Hierarchical Baysian Modeling of Hurricane Trajectories

P8160 Group Project 3 - Markov chain Monte Carlo

Jingchen Chai, Yi Huang, Zining Qi, Ziyi Wang, Ruihan Zhang

May 4, 2023

Contents

Abstract

A hurricane is a tropical storm with winds that have reached a constant speed of 74 miles per hour or more. Developing a model to predict the wind speed of the hurricane can provide early warning and help communities and individuals prepare for severe weather events. In this study we built a Bayesian model and implemented MCMC algorithm to to generate the distribution of corresponding parameters. Most parameters converged well, and the model performed good in predicting the wind speed. Our study also analyzed that there are no statistical significant differences in the seasonal difference in hurricane wind speeds, and there is no evidence to support the claim that hurricane wind speeds have been increasing over the years.

Introduction

The United States faces significant social and economic risks from hurricanes, which cause fatalities and property damage through high winds, heavy rain, and storm surges (Blendon et al, 2007). To address this, there is a growing need to accurately predict hurricane behavior, including location and speed. This project aims to forecast wind speeds by modeling hurricane trajectories using a Hierarchical Bayesian Model. The hurricane data includes specific effects unique to each hurricane, and model integration is achieved through the use of a Markov Chain Monte Carlo algorithm. Through this model, we intend to identify the key factors that influence hurricanes and draw meaningful conclusions based on the results.

Additionally, we aim to explore seasonal variations in hurricane wind speeds and investigate whether there is evidence of increasing wind speeds over time. Furthermore, we utilized our estimated model parameters and covariate values to develop a prediction model for tracking each hurricane's wind speeds at each time point. Prediction performance is compared with the actual wind speeds recorded during the hurricane.

Study Data

Variable List

ID: ID of hurricanes

Season: In which year the hurricane occurred

Month: In which month the hurricane occurred

Nature: Nature of the hurricane

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

Time: dates and time of the record

Latitude and **Longitude**: The location of a hurricane check point

Wind.kt: Maximum wind speed (in Knot) at each check point

Exploratory Data Analysis

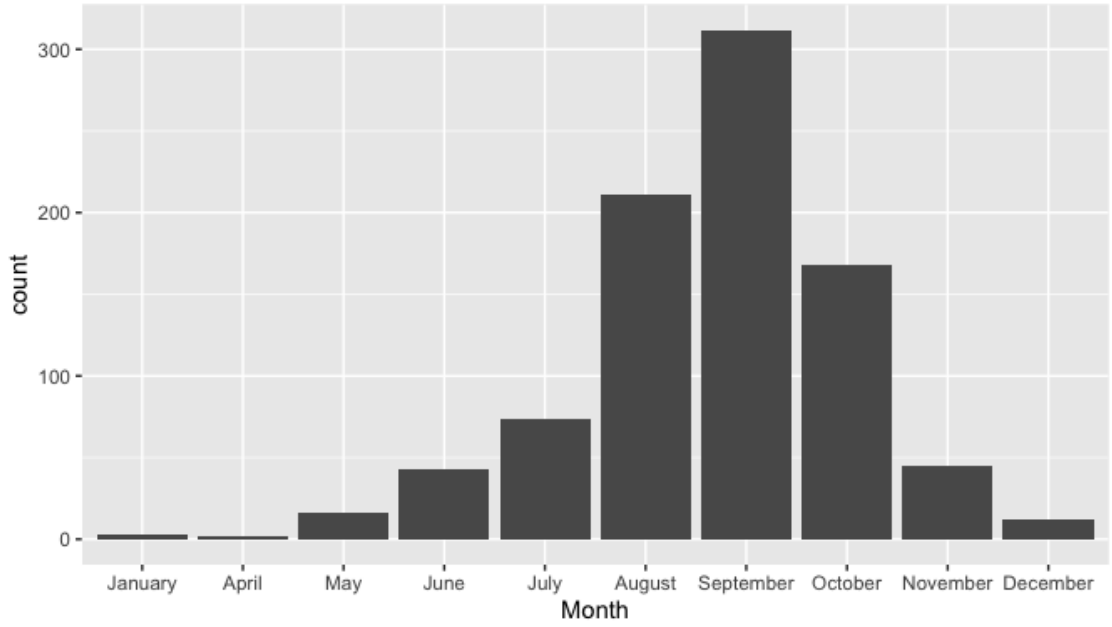

Figure 1. Count of Hurricanes in each Month

Figure 1 shows September has the highest number of hurricanes, while there are none in February and March.

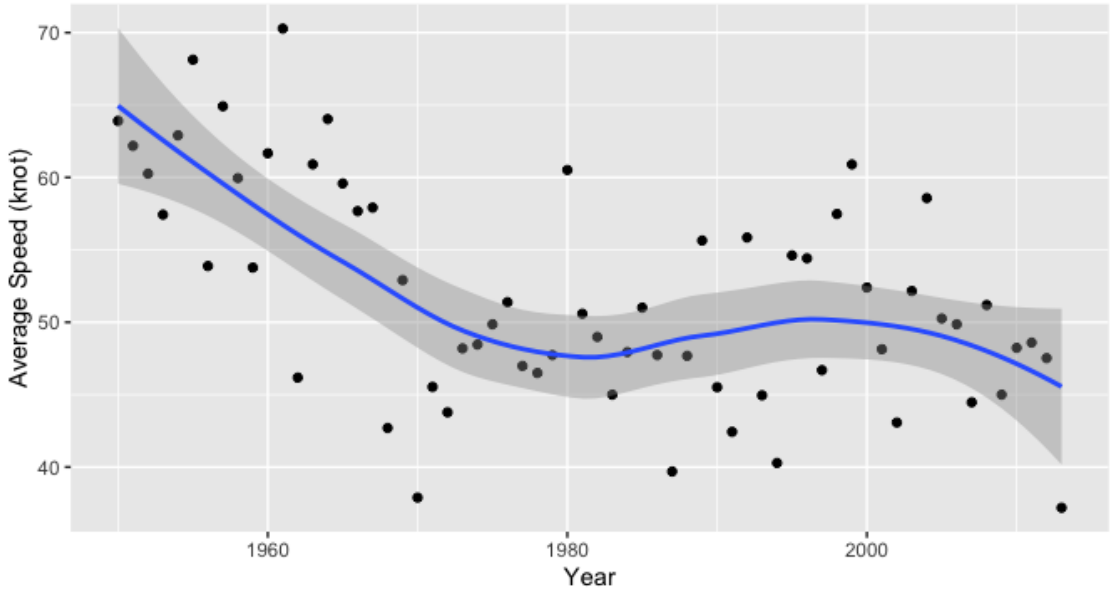

Figure 2. Average Speed of Hurricanes

Figure 2 shows as the year increases, the average speed of hurricanes first decreases, then increases a little bit, and finally decreases.

