P8160 - Project 3
Baysian modeling of hurricane

Renjie Wei, Hao Zheng, Xinran Sun
Wentong Liu, Shengzhi Luo

2022-05-09

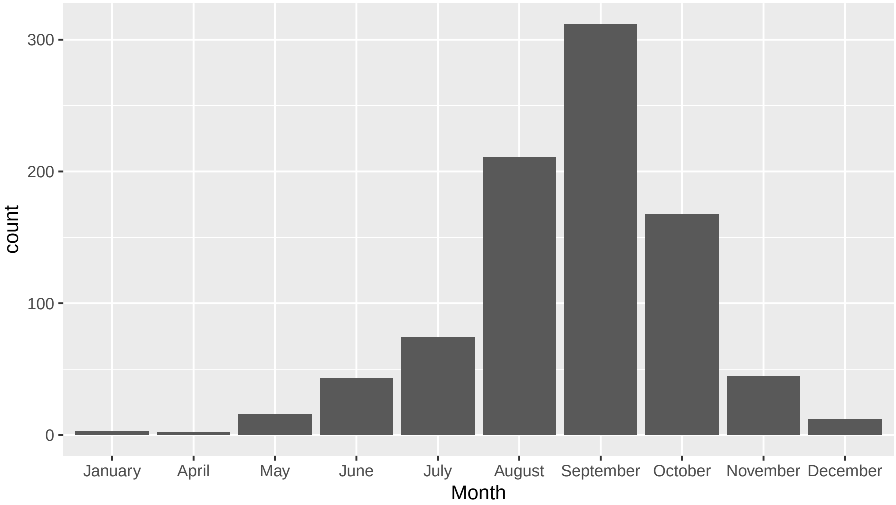
Introduction

- Hurricanes can result in death and economical damage
- There is an increasing desire to predict the speed and damage of the hurricanes
- Use Bayesian Model and Markov Chain Monte Carlo algorithm
  - Predict the wind speed of hurricanes
  - Study how hurricanes is related to death and financial loss

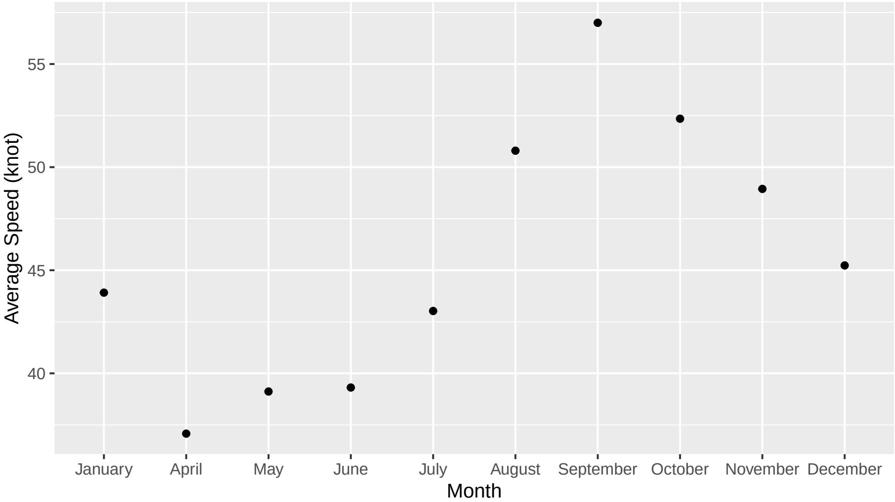
Dataset

- Hurrican703 dataset: 22038 observations $\times$ 8 variables
    - 702 hurricanes in the North Atlantic area in year 1950-2013
- Processed dataset: add 5 more variables into hurrican703
- Hurricanoutcome2 dataset: 43 observations $\times$ 14 variables

