Bayesian modeling of hurricane trajectories

P8160 Project 3 MCMC

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

Columbia University

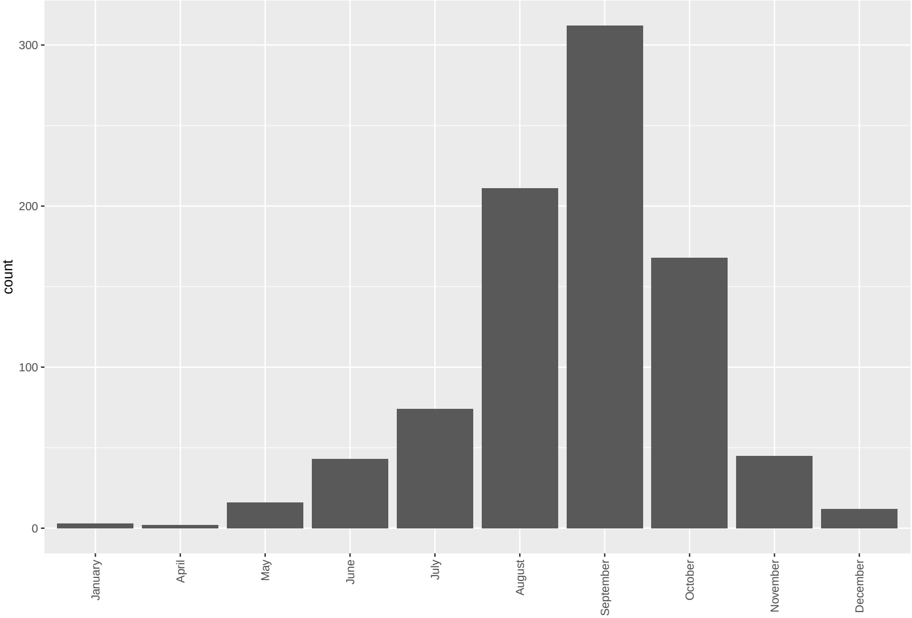
2023-05-01

Content

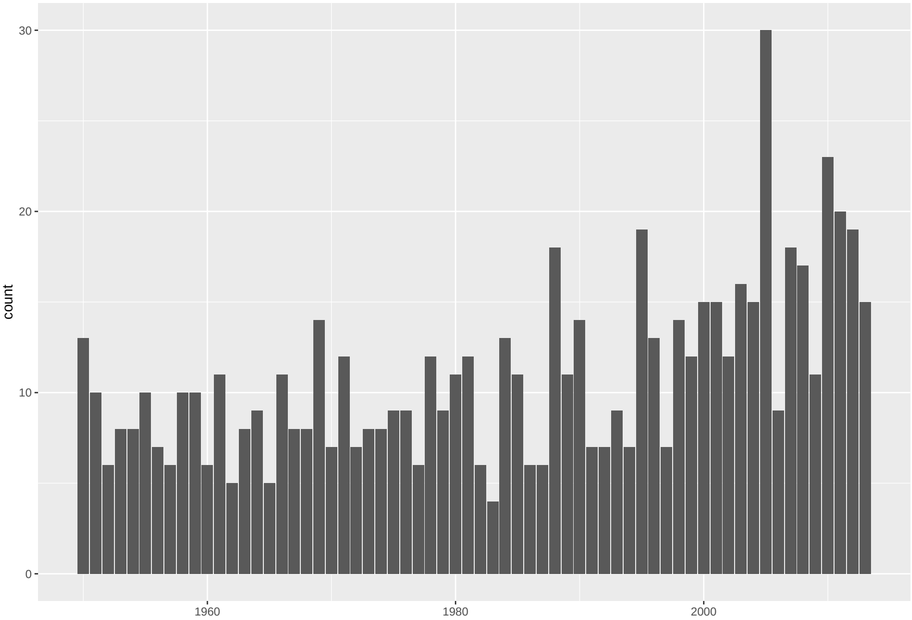
Introduction

- Hurricanes cause fatalities and property damage
- There is a growing need to accurately predict hurricane behavior, including location and speed(Taboga, 2021)
- This project aims to forecast wind speeds by modeling hurricane trajectories using a Hierarchical Bayesian Model.

