Report

Xinran Sun

5/7/2022

Introduction

Dataset

hurrican703.csv collected the track data of 702 hurricanes in the North Atlantic area from 1950 to 2013. For all the storms, their location (longitude & latitude) and maximum wind speed were recorded every 6 hours. The data includes the following variables

1. **ID**: ID of the hurricans
2. **Season**: In which **year** the hurricane occurred
3. **Month**: In which **month** the hurricane occurred
4. **Nature**: Nature of the hurricane

- ET: Extra Tropical
- DS: Disturbance
- NR: Not Rated
- SS: Sub Tropical
- TS: Tropical Storm

5. **time**: dates and time of the record
6. **Latitude** and **Longitude**: The location of a hurricane check point
7. **Wind.kt** Maximum wind speed (in Knot) at each check point

From the original dataset, we built a new dataset with contains five more variables, including:

1. **Wind_prev**: wind speed at 6 hours ago
2. **Wind_prev_prev**: wind speed at 12 hours ago
3. **Lat_change**: latitude change compared to 6 hours earlier
4. **Long_change**: longitude change compared to 6 hours earlier
5. **Wind_change**: wind speed change at 6 hours earlier compared to 12 hours earlier

These variables will help us to build the model in the following part.

The *hurricanoutcome2.csv* recorded the damages and death caused by 46 hurricanes in the U.S, and some features extracted from the hurricane records. The variables include:

1. **ID**: ID of the hurricans
2. **Season**: In which **year** the hurricane occurred
3. **Month**: In which **month** the hurricane occurred
4. **Nature**: Nature of the hurricane

- ET: Extra Tropical
- DS: Disturbance
- NR: Not Rated
- SS: Sub Tropical
- TS: Tropical Storm

5. **Damage**: Financial loss (in Billion U.S. dollars) caused by hurricanes
6. **Deaths**: Number of death caused by hurricanes
7. **Maxspeed**: Maximum recorded wind speed of the hurricane
8. **Meanspeed**: average wind speed of the hurricane
9. **Maxpressure**: Maximum recorded central pressure of the hurricane
10. **Meanpressure**: average central pressure of the hurricane
11. **Hours**: Duration of the hurricane in hours
12. **Total.Pop**: Total affected population
13. **Percent.Poor**: % affected population that reside in low GDP countres (i.e. GDP per Capita <= 10,000)
14. **Percent.USA**: % affected population that reside in the United States

EDA

We use a bar plot to examine the number of hurricanes in each month. From Figure 1, we can see that September is the month with the most hurricanes, while there are no hurricanes in February and March. Hurricanes in September also have the highest average wind speed as we can see in Figure 2.

