Biostat 625 Final Project - Post 9/11 Flight Delay

https://github.com/chelleonis/flight_delay

Ralph Jiang, Xuelin Gu, Allen Li

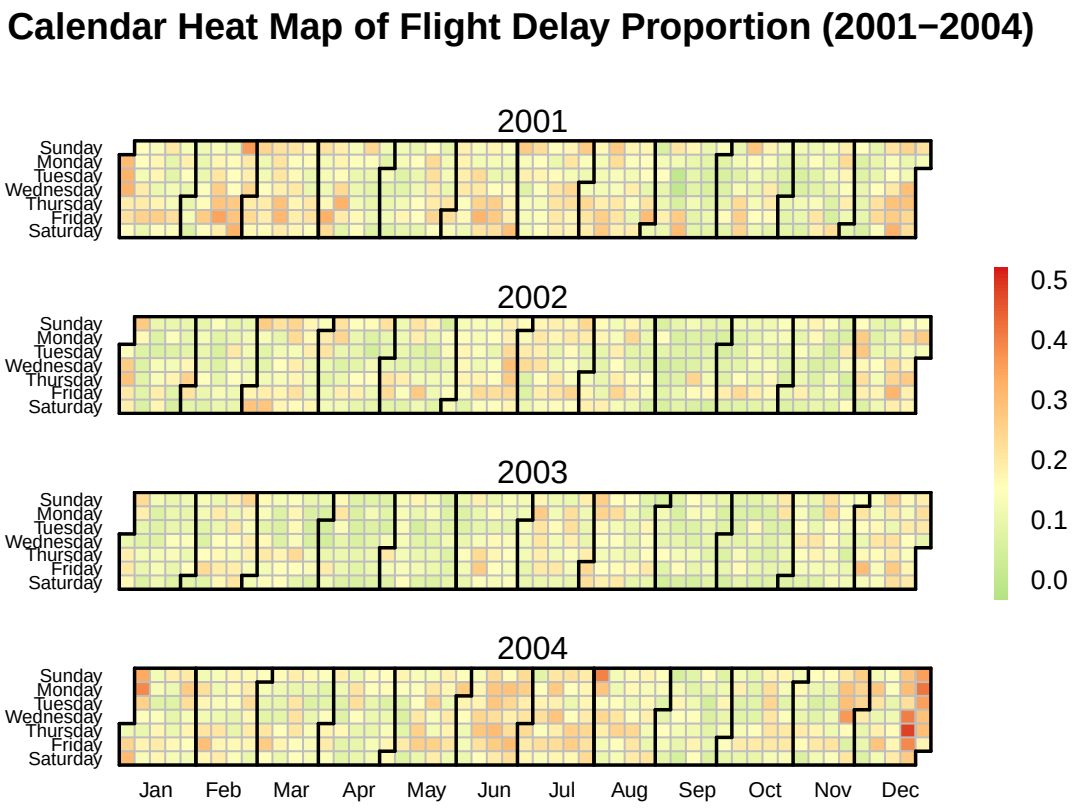
December 21, 2019

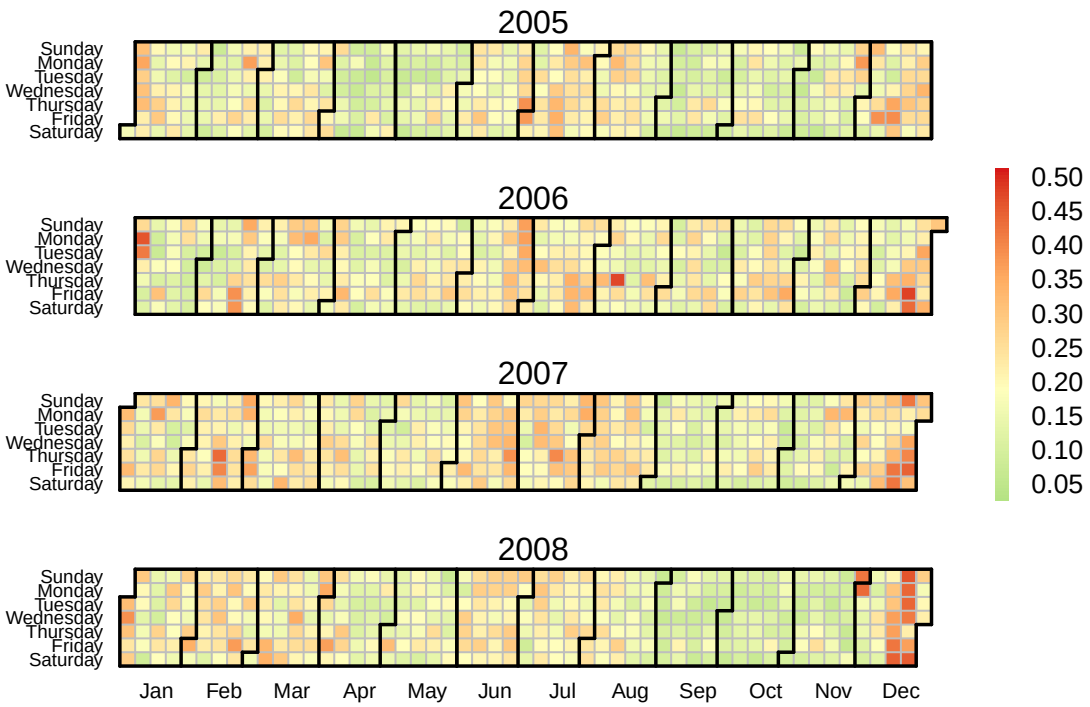
Introduction

Delayed flights are a common occurence in the airline industry, with 25 million flights being delayed (greater than 15 minutes) in the US over 20 years. The issue of delayed flights is seemingly unpredictable with so many factors preceding an on-time flight. For the customer then, it is of increased importance to be able to anticipate what may cause a delay in their flight. To try to shed some light on the issue, we analyzed a large dataset from the post 9/11 era (2001-2008), given by the Bureau of Transportation Statistics.

Calendar Heatmap

To further motivate our study of flight delays, let's take a look at the prevalence of flight delays (defined as a flight that departs more than 15 minutes past the scheduled time) from 2001-2008.





We can see that it is quite common for over 25% of flights to be delayed on any given day. The proportion of flights delayed consistently gets much worse towards the end of December / beginning of January, especially on the days right before Christmas and right after New Year's. Comparing across seasons, it appears that spring and fall are relatively delay-free, whereas in the summer, we can expect to have at least 30% of flights delayed. Unfortunately, these trends have not improved over time, and if anything, have worsened ever since 2003. Thus, we believe it is of utmost interest to investigate what factors influence the amount of time flights are delayed.

Methods

Our dataset contains, at its base, 29 variables and 130 million entries of flight data, containing variables such as flight distance, arrival delay, and calendar month. The base dataset is approximately 12 GB in size.

Data Cleaning and Pre-processing

We believe a major factor influencing flight delays is the weather. The source of our daily weather data is NCDC Cliamte Data Online. We split the dataset into managable chunks (by year). The weather datasets included 23 variables such as tempearture, precipitation, and wind speed. We merged the base data to the weather data by using a key consisting of the 3 letter airport code, year, month, and day, concatenated together. The relevant code is in the R Markdown file but not the pdf.