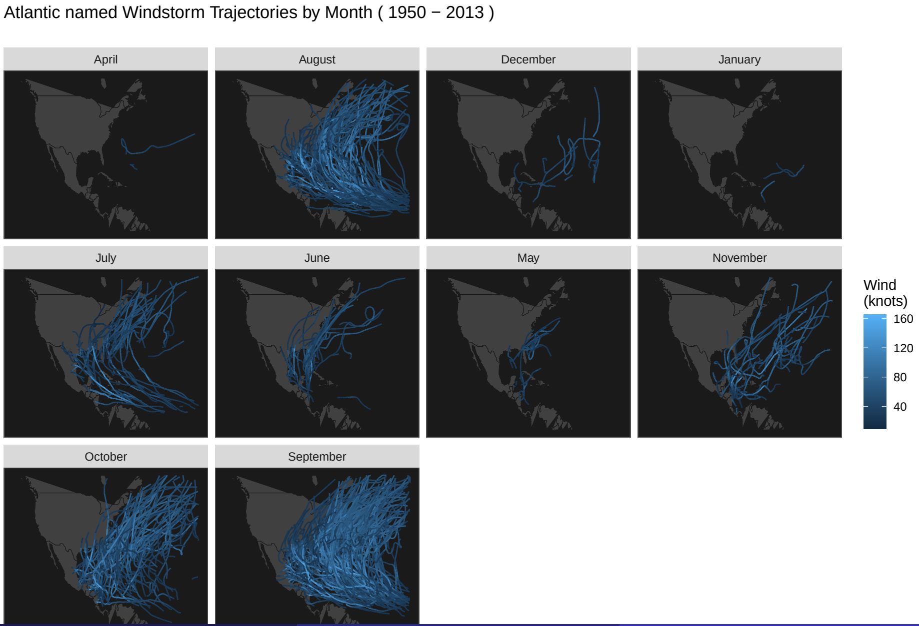
EDA-Count of Hurricanes in each Month



EDA-Count of Hurricanes in each Year



Show hurricane tracks by month



Data

ID: ID of hurricanes

Year: 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

Data Pre-processing

- We have filtered observations that occurred on a 6-hour intervals. (e.g., hour 0, 6, 12, 18)
- Calculated the lag difference for latitude, longitude and wind speed.
- After data cleaning, we obtained 20293 observations and with 699 different hurricanes.

Bayesian Model

The suggested Bayesian model is $Y_i(t+6) = \beta_{0,i} + \beta_{1,i} Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t) + X_i\gamma + \epsilon_i(t)$

- 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.
- $\beta_i = (\beta_{0,i}, \beta_{1,i}, ..., \beta_{5,i})$, we assume that $\beta_i \sim N(\mu, \Sigma)$, where d is dimension of β_i.

Prior Distribution

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

where V is a variance-covariance matrix

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

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

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

Posterior

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

Let$Z_i(t)\beta_i^\top =$
$\beta_{0,i} + \beta_{1,i}Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t) + X_i\gamma + \epsilon_i(t)$
We can find that

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

The likelihood for Y is

$$f(Y \mid B, \mu, \sigma^2, \Sigma, \gamma) = \prod_{i=1}^{N} f(Y_i | B, \mu, \Sigma, \sigma^2) =$$

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}(y_i - Z_i\beta_i - X_i\gamma_i)^\top(\sigma^2 I)^{-1}(y_i - Z_i\beta_i - X_i\gamma_i)\}$$

$$\propto (2\pi\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{n} \exp\{-\frac{1}{2}(Y_i - Z_i\beta_i - X_i\gamma_i)^\top(\sigma^2 I)^{-1}(Y_i - Z_i\beta_i - X_i\gamma_i)\}$$

where N is the total number of hurricanes.

Joint Posterior

$$\pi(\Theta|Y) = P(B,\mu,\sigma^2,\Sigma,\gamma|Y) \propto \underbrace{L(Y|B,\sigma^2)}_{\text{likelihood of } Y} \underbrace{L(B|\mu,\Sigma)}_{\text{distribution of } \mathbf{B}} \underbrace{p(\mu)p(\sigma)p(\Sigma)p(\gamma)}_{\text{priors}}$$

$$\propto \frac{1}{\sigma^N(\sigma^2+10^2)} \prod_{i=1}^{n} \exp\{-\frac{1}{2}(Y_i - Z_i\beta_i - X_i\gamma_i)^\top(\sigma^2 I)^{-1}(Y_i - Z_i\beta_i - X_i\gamma_i)\}$$

$$\times \exp\{-\frac{1}{2}\sum_i^n(\beta_i-\mu)^\top\Sigma^{-1}(\beta_i-\mu)\}|\Sigma^{-1}|^{\frac{N+d+v+1}{2}} \exp\{-\frac{1}{2}tr(S\Sigma^{-1})\}|V|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mu^\top V^{-1}\mu$$

$$\times \exp\{-200\gamma^2\}$$

where V is a variance-covariance matrix, N is the total number of hurricanes and d is the dimension of β, and v is the degree of freedom.