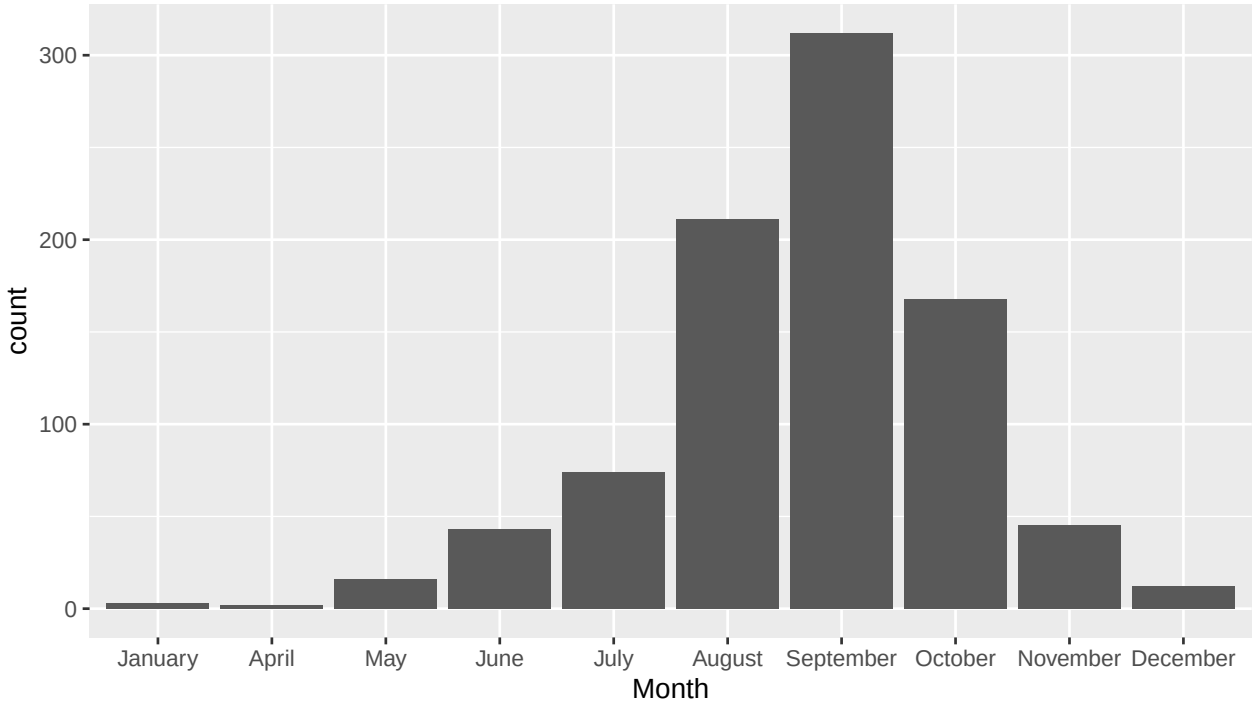


Figure 1. Count of Hurricanes in Each Month

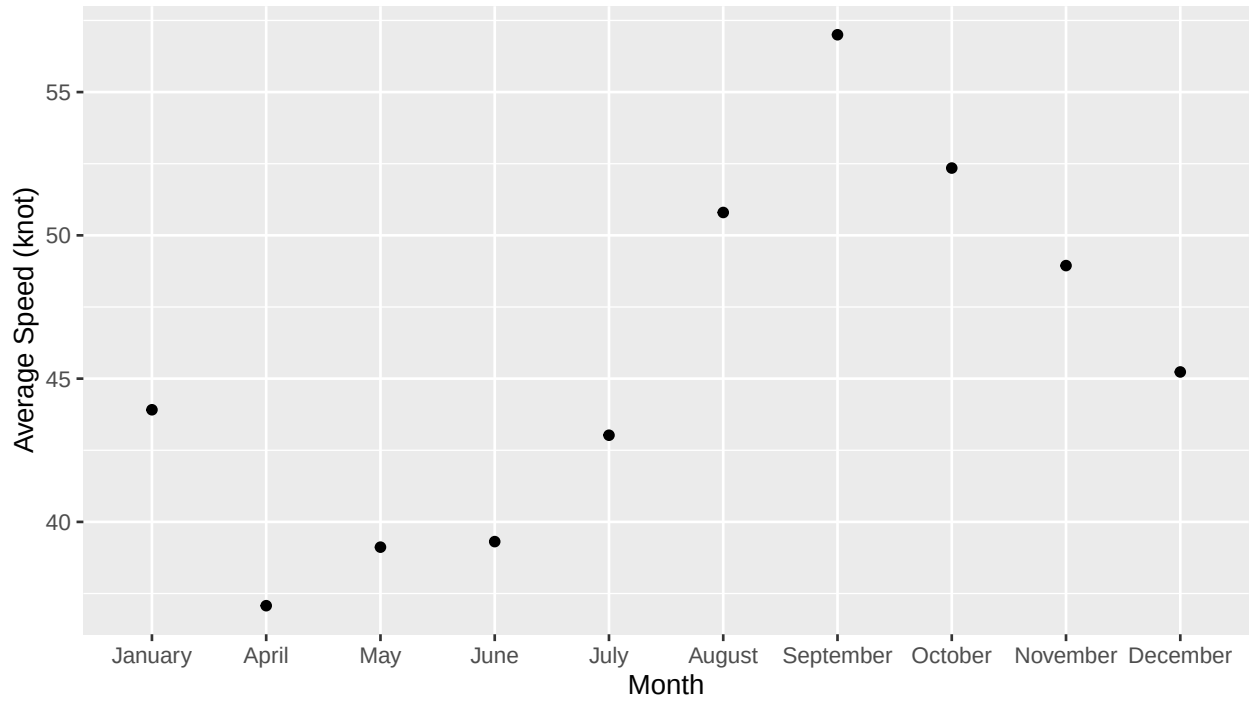


Figure 2. Average Speed (knot) of Hurricanes in Each Month

If we group the hurricanes by years, we can see in general, we have more observations in recently years compared to 50 years ago as shown in Figure 3. However, from Figure 4, the average wind speed seems to have a decreasing trend.

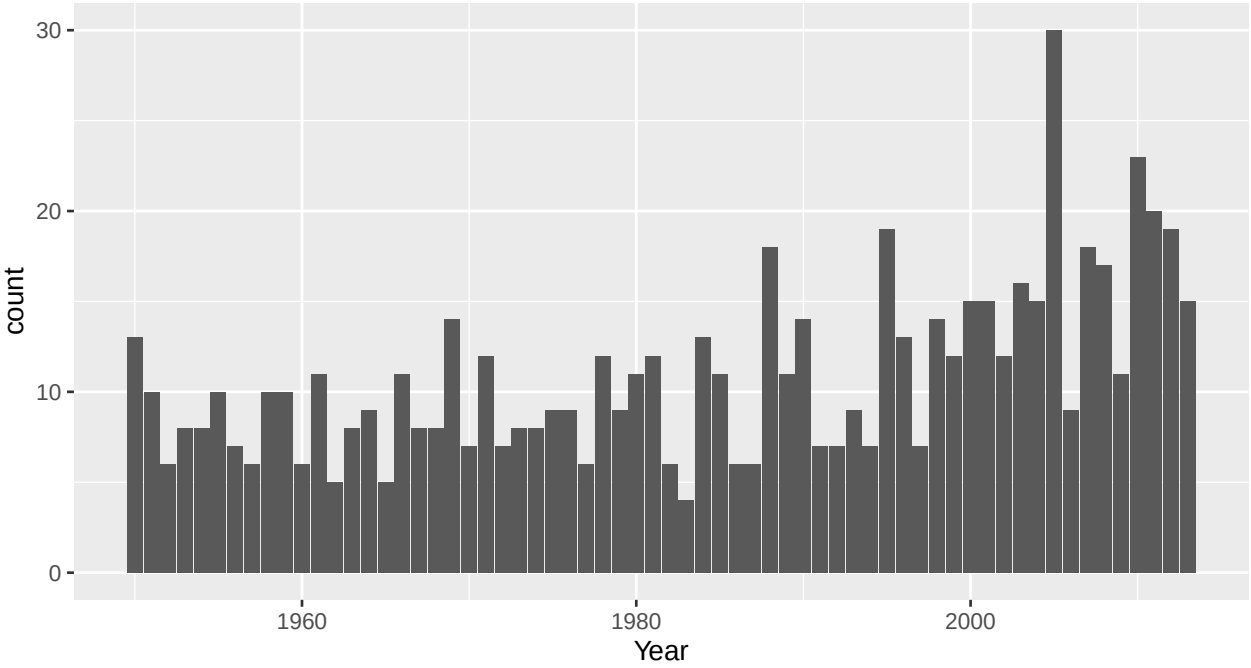


Figure 3. Count of Hurricanes in Each Year

```
## `geom_smooth()` using formula 'y ~ x'
```

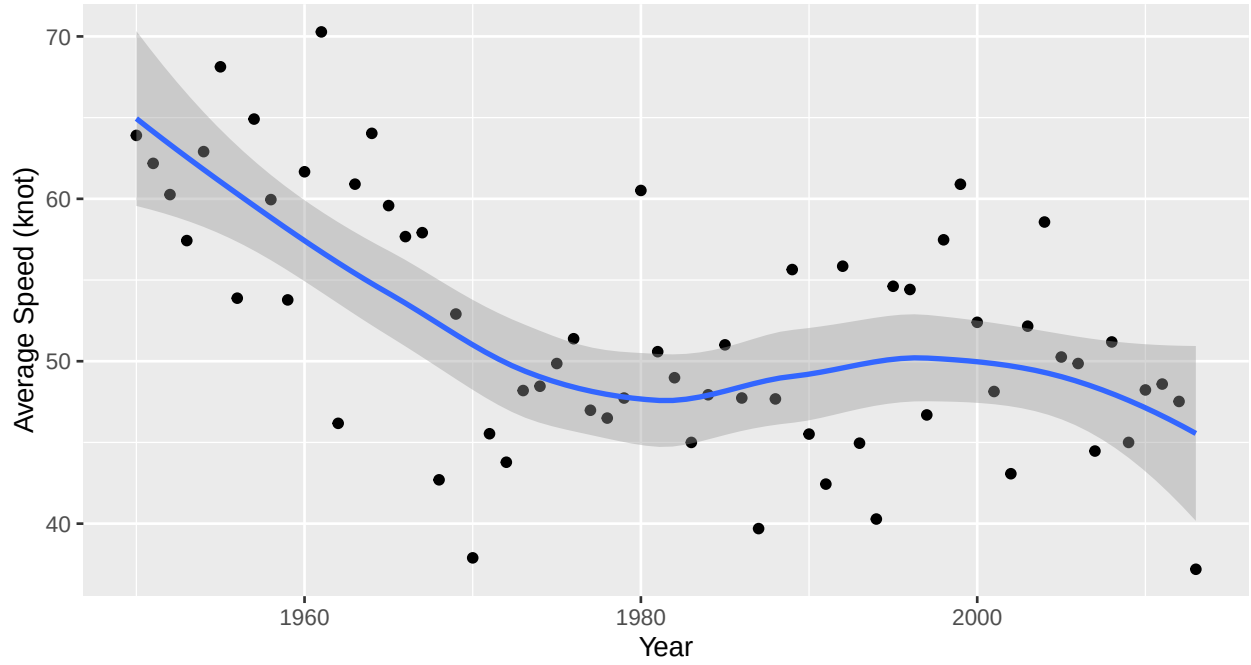


Figure 4. Average Speed (knot) of Hurricanes in Each Year

We also compare the hurricanes with different natures. In our dataset, there are 1214 different nature ratings. This number is larger than the number of hurricanes because some hurricanes are in different natures at different time. From Figure 5, we know that more than half of the natures are in Tropical Storm category. This nature also have the highest average wind speed at about 60 knot, while the disturbance and not rated hurricanes have average wind speed at round 20 knot as Figure 6 illustrates.

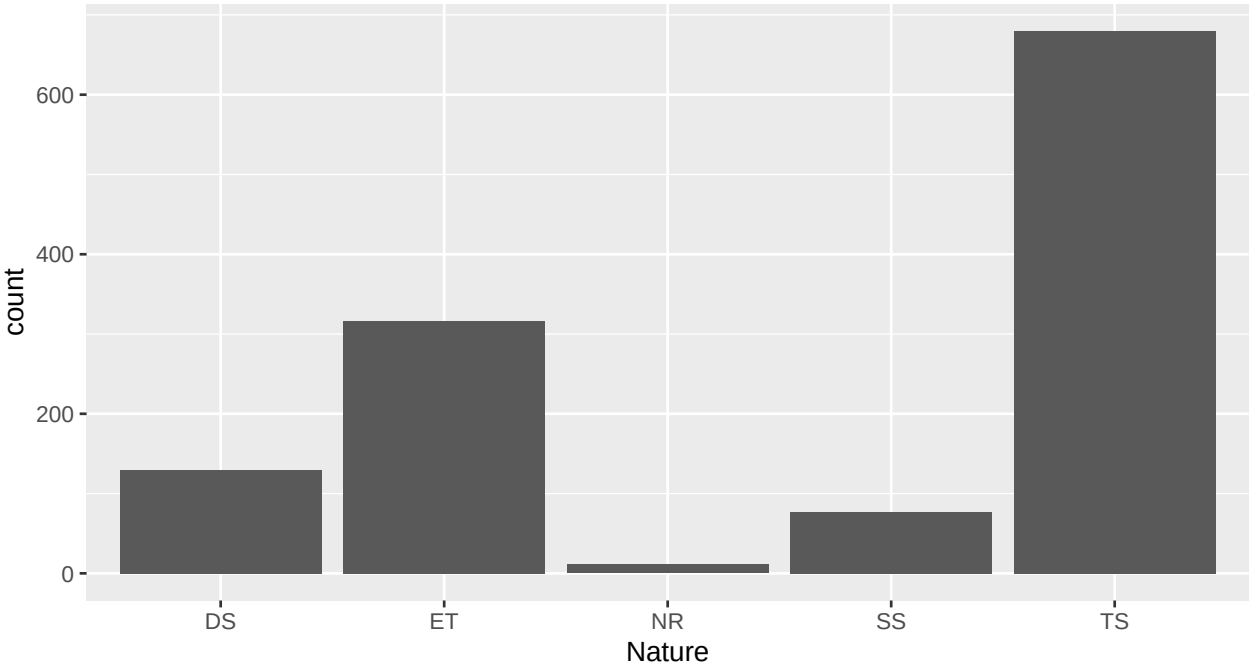


Figure 5. Count of Hurricanes in Each Nature

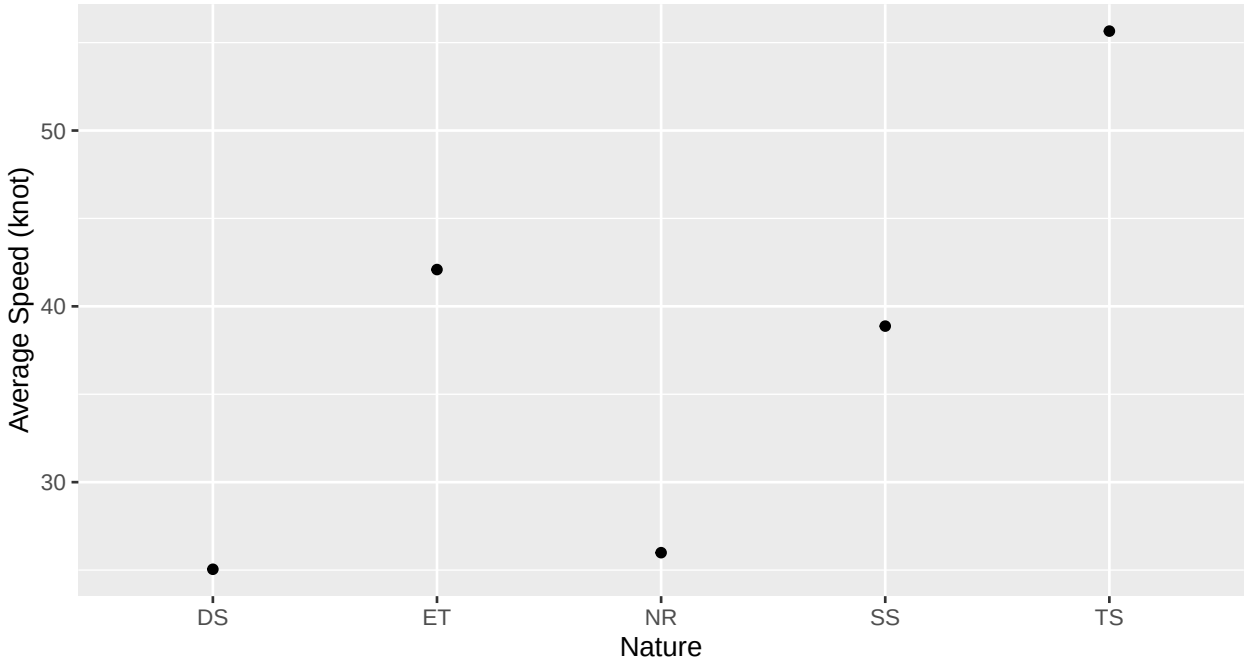


Figure 6. Average Speed (knot) of Hurricanes in Each Nature

Model Performance

We evaluate the performance of predictive ability by calculating the RMSE and the R^2 values for each hurricane. The residuals of Bayesian estimates that converged after iterations from MCMC will be used to predict the wind speed of test dataset. The overall R^2 is 0.822 and overall RMSE is 4.51. The valid R^2 is filtered with values between 0 and 1 and we get 77.5% hurriacanes (540) indicating that 22.5% of the estimated Bayesian models do not track hurricanes well and have negative R^2. One of the reason may be the limited number of observations of the hurricanes. Figure 7 shows the 10 hurriances with the least 10 RMSE. R^2 are also large enough to indicates that the estimated model track most hurricanes well and the smallest RMSE is GUSTAV.1996 with R^2 being 0.952.

ID	r_square	rmse
GUSTAV.1996	0.952	0.537
LORENZO.2001	0.914	0.733
ERIN.2013	0.878	0.823
JOSE.2011	0.970	0.872
GRETA.1970	0.980	0.876
DELTA.1972	0.825	0.904
EDITH.1967	0.826	0.983
FABIAN.1997	0.955	1.002
DEBBY.2006	0.984	1.045
CRISTOBAL.2002	0.956	1.053

Figure 7. R-square and RMSE for prediction result on test data

Figure 8 shows the actual wind speed and the estimated wind speed of randomly selected four hurricanes . We can see that most parts of the two curves overlapped indicating that the predicted values are close to the actual values. In DEBBY.2006, we can see that this is a very good model prediction with small deviation.

