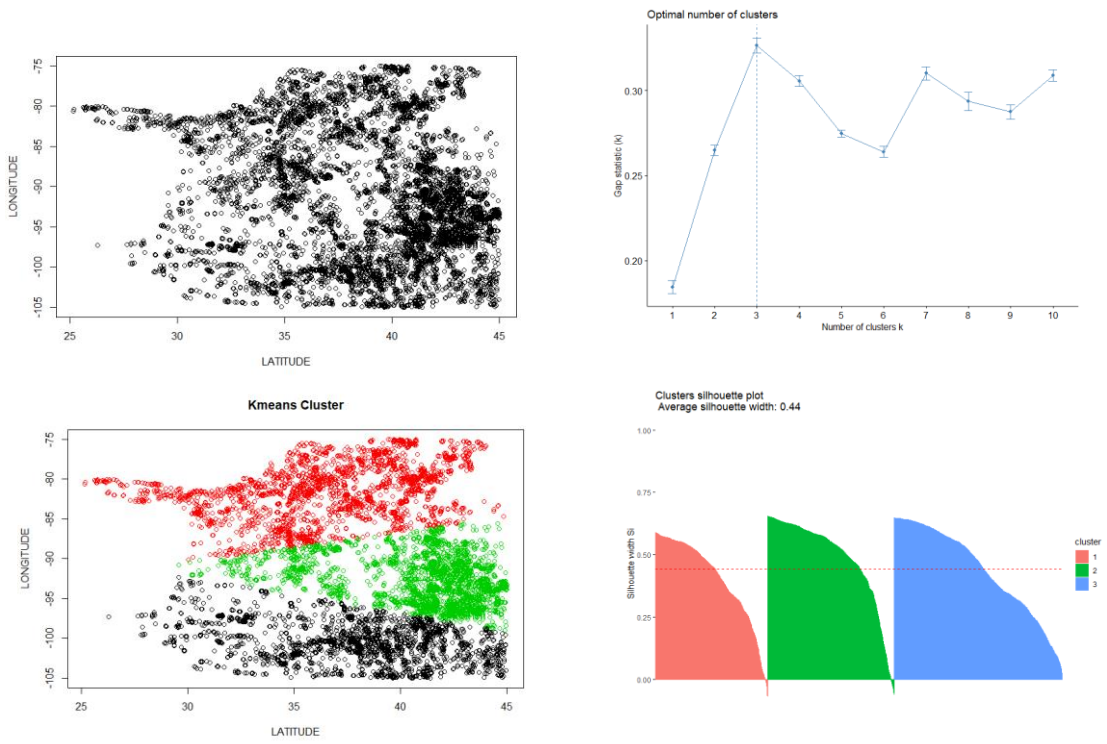

Fig.13 Four plot of the location data of 2014

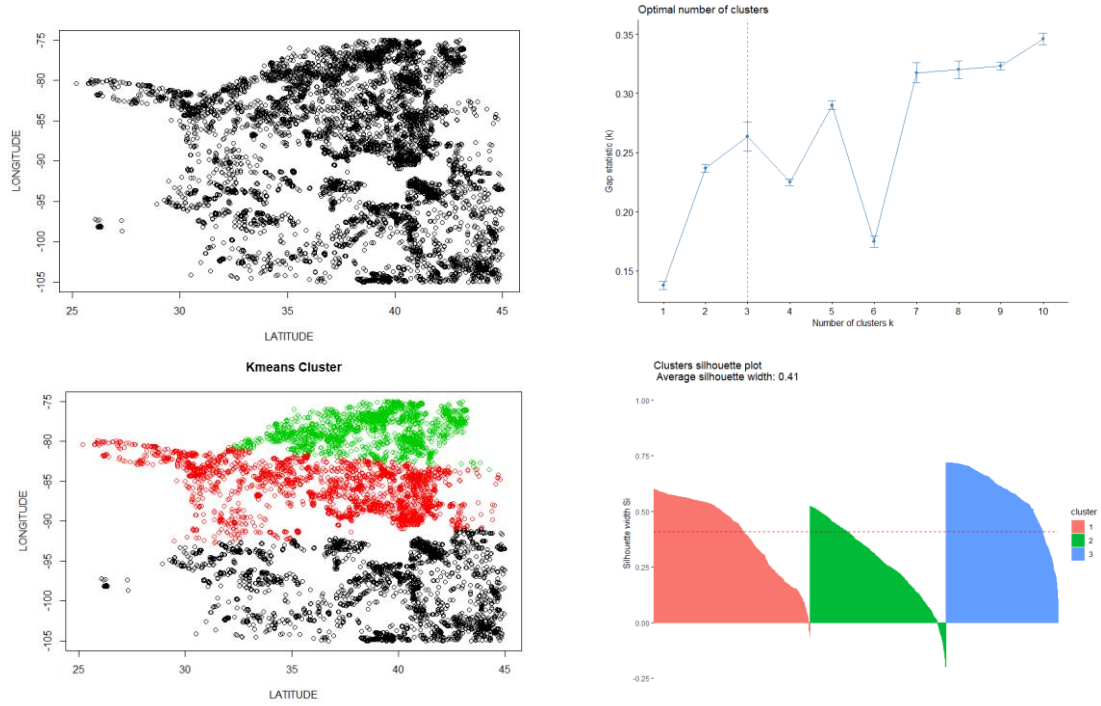

Fig.14 Four plot of the location data of 2015

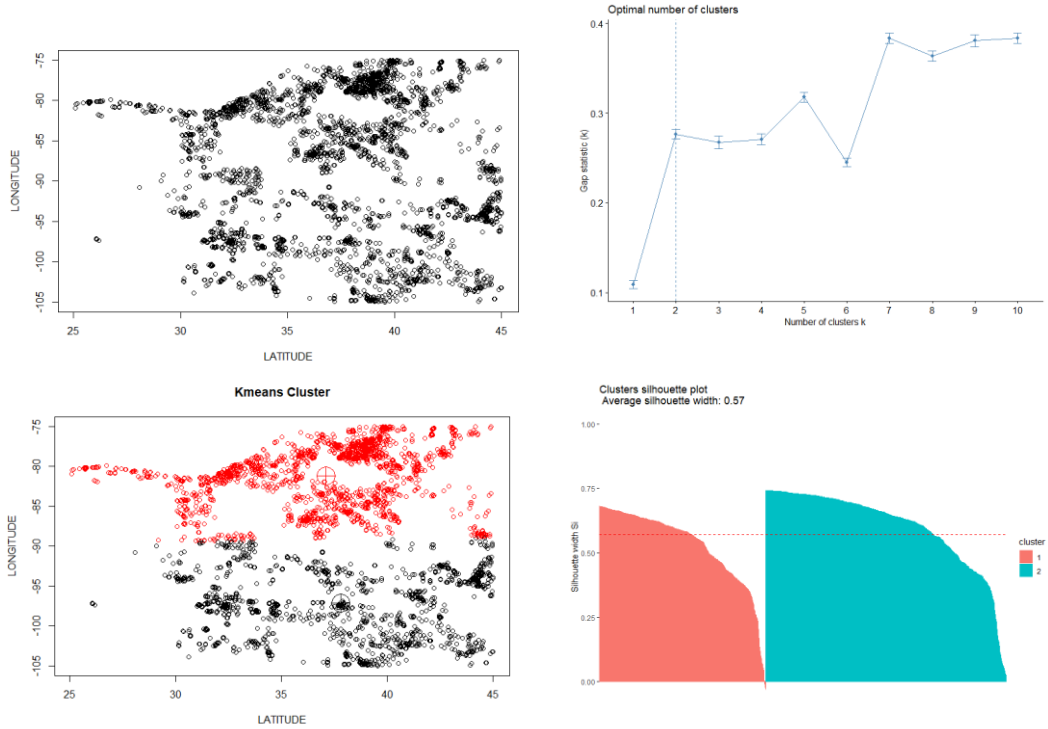

Fig.15 Four plot of the location data of 2016

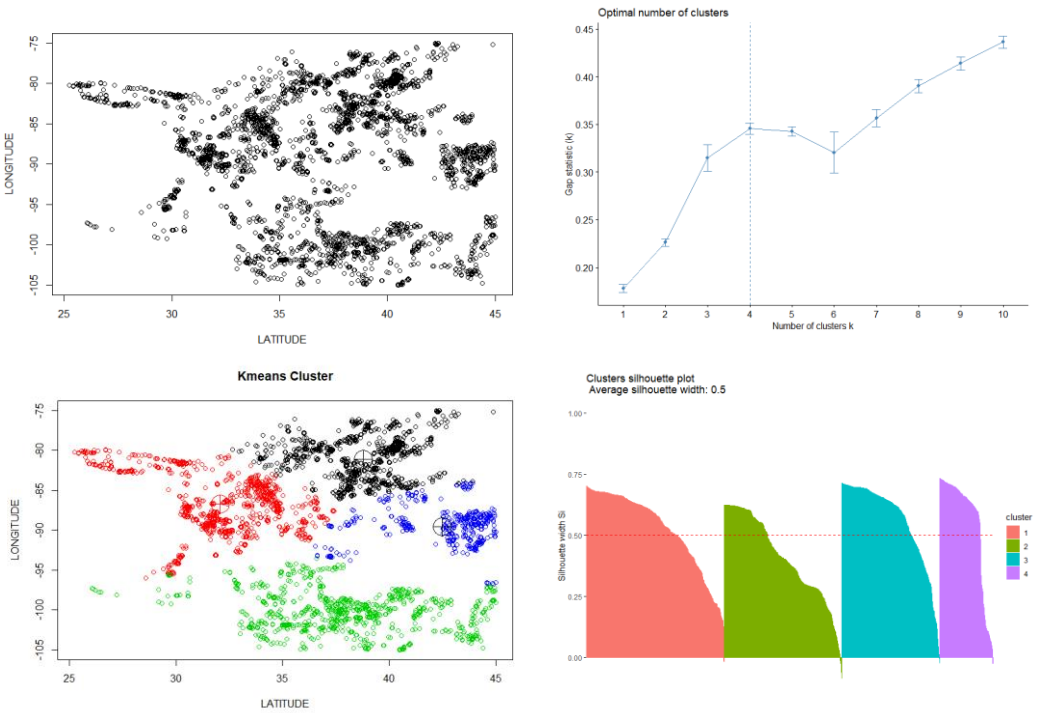

Fig.16 Four plot of the location data of 2017

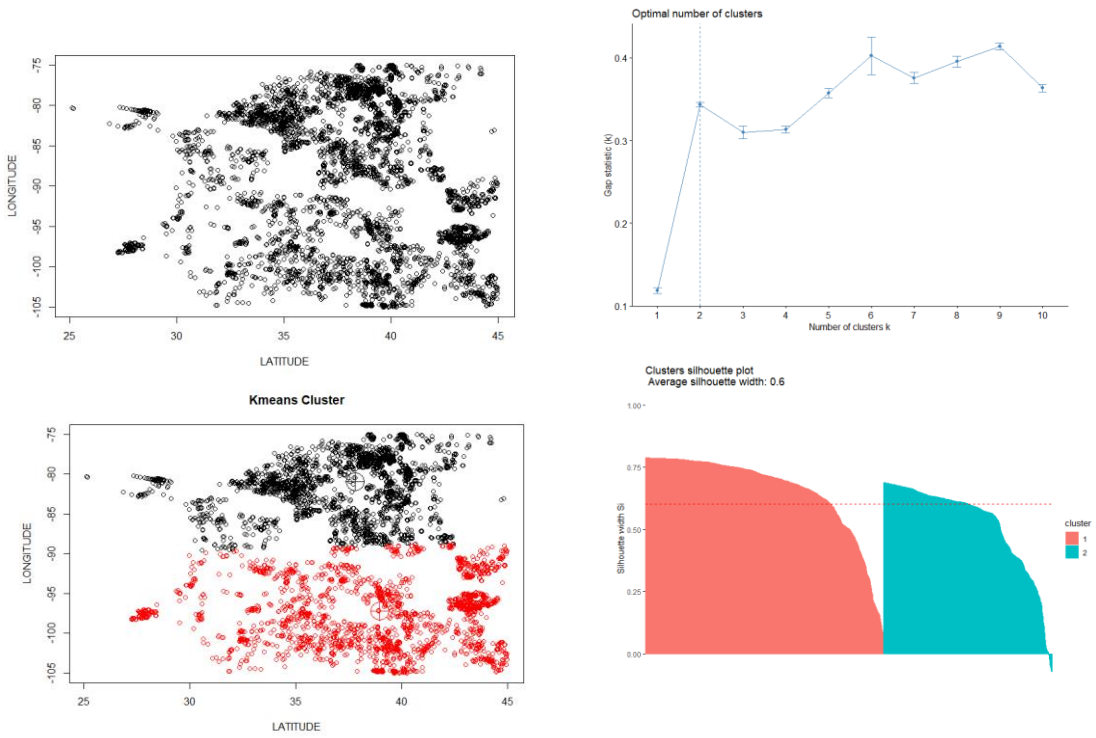

Fig.17 Four plot of the location data of 2018

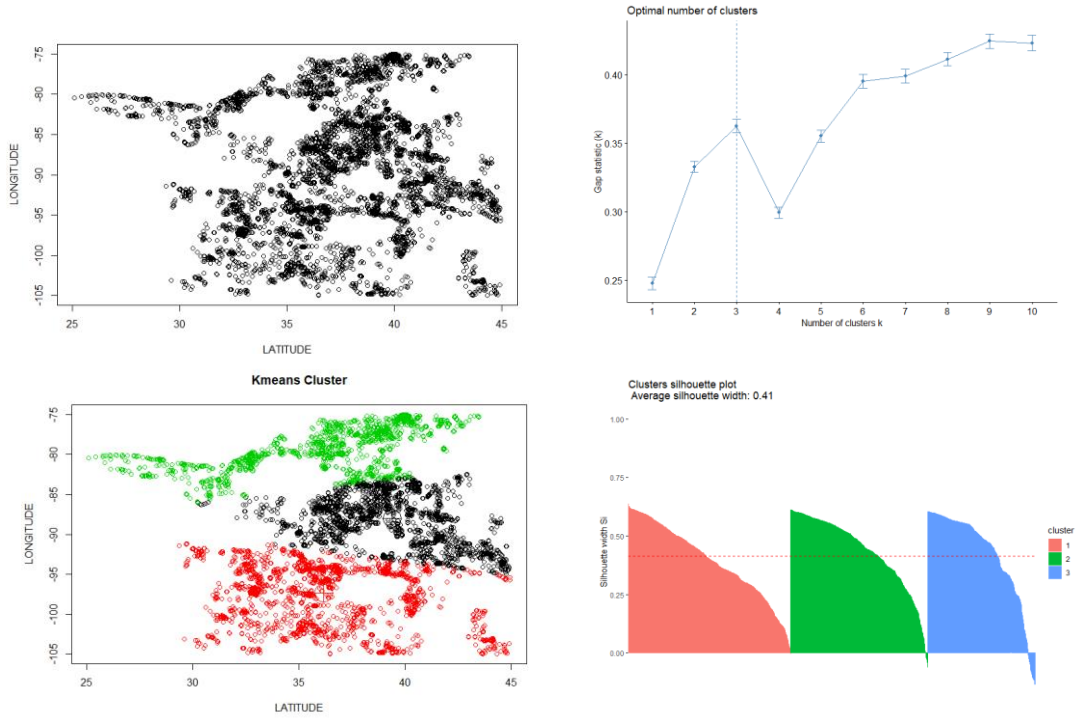

Fig.18 Four plot of the location data of 2019

Table 1 The parameters of clustering model

Data_year	Cluster number	center	size	Silhouette width	Average silhouette
2010	3	(36.65185, -83.04941)	3589	0.41	0.41
		(41.34990, -93.11074)	4487	0.48	
		(37.36356, -99.70159)	2583	0.27	
2011	3	(35.76666, -81.01120)	3420	0.47	0.42
		(38.33306, -89.70483)	3651	0.33	
		(39.91638, -98.56525)	3266	0.47	
2012	4	(38.74506, -79.12999)	1791	0.60	0.49
		(31.42750, -84.97389)	1078	0.41	
		(42.23354, -92.25361)	864	0.47	
		(36.94047, -100.63014)	1225	0.41	
2013	3	(37.60242, -79.57362)	3946	0.53	0.46
		(39.73934, -90.70168)	3566	0.41	
		(37.21955, -100.20577)	1902	0.42	
2014	3	(37.02389, -81.24997)	3322	0.49	0.44
		(41.23246, -92.74847)	4403	0.44	
		(37.64790, -100.36816)	2937	0.41	
2015	3	(38.89373, -78.42591)	2748	0.57	0.41
		(37.11346, -85.95348)	3317	0.28	
		(38.89373, -97.57223)	2748	0.57	

2016	2	(37.10599, -81.21204)	2593	0.6	0.57
		(35.79962, -97.14565)	1789	0.53	
2017	4	(38.78728, -81.16162)	1532	0.52	0.5
		(32.06778, -86.75856)	1311	0.41	
		(42.47976, -89.60717)	592	0.54	
		(37.62579, -100.09712)	1090	0.56	
2018	2	(37.79208, -80.91507)	4490	0.66	0.6
		(38.97090, -97.23462)	3181	0.52	