MCMC for Hierarchical Bayesian Model: Method

Conditional Distribution of each parameter:

- $\beta_i \sim MVN(N^{-1}M, N^{-1})$, where $N = \frac{Z_i^\top Z_i}{\sigma^2} + \Sigma^{-1}$, $M = \frac{Z_i^\top Y_i - Z_i^\top X_i \gamma}{\sigma^2} + \mu\Sigma^{-1}$
- $\mu \sim MVN(N^{-1}M, N^{-1})$, where $N = N\Sigma^{-1} - \frac{1}{V}$, $M = \sum_i^n \beta_j \Sigma^{-1}$
- $\Sigma \sim W^{-1}(S + \sum_i^n (\beta_i - \mu)(\beta_i - \mu)^\top, n + v)$
- $\gamma \sim MVN(N^{-1}M^\top, N^{-1})$, where $N = \frac{X_i^\top X_i}{\sigma^2} + 400I$, M=$M = \frac{Z_i^\top Y_i - Z_i^\top X_i \gamma}{\sigma^2} + \mu\Sigma^{-1}$
- $$\pi(\sigma | Y, \mathbf{B}^\top, \mu^\top, \Sigma, \gamma) \propto \frac{1}{\sigma^N(\sigma^2 + 10^2)} \times \prod_{i=1}^n \exp\{-\frac{1}{2(\sigma^2 I)}(Y_i - Z_i\beta_i - X_i\gamma_i)^\top (Y_i - Z_i\beta_i - X_i\gamma_i)\}$$

MCMC Algorithm - Metropolis-Hastings

*Target distribution is

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

- Choose a random walk with step size distributed as a uniform random variable
- The conditional density is $q(x|y) = \frac{1}{2a}1_{[y-a,y+a]}(x)$
- Proposed q is symmetric, thus the acceptance rate is only depend on $P(\sigma|B,\mu,A,\gamma,Y)$

MCMC Algorithm - Metropolis-Hastings

- The acceptance rate $\alpha_{XY} = \min(1, \frac{P(X|B,\mu,A,\gamma,Y)}{P(Y|B,\mu,A,\gamma,Y)})$
- Accept X if $U < \alpha_{XY}$
- Iterate over 1000 times
- New σ is the mean of last 200 values in the chain

MCMC Algorithm - Gibbs Sampling

We apply a MCMC algorithm consisting of Gibb Samping and Metropolis-Hastings steps.

Parameters are updated component-wise for each $k = 1, ..., N, N = 5000$:

- Generate $\beta_{ij}, j = 0, 1, 2, 3, 4$ for i^{th} hurricane from $\pi(\mathbf{B}|Y, \mu_{k-1}^{\top}, \sigma_{k-1}, \Sigma_{k-1}, \gamma_{k-1})$
- Generate $\mu_j, j = 0, 1, 2, 3, 4$ from $\pi(\mu|Y, \mathbf{B}_k, \sigma_{k-1}, \Sigma_{k-1}, \gamma_{k-1})$
- Generate σ_k from the Metropolis-Hastings steps
- Generate Σ_k from $\pi(\Sigma|Y, \mathbf{B}_k, \mu_k, \sigma_k, \gamma_{k-1})$
- Generate γ_k from $\pi(\gamma|Y, \mathbf{B}_k, \mu_k, \sigma_k, \Sigma_k)$

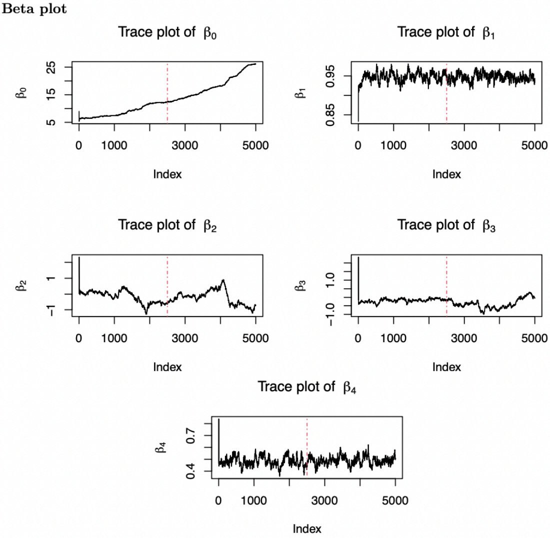
MCMC Algorithm - Initial Values

We first fit a Generalize Linear Mixed Models(GLMM)

