# Bayesian Analysis of Flight Delays

## ABSTRACT

Flight delays pose a serious challenge to individual flyers, families, and the transportation of goods. As air travel becomes more accessible in Asia, Latin America, and Africa, the impact of delays will be felt by more people and companies. A study on air travel in China has shown that flight delays negatively impact passenger satisfaction and loyalty (Du and Zhang, 2020). Additionally, the financial loss to airlines and passengers is estimated to be around $8.3 billion and $16.7 billion respectively. (Ashmith, 2023).

This study will focus on the airport-specific delays under certain weather conditions. Using a hierarchical Bayesian model, we analyzed how delays at the airport level are influenced by weather conditions. The study uses a dataset of 2019 airline delay data from Kaggle. The data were gathered from the Bureau of Transportation statistics and National Centers for Environmental Information (NOAA).[3]

## DATA EXPLORATION

This study uses the train dataset downloaded from Kaggle. The dataset has about 4.5 million rows of data and 30 columns. This study will use the following columns:

- RAIN – Inches of precipitation for the day
- SNOW – Inches of snowfall for the day at the departing airport
- TMAX – Maximum temperature for the day at the departing airport
- AWND – Maximum wind speed for the day at the departing airport
- DEP_DEL15 – Binary value if the flight departure was greater than 15 minutes. 1 means the flight was delayed.

There are 96 unique airports and 17 carrier airlines in the dataset. Atlanta, Chicago O'Hare, and Dallas Fort Worth have the highest average departing flights. Conversely, Puerto Rico International, Chicago Midway, and Newark Liberty have the highest average flights delayed by 15 minutes.

Table 1: Airports with the highest average annual flights.

| **Departing Airport** | **Average Annual Flights** |
|---|---|
| Atlanta Municipal | 32837 |
| Chicago O'Hare International | 27611 |
| Dallas Fort Worth Regional | 24826 |
| Stapleton International | 20794 |

Table 2: Airports with the highest average annual delays.

| Departing Airport | Average Annual Delays |
|---|---|
| Puerto Rico International | 0.31 |
| Chicago Midway International | 0.27 |
| Newark Liberty International | 0.25 |
| William P Hobby | 0.25 |

Regarding weather, Greater Buffalo International experiences the highest average snowfall, Minneapolis-St Paul International has the highest average rainfall, Lihue Airport records the highest average wind speed, and Pensacola Regional sees the highest average temperature.

Table 3: Airports with the highest average snowfall.

| Departing Airport | Average Snowfall (inches) |
|---|---|
| Greater Buffalo International | 0.26 |
| Kent County | 0.20 |
| Rochester Monroe County | 0.18 |

Table 4: Airports with the highest average rainfall.

| Departing Airport | Average Rainfall (inches) |
|---|---|
| Minneapolis-St Paul International | 0.25 |
| Pensacola International | 0.20 |
| Memphis Internation | 0.20 |

Table 5: Airports with the highest average windspeed.

| Departing Airport | Average Windspeed |
|---|---|
| Lihue Airport | 12.36 |
| Logan International | 11.18 |
| Tulsa International | 10.97 |

Table 6: Airports with the highest average temperature.

| Departing Airport | Average Temperature (F) |
|---|---|
| Pensacola International | 96.21 |
| Kahului Airport | 86.31 |
| Houston Intercontinental | 86.18 |

# AIRPORT DELAYS

A hierarchical Bayesian logistic regression model was constructed to estimate the log-odds of flight delays. The model uses airport-specific random effects to account for differences in baseline airport delay probabilities and fixed effects for weather. Weak non-informative priors were chosen for the fixed weather effects as normal distributions along with informative hyperpriors for airport-specific effects modelled as an exponential distribution.

The hierarchy of model introduces airport-level random effects to capture unobserved factors specific to each airport. The addition of hyperpriors ensures that these random effects borrow strength from the entire dataset thereby improving estimates for each airport with limited data.

The figure below represents the hierarchical Bayesian model built using the pymc library.

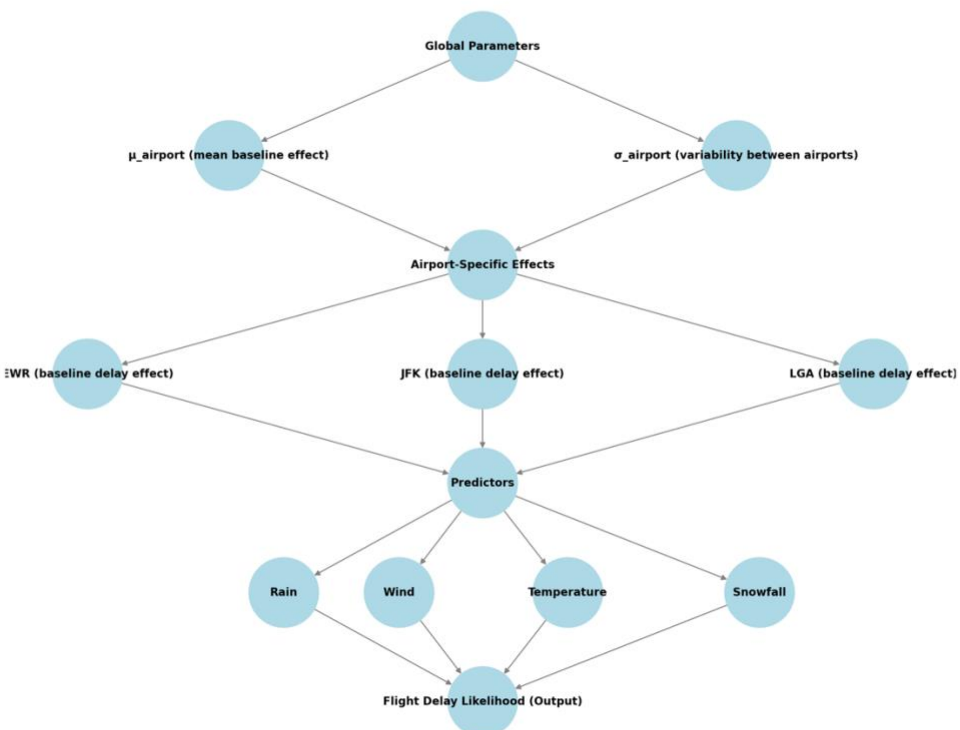


Figure 1: Representation of Hierarchical Bayesian Logistic Model.