2019	3	(36.53840, -79.10956)	1971	0.42	0.41
		(39.39215, -87.96970)	2976	0.39	
		(36.17557, -97.32304)	2519	0.43	

From the table above, it's easy to find that based on ten datasets, the model has created three clusters for six datasets, two clusters for two datasets and four clusters for two clusters. Thus, three is the best number of clusters for the storm dataset. If the datasets with three clusters are focused, then three cluster centers can be gotten, which are (37.06754, -80.56880), (39.69604, -90.36194), (37.99728, -98.90951). Thus, the government rescue center can be set near these three coordinates in June, and then when a storm happens, the rescue team can depart to rescue people in the area quickly.

But this model still has problems. In this model, function silhouette() is used to validate the model and this function produces a value called silhouette width. If the clustering model can be applied successfully, the average silhouette width should be near to 1. However, in this model, none of the average silhouette width of these the datasets are near to 1. They are near to 0.5, which is not good enough. For these ten models, their total average silhouette width is 0.46. The main reason is that form the location plot, it's easy to find that these points are gather closely, which is not a good prerequisite for the clustering model. If the details in the location dataset can be provided more, then it's possible to remove some more useless points to make the clustering model. For example, since this clustering model is aimed to help set the rescue centers, if the

dataset can provide details of property damage and life danger, then maybe some storms which caused no damage can be removed from the datasets. What's more, another reason may be that the dataset has be removed some data, like the storm with 0 range. If these data can be provided accurate values and added into the model, then the silhouette width may rise.

5. Model: Decision Tree

In this model, 20-year dataset will be used, from 2000 to 2019. For the fatality dataset, the variables include fatality_year, fatality_month, fatality_day, fatality_type, fatality_age and fatality_location. In this model, fatality_location is chosen to be the dependent value and the rest are chosen as independent value. In this model, function rpart() is used to create the model and function prune() is used to avoid overfitting in the decision trees. The plot below has shown the decision tree.