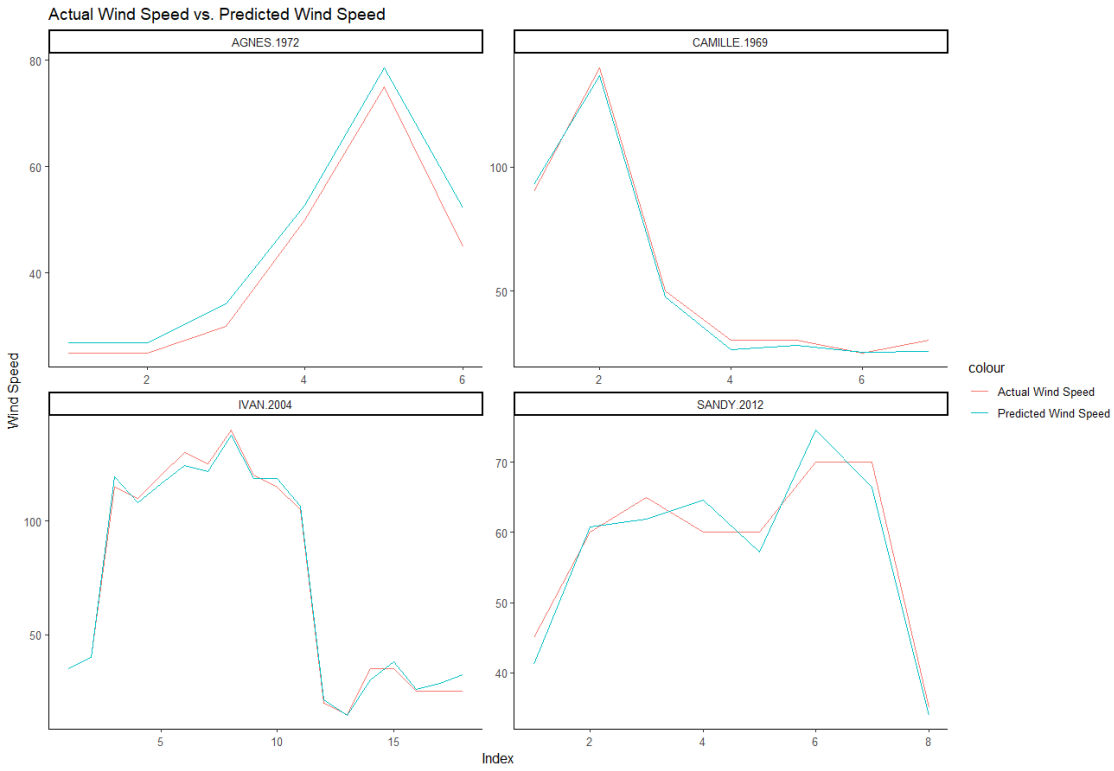
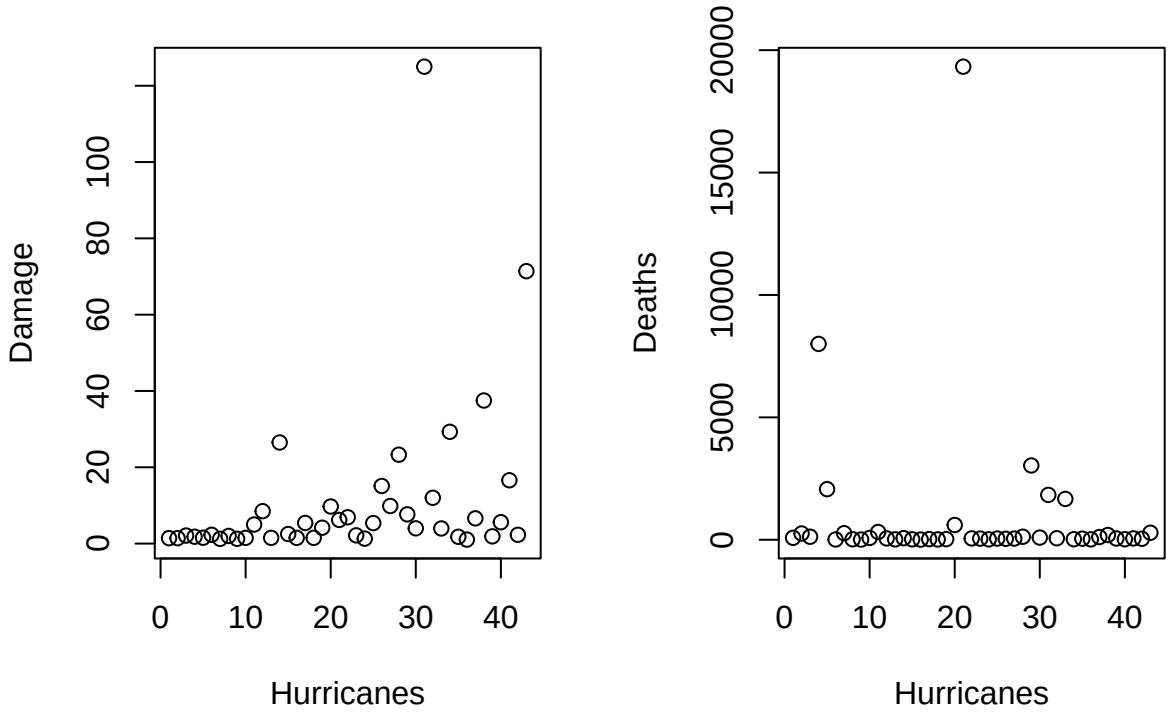