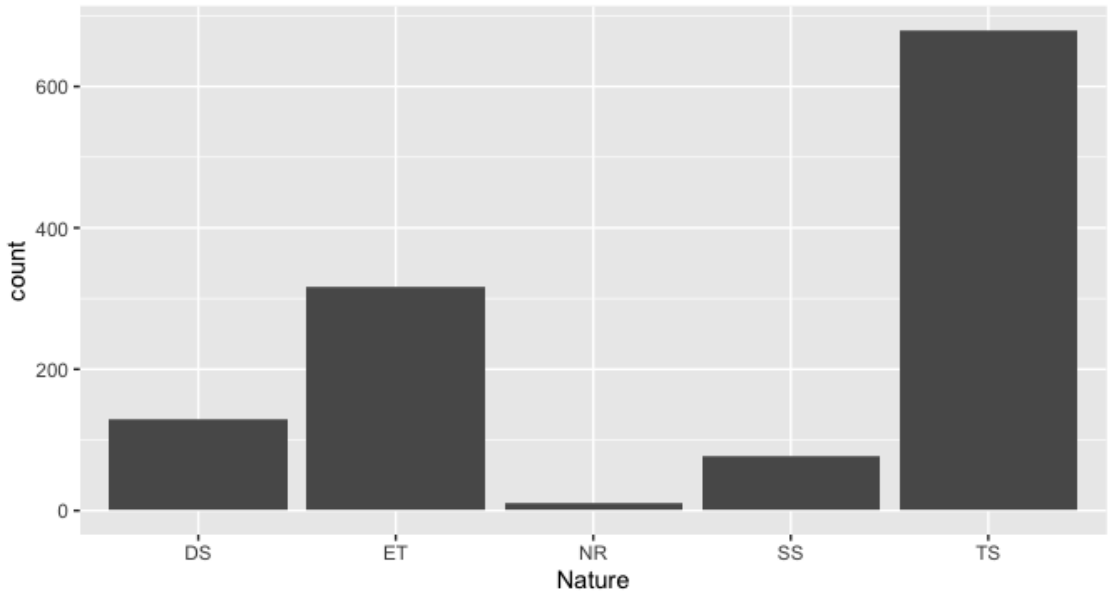

Figure 3. Count of Hurricanes in each Nature

Figure 3 shows over 50% of the nature ratings are Tropical Storm, and TS has the highest average wind speed (approximately 60 knots). Note that some hurricanes have different nature ratings at different times.

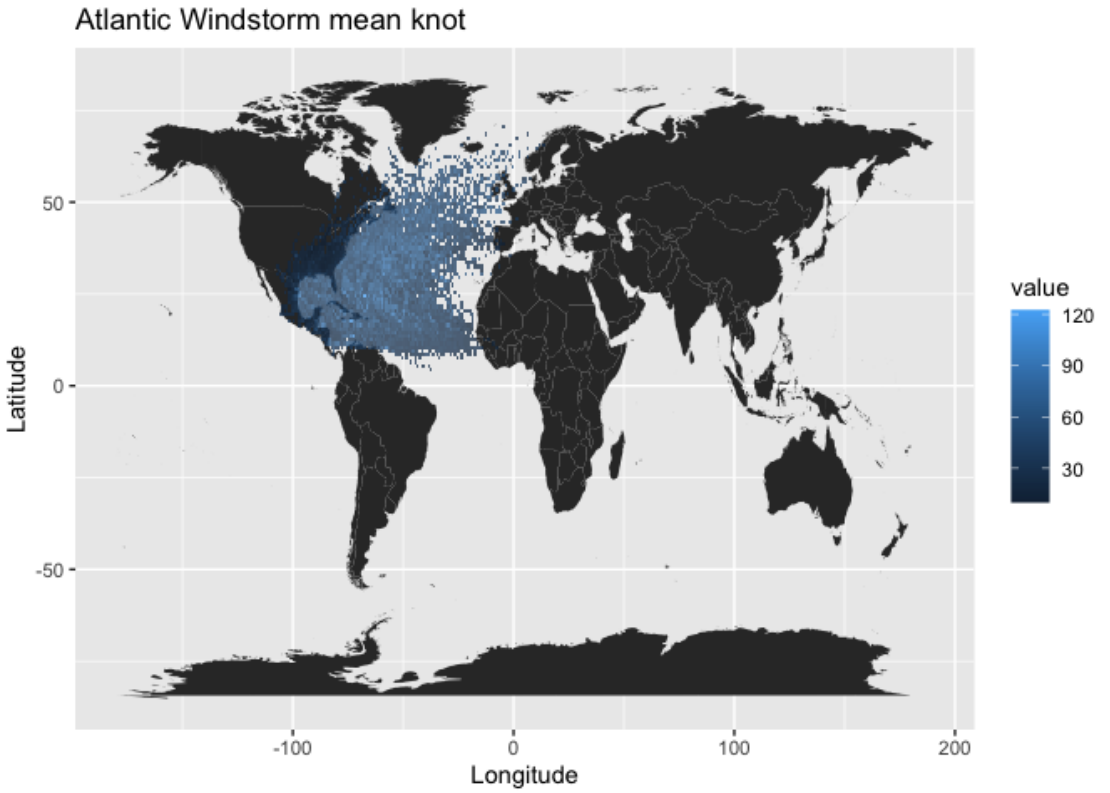

Figure 4. Atlantic Windstorm Mean Knot

Figure 4 shows the Atlantic windstorm mean knot, and the windstorm mainly concentrates on the US.