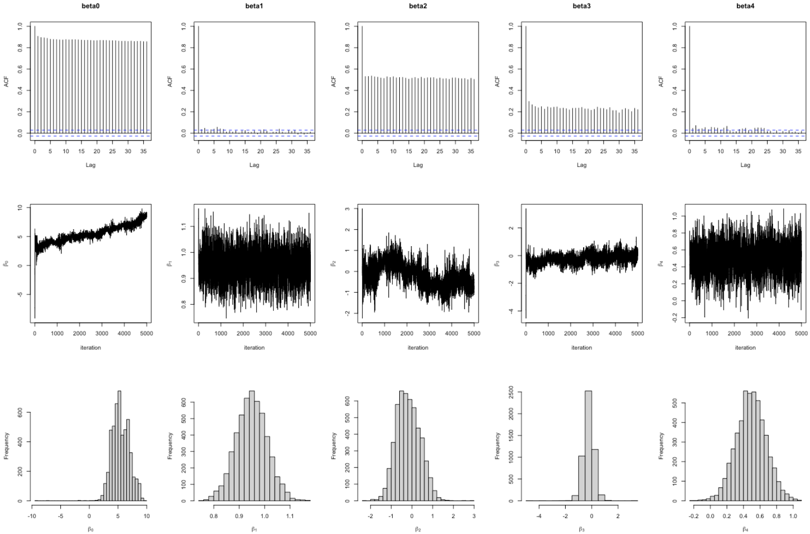
- $\beta_i^{(0)}$: The random effect for i^{th} hurricane from GLMM as start values
- $\mu^{(0)}$: Average over $\beta_i^{(0)}$
- $\sigma^{(0)}$: Residuals from the GLMM
- $\Sigma^{(0)}$: Variance-Covariance matrix of $\beta_i^{(0)}$
- $\gamma^{(0)}$: Fixed effects from the GLMM

MCMC Results - Convergence Plots of B

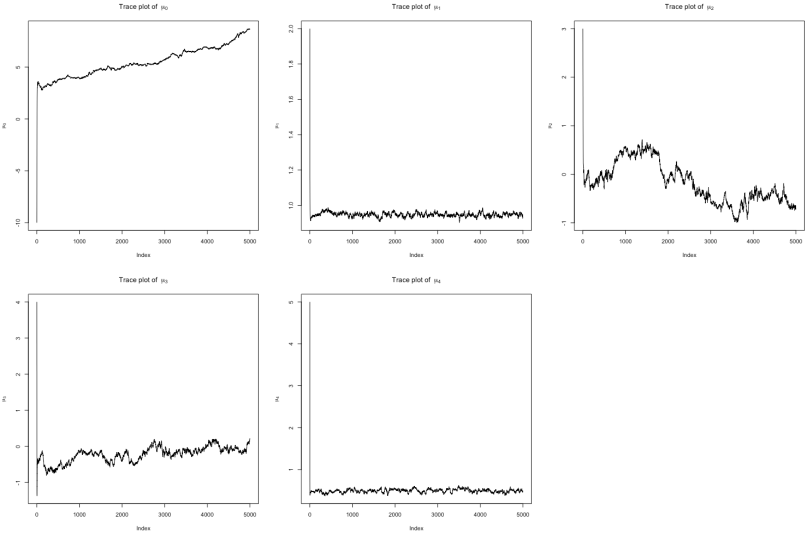
- Trace plots based on 5000 MCMC sample.