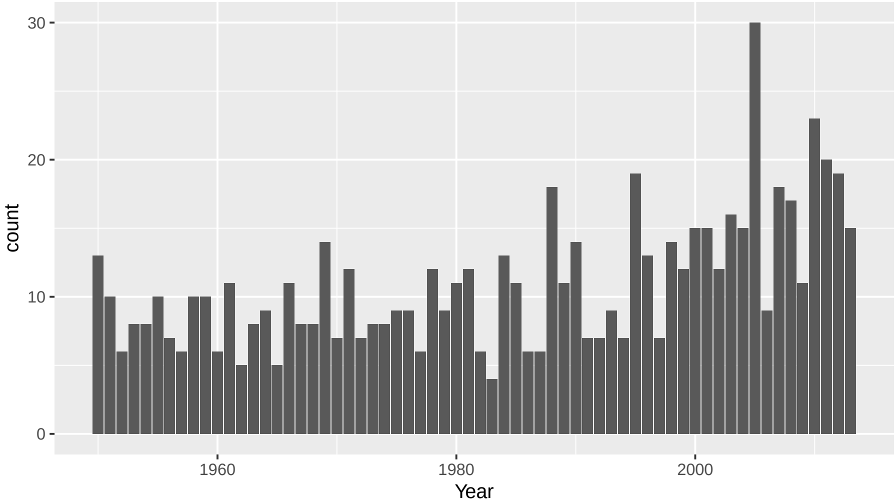
EDA - Count of Hurricanes in Each Month



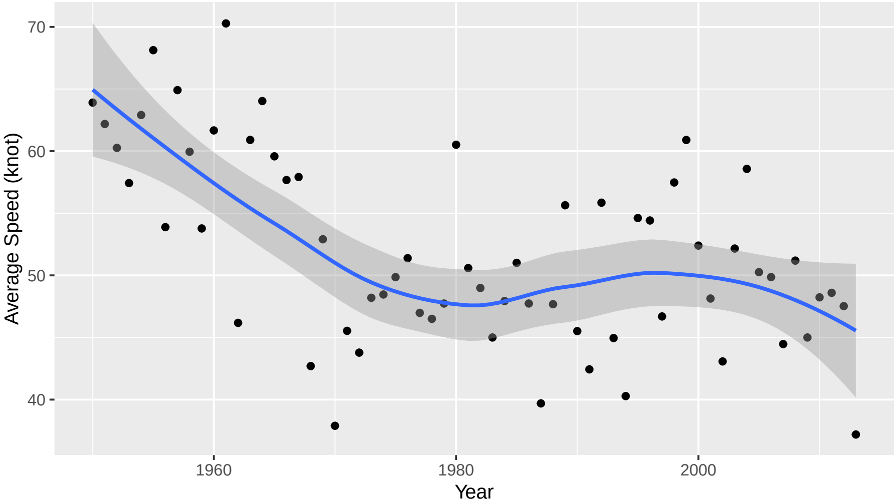
EDA - Average Speed (knot) of Hurricanes in Each Month



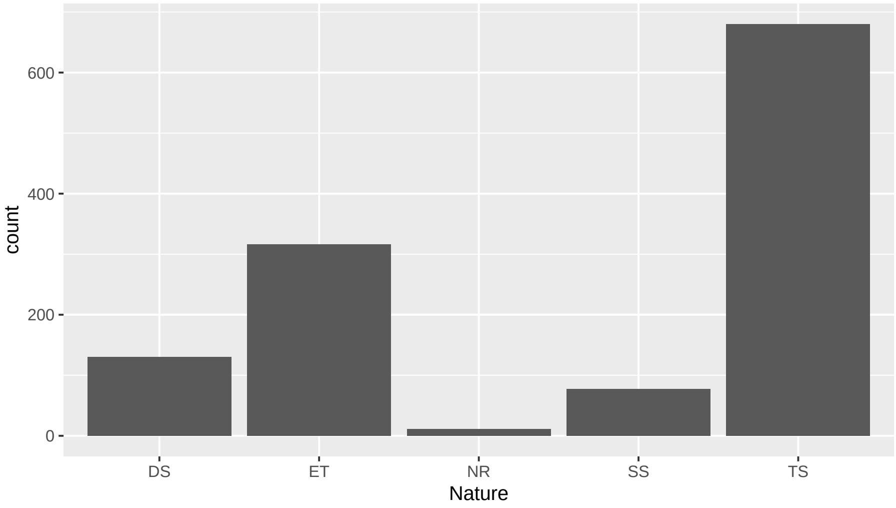
EDA - Count of Hurricanes in Each Year



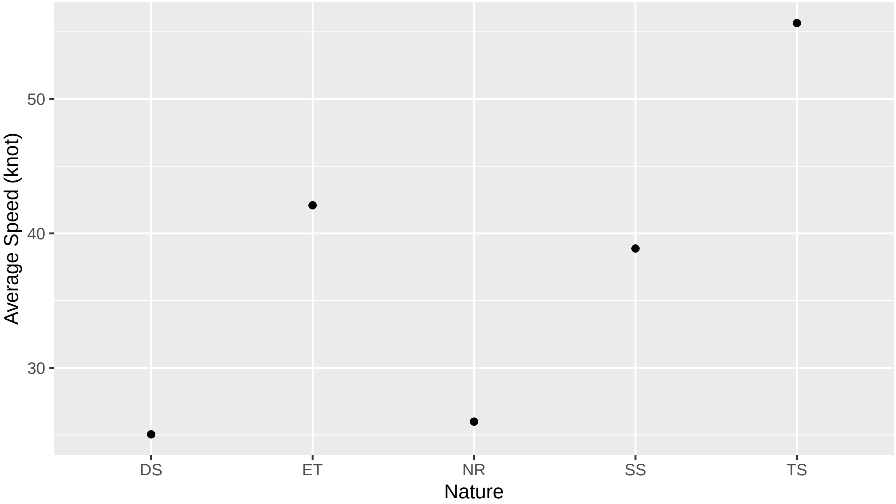
EDA - Average Speed (knot) of Hurricanes in Each Year