The model uses hyperpriors to account for variability across airports. An exponential distribution was selected as a hyperprior because it helps regularize the model by assuming smaller variability unless strongly supported by the data. This prevents extreme values and aligns with the assumption that most airports have similar baseline behaviors unless influenced by unique factors; like in this case weather events.

Weather predictors (rain, snow, wind, and temperature) were modelled with weak non-informative priors. A normal distribution with a mean of 0 and a large variance of 10 was used.

This ensures that the model allows the data to drive the estimates without imposing strong assumptions.

The likelihood of flight delays was modelled using the equation below. The log-odds of delays is represented as combination of weather effects and the airport-specific baseline effects. This allows us to generalize the model and understand local variations.

$$\begin{aligned} logit\left(Prob_{delay}\right) &= \alpha_{airport\_effects} + \beta_{rain} \times inches\ of\ rain \\ &+ \beta_{temperature} \times maximum\ daily\ temperature\ + \beta_{snow} \times inches\ of\ snow \\ &+ \beta_{wind} \times windspeed \end{aligned}$$

## MODEL EVAULATION

The model shows strong statistical convergence with a well-fitted structure that explains airport-specific delays influenced by weather conditions. The r-hat values for the predictors are consistently 1.0, suggesting convergence across all parameters. Additionally, the bulk and tail effective sample sizes are sufficiently high, confirming the reliability of the posterior distribution. The model has a Widely Applicable Information Criterion (WAIC) of -154674.47 and a standard error (SE) of 286.36 from 297,469 observations. These values suggests a relatively good fit with a model's complexity of 7.13 represented by p_waic.

These results are available in the attached jupyter notebook.

## MODEL RESULTS

The hierarchical Bayesian logistic regression model was trained on data from the 3 major airports serving the New York City area: Newark International (EWR), John F. Kennedy Airport (JFK), and LaGuardia Airport (LGA). Using a smaller subset of data significantly reduced the training time for the model while allowing us to analyze the effects of weather on airports in the same geographical area.

### AIRPORT EFFECTS

The airport effects revealed notable differences in the likelihood of delays. EWR exhibited the most negative baseline log-odds, -1.5, followed by JFK, -1.3, and finally LGA, − 1.1. These results suggest that EWR experiences the fewest delays among the 3 airports. The standard deviation of airport effects, 0.44, suggests moderate variability between all 3 airports.

### WEATHER EFFECTS

All weather conditions modelled are shown to increase the chance of airport delays. Rain is seen to be strongest predictor of airport delays, increasing the log-odds of a delay by 0.22. Wind exhibited a similarly strong effect, with a chance of increasing the log-odds of airport delays by

0.20. Temperature slightly reduced impact on causing delays, with a beta of 0.12. Snowfall had the least impact on delays with a log-odds of 0.11.

Table 7: Weather Effects on Airport Delays.

| **Weather Predictor** | **Beta Coefficient** | **Percentage Change (%)** |
|---|---|---|
| Rain | 0.22 | 24.6 |
| Wind | 0.20 | 22.1 |
| Temperature | 0.12 | 12.7 |
| Snowfall | 0.11 | 11.6 |

## PRACTICAL APPLICATION

The model highlights that rain and wind significantly impact delays lasting longer than 15 minutes. These insights can enable passengers, airport operators, and airlines to proactively plan for potential disruptions caused by heavy thunderstorms or high winds in the area surrounding New York area. Proactively addressing delay factors in advance can improve customer satisfaction for the airlines and airport, minimize rebooking costs, and reduce financial penalties associated with delays. Additionally, enhanced delay predictions facilitate smoother air traffic management, alleviating congestion and optimizing runway and gate usage. The model can be used to compare airports, identifying those that perform better under adverse weather conditions. High-performing airports can serve as benchmarks, offering valuable operational strategies that can be adopted to improve performance at other airports.

## FUTURE CONSIDERATONS

This study can be improved by include more airports to allow for broader regional and national comparisons and insights. A more expansive study will require greater computing resources. Additionally, adding time dimensions, such as seasonality and time-of-day effects, can offer an understanding of how time patterns can influence airport operations. Furthermore, including route-specific factors and airline-specific variables, such as origin-destination pairs or fleet operations, may provide deeper insights into delay likelihood.

Lastly, incorporating the economic impacts of weather-delays and non-weather delays into the model can helping prioritize interventions that maximize efficiency and minimize costs. These directions would enhance the applicability and value of the model for aviation stakeholders.

## References


1. Y. Du, L. Zhang (2020), The current situation and countermeasures of flight delay in China, China Sci. Technol. Inform., 21 (2020), pp. 35-37
2. Anupkumar, Ashmith, "INVESTIGATING THE COSTS AND ECONOMIC IMPACT OF FLIGHT DELAYS IN THE AVIATION INDUSTRY AND THE POTENTIAL STRATEGIES FOR REDUCTION" (2023). Electronic Theses, Projects, and Dissertations. 1653.
3. https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations