CS-BIGS 5(1): 51-57
http://www.bentley.edu/centers/csbigs/crotty.pdf

Visualizing More Than Twenty Years of Flight Data for the Raleigh-Durham International Airport

Michael T. Crotty
SAS Institute, USA

Over two decades of airline flight data for the Raleigh-Durham International airport (RDU) are examined. Visualizing the data reveals that there are multiple phases of air traffic activity at RDU, corresponding to the transition from being an American Airlines hub airport to being a non-hub airport serving a greater variety of airlines. The changing patterns involve the daily number of flights as well as the locations of the reciprocal flights flying in and out of RDU. An analysis of arrival and departure delay data is also presented.

Keywords: *large data, visualization, data exploration*

1. Introduction

In this paper, we investigate and visualize data for domestic flights that originated or terminated at Raleigh Durham International airport (RDU). The data set contains over 2 million flights from October 1987 to December 2008. Some of the questions we used to direct the visualization project were:

- Has the fast population growth of the Raleigh-Cary metropolitan area been reflected in the number of flights in and out of RDU?
- How has the amount of air traffic changed over the 21 years of data?
- What trends in RDU's air traffic can be found in the airline data provided?
- How has the geographic distribution of airports with reciprocal traffic with RDU changed over time?
- Has the on-time performance of RDU flights improved or worsened over the 21 years of data?

Starting from the questions listed above, our analysis split into four categories which correspond to the next four sections of this paper. Section 2 describes the overall trends in the data, including descriptions of the four distinct phases of air traffic at RDU over the 21 year period. Section 3 describes the analysis of daily scheduled flights in and out of RDU. Section 4 shows the changes in geographic distribution of RDU's reciprocal flights. Section 5 gives a brief analysis of flight delays at RDU over the 21 year period. Finally, Section 6 contains some final thoughts on this visualization project and possible future extensions to it.

With the exception of Section 5 focused on (observed) arrival and departure delays, this analysis looks only at *scheduled* flights. Therefore, there is no direct focus on the effect on RDU air travel of the September 11, 2001 terrorist attacks. However, there is some evidence of lengthy delays decreasing in 2001 before increasing dramatically from 2002 onward. This will be discussed in further detail in Section 5.

This analysis was originally presented at the 2009 Joint Statistical Meetings as an entry to the Data Expo 2009 poster competition sponsored by the American Statistical

Association's Sections on Statistical Computing and Statistical Graphics. At the 2010 Joint Statistical Meetings, this analysis was presented in a topic contributed session entitled "Data Expo 2009: A Second Look at Flight Delays".

While some initial data preparation was done using the DATA step in SAS®, the bulk of the work (including all the visualizations) was performed using JMP®. Since JMP stores data in memory, it was necessary to subset the overall airline data set. Planes fly to and from RDU over the author's house every day, so confining the data to all flights in and out of RDU was a natural choice. JMP has recently made it easier to add maps to graphs, and the interactive nature of JMP makes it ideal for a data exploration project such as this one. The data and scripts for this analysis are available in the supplementary materials for this article.

2. Overall Trends Analysis

The population of the Raleigh-Cary metropolitan area grew steadily during the 21 year period; in fact, in 2009, Forbes.com named it the fastest growing metropolitan area in the country. We started our analysis by relating the yearly population of the area (using Wake County data as a proxy for the metropolitan area) to the number of scheduled flights per year and the total number of passengers going through RDU per year. Figure 1 displays this three-way comparison. We noted that the flight and passenger trends relate to each other fairly closely in the 1990s, but less so during the rest of the 21 year period.

Also, the population trend does not correlate with either the flight or passenger trends until about 1995, after which point, the flight and passenger trends track the population data better. This observation led to an investigation of the history of RDU and the airlines serving it.

After investigating the history of RDU over the 21 year period, four phases of air traffic patterns began to emerge. American Airlines (AA) used RDU as their central east coast hub up until late 1993; between late 1993 and early 1996, there was a transition period during which the hub was slowly closed down. After early 1996, there was a gradual recovery of flights and passengers. Finally, the years 2004 through 2008 were relatively stable, although there was still a fair amount of variation during these years. These four phases and their dates are listed in Table 1 and are used in the analyses of daily scheduled flights (Section 3) and of flight distribution (Section 4).

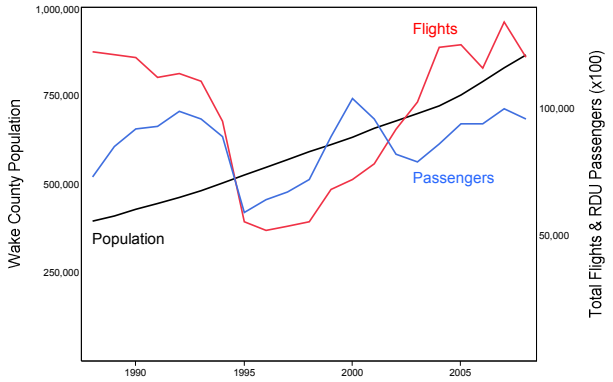

Figure 1. Comparison of annual trends in the population of Wake County (home of RDU), the number of flights at RDU and the number of passengers going through RDU. The number of passengers values are 100 times the values on the right axis. After the closing of the AA hub, the flights and passengers data track the population data better than during the years of RDU being an AA hub.

Figure 2 gives a plot of monthly trends in total flights in/out of RDU, the distinct number of airlines operating at RDU, and the number of AA flights. This plot shows the effects of the closing of the AA hub at RDU between September 1993 and April 1996. There is also a noticeable increase in the number of airlines in the Recovery and Stability phases of the data. One possible explanation for this is that after the AA hub closed, there was more room at the airport for other carriers to operate.

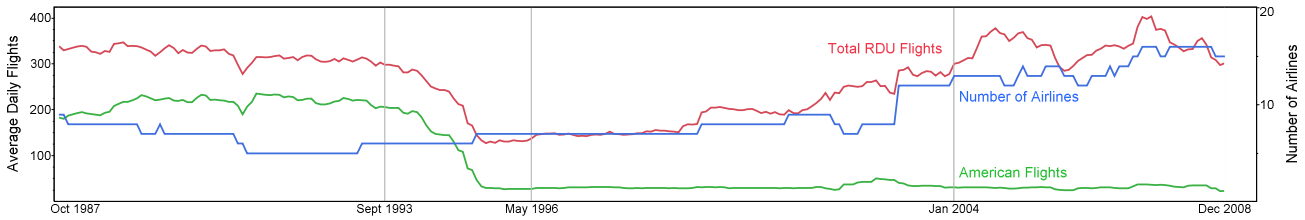

Figure 2. Time plot of the average number of daily scheduled flights per month in/out of RDU over the entire period of data as well as the average number of daily scheduled AA flights in/out of RDU. This clearly shows the initial effects of AA discontinuing its hub service. Also plotted is the number of airlines serving RDU each month. As the recovery of air traffic following the AA hub closing led to more stabilized current levels, the diversity of airlines serving RDU steadily increased.

Note that no comparison to national trends in carrier diversity was performed, so it could be that the number of airline carriers was increasing nationally over this time as well.

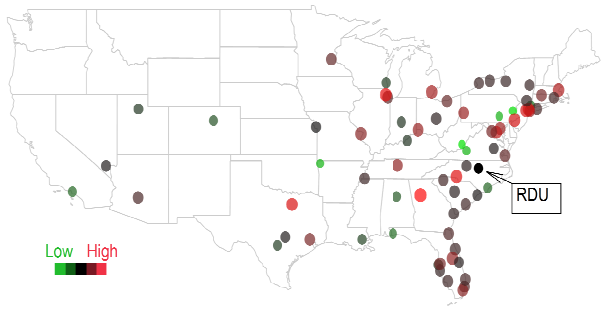

Figure 3. Map of airports that served as either a destination from RDU or an origin to RDU, sized and colored by the number of flights from 1987 to 2008. Airports with fewer than 100 flights and airports located outside the continental US were excluded from this map. See Table 2 for the number of flights that were affected by this exclusion.

The final analysis that was performed on the entire data set of scheduled flights was to look at the geographic distribution of the airports with flights to/from RDU in the 21 year period. Figure 3 shows this distribution for airports within the continental US; this restriction removed 13,249 flights to/from Puerto Rico and the US Virgin Islands. Any airports with fewer than 100 flights are excluded as well, which removed only 52 flights. Also, the original data set only contained flights entirely contained in the United States, so any international flights (to/from Canada or Europe) were excluded prior to obtaining the data.

As one might expect, the bulk of the flights to/from RDU are with airports in the eastern part of the US. Section 4 explores how the geographic distribution of reciprocal airports changed over the 21 year period, broken down into the four phases listed in Table 1.

Table 1. The four phases of the RDU airline data and their dates.

Phase	Start Date	End Date
American Airlines Hub	October 1987	August 1993
Closing the Hub	September 1993	April 1996
Recovery	May 1996	December 2003
Stability	January 2004	December 2008

3. Daily Flights Analysis

This section investigates patterns in daily scheduled flights, one phase of the data at a time. For each of the following four subsections, a *compressed time plot* was produced showing the number of scheduled flights for every day in the time period of the particular phase. The days are color coded by day of the week and grouped in vertical columns by month. There are also vertical reference lines to denote years. While this graphical technique might obscure the day-to-day time series appearance, the goal was to visualize the trends over larger temporal scales. At the same time, all the data points are still shown in a concise graph. Note that all of the graphs in this section use the same vertical scale and range; this is to allow for comparison across the four time periods.

A smoothing spline was fit to the data in each phase, to help visualize the trend over time. The spline was fit to the data as it is represented graphically. All data points in a single month are treated as having the same time value for the purposes of the spline fit.

3.1. American Airlines Hub (October 1987 – August 1993)

Figure 4 displays the daily scheduled flights for the first phase, corresponding to the time period when AA used RDU as one of their hubs. During this time, the number of flights is fairly steady. There is an unexplained dip and recovery in early 1991.

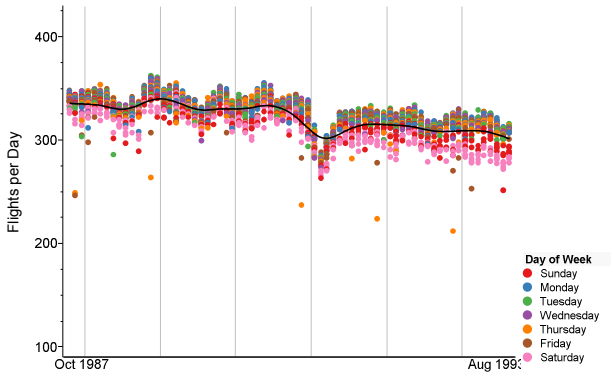

Figure 4. AA Hub (Oct 1987 – Aug 1993). Compressed time plot of number of flights per day by month with a smoothing spline fit. Points are colored by day of the week.

An interesting trend that shows up in this plot is that for most years, there is a very low data point, which is for a Thursday in November – Thanksgiving Day in most years has many fewer scheduled flights.

Because the data are grouped by month, it is not easy to view the days around Thanksgiving, although for some years, the Friday after Thanksgiving is visible as being below the rest of the data points in November.

Finally, weekdays consistently have more scheduled flights than Saturdays and Sundays. This is not surprising, since there would be a lot of business travel during the week. This could also be shown with a monthplot, but then the variability of a particular day of the week within a month would be lost.

3.2. Closing the Hub (September 1993 – April 1996)

Figure 5 displays the daily scheduled flights for the second phase, corresponding to the time period when AA closed its hub at RDU. The most obvious aspect of the daily flights analysis for this phase is the dramatic decrease in the number of flights. This decrease corresponds with the closing of the AA hub at RDU, which was a gradual process that took about 2 years. By early 1996, the number of daily flights had bottomed out at between 100 and 150 flights per day. Throughout this period, the higher number of flights on weekdays continues to hold. However, the Thanksgiving Day effect is only strong in 1993, and is much less dramatic in 1994 and 1995. This is also the shortest phase of the data.

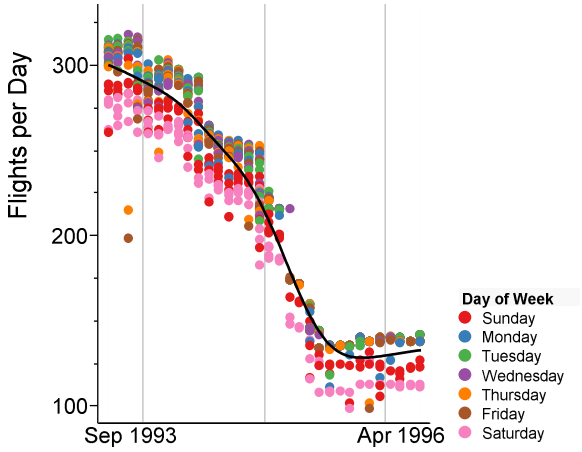

Figure 5. Closing the Hub (Sep 1993 – Apr 1996). Compressed time plot of number of flights per day by month with a smoothing spline fit. Points are colored by day of the week.

3.3. Recovery (May 1996 – December 2003)

From the shortest phase of the data, we move to the longest phase, which is a gradual recovery of flights over 8 years. In this phase, shown in Figure 6, we start to see the weekend effect change a bit. Sunday is still lower than weekdays, but Saturday is now even lower than Sunday. This trend also seems to intensify toward the end of this phase. The Thanksgiving Day effect is still present in this phase as well, although there is less difference between Thanksgiving Day and other very low days of other months of the year now.

By the end of this third phase of the data, we see that the number of daily flights is back to around 300, which is near where it was for the AA Hub phase. Finally, as mentioned in Section 1, since this is an analysis of scheduled flights, it is difficult to see any noticeable effect of the September 11, 2001 terrorist attacks.

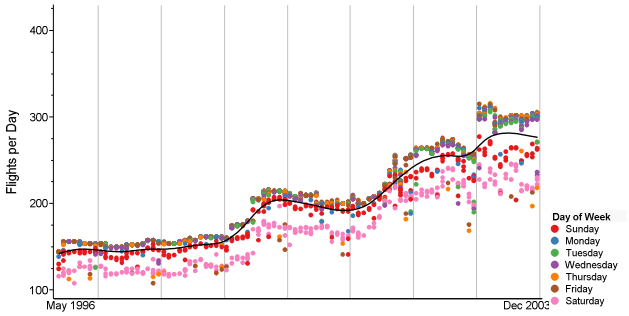

Figure 6. Recovery (May 1996 – Dec 2003). Compressed time plot of number of flights per day by month with a smoothing spline fit. Points are colored by day of the week.

3.4. Stability (January 2004 – December 2008)

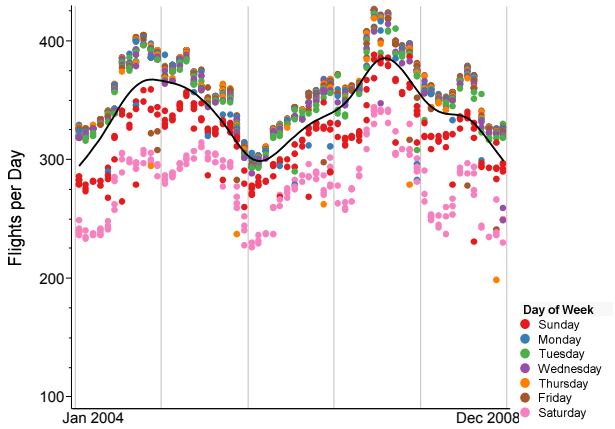

Figure 7. Stability (Jan 2004 – Dec 2008). Compressed time plot of number of flights per day by month with a smoothing spline fit. Points are colored by day of the week.

The final phase of the data, shown in Figure 7, shows a lot of variability, but the daily number of flights continues to average between 300 and 400. This is actually slightly higher than the daily number of flights in the first phase when RDU was an AA hub.

Interestingly, the weekday versus weekend difference is greatest in this phase, while the Thanksgiving Day trend is much harder to identify in many of the years in this

phase. The Saturday flights being less than the Sunday flights trend is even greater in this phase than in the Recovery phase.

We do not have any theories on why there seems to be a 2 to 3 year cycle in the Stability phase, but the decline at the end of 2008 could be related to the start of the economic recession of the late 2000s in the United States.

4. Flight Distribution Analysis

This section breaks down the overall geographic distribution of flights to/from RDU that was presented in Figure 3 for each of the four phases of the data. Just as in Figure 3, airports with fewer than 100 flights were excluded from the maps, as were flights to/from Puerto Rico and the US Virgin Islands. See Table 2 for details on how many flights were excluded from each map based on these criteria.

Table 2. Number of excluded flights for Figures 8 through Figures 11.

Phase	Under 100 flights	Outside Continental US
American Airlines Hub	4	9889
Closing the Hub	1	3156
Recovery	10	204
Stability	37	0
Total	52	13249

At the beginning of the 21 year period when RDU was an AA hub (shown in Figure 8), all of the reciprocal airports are east of the Mississippi, with the exceptions of St. Louis and Dallas. Much of the traffic is to the north or south of RDU, consistent with the status of RDU being AA's central east coast hub between New York and Miami. Chicago, Dallas and Atlanta also have a large number of flights in the AA Hub phase.

