Analysis of geographical patterns in weather-related flight delays

Mathew Edwards, Nandhita Narendra Babu, Wanli Zhang

May 10, 2014

Contents

Data Information

- Airline on time data from the Bureau of Transportation Statistics for all US States

- Weather-related flight delays data from June 2003 to December 2013

Question of Interest

Are there geographical patterns in weather-related flight delays, and do these change over time?

Assumptions

- NA $\rightarrow$ No Delay
- For Proportion
    - Weather delay value greater than 0 $\rightarrow$ 'Delayed' regardless of severity
    - Denominator $\rightarrow$ count of all flights including NA
- Weather within a month is 'reasonably consistent'
- Defined regions using National Oceanic and Atmospheric Administration (NOAA) map
    - Added Alaska and Hawaii as their own region
    - Lumped unlinked airports into an 'Other' region
- Weather is similar in climate regions
- Analysis done on origin airport. Assumes delay on origin side.

NOAA Map of Regions
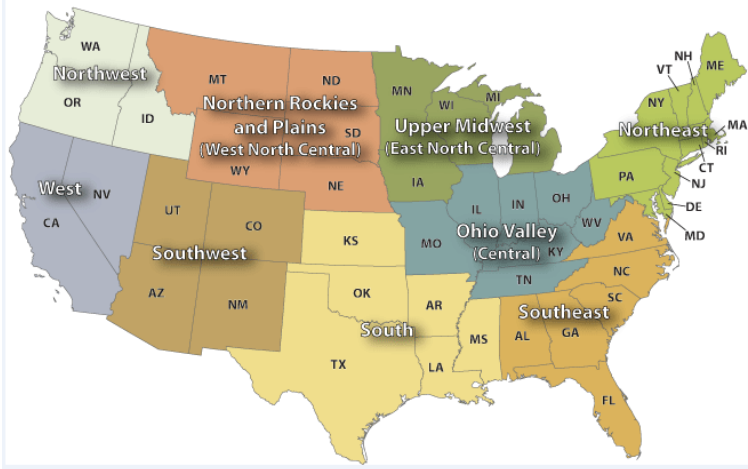
Population-based Results

- Aggregate to monthly basis
- Higher percentages of delayed flights are in winter
- Highest winter proportions are in Central region
  - Chicago: Major hub plus cold winters
- No month/region exceeds 4% in the proportion of delayed flights
- Bump in Jun, Jul and August for the South, Southeast, Northeast and Upper Midwest
  - Tornado season?
- South is consistently above 1% across all months
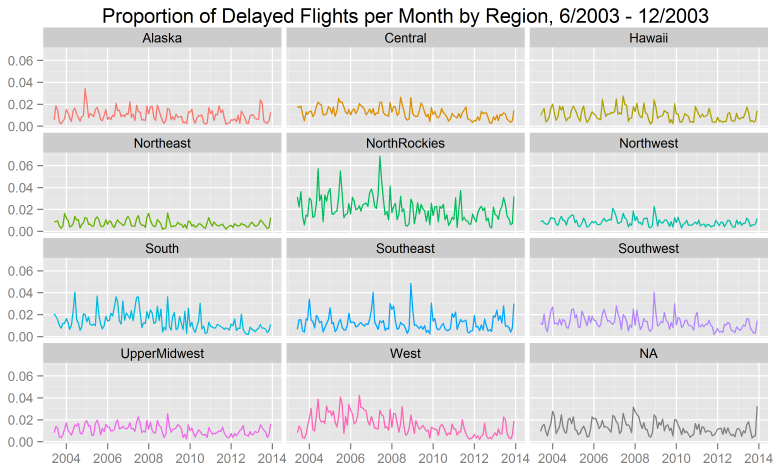
Proportion of Weather Delayed Flights, by Month

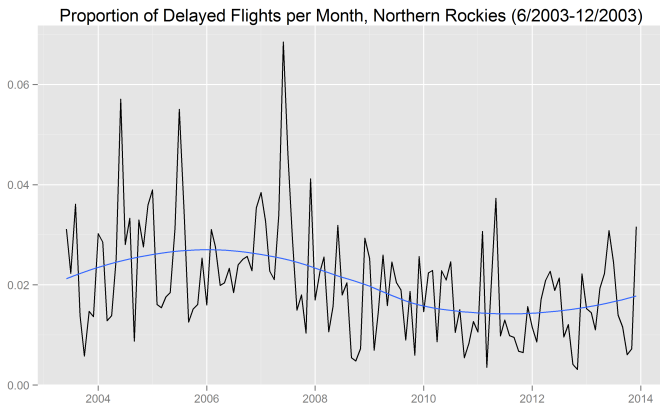|	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Alaska	1.832%	1.353%	1.175%	0.517%	0.347%	0.319%	0.455%	0.661%	0.698%	0.865%	1.529%	1.768%
Central	3.155%	2.849%	1.740%	1.349%	1.614%	1.957%	1.733%	1.450%	0.809%	0.802%	1.129%	3.385%
Hawaii	0.108%	0.114%	0.113%	0.029%	0.056%	0.021%	0.030%	0.116%	0.013%	0.051%	0.091%	0.344%
Northeast	2.007%	1.853%	1.422%	1.038%	1.185%	1.992%	2.011%	1.710%	0.872%	0.963%	0.971%	2.409%
NorthRockies	1.641%	1.321%	1.000%	0.725%	0.708%	0.830%	0.739%	0.707%	0.523%	0.636%	0.830%	2.071%
Northwest	1.367%	0.505%	0.422%	0.179%	0.151%	0.210%	0.202%	0.230%	0.182%	0.293%	0.667%	1.513%
Other	0.368%	0.270%	0.218%	0.275%	0.421%	0.514%	0.762%	0.612%	0.393%	0.510%	0.328%	0.329%
South	1.808%	1.939%	1.988%	1.938%	2.472%	3.236%	2.877%	2.173%	1.232%	1.494%	1.285%	2.356%
Southeast	1.177%	1.220%	1.133%	1.028%	1.252%	2.464%	2.475%	1.981%	1.065%	0.867%	0.783%	1.217%
Southwest	1.117%	0.741%	0.617%	0.540%	0.589%	0.961%	1.079%	0.904%	0.446%	0.519%	0.457%	1.378%
UpperMidwest	1.653%	1.913%	1.349%	1.045%	1.489%	1.965%	1.728%	1.262%	0.977%	0.914%	1.069%	2.266%
West	0.651%	0.497%	0.412%	0.303%	0.276%	0.319%	0.403%	0.383%	0.277%	0.430%	0.433%	0.729%

Population-based Results

- Across all years

- Northern Rockies show greatest fluctuation over the years

- Consistently low percentage of delayed flights in northeastern area

- Northern Rockies alone:
  - No obvious seasonal patterns
  - Gradual change of mean level over the years

All Regions

Proportion of Delayed Flights per Month by Region, 6/2003 - 12/2003

Northern Rockies

Proportion of Delayed Flights per Month, Northern Rockies (6/2003-12/2003)

Assumptions

Same as that of Population Based Analysis

Sampling Technique

- Strata $\rightarrow$ Single region for single month

- Number of flights per month ranging from an average of 3,000 in Alaska to an average of 115,000 in the Southwest

- Proportional sampling $\rightarrow$ sampled approximately 2.5% of the flights in each stratum

Sample-based Results

- Very similar to the population results

- Heat map of monthly proportions show the same pattern
    - Nothing over 4%
    - Same bump in the summer in the south

- Zero proportion in some strata, with very low population proportions

- Standard errors all below 0.5%, and generally below 0.25%
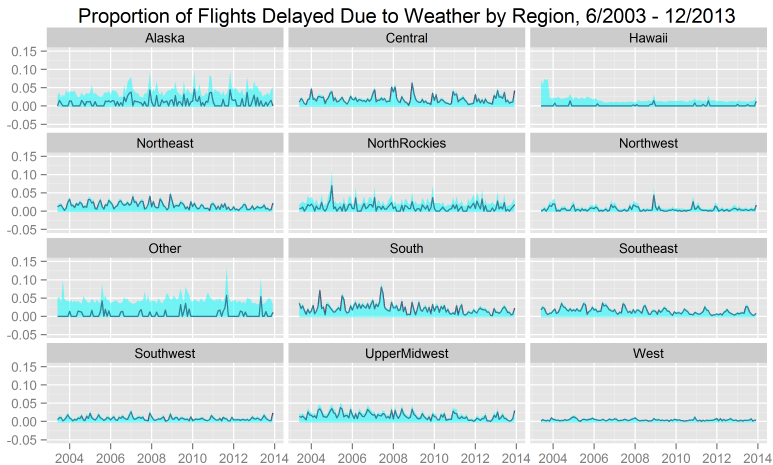
Proportion of Weather Delayed Flights, by Month

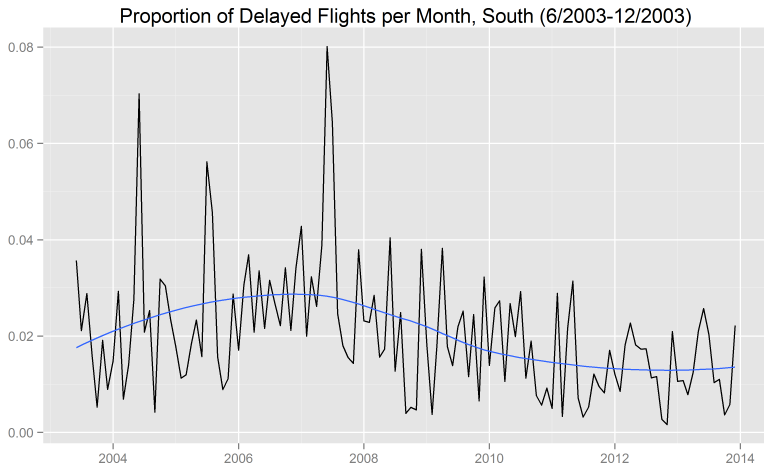| p (se(p))	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Alaska	1.711%	0.939%	0.941%	0.650%	0.720%	0.334%	0.718%	0.649%	0.933%	0.728%	1.312%	1.820%
	(0.46%)	(0.38%)	(0.35%)	(0.29%)	(0.29%)	(0.16%)	(0.24%)	(0.23%)	(0.31%)	(0.27%)	(0.41%)	(0.46%)
Central	3.147%	2.795%	1.762%	1.349%	1.634%	1.963%	1.660%	1.468%	0.811%	0.870%	1.200%	3.323%
	(0.12%)	(0.12%)	(0.09%)	(0.08%)	(0.08%)	(0.09%)	(0.08%)	(0.08%)	(0.06%)	(0.06%)	(0.07%)	(0.12%)
Hawaii	0.047%	0.103%	0.048%	0.000%	0.045%	0.000%	0.039%	0.127%	0.000%	0.047%	0.137%	0.373%
	(0.05%)	(0.07%)	(0.05%)	(0.00%)	(0.04%)	(0.00%)	(0.04%)	(0.07%)	(0.00%)	(0.05%)	(0.08%)	(0.12%)
Northeast	2.029%	1.881%	1.382%	1.104%	1.065%	1.930%	2.159%	1.631%	0.799%	0.935%	0.982%	2.546%
	(0.10%)	(0.10%)	(0.08%)	(0.07%)	(0.07%)	(0.09%)	(0.10%)	(0.08%)	(0.06%)	(0.07%)	(0.07%)	(0.11%)
NorthRockies	1.519%	1.527%	0.758%	0.604%	0.799%	0.862%	0.763%	0.798%	0.474%	0.637%	1.372%	2.015%
	(0.31%)	(0.32%)	(0.22%)	(0.20%)	(0.22%)	(0.21%)	(0.19%)	(0.19%)	(0.17%)	(0.19%)	(0.29%)	(0.33%)
Northwest	1.079%	0.735%	0.519%	0.187%	0.086%	0.262%	0.219%	0.291%	0.179%	0.426%	0.885%	1.343%
	(0.16%)	(0.14%)	(0.11%)	(0.07%)	(0.04%)	(0.07%)	(0.06%)	(0.07%)	(0.06%)	(0.09%)	(0.14%)	(0.16%)
Other	0.135%	0.141%	0.000%	0.265%	0.878%	0.617%	0.711%	1.167%	0.864%	0.160%	0.146%	0.360%
	(0.13%)	(0.14%)	(0.00%)	(0.18%)	(0.35%)	(0.27%)	(0.29%)	(0.36%)	(0.38%)	(0.16%)	(0.14%)	(0.21%)
South	1.797%	1.948%	1.919%	1.979%	2.568%	3.235%	2.767%	2.229%	1.321%	1.387%	1.214%	2.477%
	(0.09%)	(0.10%)	(0.10%)	(0.10%)	(0.11%)	(0.12%)	(0.11%)	(0.10%)	(0.08%)	(0.08%)	(0.08%)	(0.11%)
Southeast	1.232%	1.114%	1.122%	1.038%	1.227%	2.568%	2.484%	1.944%	1.043%	0.845%	0.921%	1.235%
	(0.06%)	(0.06%)	(0.06%)	(0.06%)	(0.06%)	(0.09%)	(0.08%)	(0.08%)	(0.06%)	(0.05%)	(0.05%)	(0.06%)
Southwest	1.107%	0.695%	0.645%	0.560%	0.594%	0.852%	0.942%	0.939%	0.442%	0.354%	0.561%	1.569%
	(0.09%)	(0.07%)	(0.07%)	(0.06%)	(0.07%)	(0.07%)	(0.08%)	(0.08%)	(0.06%)	(0.05%)	(0.06%)	(0.10%)
UpperMidwest	1.645%	1.800%	1.190%	0.893%	1.765%	2.008%	1.795%	1.080%	0.877%	1.008%	1.155%	2.436%
	(0.14%)	(0.15%)	(0.12%)	(0.10%)	(0.14%)	(0.14%)	(0.13%)	(0.11%)	(0.10%)	(0.10%)	(0.11%)	(0.16%)
West	0.656%	0.475%	0.435%	0.214%	0.288%	0.276%	0.385%	0.351%	0.327%	0.388%	0.396%	0.716%
	(0.06%)	(0.05%)	(0.05%)	(0.03%)	(0.04%)	(0.03%)	(0.04%)	(0.04%)	(0.04%)	(0.04%)	(0.04%)	(0.06%)

Sample-based Analysis

- Proportions with point-wise confidence intervals

- Regions Alaska, Hawaii and Other have greater margins of error
    - Small sample sizes

- No apparent trend over the years for most regions, except for the South

- South alone:
    - No obvious seasonal patterns
    - Bump at the end of 2006/beginning of 2007

All Regions

Proportion of Flights Delayed Due to Weather by Region, 6/2003 - 12/2013

Southern Area

Proportion of Delayed Flights per Month, South (6/2003-12/2003)

Discussion

- **Expecting**: seasonal differences (more delays in winter)

- **Not expecting**: large amount of summer delays, especially in South and Southeast

- Observed regional differences in amount of delays

- No significant shift in regional differences over time

Population and Sample Monthly Proportions

Population Data: Monthly Proportion of Delayed Flights by Region (6/2003 - 12/2013)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Alaska	1.832%	1.353%	1.175%	0.517%	0.347%	0.319%	0.455%	0.661%	0.698%	0.865%	1.529%	1.768%
Central	3.155%	2.849%	1.740%	1.349%	1.614%	1.957%	1.733%	1.450%	0.809%	0.802%	1.129%	3.385%
Hawaii	0.108%	0.114%	0.113%	0.029%	0.056%	0.021%	0.030%	0.116%	0.013%	0.051%	0.091%	0.344%
Northeast	2.007%	1.853%	1.422%	1.038%	1.185%	1.992%	2.011%	1.710%	0.872%	0.963%	0.971%	2.409%
NorthRockies	1.641%	1.321%	1.000%	0.725%	0.708%	0.830%	0.739%	0.707%	0.523%	0.636%	0.830%	2.071%
Northwest	1.367%	0.505%	0.422%	0.179%	0.151%	0.210%	0.202%	0.230%	0.182%	0.293%	0.667%	1.513%
Other	0.368%	0.270%	0.218%	0.275%	0.421%	0.514%	0.762%	0.612%	0.393%	0.510%	0.328%	0.329%
South	1.808%	1.939%	1.988%	1.938%	2.472%	3.236%	2.877%	2.173%	1.232%	1.494%	1.285%	2.356%
Southeast	1.177%	1.220%	1.133%	1.028%	1.252%	2.464%	2.475%	1.981%	1.065%	0.867%	0.783%	1.217%
Southwest	1.117%	0.741%	0.617%	0.540%	0.589%	0.961%	1.079%	0.904%	0.446%	0.519%	0.457%	1.378%
UpperMidwest	1.653%	1.913%	1.349%	1.045%	1.489%	1.965%	1.728%	1.262%	0.977%	0.914%	1.069%	2.266%
West	0.651%	0.497%	0.412%	0.303%	0.276%	0.319%	0.403%	0.383%	0.277%	0.430%	0.433%	0.729%

Data was grouped by Month & Region, (n delay flights)/(n flights) for each of the 144 combinations.

p
(se(p))

Sample Data: Monthly Proportion of Delayed Flights by Region (6/2003 - 12/2013)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Alaska	1.711%	0.939%	0.941%	0.650%	0.720%	0.334%	0.718%	0.649%	0.933%	0.728%	1.312%	1.820%
Central	3.147%	2.795%	1.762%	1.349%	1.634%	1.963%	1.660%	1.468%	0.811%	0.870%	1.200%	3.323%
Hawaii	0.047%	0.103%	0.048%	0.000%	0.045%	0.000%	0.039%	0.125%	0.000%	0.047%	0.137%	0.373%
Northeast	2.029%	1.881%	1.382%	1.104%	1.065%	1.930%	2.159%	1.631%	0.799%	0.935%	0.982%	2.546%
NorthRockies	1.519%	1.527%	0.758%	0.604%	0.799%	0.862%	0.763%	0.798%	0.474%	0.637%	1.372%	2.015%
Northwest	1.079%	0.735%	0.519%	0.187%	0.086%	0.262%	0.219%	0.291%	0.179%	0.426%	0.885%	1.343%
Other	0.135%	0.141%	0.000%	0.265%	0.878%	0.617%	0.711%	1.167%	0.864%	0.160%	0.146%	0.360%
South	1.797%	1.948%	1.919%	1.979%	2.568%	3.235%	2.767%	2.229%	1.321%	1.387%	1.214%	2.477%
Southeast	1.232%	1.114%	1.122%	1.038%	1.227%	2.568%	2.484%	1.944%	1.043%	0.845%	0.921%	1.235%
Southwest	1.107%	0.695%	0.645%	0.560%	0.594%	0.852%	0.942%	0.939%	0.442%	0.354%	0.561%	1.569%
UpperMidwest	1.645%	1.800%	1.190%	0.893%	1.765%	2.008%	1.795%	1.080%	0.877%	1.008%	1.155%	2.436%
West	0.656%	0.475%	0.435%	0.214%	0.288%	0.276%	0.385%	0.351%	0.327%	0.388%	0.396%	0.716%

Obstacles and Solutions

- Defining regions & assigning regions to flights
    - left_join of DPLYR was very helpful
    - Saved the IATA, State and Region for all airports in a .csv file for easy use

- Five IATA codes in the data not documented in the Data Expo .csv file
    - Googled it!

Obstacles and Solutions

- Figuring out how to sample by our strata (region/month/year)

  - Define list of vectors:
    ```
    ak.list <- [All Alaska Airport Codes]
    c.list <- [All Central Airport Codes]
    ...
    r.list <- list(ak.list, c.list,...)
    ```

  - Within loop, assign one vector from list to a variable:
    ```
    o.list = r.list[[k]]
    ```

  - Within filter, use %in% statement and list variable:
    ```
    filter(...,origin %in% o.list,...)
    ```

Obstacles and Solutions

- Confidence intervals for zero proportions
  - Rule of Three: $3/n_h$ gives upper bound of 95% CI

- "Other" region: military bases and protectorates, not geographically consistent
  - Included in summary but no conclusions drawn from it