EDA - Count of Hurricanes in Each Nature



EDA - Average Speed (knot) of Hurricanes in Each Nature



Bayesian Model Setting

Model

The suggested Bayesian model is

$Y_i(t+6) = \beta_{0,i} + \beta_{1,i}Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t) + \epsilon_i(t)$

- where $Y_i(t)$ the wind speed at time t (i.e. 6 hours earlier), $\Delta_{i,1}(t)$, $\Delta_{i,2}(t)$ and $\Delta_{i,3}(t)$ are the changes of latitude, longitude and wind speed between t and $t-6$, and $\epsilon_{i,t}$ follows a normal distributions with mean zero and variance σ^2, independent across t.
- $\beta_i = (\beta_{0,i}, \beta_{1,i}, ..., \beta_{5,i})$, we assume that $\beta_i \sim N(\mu, \Sigma_{d\times d})$, where d is dimension of β_i.

Priors

$$P(\sigma^2) \propto \frac{1}{\sigma^2}; \quad P(\mu) \propto 1; \quad P(\Sigma^{-1}) \propto |\Sigma|^{-(d+1)} \exp(-\frac{1}{2}\Sigma^{-1})$$

Posterior

- Derive $\pi(\Theta|Y)$, where $\Theta = (\mathbf{B}^\top, \mu^\top, \sigma^2, \Sigma)$, $\mathbf{B} = (\beta_1^\top, ..., \beta_n^\top)^\top$

Joint posterior

Notations

- $X_i(t)\beta_i^\top = \beta_{0,i} + \beta_{1,i}Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t)$
- For i^{th} hurricane, there may be m_i times of record (excluding the first and second observation), let

$$Y_i = \begin{pmatrix} Y_i(t_0+6) \\ Y_i(t_1+6) \\ \vdots \\ Y_i(t_{m_i-1}+6) \end{pmatrix}_{m_i \times 1}$$

- Hence, $Y_i \mid X_i, \beta_i, \sigma^2 \sim N(X_i\beta_i^\top, \sigma^2 I)$
- Where, X_i is a $m_i \times d$ dimensional matrix

$$X_i = \begin{pmatrix} 1 & Y_i(t_0) & \Delta_{i,1}(t_0) & \Delta_{i,2}(t_0) & \Delta_{i,3}(t_0) \\ 1 & Y_i(t_1) & \Delta_{i,1}(t_1) & \Delta_{i,2}(t_1) & \Delta_{i,3}(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Y_i(t_{m_i-1}) & \Delta_{i,1}(t_{m_i-1}) & \Delta_{i,2}(t_{m_i-1}) & \Delta_{i,3}(t_{m_i-1}) \end{pmatrix}$$

Joint posterior

Posterior

$$\begin{aligned}
\pi(\Theta|Y) &= \pi(\mathbf{B}^\top, \mu^\top, \sigma^2, \Sigma \mid Y) \\
&\propto \underbrace{\prod_{i=1}^{n} f(Y_i \mid \beta_i, \sigma^2)}_{\text{likelihood of } Y} \underbrace{\prod_{i=1}^{n} \pi(\beta_i \mid \mu, \Sigma)}_{\text{distribution of } \mathbf{B}} \underbrace{P(\sigma^2)P(\mu)P(\Sigma^{-1})}_{\text{priors}} \\
&\propto \prod_{i=1}^{n} \Big\{ (2\pi\sigma^2)^{-m_i/2} \exp\big\{ -\frac{1}{2}(Y_i - X_i\beta_i^\top)^\top (\sigma^2 I)^{-1} (Y_i - X_i\beta_i^\top) \\
&\times \prod_{i=1}^{n} \Big\{ \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\big\{ -\frac{1}{2}(\beta_i - \mu)\Sigma^{-1}(\beta_i - \mu)^\top \big\} \Big\} \\
&\times \frac{1}{\sigma^2} \times \det(\Sigma)^{-(d+1)} \exp\big\{ -\frac{1}{2}\Sigma^{-1} \big\}
\end{aligned}$$

MCMC Algorithm

- Monte Carlo Method
  - Random sampling method to estimate quantity
- Markov Chain
  - Generates a sequence of random variables where the current state only depends on the nearest past
- Example: Gibbs Sampler
  - MCMC approaches with known conditional distributions
  - Samples from each random variables in turn given the value of all the others in the distribution

Conditional Posterior

- To apply MCMC using Gibbs sampling, we need to find conditional posterior distribution of each parameter, then we can implement Gibbs sampling on these conditional posterior distributions.
  - $\pi(\mathbf{B}|Y, \mu^\top, \sigma^2, \Sigma)$
  - $\pi(\sigma^2|Y, \mathbf{B}^\top, \mu^\top, \Sigma)$
  - $\pi(\Sigma|Y, \mathbf{B}^\top, \mu^\top, \sigma^2)$
  - $\pi(\mu|Y, \mathbf{B}^\top, \sigma^2, \Sigma)$