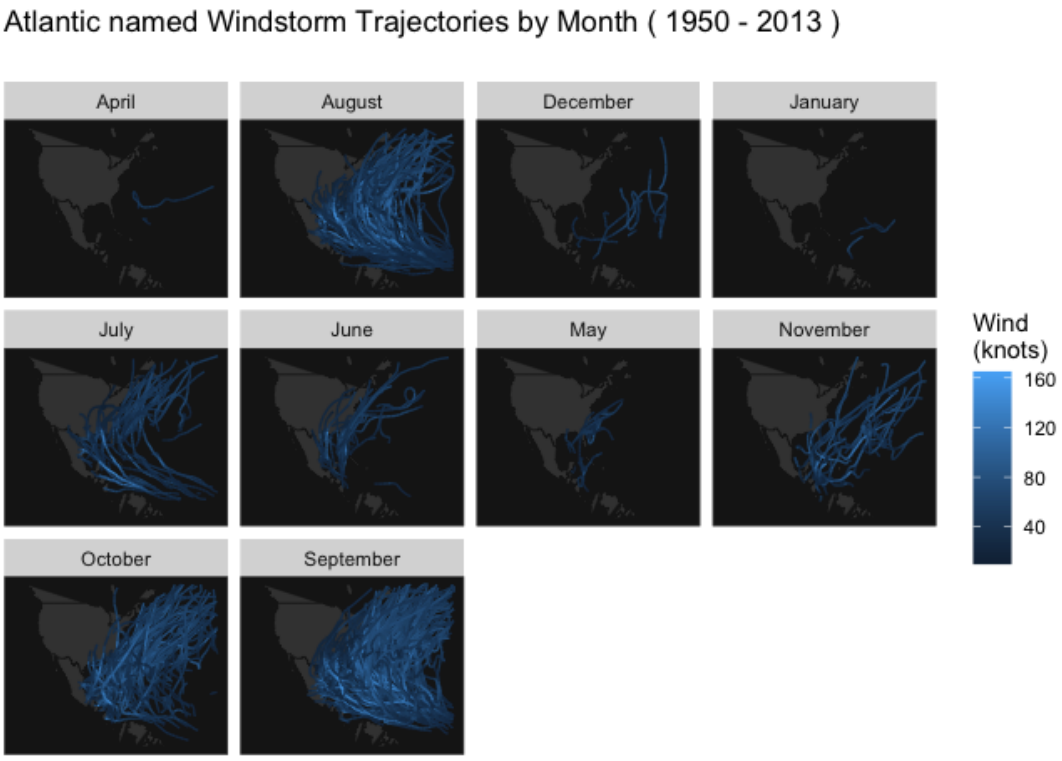

Figure 5. Atlantic Windstorm Trajectories by Months

Figure 5 shows the Atlantic windstorm trajectories, and the windstorm mainly occurs in August and September.

Data Pre-processing

We selected observations that occurred on a 6-hour intervals (e.g., hour $0, 6, 12, 18$). Three new variables including the changes of latitude, the changes of longitude, and the changes of wind speed between the time t and $t-6$ is created for analysis purpose. After data cleaning, we obtained 20293 observations and with 699 different hurricanes.

Methodology

Markov chain Monte Carlo

In our project, we employed a Markov chain Monte Carlo (MCMC) simulation to estimate the parameters of a model that predicts wind speed based on velocity trajectory data. The MCMC algorithm generates samples from the Markov Chain in a way that leads us closer to the desired posterior. In our study, we used two MCMC techniques: the Metropolis-Hastings algorithm and Gibbs sampling.

Hierarchical Bayesian Model

Bayesian hierarchical modeling is a statistical approach that involves writing a model in multiple levels or a hierarchical form to estimate the parameters of the posterior distribution using Bayesian methodology. This technique assumes that the observed data are generated from a hierarchy of unknown parameters, and it estimates the posterior distribution of these parameters using a Bayesian approach. In other words, Bayesian hierarchical modeling is a way of modeling complex data structures by breaking them down into smaller, more manageable components and using Bayesian analysis to estimate the unknown parameters in each component.

From the Bayes' theorem:

$$posterior\ distribution \propto likelihood \times prior\ distribution$$

$$\pi(\theta|X) \propto \pi(X|\theta) \times \pi(\theta)$$

The Hierarchical Bayes

$$\pi(\theta, \alpha|X) \propto \pi(X|\theta) \times \pi(\theta|\alpha) \times \pi(\alpha)$$

Our suggested Bayesian model is

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

- 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.
- $X_i = (x_{i,1}, x_{i,2}, x_{i,3})$ are covariates with fixed effect γ, where $x_{i,1}$ be the month of year when the i-th hurricane started, $x_{i,2}$ be the calendar year of the i hurricane, and $x_{i,3}$ be the type of the i-th hurricane.
- $\boldsymbol{\beta}_i = (\beta_{0,i}, \beta_{1,i}, ..., \beta_{5,i})$, we assume that $\boldsymbol{\beta}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Prior Distribution

$$P(\boldsymbol{\mu}) = \frac{1}{\sqrt{2\pi}|\boldsymbol{V}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\mu^\top \boldsymbol{V}^{-1}\mu\} \propto |\boldsymbol{V}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mu^\top \boldsymbol{V}^{-1}\mu\}$$

where V is a variance-covariance matrix

$$P(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{(\nu+d+1)}{2}} \exp(-\frac{1}{2}tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}))$$

where d is the dimension of $\boldsymbol{\beta}_i$, S is the scale matrix

$$P(\gamma) \propto exp(-\frac{\gamma^2}{2 \times (0.05)^2}) = e^{-200\gamma^2}$$

$$P(\sigma) = \frac{2\alpha}{\sigma^2 + \alpha^2} \propto \frac{1}{\sigma^2 + \alpha^2} = \frac{1}{\sigma^2 + 100}$$

Joint Posterior Distribution

Let $\boldsymbol{B} = (\beta_1^\top, ..., \beta_n^\top)^\top$, derive the posterior distribution of the parameters $\Theta = (\boldsymbol{B}^\top, \boldsymbol{\mu}^\top, \sigma^2, \boldsymbol{\Sigma}, \gamma^\top)$.

Let $Z_i(t)\beta_i^\top = \beta_{0,i} + \beta_{1,i}Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t)$

Where Z_i is the n_ixd covariate matrix for hurricane i.

We can find that

$$Y_i \sim MVN(Z_i\beta_i^T, \sigma^2 I)$$

The likelihood for Y is:

$$\begin{aligned} f(Y \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma, \boldsymbol{\Sigma}, \gamma) &= \prod_{i=1}^{n} f(Y_i|\boldsymbol{B}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma, \gamma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}(Y_i - Z_i\beta_i^T - X_i\gamma)^\top(\sigma^2 I)^{-1}(Y_i - Z_i\beta_i^T - X_i\gamma)\} \\ &\propto (\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{n} \exp\big\{ -\frac{1}{2}(Y_i - Z_i\beta_i^T - X_i\gamma)^\top(\sigma^2 I)^{-1}(Y_i - Z_i\beta_i^T - X_i\gamma)\big\} \end{aligned}$$

For simple notation, let $N{=}(\sum_i^n n_i)$, representing the total number of unique hurricanes.