During the Closing the Hub phase (shown in Figure 9), there is not much change in the distribution of flights, with the exception that it becomes a bit sparser overall.

The Recovery and Stability phases are perhaps the most interesting in terms of geographic distribution. They show the rise of geographic diversity of reciprocal airports, primarily west of the Mississippi. The introduction of Southwest Airlines to RDU may account for many of these new destinations in the western US, including Chicago's Midway airport. Between Figures 10 and 11, there is a general trend of more flights to/from airports west of RDU and fewer east coast airports. However, in the Stability phase, the share of flights to/from the New York airports does seem to increase.

5. Delay Data Analysis

Our final analysis uses a different aspect of the data than the previous sections. Instead of looking at scheduled flights, we shift our focus to actual delays for both arrivals and departures. Specifically, we are interested in answering the question of whether delays at RDU have changed over the 21 year period.

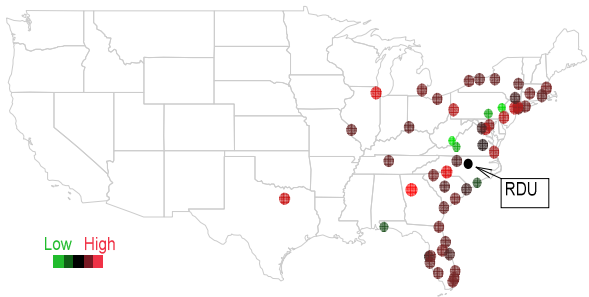

Figure 8. AA Hub (Oct 1987 – Aug 1993). Analog map to Figure 3 for the period that RDU served as an AA hub.

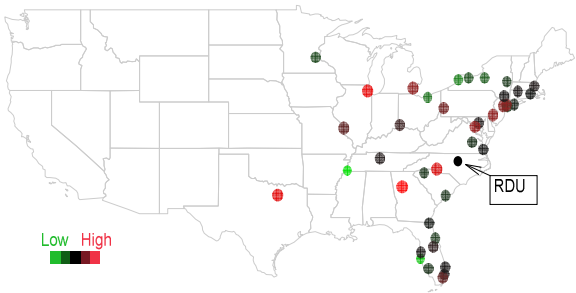

Figure 9. Closing the Hub (Sep 1993 – Apr 1996). Analog map to Figure 3 for the period that AA scaled back hub service at RDU.

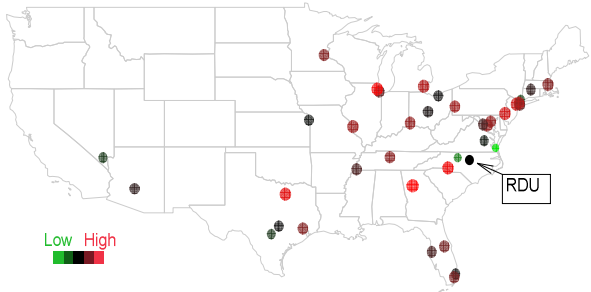

Figure 10. Recovery (May 1996 – Dec 2003).Analog map to Figure 3 for the period that RDU traffic increased to 2008 levels.

The distribution of delays (measured in minutes) is necessarily a skewed distribution; flights can only arrive (or depart) so many minutes early, but they can be many

minutes, or even hours, late in arriving or departing. To account for this skewness in the data, we approached the problem by looking at medians of the delays per month. The median delay is represented by the black line in Figures 12 and 13 for arrivals and departures, respectively.

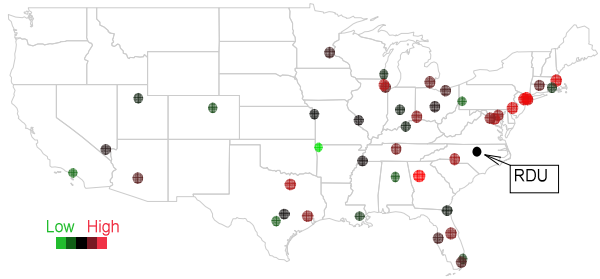

Figure 11. Stability (Jan 2004 – Dec 2008). Analog map to Figure 3 for the last five years where RDU traffic has leveled off.

