STAT 599 - Statistical Computing and Big Data

PROJECT 2
Bureau of Transportation Statistics

Presented and submitted by
Mathew Edwards
Nandhita Narendra Babu
Wanli Zhang

Date of Submission : 12 May 2014

1 Introduction & Question of Interest

Our question of interest is: **Are there geographical patterns in weather-related flight delays, and do these change over time?**

The data we used in this project is Airline On Time Statistics from the Bureau of Transportation Statistics, available from January 1995 through February 2014. The population used in this project includes data from June 2003 to December 2013 as weather-related flight delays were not recorded prior to June 2003.

For calculating our proportion, any flight with a weather delay value greater than 0 was counted as "delayed," regardless of severity. The denominator was a count of all flights for the strata, including those with a weather delay value of "NA".

2 Analysis

2.1 Population Assumptions

Not counting "NA" from before June 2003, we made the assumption that "NA" indicated no delay, but that may not be accurate.

Because our question involved geographic differences, we made the assumption that weather would be similar in climate regions. (This also gave us less strata to deal with.) We used data from NOAA to define the regions.[1] We added Alaska and Hawaii as their own region, and lumped any airport we could not link to a state into an "Other" region. Despite the NOAA endorsement, this assumption could be problematic, as we know that climates vary within state (see Eastern vs. Western Oregon or Washington, for example.)

For the purpose of this analysis, regions are determined based on the origin of the flight, not the destination. This decision assumes that the weather delay is on the side of origin, which it may or may not be.

We also assumed that the weather within a month would be "reasonably consistent", and aggregated to the month level.

2.2 Population Results

When aggregated to a monthly basis, by region, we see higher percentages of delayed flights in the winter months, as we would expect. The highest winter proportions are in the Central region, which encompasses the Ohio Valley. We suspect this is due in part to Chicago being a major hub for many airlines, and having cold/snowy winters. (For space, only the table of sample data is shown, but the percentages were similar for the population.)

No month or region exceeds 4% in the proportion of delayed flights.

Interestingly, there is a bump in weather-delayed flights in Jun, Jul and August for the South, Southeast, Northeast and Upper Midwest. This may coincide with tornado season. [2] We were surprised that the South was consistently above 1%, (only 3 months below 1.5%) in contrast to other regions, including the Northern Rockies and the Northeast.

2.3 Sampling Assumptions and Methods

We carried the assumptions from our population analysis through to the sampling analysis.

Our strata was a single region, for a single month. Each strata had a different number of flights per month, ranging from an average of 3,000 in Alaska to an average of 115,000 in the Southwest. We decided to use proportional sampling, and sampled approximately 2.5% of the flights in each strata. This gave us a reasonable amount in the smaller strata (75 per month out of 3000) and a very sufficient amount in the larger strata.

2.4 Sampling Results

The sampling results were very similar to the population results. Nothing is over 4%, there is the same bump in delays in the summer in the south. There are a few strata where we observed no delayed flights in the sample, but these correspond to areas in the population where there was a very low percentage of delayed flights.

The standard errors for our sample were all below 0.5%, and generally below 0.25%.

2.5 Conclusion

As one would expect, there are seasonal differences in the amount of flights delayed by weather. Surprisingly, there are a large amount of summer delays, especially in the South and Southeast.

We do see regional differences in the amount of delays.

[1] http://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php

[2] http://www1.ncdc.noaa.gov/pub/data/cmb/images/tornado/clim/tornadoes_bymonth.png

Figure 1: Sample Data: Monthly Proportion of Delayed Flights by Region (6/2003 - 12/2013)

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%)

Over the long term, there does not appear to be any significant shift in regional differences. (Graphs are in presentation due to space.)

3 Obstacles & Solutions

Defining our regions and assigning regions to flights. The NOAA map gave us the regions we wanted, and left-join was very helpful in associating the database data with the regions. We saved the IATA, State and Region for the 360 unique airports in a .csv file so we could grab it when needed.

Figuring out how to sample by our strata (region/month/year). SQL allows a WHERE(<var>IN <list>) syntax, and we were able to accomplish this by storing the airport IATA designations in a list for each region, and then plugging the list in as a variable in our filter statement: filter(origin %in% list)

Dealing with "NA" answers: Prior to June 2003, "NA" was due to the variable not being tracked. It was tracked consistently from 6/2003 to late 2007. In late 2007, many flights once again had an "NA" value. We could not find an official reason. We dealt with this by assuming "NA" meant "no delay."

There were 5 IATA codes in our data not represented in the airports.csv file. Once we identified the issue, it was a simple matter to look them up on Wikipedia.

Confidence intervals for 0 proportions: We used the rule of 3: 3/n.h gives upper bound of 95% CI

"Other" Region: not geographically consistent, small number of airports, solved by generally ignoring it.

Thinking about how to summarize data for 12 months for 11.5 years for 12 regions in a way that was concise and made sense and didn't aggregate too much. For graphing, used facets, colors, aggregated when we needed, geom-ribbon creates shades between lines.