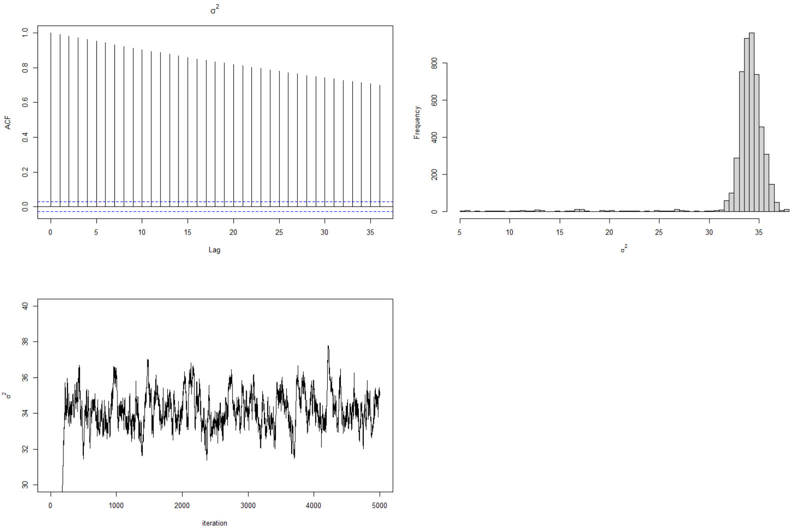
Selected B: Hurricane GEORGE.1951



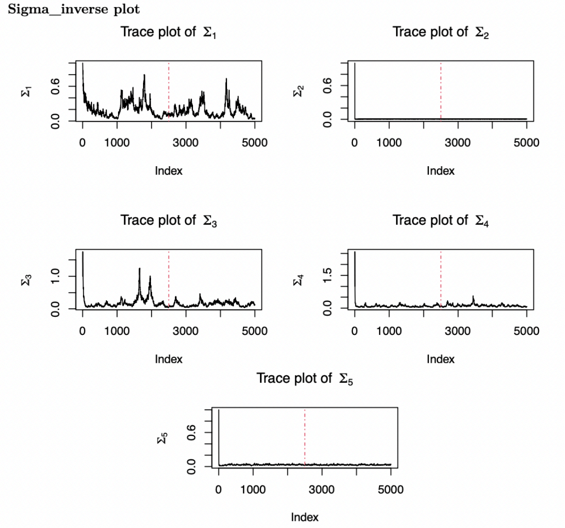
MCMC Results - Convergence Plots of μ



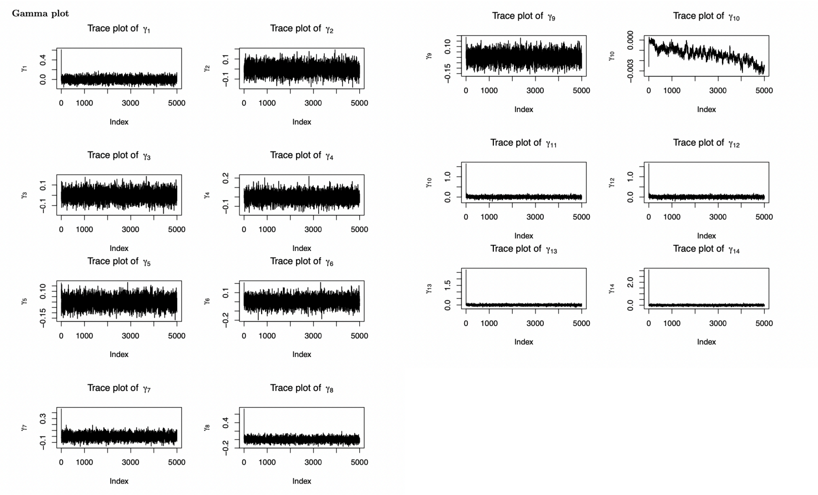
MCMC Results - Convergence and Distribution of σ^2



MCMC Results - Convergece Plots of Σ



MCMC Results - Convergece Plots of γ

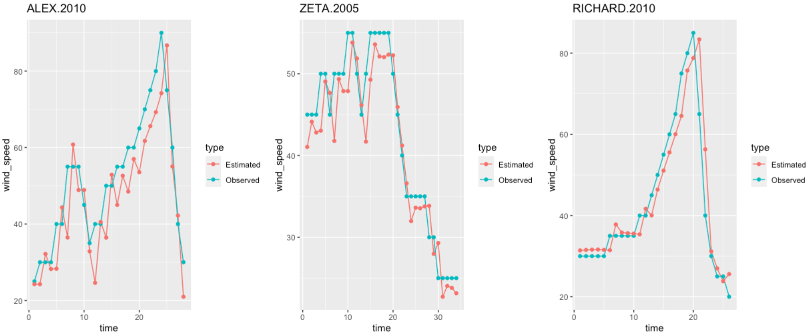


Bayesian Model Performance

- The overall mean RMSE is 6.467.

	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

Bayesian Model Performance



Estimated Wind Speed vs. Predicted Wind Speed

Limitations

-Different initial values

-Low performance on hurricanes without enough observations

Conclusion

- Our MCMC algorithm successfully estimates the high-dimensional parameters
  - All the parameters converges under a good initial values setting
  - The overall R^2 is relatively large, and the overall RMSE is relatively small, so our model fits the data well
- There are no discernible variations between the months. The impact of the wind speed from six months ago on the current wind speed may gradually diminish over time.
- When it comes to foretelling the harm and fatalities brought on by storms, the β_i coefficients calculated from the Bayesian model are effective.

Reference

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

Q&A

Thank you for your attention. Any questions?