MCMC Algorithm - Conditional Posterior

- β_i: $\pi(\beta_i | Y, \mu^\top, \sigma^2, \Sigma) \sim \mathcal{N}(\hat{\beta}_i, \hat{\Sigma}_{\beta_i})$
  - where $\hat{\beta}_i = (\Sigma^{-1} + X_i^\top (\sigma^2 I)^{-1} X_i)^{-1} Y_i^\top (\sigma^2 I)^{-1} X_i + \mu \Sigma^{-1}, \hat{\Sigma}_{\beta_i} = (\Sigma^{-1} + X_i^\top (\sigma^2 I)^{-1} X_i)^{-1}$
- σ^2: $\pi(\sigma^2 | Y, \mathbf{B}^\top, \mu^\top, \Sigma) \sim IG(\frac{1}{2} \sum_{i=1}^{n} m_i, \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i \beta_i^\top)^\top (Y_i - X_i \beta_i^\top))$
- Σ: $\pi(\Sigma | Y, \mathbf{B}^\top, \mu^\top, \sigma^2) \sim IW(n + d + 1,\ I + \sum_{i=1}^{n} (\beta_i - \mu)(\beta_i - \mu)^\top)$
- μ: $\pi(\mu | Y, \mathbf{B}^\top, \sigma^2, \Sigma) \sim \mathcal{N}(\frac{1}{n} \sum_{i=1}^{n} \beta_i, \frac{1}{n} \Sigma)$

MCMC Algorithm - Parameter Updates

The update of parameters is component wise, at $(t+1)^{\text{th}}$ step, updating parameters in the following the order:

1. Sample $\mathbf{B}^{(t+1)}$, i.e., sample each $\beta_i^{(t+1)}$ from $\mathcal{N}(\hat{\beta}_i^{(t)}, \hat{\Sigma}_{\beta_i}^{(t)})$
2. Then, sample σ^2 from
$IG(\frac{1}{2}\sum_{i=1}^{n} m_i, \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\beta_i^{(t+1)^\top})^\top(Y_i - X_i\beta_i^{(t+1)^\top}))$
3. Next, sample $\Sigma^{(t+1)}$ from
$IW(n+d+1,\ I + \sum_{i=1}^{n}(\beta_i^{(t+1)} - \mu^{(t)})(\beta_i^{(t+1)} - \mu^{(t)})^\top)$
4. Finally, sample $\mu^{(t+1)}$ from $\mathcal{N}(\frac{1}{n}\sum_{i=1}^{n}\beta_i^{(t+1)}, \frac{1}{n}\Sigma^{(t+1)})$

MCMC Algorithm - Train-Test split and Inital Values

Train-test split

- Drop the data of hurricane with less than 3 observations. Results in 697 hurricanes
- Within each hurricane's data, randomly 80% train, 20% test

Initial Values

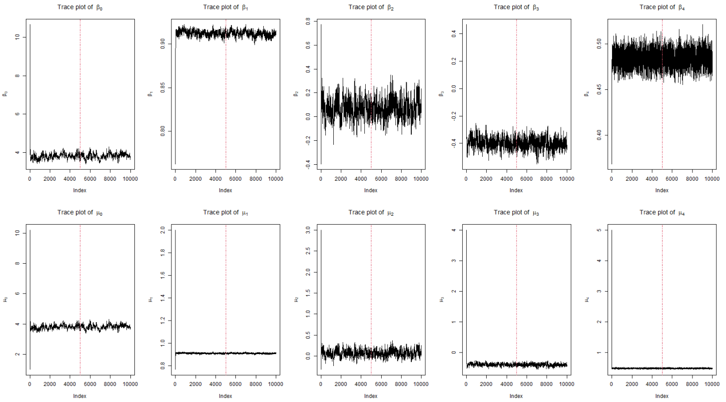
1. For initial value of $\mathbf{B}$, we run multivariate linear regressions for each hurricane and use the regression coefficients β_i^{MLR} as the initial value for β_i. Then, the initial value of $\mathbf{B}$ can be represented as $\mathbf{B}_{init} = (\beta_1^{MLR^\top}, \dots, \beta_n^{MLR^\top})^\top$.
2. For initial value of μ, we take the average of β_i^{MLR}, that is $\mu_{init} = \frac{1}{n}\sum_{i=1}^{n} \beta_n^{MLR}$
3. For initial value of σ^2, we take the average of the MSE for i hurricanes.
4. For initial value of Σ, we just set it to a simple diagonal matrix, i.e. $\Sigma_{init} = diag(1, 2, 3, 4, 5)$