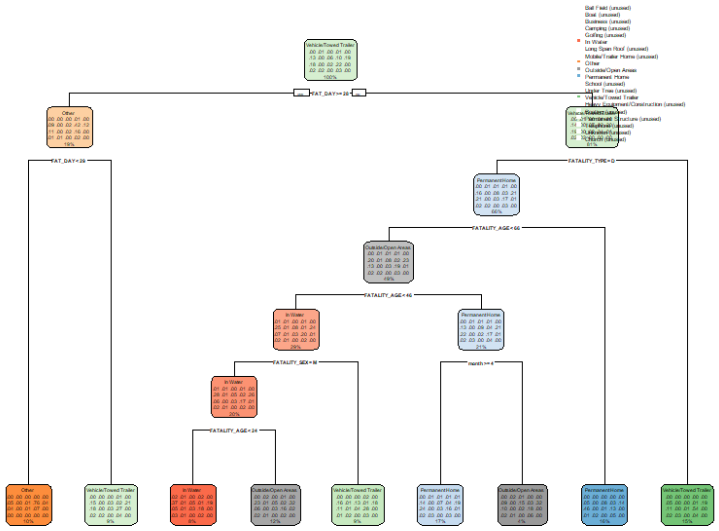

Fig.19 Decision tree based on the fatality data of 20 years from 2000 to 2019

The result has shown that root node error is relatively big, about 77.56%. The reason may be because only five kinds of fatality_location are used in the decision tree. They are "Vehicle/Towed Trailer", "Outside/Open Areas", "Permanent Home", "In Water" and "Other". These five kinds of fatality_location has occupied about 82% of all kinds of fatality_location. Thus, data with fatality_location which are not those five kinds will be removed and a new decision tree will be made.

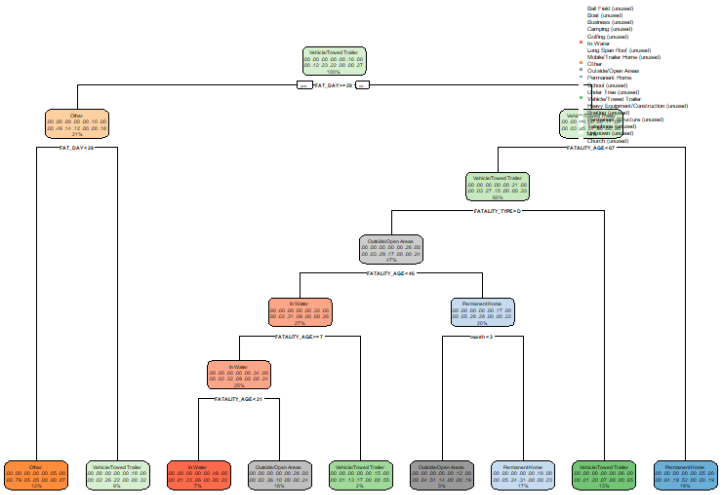

Fig.19 Decision tree based on the fatality data of 20 years from 2000 to 2019 after removing other fatality_location

After pruning the branch, this is the new decision tree. The root node error has improved, though it's still not so good. Now the root node error is about 72.67%.

Then, K-fold cross validation will be used to validate this model. The value of k in this validation is set to be 5. The decision tree models after pruning are shown below.