After the merge, we combined these separate datafiles as follows:

```
## read separate files in the dataset folder and conbine their rows
dataFiles = list.files(pattern = "*.csv") %>% lapply(read.csv, stringsAsFactors = F) %>%
```

```
    bind_rows
write.csv(dataFiles, file = "out.csv")
```

The operation above was performed in the biostatisitics computing cluster using bash scripts, as the operation was too big to perform in Rstudio due to it's memory limits.

All in all, our dataset contains 72 variables, which we then trimmed in the subsequent steps. First, we deleted the cancelled and diverted flights that may have different situations with other common delayed flights. Second, to make the analysis more efficient, we removed variables meeting the following criteria:
1. Provides little information on flight delays 2. Provides information contained in other variables 3. Not included in this project objectives 4. Variable has columns with only 'NA' 5. Variables with highly missing data

Third, we add a covariate named "Season" (i.e. Fall, Spring) based on the "Month" value.
Snippets of our data cleaning can be found in the Rmarkdown file.

Data Importing

Data was loaded in a variety of manners. For the parts we could break down by year, either read.table or read.csv were used.

```
fpath <- file.path(path, "flight_weather_cleaned.csv")
tic("fread 6gb data import")
flight_data <- data.table::fread(fpath)
toc()
# fread 6gb data import: 180.45 sec elapsed
```

For the cleaned data, it takes approximately 3 minutes to import the data for 32 million observations.

```
tic("test for read.csv")
x_test <- read.csv("D:/bios625data/flight_weather_cleaned.csv")
toc()
# test for read.csv: 2654.62 sec elapsed
```

Whereas for read.csv, it took 40 minutes to import the data.

The bigmemory package had troubles with loading a 9 GB file on a intel i7-6600U, 2.40 GHz, with 8GB memory, where it would crash Rstudio midway through. The analyses including that package and biganalytics were performed in the biostatistics computing cluster.