MCMC Results

Details

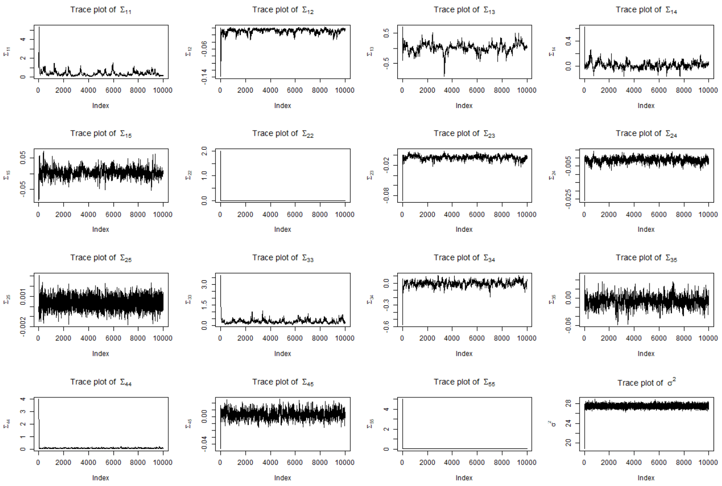
- 10000 iterations
- First 5000 iterations as burn-in period
- Estimates and inferences based on last 5000 MCMC samples

MCMC Results - Trace Plots 1



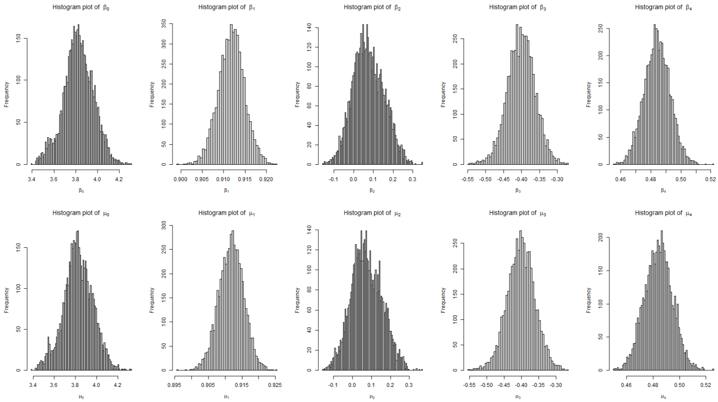
Trace plots of model parameters, based on 10000 MCMC sample

MCMC Results - Trace Plots 2



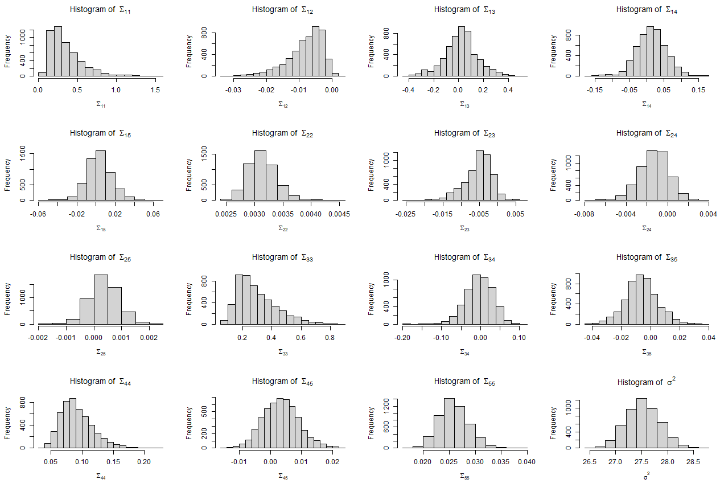
Trace plots of variance parameters, based on 10000 MCMC sample

MCMC Results - Histograms 1



Histograms of model parameters, based on last 5000 MCMC sample

MCMC Results - Histograms 2



Histograms of variance parameters, based on last 5000 MCMC sample

MCMC Results - Model Parameter Estimations and Inferences

Variables	$\bar{\beta}_i$	Var($\bar{\beta}_i$)	95% CI of $\bar{\beta}_i$	$\bar{\mu}$	Var($\bar{\mu}$)	95% CI of $\bar{\mu}$
intercept	3.8252	0.0185	(3.5588,4.0916)	3.8166	0.0190	(3.5468,4.0865)
Wind_prev	0.9118	0.0000	(0.9059,0.9177)	0.9121	0.0000	(0.9049,0.9194)
Lat_change	0.0744	0.0060	(-0.0776,0.2264)	0.0720	0.0065	(-0.0857,0.2298)
Long_change	-0.4014	0.0015	(-0.4771,-0.3257)	-0.3968	0.0016	(-0.4759,-0.3177)
Wind_change	0.4841	0.0001	(0.4674,0.5009)	0.4847	0.0001	(0.464,0.5053)