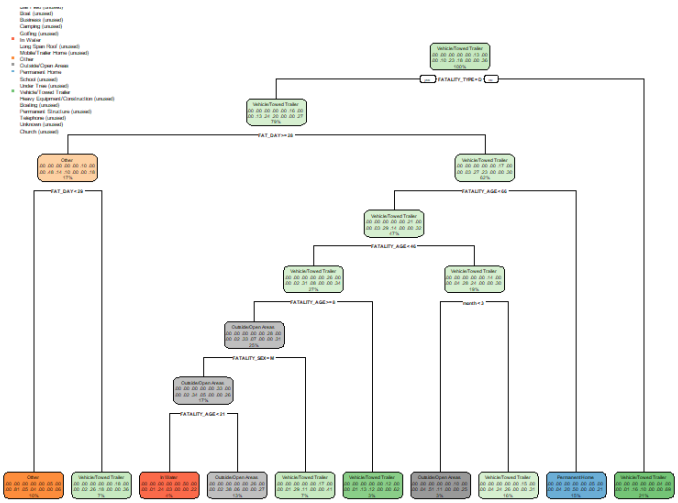

Fig.20 Decision tree based on training dataset1

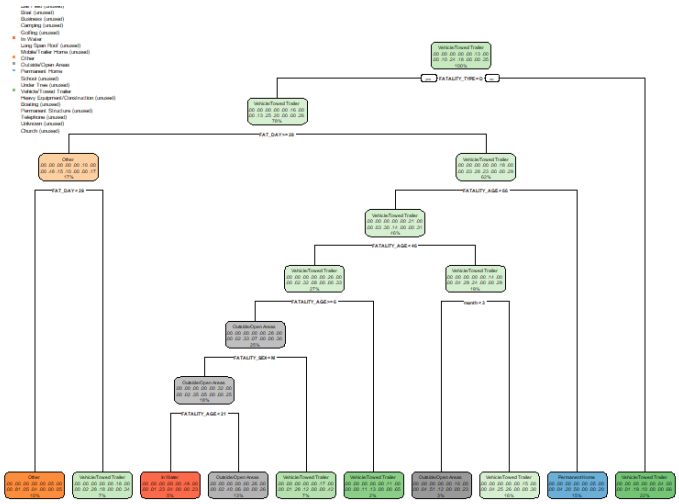

Fig.21 Decision tree based on training dataset2

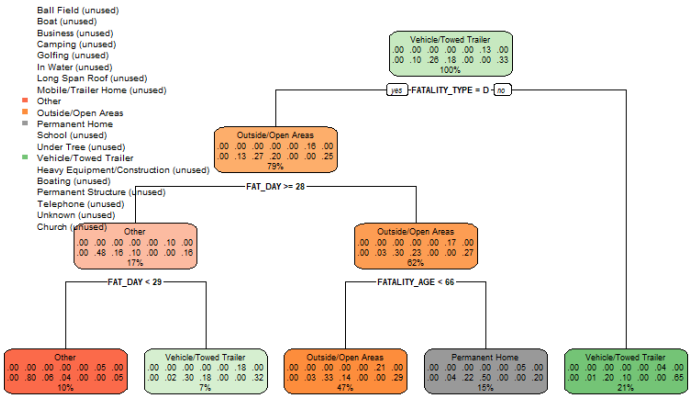

Fig.22 Decision tree based on training dataset3

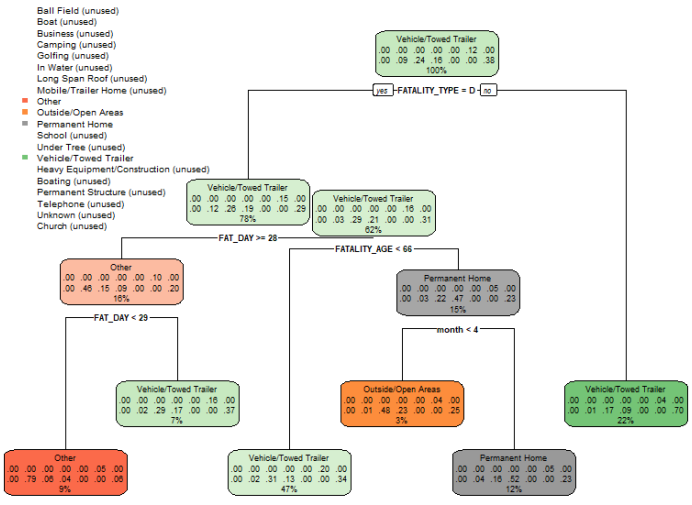

Fig.23 Decision tree based on training dataset4

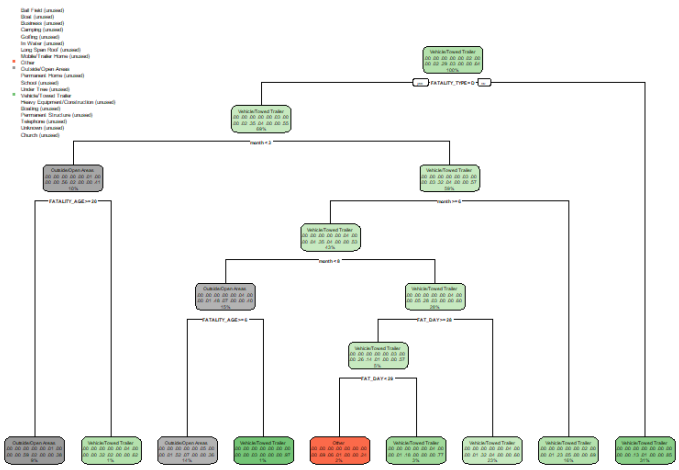

Fig.24 Decision tree based on training dataset5

Based on the training dataset1, 2, 5, the model gives the similar decision tree as the model based on the total data. The reason that the model based on training dataset3, 4 is difficult may be the particularity of dataset3 and 4. The five training datasets are created randomly by sample, so it's possible to create some special dataset.

The parameters of these models based on the five training datasets are shown below.

Table 2 The parameters of decision tree model

Dataset	Root node error	Prediction accuracy
1	64.07%	54.74%
2	65.17%	58.2%
3	67.04%	49.01%
4	62.09%	39.31%
5	35.7%	42.94%

The root node error is similar to the model based on the total data. The average root node error is about 58.81% and average prediction accuracy is about 48.84%. Both root node error and prediction accuracy are not so good. The reason may be the previous cleaning. Some data with N.A. fatality_age, zero fatality_age and N.A. fatality_sex is removed. If these data are not removed and provides some useful value, then the model may be more accurate.

This decision tree model is aimed to help rescue team to determine the fatality_location quickly based on the known information, so they can prepare certain equipment to rescue victims. However, this model doesn't work so well. So, the application of this model needs to be postponed until the model is improved.

6. Conclusion

In this project, I was proposed to use linear regression to predict the number of storms happening in the future. However, after trying, I found that the variables in the dataset have little relation with each other and after applying the linear regression model, the multiple R-squared are less than 0.01, far away from 1, which means linear regression is not fit to this dataset. So, I decided to use clustering and decision tree model to do the project.