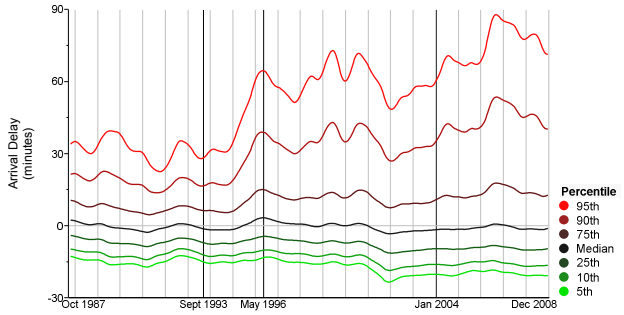

Figure 12. Arrival Delays.Plot of percentiles of the monthly arrival delays.

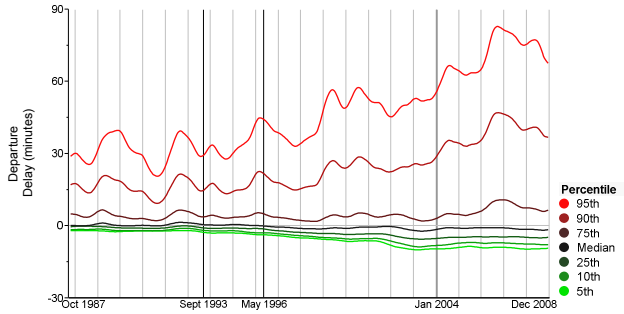

Figure 13. Departure Delays.Plot of percentiles of the monthly departure delays.

After noticing that the median delay was fairly constant over the 21 year period for both arrivals and departures, we delved deeper into the data and looked at the 5th, 10th, and 25th percentiles on both the high and low side of the median. Because of the skewness of the data, the three percentiles below the median are fairly close together and not terribly interesting. The three percentiles above the median, however, show that there has been an increase over time in the length of the longest delays for both arrivals and departures.

There are a few possible explanations for the increased length of longer delays. One is that there is more security to go through after September 11, 2001 that could be slowing down the system in general. Also, as can be inferred from Figure 1, after 2000, either flights operating out of RDU are simply using larger planes or they are filling the planes more efficiently; it is plausible that either of these could increase lengthy delays.

We fit a smoothing spline to the monthly percentiles as a way to clearly visualize the trend over time. Also, flights that were recorded as over 1 hour early or over 6 hours late were excluded as being suspect data. This only removed 47 arrivals and 33 departures out of over 2 million flights. As with other parts of the project, we have not compared any of these RDU trends to overall national trends.

6. Conclusions

Changes in the patterns of air traffic at RDU over 21 years (from 1987 to 2008) are explored. Most notably, American Airlines utilized RDU as a hub on the central east coast of the US until 1993. By 1996, the hub was completely gone, but a slow recovery started to take place. The levels of air traffic were more or less stable over the five year period ending in 2008. The population growth of the surrounding area only seems to correlate with air traffic levels in the post-hub time period. Also, as more airlines served the airport, more destinations were added. However, delays for both arrivals and departures increased as well.

There are many possible future investigations that could be pursued. One would be to compare RDU's trends with national trends. The analysis performed here could even be replicated for other airports around the country.

It would also be interesting to get more data from the RDU airport aside from just flight data that could help shed light on the validity of the speculative explanations for the trends presented here. With regards to the delay data, adding weather data could be very interesting in helping explain some of the variation in delays. There are also more data for RDU contained in the flight data provided by the Data Expo 2009 competition that could be further explored, especially in the on-time performance area. What airline, time of day, day of the week, month of the year should you choose to travel on to minimize the chance of being delayed at RDU?

Finally, it is clear that on-time performance is only one criterion when people book an airline ticket. It is tempting to think about combining other factors, namely ticket price, to this data set to better assess various regular flights in and out of a particular airport.

Acknowledgements: The author wishes to thank the editor and an anonymous reviewer for their helpful comments.

References

Data Expo 2009. Available at http://stat-computing.org/dataexpo/2009

City growth. Available at http://www.forbes.com

Passenger data. Available at http://en.wikipedia.org/wiki/Airline_hubs_at_RDU#Passenger_statistics

Population data. Available at http://www.google.com/publicdata

Flight data. Available at http://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

Airport data. Available at http://www.faa.gov

Airline data. Available at http://stat-computing.org/dataexpo/2009/supplemental-data.html

RDU history. Available at http://www.rdu.com/aboutrdu/history.htm

RDU history. Available at http://wikibin.org/articles/airline-hubs-at-rdu.html

Correspondence : Michael.Crotty@sas.com