Bayesian posterior estimates for model parameters

MCMC Results - Variance Parameter Estimations and Inferences

$$\Sigma = \begin{pmatrix} 0.349 & -0.008 & 0.020 & 0.013 & 0.004 \\ -0.008 & 0.003 & -0.005 & -0.001 & 0.0004 \\ 0.020 & -0.005 & 0.296 & -0.003 & -0.006 \\ 0.013 & -0.001 & -0.003 & 0.092 & 0.003 \\ 0.004 & 0.0004 & -0.006 & 0.003 & 0.026 \end{pmatrix}$$

$$\rho = \begin{pmatrix} 1 & -0.245 & 0.063 & 0.073 & 0.037 \\ -0.245 & 1 & -0.174 & -0.078 & 0.041 \\ 0.063 & -0.174 & 1 & -0.019 & -0.069 \\ 0.073 & -0.078 & -0.019 & 1 & 0.070 \\ 0.037 & 0.041 & -0.069 & 0.070 & 1 \end{pmatrix}$$

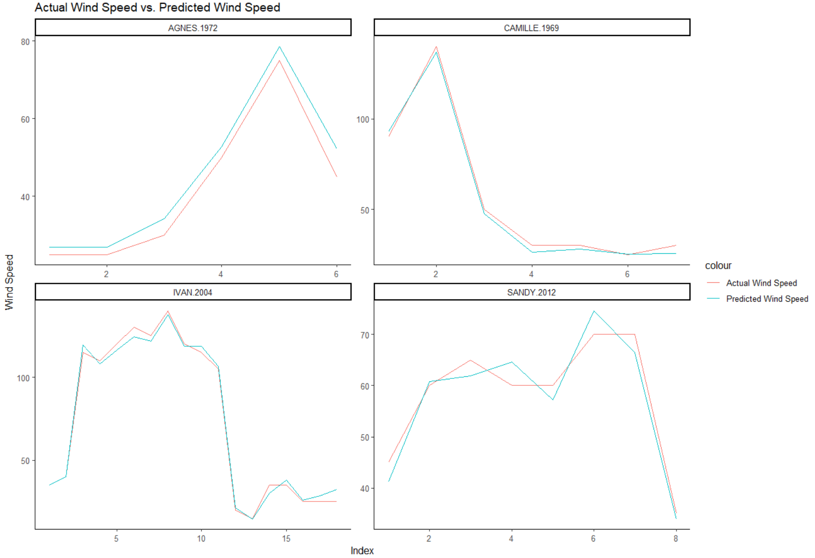
Bayesian Model Performance

- The overall mean R^2 is 0.82245
- The overall mean RMSE is 4.51023

Table 1: R-square and RMSE for prediction result on test data

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

Bayesian Model Performance



Actual Wind Speed vs. Predicted Wind Speed

Seasonal Difference Exploration and Wind Change against Year

- Bayesion model:

$$Y_i(t+6) = \beta_{0,i} + \beta_{1,i} Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t) + \epsilon_i(t)$$

- Model 1: $\beta_j \sim Month + Year + Nature$
- Model 2: $\beta_j \sim Season$
- Model 3: $\beta_j \sim Year$
- β_j corresponds to $\beta_0 \sim \beta_4$ in the Bayesian model

Seasonal Difference Exploration - Model 1

	$beta_0$		$beta_1$		$beta_2$		$beta_3$		$beta_4$	
	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
(Intercept)	4.481	0.000	1.343	0.000	0.041	0.951	-0.834	0.019	0.289	0.448
monthApril	0.023	0.835	0.015	0.670	0.017	0.931	0.042	0.680	0.036	0.739
monthMay	0.026	0.783	0.000	0.997	0.071	0.660	0.063	0.458	-0.016	0.859
monthJune	0.028	0.765	0.005	0.851	-0.007	0.964	0.056	0.505	0.024	0.792
monthJuly	0.013	0.891	0.015	0.590	-0.009	0.954	0.036	0.664	0.013	0.884
monthAugust	-0.020	0.828	0.023	0.412	-0.052	0.738	0.012	0.881	0.031	0.726
monthSeptember	-0.007	0.938	0.026	0.359	-0.036	0.817	0.021	0.797	0.044	0.618
monthOctober	0.009	0.919	0.021	0.459	-0.029	0.855	0.034	0.680	0.035	0.694
monthNovember	0.015	0.875	0.025	0.393	0.024	0.879	0.026	0.753	0.021	0.817
monthDecember	0.006	0.953	0.009	0.772	-0.054	0.745	0.042	0.633	0.011	0.905
year	0.000	0.072	0.000	0.000	0.000	0.910	0.000	0.203	0.000	0.625
natureET	0.001	0.977	0.004	0.688	-0.070	0.169	-0.026	0.329	-0.021	0.473
natureNR	0.001	0.987	-0.015	0.333	0.006	0.943	0.003	0.944	-0.022	0.646
natureSS	0.014	0.490	-0.003	0.602	-0.001	0.969	0.013	0.496	-0.024	0.234
natureTS	0.012	0.479	-0.006	0.249	-0.015	0.588	-0.023	0.126	-0.017	0.283