In the clustering model, I use past-ten-year datasets and apply the model separately. It shows that it's better to make three clusters and the cluster center should be (37.06754, -80.56880), (39.69604, -90.36194), (37.99728, -98.90951). The problem is that after using function silhouette() to validate the model, the average silhouette width is far from 1, about 0.46. The reason may be because that the points are located too close, which makes the clustering model difficult to apply.

In the decision tree model, I use the past-20-year datasets. I apply the model with the dependent variable, fatality_location. First, I apply all data into the model and found only five kinds of fatality_location are used. The percent of five kinds of fatality_location is about 82%. The root node error is 77.56%. So, then I removed the other kinds of fatality_location and apply the model again. It shows the root node error has fallen to 72.67%. Based on the new model, I use K-fold cross validation to validate the model. It shows that average root node error is about 58.81% and prediction accuracy is about 48.84%. Both of them are not so well. The reason may be because I have removed some data with N.A. fatality_age, zero fatality_age and N.A. fatality_sex. Maybe those data with right value, the decision tree can be improved.

To improve the model, I want to find more information. I have a hypothesis. Now I can get two kinds of dataset. One includes the parameters of storms and the other includes the information of victims due to the storm. If NCDC can provide one dataset with the parameters of storms and the victims due to that storm together, maybe I can remove some storms which causes no loss of property and life. Then, I can apply the clustering model better and rescue team can focus on the storms which is danger to property and people.

7. Reference

[1] https://catalog.data.gov/dataset/ncdc-storm-events-database
[2] https://en.wikipedia.org/wiki/Storm
[3] https://uc-r.github.io/kmeans_clustering
[4] https://www.rdocumentation.org/packages/cluster/versions/2.1.0/topics/silhouette
[5] https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/

8. Github URL

https://github.com/nico37alonso/DataAnalyticsSpring2020Project