The likelihood for B is:

$$\begin{aligned} f(\boldsymbol{B} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} f(\boldsymbol{B}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\propto (|\boldsymbol{\Sigma}|)^{-\frac{N}{2}} \prod_{i=1}^{n} \exp\big\{ -\frac{1}{2}((\beta_i - \mu)^\top \boldsymbol{\Sigma}^{-1}(\beta_i - \mu))\big\} \end{aligned}$$

Joint Posterior

By using the Bayesian rule, we can show the posterior distribution for $\boldsymbol{\Theta}$ is:

$$\begin{aligned} \pi(\boldsymbol{\Theta}|Y) &= P(\boldsymbol{B}, \boldsymbol{\mu}, \sigma, \boldsymbol{\Sigma}, \gamma|Y) \\ &\propto \underbrace{L(Y|\boldsymbol{B}, \sigma)}_{\text{likelihood of } Y} \underbrace{L(\boldsymbol{B}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\text{likelihood of } \mathbf{B}} \underbrace{p(\boldsymbol{\mu})p(\sigma)p(\boldsymbol{\Sigma})p(\gamma)}_{\text{priors}} \\ &\propto \frac{1}{\sigma^N(\sigma^2 + 10^2)} \prod_{i=1}^{n} \exp\big\{ -\frac{1}{2}(Y_i - Z_i\beta_i^T - X_i\gamma)^\top(\sigma^2 I)^{-1}(Y_i - Z_i\beta_i^T - X_i\gamma)\big\} \\ &\times \exp\{-\frac{1}{2}\sum_i^n(\beta_i - \mu)^\top\Sigma^{-1}(\beta_i - \mu)\}|\boldsymbol{\Sigma}^{-1}|^{\frac{N+d+v+1}{2}} \exp\{-\frac{1}{2}tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})\}|\boldsymbol{V}|^{-\frac{1}{2}} \\ &\times \exp\{-\frac{1}{2}\mu^\top\boldsymbol{V}^{-1}\mu\} \\ &\times \exp\{-200\gamma^2\} \end{aligned}$$

Conditional Posterior Distribution

1. The posterior distribution of B

$$\begin{aligned}
\pi(\boldsymbol{B}|Y,\boldsymbol{\mu}^\top,\sigma,\boldsymbol{\Sigma}) &\propto \prod_{i=1}^{n}\exp\big\{-\frac{1}{2}(Y_i-Z_i\beta_i^T-X_i\gamma)^\top(\sigma^2I)^{-1}(Y_i-Z_i\beta_i^T-X_i\gamma)\big\}\\
&\times\exp\{-\frac{1}{2}\sum_i^n(\beta_i-\mu)^\top\Sigma^{-1}(\beta_i-\mu)\}\\
&\propto \prod_{i=1}^{n}\exp\big\{-\frac{1}{2}(Y_i-Z_i\beta_i^T-X_i\gamma_i)^\top(\sigma^2I)^{-1}(Y_i-Z_i\beta_i^T-X_i\gamma)\big\}+(\beta_i-\mu)^\top\boldsymbol{A}(\beta_i-\mu)\}\\
&\propto \prod_{i=1}^{n}\exp\big\{-\frac{1}{2}(\beta_i(Z_i^\top(\sigma^2I)^{-1}Z_i+\boldsymbol{A})\beta_i^\top-2(Z_i^\top(\sigma^2I)^{-1}Y_i-Z_i^\top\gamma X_i(\sigma^2I)^{-1}+\mu\boldsymbol{A})\beta_i\\
&\propto \prod_{i=1}^{n}\exp\big\{-\frac{1}{2}[\beta_i-(Z_i^\top(\sigma^2I)^{-1}Z_i+\boldsymbol{A})^{-1}(Z_i^\top(\sigma^2I)^{-1}Y_i-Z_i^\top\gamma X_i(\sigma^2I)^{-1}+\mu\boldsymbol{A})^\top]^\top\\
&\times(Z_i^\top(\sigma^2I)^{-1}Z_i+\boldsymbol{A})[\beta_i-(Z_i^\top(\sigma^2I)^{-1}Z_i+\boldsymbol{A})^{-1}(Z_i^\top(\sigma^2I)^{-1}Y_i-Z_i^\top\gamma X_i(\sigma^2I)^{-1}+\mu\boldsymbol{A})^\top]\big\}\\
&\propto MVN(N^{-1}M,N^{-1})
\end{aligned}$$

where $\boldsymbol{A}=\boldsymbol{\Sigma}^{-1}$, $N=\frac{Z_i^\top Z_i}{\sigma^2}+\boldsymbol{A}$, $M=\frac{Z_i^\top Y_i-Z_i^\top X_i\gamma}{\sigma^2}+\mu\boldsymbol{A}$

2. The posterior distribution of μ

$$\begin{aligned}
\pi(\boldsymbol{\mu}|\boldsymbol{B},\sigma,\boldsymbol{A},\gamma) &\propto \exp\{-\frac{\mu^\top\boldsymbol{V}^{-1}\mu}{2}\}\prod_{i=1}^{N}\exp\{-\frac{(\beta_i-\mu)^\top\boldsymbol{A}(\beta_i-\mu)}{2}\}\\
&=\exp\{\sum_i^N-\frac{1}{2}(\mu^\top(\boldsymbol{A}-\frac{1}{N}\boldsymbol{V}^{-1})\mu-2\mu^\top\boldsymbol{A}\beta_i+\beta_i^\top\boldsymbol{A}\beta_i)\}\\
&=\exp\{-\frac{1}{2}(\mu^\top(N\boldsymbol{A}-\boldsymbol{V}^{-1})\mu-2\mu^\top\sum_i^N(\boldsymbol{A}\beta_i)+\beta_i^\top\boldsymbol{A}\beta_i)\}\\
&\propto MVN(M^{-1}N,M^{-1})
\end{aligned}$$

where $M=NA-V^{-1}$ and $N=\sum_i^N(\boldsymbol{A}\beta_i)$

3. The posterior distribution of Σ

$$\begin{aligned}
\pi(\boldsymbol{\Sigma}|\boldsymbol{B},\boldsymbol{\mu},\gamma,\sigma,Y) &\propto |\boldsymbol{\Sigma}|^{-\frac{(N+v+d+1)}{2}}\exp\{-\frac{1}{2}(\sum_i^N(\beta_i-\mu)^\top\boldsymbol{\Sigma}^{-1}(\beta_i-\mu)+tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}))\}\\
&\propto |\boldsymbol{\Sigma}|^{-\frac{(N+v+d+1)}{2}}\exp\{-\frac{1}{2}tr(\boldsymbol{S}+\sum_i^N(\beta_i-\mu)(\beta_i-\mu)^\top)\boldsymbol{\Sigma}^{-1}\}\\
&\propto w^{-1}(\boldsymbol{S}+\sum_i^N(\beta_i-\mu)(\beta_i-\mu)^\top,N+v)
\end{aligned}$$

where $w^{-1}(\boldsymbol{S}+\sum_i^N(\beta_i-\mu)(\beta_i-\mu)^\top,N+v)$ is the inverse-Wishart distribution with degrees of freedom $N+v$ and scale matrix $\boldsymbol{S}+\sum_i^N(\beta_i-\mu)(\beta_i-\mu)^\top$.