- The effect the previous wind speed has will decrease over years

Model 2

- Regress β_j only using Season as predictor

response	coefficient	Estimate	Pr(>|t|)
β_0	Intercept	3.837	0.000
β_0	seasonSummer	-0.031	0.205
β_0	seasonAutumn	-0.024	0.325
β_0	seasonWinter	-0.019	0.654
β_1	Intercept	0.894	0.000
β_1	seasonSummer	0.015	0.044
β_1	seasonAutumn	0.021	0.005
β_1	seasonWinter	0.003	0.794
β_2	Intercept	0.161	0.000
β_2	seasonSummer	-0.098	0.017
β_2	seasonAutumn	-0.091	0.025
β_2	seasonWinter	-0.098	0.164
β_3	Intercept	-0.350	0.000
β_3	seasonSummer	-0.047	0.034
β_3	seasonAutumn	-0.043	0.046
β_3	seasonWinter	-0.009	0.802
β_4	Intercept	0.442	0.000
β_4	seasonSummer	0.036	0.120
β_4	seasonAutumn	0.049	0.035
β_4	seasonWinter	0.015	0.711

- Effects $Y_{i,t}$ and $\Delta_{i,3}(t)$ has on the wind speed change across seasons

Model 3

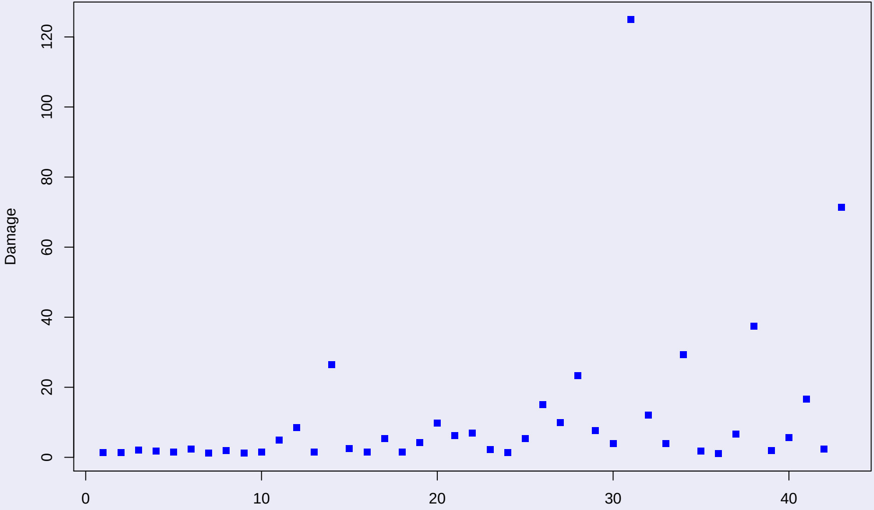
- Regress β_j only using Year as predictor

response	coefficient	Estimate	Pr(>|t|)
β_0	Intercept	4.514	0.000
β_0	year	0.000	0.050
β_1	Intercept	1.345	0.000
β_1	year	0.000	0.000
β_2	Intercept	-0.106	0.863
β_2	year	0.000	0.776
β_3	Intercept	-1.027	0.002
β_3	year	0.000	0.053
β_4	Intercept	0.305	0.382
β_4	year	0.000	0.607

- The impact of the nearest past wind speed has on current wind speed will decrease across years

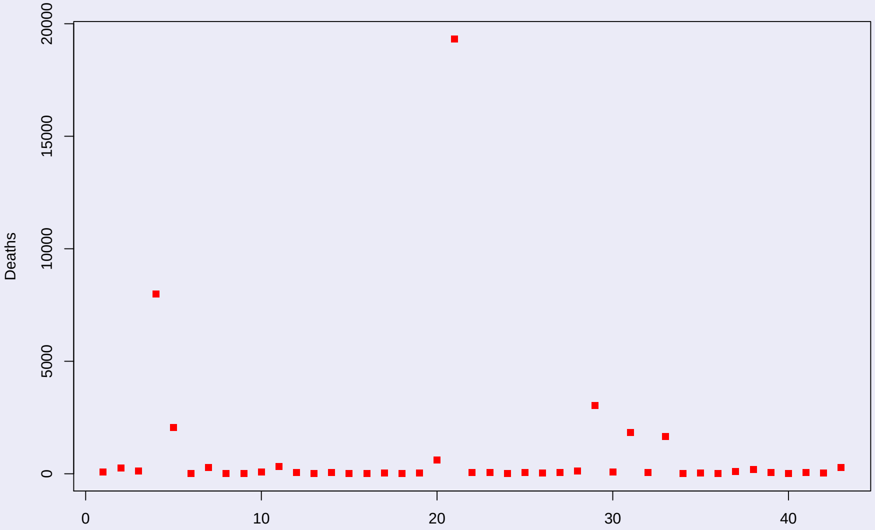
Predictions of Damage and Deaths

Basic plot of Damage and Deaths



Predictions of Damage and Deaths

Basic plot of Damage and Deaths



Generalized Linear Model - Poisson

The poisson model used in predicting deaths and damage is:

$$log(Damage * 1000 \text{ or } Deaths) = \beta_i X_i$$

- where X_i includes $\beta_0 \sim \beta_4$ and the predictors in new data
- convert `Damage` units from billions to millions to get integer data

Coefficient Table

Table 2: Coefficient estimates table from Bayesian model

ID	β_0	β_1	β_2	β_3	β_4
agnes.1972	3.951	0.922	0.006	-0.310	0.545
alex.2010	3.799	0.937	0.070	-0.394	0.540
alicia.1983	3.897	0.904	-0.075	-0.399	0.548
allen.1980	3.687	0.966	0.131	-0.546	0.547
andrew.1992	3.676	0.938	-0.284	-0.578	0.537
betsy.1965	3.808	0.951	-0.450	-0.389	0.424
bob.1991	3.629	0.923	0.028	-0.575	0.438
camille.1969	3.994	0.936	0.073	-0.573	0.670
charley.2004	3.639	0.948	-0.180	-0.696	0.182
david.1979	3.790	0.958	-0.046	-0.382	0.685

Predict Damage

Table 3: Coefficients of damage prediction model

term	estimate	std.error	statistic	p.value
(Intercept)	-211.035	2.017	-104.623	0
β_0	5.045	0.028	182.820	0
β_1	62.835	0.444	141.656	0
β_2	-1.096	0.013	-81.665	0
β_3	3.378	0.026	130.910	0
β_4	-1.393	0.034	-41.399	0
nobs	0.049	0.000	193.646	0
Season	0.075	0.000	187.765	0
MonthJuly	0.548	0.019	29.460	0
MonthJune	-3.416	0.024	-141.750	0
MonthNovember	-1.902	0.025	-76.221	0
MonthOctober	-1.291	0.009	-136.870	0
MonthSeptember	-1.764	0.008	-229.409	0
NatureNR	-4.317	0.036	-121.180	0
NatureTS	-2.038	0.014	-142.332	0
Maxspeed	0.050	0.000	235.831	0
Meanspeed	-0.066	0.000	-134.784	0
Maxpressure	-0.007	0.001	-5.368	0
Meanpressure	0.000	0.000	-3.818	0
Total.Pop	0.000	0.000	49.870	0
Percent.Poor	-0.038	0.000	-206.165	0
Percent.USA	-0.005	0.000	-63.246	0

Predict Deaths

Table 4: Coefficients of death prediction model

term	estimate	std.error	statistic	p.value
(Intercept)	116.498	12.580	9.261	0.000
β_0	11.675	0.256	45.530	0.000
β_1	114.119	2.200	51.869	0.000
β_2	5.529	0.123	45.084	0.000
β_3	8.562	0.285	30.007	0.000
β_4	-10.492	0.306	-34.307	0.000
nobs	0.003	0.001	3.073	0.002
Season	0.006	0.002	2.914	0.004
MonthJuly	-1.184	0.145	-8.171	0.000
MonthJune	-1.292	0.090	-14.402	0.000
MonthNovember	-2.533	0.155	-16.323	0.000
MonthOctober	-1.547	0.065	-23.918	0.000
MonthSeptember	-0.275	0.046	-5.995	0.000
NatureNR	2.349	0.129	18.205	0.000
NatureTS	3.563	0.121	29.451	0.000
Meanspeed	-0.037	0.003	-11.696	0.000
Maxpressure	-0.269	0.010	-27.775	0.000
Meanpressure	0.005	0.000	26.759	0.000
Total.Pop	0.000	0.000	36.369	0.000
Percent.Poor	0.036	0.001	44.860	0.000
Percent.USA	-0.007	0.001	-12.950	0.000

Conclusions

- Based on posterior estimates of μ, an increase in current wind speed and the change in wind speed is associated with increase in the wind speed in the upcoming future.
- Our MCMC algorithm successfully estimates the high-dimensional parameters
  - All the parameters converges quickly under a good initial values setting
  - The overall R^2 is relatively large, our model fits the data well
- For different months, there is no significant differences observed. Over years, the effect the wind speed 6 months ago has on the current wind speed may decrease a little.
- Th β_i coefficients estimated from the Bayesian model is powerful when predicting the damage and deaths caused by hurricanes

Limitations

- Different initial values
- Low performance on hurricanes with few observations