Figure 8. Actual Wind Speed vs. Predicted Wind Speed

Predict the hurricane-induced damage and deaths

Firstly, we plot deaths and financial loss separately. Figure 99. shows the distributions of deaths and damage. We could easily find that a few points which are far away from most of points indicate serious damage of society. In predictions of disasters, these extreme points are important because they enables the model the ability to predict the worst outcome. Therefore, we keep these points in model building.

Figure 99. Distributions of Damage and Deaths



In order to build a model that combines information in original data and the estimated coefficients from the Bayesian model, we extract the coefficients from the previous results. By taking the average of β_i at different time points, we obtain $\beta_0 \sim \beta_4$ of each hurricane. Part of the results is shown in Table 99.

Table 99. Coefficients of Each Hurricane

id	intercept	beta1	beta2	beta3	beta4
agnes.1972	3.950974	0.9224097	0.0059532	-0.3103372	0.5453543
alex.2010	3.798737	0.9370333	0.0698849	-0.3937358	0.5400187
alicia.1983	3.897408	0.9036878	-0.0748341	-0.3994486	0.5477718
allen.1980	3.687070	0.9655304	0.1306393	-0.5460144	0.5466129
andrew.1992	3.676279	0.9375384	-0.2843257	-0.5782973	0.5370158
betsy.1965	3.808396	0.9513766	-0.4500720	-0.3890718	0.4244575
bob.1991	3.629466	0.9232143	0.0279527	-0.5751636	0.4382048
camille.1969	3.994355	0.9355674	0.0729188	-0.5734830	0.6703910
charley.2004	3.638829	0.9482764	-0.1797332	-0.6955016	0.1818395
david.1979	3.789678	0.9579657	-0.0461134	-0.3823658	0.6853938

Fortunately, 43 hurricanes recorded in *hurricanoutcome*2.*csv* are also in *hurrican*703.*csv*. Thus, we merge two data frame by hurricane id to predict the deaths and damage caused by hurricanes.

The death variable is a count variable, so we decided to use Poisson regression to analysis relationship between death and other variables excluding damage. We use `Total.Pop` and `Hours` as the offset, since the outcome of deaths is proportional and the results would be different in some dimension (different populations, different duration). The Poisson regression is:

$$log(E(Deaths)) = \beta_i X_i + offset$$

Where X_i indicates all predictors included in the model. We use `glm` function to achieve the Poisson model. The coefficients result is in Table 99.

Table 99. Coefficients of Deaths Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	272.6548414	11.9772996	22.764300	0.0000000
intercept	12.8670811	0.2681866	47.978086	0.0000000
beta1	139.3302799	2.2527075	61.850143	0.0000000
beta2	6.2673041	0.1222299	51.274709	0.0000000
beta3	11.6385648	0.3090069	37.664416	0.0000000
beta4	-15.5433361	0.3093214	-50.249788	0.0000000
nobs	-0.0259659	0.0011052	-23.494615	0.0000000
Season	-0.0101636	0.0022226	-4.572917	0.0000048
MonthJuly	-2.6409629	0.1476034	-17.892292	0.0000000
MonthJune	-0.3043892	0.0911374	-3.339894	0.0008381
MonthNovember	-3.0007313	0.1534637	-19.553361	0.0000000
MonthOctober	-2.1031926	0.0625508	-33.623762	0.0000000
MonthSeptember	-0.6480044	0.0473666	-13.680606	0.0000000
NatureNR	2.1051263	0.1265128	16.639630	0.0000000
NatureTS	4.2677929	0.1266627	33.694166	0.0000000
Maxspeed	0.0099355	0.0010283	9.661642	0.0000000
Meanspeed	-0.0634120	0.0033070	-19.175107	0.0000000
Maxpressure	-0.4345774	0.0087466	-49.685178	0.0000000
Meanpressure	0.0092073	0.0002035	45.253527	0.0000000
Percent.Poor	0.0520424	0.0008541	60.929497	0.0000000
Percent.USA	-0.0206136	0.0005851	-35.231400	0.0000000

From the results, $\beta_0 \sim \beta_4$ indicate the relatively strong association. Especially, β_1, which represents the earlier wind speed has the biggest coefficient. We could conclude that high wind speed of hurricane more easily leads to serious casualties. Also, months seem to be an important factor in prediction. Comparing to June and September, July, November and October have lower proportion of death given all other variables constant.

In order to obtain the integer data, we transform the units of `Damage` from billion to million. Thus, `Damage` could be regarded as a count variable which could also be fitted by Poisson regression. In order to adjust the exposure, we use `Hours` as the offset.

$$log(E(Damage * 1000)) = \beta_i X_i + offset$$

Where X_i presents all predictors included in the model. We use `glm` function to achieve the Poisson model. The coefficients results is in Table 99.

term	estimate	std.error	statistic	p.value
(Intercept)	-206.5342295	2.0185627	-102.317470	0.00e+00
intercept	4.7792873	0.0279968	170.708107	0.00e+00
beta1	60.3768193	0.4513469	133.770326	0.00e+00
beta2	-1.0909958	0.0132414	-82.392801	0.00e+00
beta3	3.6396116	0.0259580	140.211430	0.00e+00
beta4	-1.6090883	0.0335258	-47.995549	0.00e+00
nobs	0.0308929	0.0002582	119.662746	0.00e+00
Season	0.0757660	0.0003973	190.691693	0.00e+00
MonthJuly	0.4806325	0.0187755	25.598871	0.00e+00
MonthJune	-3.2641224	0.0241511	-135.154056	0.00e+00
MonthNovember	-1.8378419	0.0249958	-73.526098	0.00e+00
MonthOctober	-1.3044789	0.0094414	-138.165832	0.00e+00
MonthSeptember	-1.7755988	0.0077805	-228.211660	0.00e+00
NatureNR	-4.2815939	0.0356423	-120.126833	0.00e+00
NatureTS	-1.9548375	0.0143805	-135.936287	0.00e+00

term	estimate	std.error	statistic	p.value
Maxspeed	0.0507898	0.0002142	237.112410	0.00e+00
Meanspeed	-0.0644452	0.0004911	-131.235042	0.00e+00
Maxpressure	-0.0140045	0.0012224	-11.456463	0.00e+00
Meanpressure	-0.0001769	0.0000407	-4.351453	1.35e-05
Total.Pop	0.0000003	0.0000000	62.112535	0.00e+00
Percent.Poor	-0.0384685	0.0001886	-203.998906	0.00e+00
Percent.USA	-0.0049715	0.0000729	-68.209623	0.00e+00

The results of coefficients in predicting damage also show the importance of β_1. From the model, we can see that serious casualties are also accompanied by serious financial losses which are strongly influenced by earlier wind speed and are slightly affected by months, latitude change, longitude change and wind speed change. $\beta_0 \sim \beta_4$ are generally powerful in damage and deaths prediction.

Appendix

Model Performance

```
beta.res.postmean = beta.res.postmean %>% rename(beta_0 = intercept,
                                                 beta_1 = Wind_prev,
                                                 beta_2 = Lat_change,
                                                 beta_3 = Long_change,
                                                 beta_4 = Wind_change)

dt_res = merge(dt_test_id, beta.res.postmean, by = "ID")

dt_res = dt_res %>%
  mutate(Wind_kt_pred = beta_0*intercept+beta_1*Wind_prev
         +beta_2*Lat_change+beta_3*Long_change+beta_4*Wind_change) %>%
  group_by(ID) %>%
  mutate(r_square = 1-(sum((Wind_kt_pred-Wind.kt)^2))/(sum((Wind.kt-mean(Wind.kt))^2)),
         rmse = rmse(Wind.kt,Wind_kt_pred))
```

```
  dt_rmse=
  dt_res %>%
  dplyr::select(ID, r_square, rmse) %>%
  distinct() %>%
  mutate(r_square = round(r_square, 3),
         rmse = round(rmse,3)) %>%
  filter(r_square > 0 && r_square < 1) %>%
  arrange(rmse)
mean(dt_rmse$r_square)
mean(dt_rmse$rmse)
```