4. The posterior distribution of γ

$$\begin{aligned}\pi(\gamma|\boldsymbol{B},\boldsymbol{\mu},\boldsymbol{\Sigma},\sigma,Y) &\propto \prod_{i=1}^{N}\exp\{-\frac{1}{2}(Y_i - Z_i\beta_i^T - X_i\gamma)^\top(\sigma^2 I)^{-1}(Y_i - Z_i\beta_i^T - X_i\gamma)\}\times\exp\{-\frac{400\gamma^\top\gamma}{2}\}\\ &\propto \exp\{-\frac{1}{2}\sum_i^N \gamma^\top(X_i^\top\sigma^{-2}IX_i + 400N^{-1}I)\gamma - 2\gamma^\top(X_i^\top\sigma^{-2}IY_i - X_i^\top\sigma^{-2}IZ_i\beta_i)\\ &+ Y_i^\top\sigma^{-2}IY_i - 2Y_i^\top\sigma^{-2}IZ_i\beta_i^T + \beta_i^\top Z_i^\top\sigma^{-2}IZ_i\beta_i^T\}\\ &\propto MVN(M^{-1}N, M^{-1})\end{aligned}$$

where $M = \frac{\sum_i^N X_i^\top X_i}{\sigma^2} + 400I$ and $N = \frac{\sum_i^N (X_i^\top Y_i - X_i^\top Z_i\beta_i^T)}{\sigma^2}$

5. The posterior distribution of σ

$$\begin{aligned}\pi(\sigma|Y,\boldsymbol{B},\boldsymbol{\mu},\boldsymbol{\Sigma},\gamma) &\propto \frac{1}{\sigma^N(\sigma^2+10^2)}\\ &\times\prod_{i=1}^{n}\exp\big\{-\frac{1}{2(\sigma^2 I)}(Y_i - Z_i\beta_i^T - X_i\gamma)^\top(Y_i - Z_i\beta_i^T - X_i\gamma)\big\}\end{aligned}$$

σ does not have a closed distribution.

MCMC Algorithm

After deriving the conditional posterior of parameter that we want to estimate, the next step is to apply these conditional posterior to the MCMC Algorithm. Our MCMC algorithm is a hybrid of Metropolis-Hastings and Gibb Sampling.

Metropolis-Hastings

From the conditional posterior of σ, it is hard to find a closed form distribution for it, unlike other parameters. Here, we apply Metropolis-Hastings to generate new σ. The detailed steps of Metropolis-Hasting is shown below:

Algorithm 1 MCMC: Metropolis-Hastings

Require: Target distribution $\pi(\sigma)$
for $i = 1$ to 1000 **do**
1. Proposed $\sigma_{proposed} = \sigma^{(i-1)} + (U - 0.5) * 2 * a$, where $U \sim$ Uniform(0,1), a is step size
2. Calculate acceptance rate$\alpha_{XY} = \min(0, \frac{\pi(\sigma_{proposed})}{\pi(\sigma^{(i-1)})})$
3. If $U < \alpha_{XY}$: $\sigma^{(i)} = \sigma_{proposed}$, else $\sigma^{(i)} = \sigma^{(i-1)}$

end for
$\sigma_k = \sum_{i=801}^{1000} \frac{\sigma^{(i)}}{200}$, where k is the iteration of Gibb Sampling

The target distribution is the conditional posterior of σ. By setting the step size to 0.5, the acceptance rate reaches 43.5%, which is acceptable. The new sigma generated for Gibb Sampling will be the mean of next 200 values in the chain.

Gibb Sampling

After defining the Metropolis-Hastings algorithm to generate σ, we combine the Metropolis-Hastings with Gibb Sampling. We first initialized the parameters to start the algorithm. The parameters in Gibb Sampling will be updated component-wise. For each parameter to be updated, it always conditioned on the most recent values of other parameters. More precisely,

Algorithm 2 MCMC: Gibb Sampling

Require: Initialize $\boldsymbol{B}, \boldsymbol{\mu}, \sigma, \boldsymbol{\Sigma}, \boldsymbol{\gamma}$
for $k = 1$ to 5000 **do**
1. Generate $\boldsymbol{\beta}_i^k$ for i^{th} hurricane from $\pi(\mathbf{B}|\boldsymbol{Y}, \boldsymbol{\mu}^{k-1}, \sigma^{k-1}, \boldsymbol{\Sigma}^{k-1}, \boldsymbol{\gamma}^{k-1})$
2. Generate $\boldsymbol{\mu}^k$ from $\pi(\boldsymbol{\mu}|\boldsymbol{Y}, \mathbf{B}^k, \sigma^{k-1}, \boldsymbol{\Sigma}^{k-1}, \boldsymbol{\gamma}^{k-1})$
3. Generate σ^k from the Metropolis-Hastings steps
4. Generate $\boldsymbol{\Sigma}^k$ from $\pi(\boldsymbol{\Sigma}|\boldsymbol{Y}, \mathbf{B}^k, \boldsymbol{\mu}^k, \sigma^k, \boldsymbol{\gamma}^{k-1})$
5. Generate $\boldsymbol{\gamma}^k$ from $\pi(\boldsymbol{\gamma}|\boldsymbol{Y}, \mathbf{B}^k, \boldsymbol{\mu}^k, \sigma^k, \boldsymbol{\Sigma}^k)$

end for

We have tested different start values for MCMC algorithm, the result chain behave similarly. We finally decide to initialize the parameters by using the results from fitting generalized linear mixed model in R. $\boldsymbol{B}$ is a 5 x 699 matrix, $\boldsymbol{\mu}$ is a 5 x 1 matrix, σ is a number, $\boldsymbol{\Sigma}$ is a 5 x 5 matrix, and $\boldsymbol{\gamma}$ is a 14 x 1 matrix.