Analytics

Descriptive Statistics

To preview our data, we were interested in obtaining descriptive statistics for each categories. We investigated variables such as the mean, variance, frequency, and correlation, the code of which can be found in the Rmarkdown file.

From our results, most variables had low correlation with each other. Some related factors such as raw temperature and maximum temperature were inherently highly correlated. The only other noticably highly correlated pair was temperature and dew point, which we deemed to be not a large issue.

Linear regression

Linear Regresion was performed with the biganalytics package, where we analyzed delays for departures and arrivals in their own models. The code for arrival delays is included below.

```
# certain variables were chosen to be categorical via as.factor() Departure
# Delays was calculated similarly, but with weather variables coded as .x (the
# starting destination) instead of .y

Arr_result = biglm.big.matrix(ArrDelay ~ Year + Month + DayofMonth + DayOfWeek +
    Distance + TEMP.y + DEWP.y + SLP.y + VISIB.y + WDSP.y + MXSPD.y + PRCP.y + SNDP.y +
    UniqueCarrier + Origin + Dest + Fog.y + Rain.y + Snow.y + Hail.y + Thunder.y +
    Tornado.y + season, data = dat)
```

Full results of the output can be found in the xlsx file "LM result for all covariates".

We additionally performed diagnoses for our model for issues with multicolinearity and overfitting (cross-validation) and there is no existing package for class obtained from biglm.big.matrix(). The method can be found in the Rmarkdown file or the file "residual_diagnostics.R". Our GVIF and PRESS results can be found in the file "lm VIF results.txt"

From our diagnosis, we are interested in VIF values greater than 10 (the typical cutoff). After scaling the values to account for degrees of freedom in categories, we found that none of the covariates had an overwhelmingly large variance inflation factor. Thus, for our linear model we are not concerned with issues of multicolinearity.

Linear Mixed Models

We encountered a memory issue with running the lme4 package on our imported data: > "Error: cannot allocate vector of size 256.0 Mb" on Rstudio

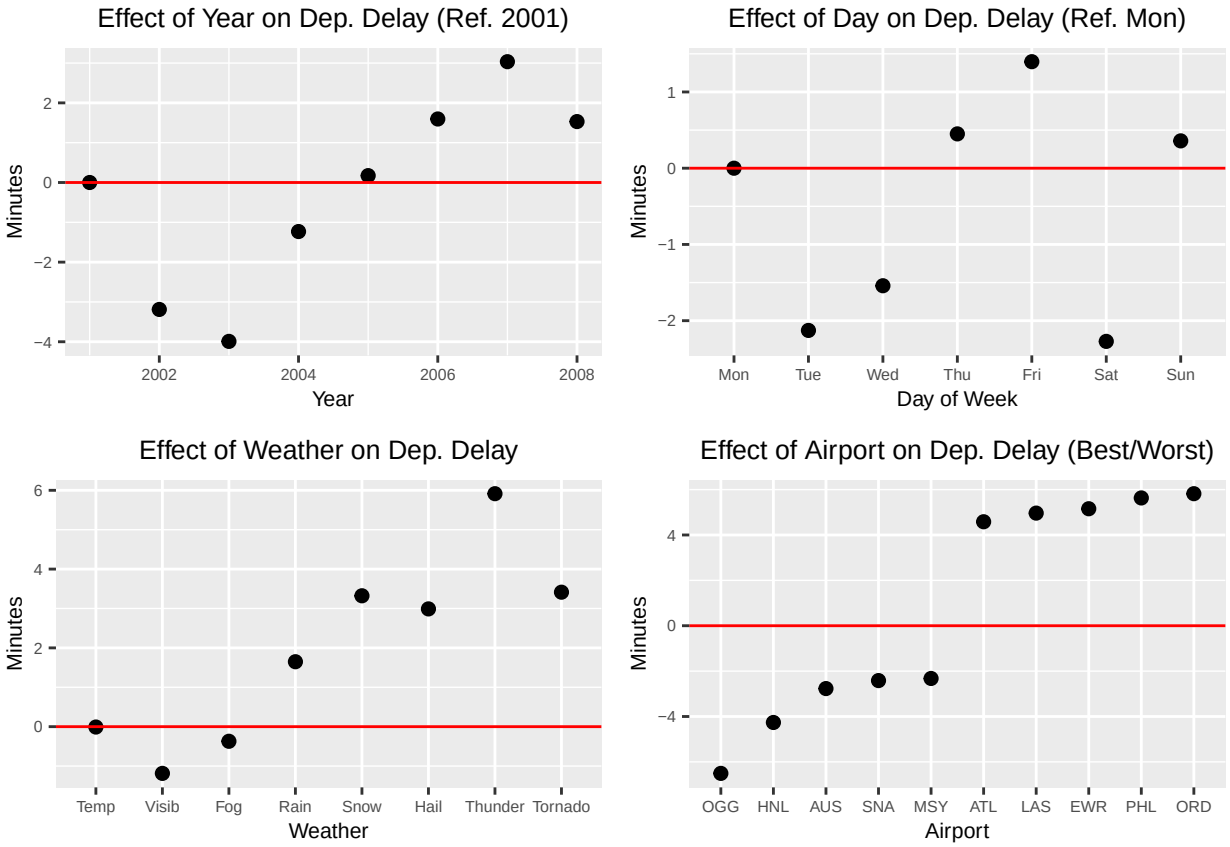
We ran the analysis on the biostatistics computing cluster for both arrival and departure models on a select few variables we had interest in.

```
# Departure Delay is calculated similarly
library(car)
flight_model2 <- lmer(ArrDelay ~ Year + Fog.x + Rain.x + Snow.x + Hail.x + Thunder.x +
    TEMP.x + (1 | Year), dat)
vif(flight_model2)
# LMM model runtime: 671.962 sec elapsed
```