Results

MCMC Convergence Diagnostics

In Markov Chain Monte Carlo, determining the appropriate number of iterations can depend on many factors such as the complexity of the model, the size of the dataset, the convergence rate, etc. Therefore, it is difficult to make a general statement about a specific number of iterations that will be sufficient for all MCMC simulations. In our algorithm, we believe for most of the parameters, 5000 iterations reached the stationary of their posterior distribution. For convergence diagnostics, we generate trace plots. We also randomly choose a Hurricane George 1951 from the data to check its β trace plot, autocorrelation, and distribution.

Random Effect Parameter $\boldsymbol{B}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, $\boldsymbol{\Sigma}$

In Figure 6, the first row is the trace plots of $\boldsymbol{B}$, it shows the history of our parameter β_1 across iterations of the chain. β_1 takes only a few steps to reach stationary. This chain appears most likely to converge with an average value of about 0.95. Similarly for β_4, the chain appears most likely to converge with an average value of about 0.48. β_2 and β_3 need to take more iterations to achieve convergence. In regression model, the intercept β_0 represents the expected value of the response variable when all predictor variables are equal to zero. Its convergence is not as informative for diagnosing the convergence of the MCMC algorithm as the convergence of the other coefficients. Similarly, the second row is the trace plots of $\boldsymbol{\mu}$, all $\mu's$ converge very quickly. $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\boldsymbol{B}$, we plot the trace plot of its diagonal, which is the variance of $\boldsymbol{B}$. The result implies, all variance of $\boldsymbol{B}$ converge very quickly, the result suggests that our algorithm has produced reliable estimates of the posterior distribution of the beta coefficients. This is important for making inferences and predictions based on the model.

In Figure 7, we extract Hurricane George 1951 from the data to check its β convergence plots and distribution. We can see on trace plot of 5000 iterations for the selected parameter, each of the distributions are relatively normal with some heavy tails in β_0. In Figure 8, the convergence plot of $\boldsymbol{\sigma}^2$ suggests that the chain is mixing well and that the algorithm is converging to its posterior distribution of sigma^2. After iteration 500, this

chain appears to converge with an average value of 34.5 This also indicates that the estimated values of $\boldsymbol{\sigma}^2$ are becoming more independent and less influenced by their past values as iterations increase.

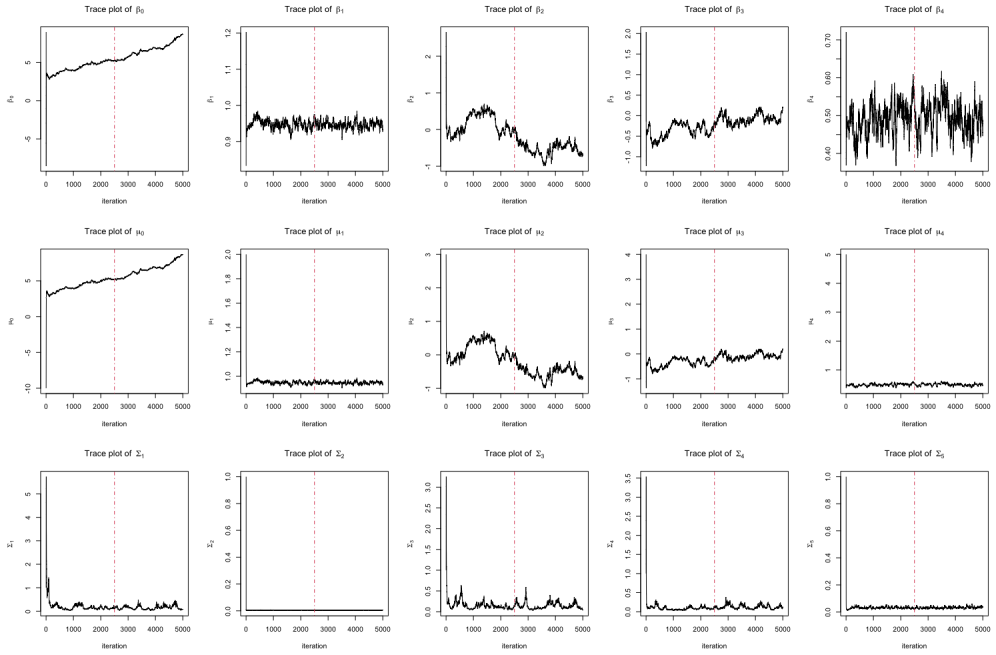

Figure 6 Convergence Plot

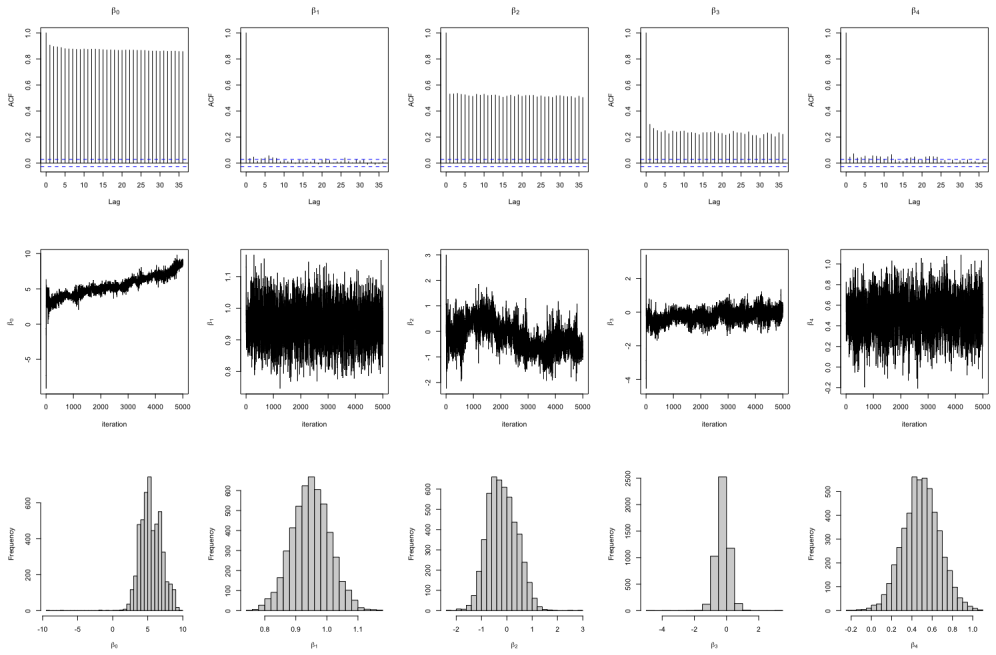

Figure 7 The ACF, Convergence, Histogram of $\beta_{HurricaneGeorge1951}$

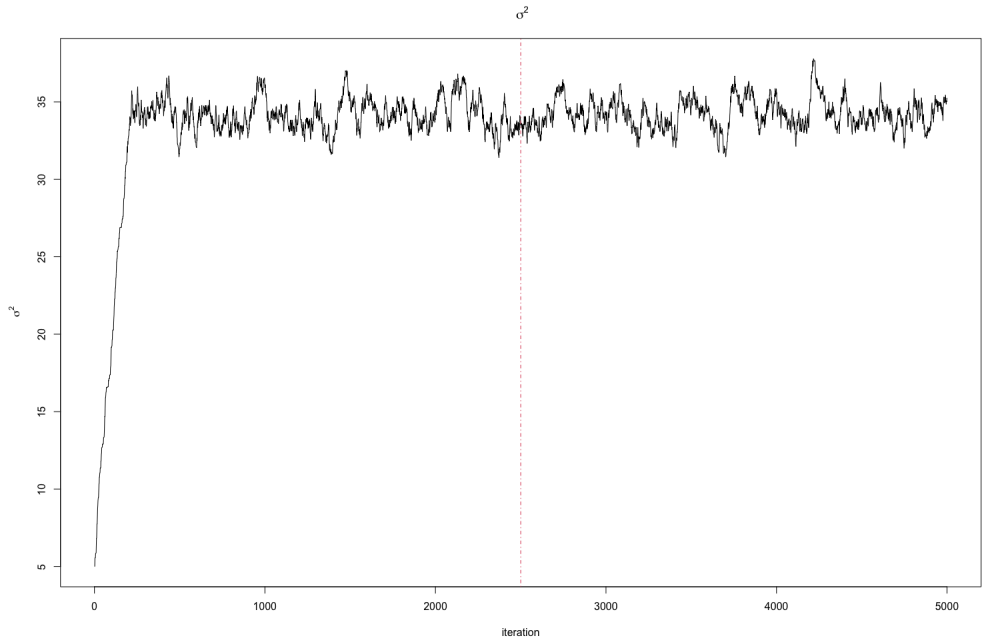

Figure 8 Trace Plot of σ^2

Analysis of the Fix Effect $\boldsymbol{\gamma}$

MCMC convergence

The trace plots of $\boldsymbol{\gamma}$ shows the majority of $\gamma's$ are stationary base on the trends, except γ_{10}. To solve this issue, one possible method is to increase the number of iterations or change the year variable to a smaller scale. However, if the year variable have a weak relationship with the response variable, it is difficult to estimate its coefficient accurately, then it is reasonable that γ_{10} is not converge in our algorithm.

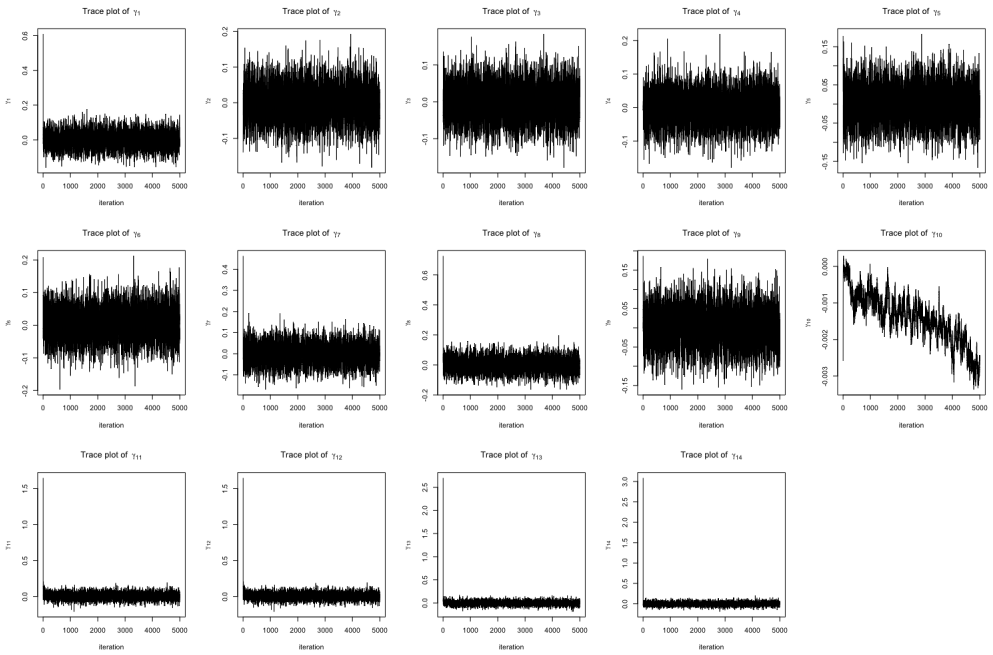

Figure 9 Trace Plot of γ

95% credible intervals of Gamma

$95\%CI$	2.5%	97.5%
April	−0.09922543	0.09022636
May	−0.10038368	0.09254019
June	−0.09761796	0.09437084
July	−0.09928455	0.10249011
August	−0.1028288	0.0916931
September	−0.0931095	0.1077383
October	−0.09762039	0.09269326
November	−0.1097137	0.1001810
December	−0.09799494	0.09315765
Year	−0.003099822	−0.001501638
TS	−0.08982878	0.10641711
ET	−0.10332349	0.09418438
SS	−0.09844162	0.09384343
NR	−0.09398186	0.09398186

Seasonal differences in hurricane wind speeds

- Summer season: July, August, September, October
- Non-summer season: April, May, June, November, December
- H_0: $\gamma_{summer} = 0$; vs H_1: $\gamma_{summer} \neq 0$

$95\%CI$	2.5%	97.5%
	−0.05845757	0.06854177

The 95% credible interval for calculated gamma associated with Summer season contains 0, thus fail to reject H_1. No evidence to support the claim that there are seasonal differences in hurricane wind speeds.

Hurricane wind speeds increasing over the years

- H_0: $\gamma_{10} \leq 0$; vs H_1: $\gamma_{10} > 0$

$95\%CI$	2.5%	97.5%
	−0.003099822	−0.001501638

The gamma associated with type (Nature of the hurricane) is smaller than 0, thus fail to reject H_1. No evidence to support the claim that hurricane wind speeds have been increasing over the years.

Model Prediction Performance

To assess how well the model predicts hurricane wind speeds, we calculate the RMSE and R^2 values for each hurricane, using the residuals of Bayesian estimates that have converged after iterations from MCMC to predict wind speeds for a test dataset. The overall RMSE is 6.647. We filtered the valid R^2 values between 0 and 1 and find that most of the hurricanes have positive R^2 values, indicating that the model performs well for most hurricanes. However, a few of the estimated Bayesian models have negative R^2 values, which may be due to the limited number of observations of hurricanes. Table 1 displays the 20 hurricanes chosen randomly. These hurricanes indicate that the estimated model accurately predicts wind speeds for most hurricanes.

	ID	r_square	rmse
1	SUBTROP:UNNAMED.1974	0.655	4.867
2	JEANNE.1980	0.921	5.437
3	FRANCES.2004	0.978	5.628
4	CHANTAL.1995	0.947	2.388
5	ETHEL.1960	0.473	27.218
6	PHILIPPE.2011	0.843	5.598
7	JOSEPHINE.1984	0.956	4.095
8	FRANCES.1976	0.895	6.114
9	BEULAH.1963	0.930	3.873
10	HOLLY.1969	0.873	5.670
11	ISAAC.2000	0.957	5.631
12	DAVID.1979	0.949	7.899
13	ALMA.1966	0.913	6.557
14	ERIN.1995	0.883	8.036
15	ANA.1997	0.880	2.156
16	DEBBIE.1969	0.851	8.869
17	HARVEY.2005	0.941	2.836
18	ALLISON.1995	0.768	4.339
19	LAURA.1971	0.967	2.112
20	EDNA.1968	0.957	2.006

Table1. R-squared values and RMSE for prediction results on test data

Figure 10 displays the actual wind speed and the estimated wind speed for a random selection of three hurricanes. It seems that there is a relatively high degree of overlap between the two curves for most parts, indicating that most of the predicted wind speeds are in close agreement with the actual wind speeds. For the hurricane Alex.2010, the RMSE is 8.879. For the hurricane Zeta.2005, the RMSE is 3.46. For the hurricane Richard.2010, the RMSE is 6.085456.

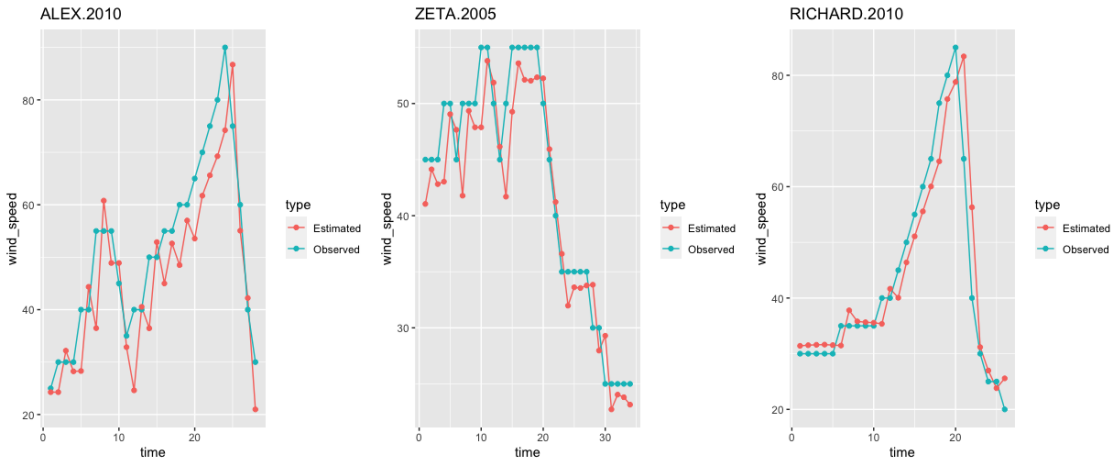

Figure 10

Limitation and Discussion

During model training and testing, although the overall R^2 and RMSE are good, but there are still a few negative R^2, and a few RMSE values are large, because of the lack of data. In the further study, it's better to use a dataset with the larger size.

Furthermore, our goal was to obtain similar results using both MCMC and ordinary linear regression methods. However, we faced some challenges with the limited amount of data for specific hurricanes and highly correlated predictor values within a limited time span. When two or more predictors for a particular hurricane were highly correlated, the determinant of the predictor matrix for that hurricane became zero, indicating that the coefficients were not unique. In such cases, NAs are returned so that it's more difficult to compare the results with the MCMC predictors.

Finally, the prior assumption in Bayesian MCMC could have resulted in a bias towards predicting larger and more destructive hurricanes that last for a longer duration.

Conclusion

Our Markov Chain Monte Carlo (MCMC) technique accurately calculates the parameters in high-dimensional settings. Most parameters converge with 5000 iterations when initialized with good values. Additionally, the model provides a good fit for the data as evidenced by a relatively high overall R^2 and a relatively low overall RMSE. Also, the predicted wind speed is highly close to the actual wind speed.

During our investigation into how the year, month, and nature of hurricanes impact wind speed, we discovered that there are no significant differences observed between wind speeds in different seasons and years. Furthermore, we found that the nature of hurricanes does play an important role in wind speed, and the most prominent nature is Tropical Storms.

References

Blendon, R., Benson, J., DesRoches, C., Lyon-Daniel, K., Mitchell, E., Pollard, W.(2007). "The Public's Preparedness for Hurricanes in Four Affected Regions". National Library of Medicine, 122(2):167-176. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1820441/

Taboga, Marco (2021). "Markov Chain Monte Carlo (MCMC) diagnostics", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix.

Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter.

Zhang, Z. (2021). A Note on Wishart and Inverse Wishart Priors for Covariance Matrix. Journal of Behavioral Data Science, 1(2), 119–126.

Contributions

Jingchen Chai, Yi Huang, and Zining Qi worked collectively on tasks 1-2, especially completing the posterior distribution derivation, the Gibbs sampler, and running the algorithm. Jingchen and Yi were responsible for Gibbs sampling and generating the LaTeX expressions of the likelihood and posterior distribution of the parameters, as well as creating convergence trace plots, histograms, and charts of the estimates. Zining designed the MCMC-Metropolis-Hastings algorithm to obtain the σ^2 parameter estimates and completed task 4. Ziyi Wang coordinated our meetings, completed task 3, checked and debugged the code, compiled references, and finalized the report. Ruihan Zhang was responsible for exploratory data analysis, help with LaTex expression, and creating the introduction for the presentation and report.

All team members worked on the presentation slides and report write-up.