Our models took approximately 10 minutes to run on the biostatistics cluster.

Results

Linear Regression



Flight delays were at a minimum in 2003, but have been steadily increasing ever since. Interestingly, the best days of the week to fly are Tuesday, Wednesday, and Saturday, with Friday being the worst.

For weather related delays in departures, it seems as if non-clear weather conditions delay the departure of planes, as expected. Temperature appears to have no adjusted linear effect on departure delay, perhaps because the true relation is quadratic or more complex.

Of our 58 selected airports, we displayed the 10 most significant results. The best airports to fly from were the two in Hawaii. The airports most susceptible to delays were large airports such as Atlanta, Newark, Philadelphia, and Chicago being the worst.

Departure delays were largely in part similar to arrival delays, with differences in the storm (weather) related variables being slightly more pronounced for the arrival delays. We suspect that flights that are delayed on arrival will affect the departure times of the next flight, creating a chain reaction of sorts for delays.

Linear Mixed Models

For our choice of model, we decide to model , with year and intercept as our random effects. The full output is given in "mixed model results and VIFs.txt", but select varirables with high t-values are shown below:

Variable	Estimate	Standard Error	t-value
Snow	9.932	0.029	340.090
Fog.x1	3.6959687	0.0171	215.239

Variable	Estimate	Standard Error	t-value
Rain.x1	4.5675803	0.0152	299.109
Thunder.x1	8.8340962	0.0243466	362.847
TEMP.x	0.0192699	0.0004031	47.810

The year trend is similar to that of the simple linear regression model, where for our year range, flights on average are more delayed as the years progress. While this trend is not necessarily linear, it indicates to us that

Additionally, the adverse weather conditions (such as thunder or fog) seem to be a significant factor in causing delays in flights.

For our random effects, our results for our year variable are quite low, 0.3825, which inidcates to us that our model results don't vary much between the years.

In general, the model has similar estimates for the year terms for both arrivals and delays. This model however, has much larger effects for weather on arrival delays compared to departure delays. We did not encounter problems with multicolinearity with our model (VIF < 10).

Conclusion and Further Work

Flight delays can usually be indicated by adverse weather conditions and the timing of the week. While these results are not absolute, customers should double check the weather on the day of the flight to predict delays and plan on flying on Tuesday, Wednesday, or Saturday to reduce chances of delays. As our model contains an excessive amount of variables, we may have ran into a problem with overfitting, as inidicated by our cross-validation results. We believe an implementation of model fitting techniques would be helpful for the inference of these models and moving forward with investigation on a larger more recent dataset of the current decade.

Our work was comprehensive, but not exhaustive, with the following topics of interest for future investigation
* Pre 9/11 Era comparison
+ As our analysis contains information on post-9/11, where security measures have been tightened signficantly, we would like to see if effects that are signficant in causing delays in this era are more pronounced before 9/11
* Compilation of more recent Data (2009-present)
* Full dataset (1987-2008) would be too large for our methods used (greater than 20 GB)
+ Having computation troubles as seen above, may need different forms
+ Would require different forms of data storage
+ Look into using RHadoop
+ Not limited to cluster computing
* Investigation of different models
* Dealing with missing data on a large scale
* Expanded Linear Mixed Model Analysis

Special Acknowledgements

Special Thanks to Dan Barker for his help and availibility in teaching us how to use the computing clusters.