An Analysis of Airline Delays with SAS/IML® Studio

Rick Wicklin
SAS Institute Inc.

November 24, 2009

1 Introduction

The Data Expo is a biannual poster session usually sponsored jointly by the ASA Sections on Statistical Graphics and Statistical Computing. The purpose of the poster session is to distribute an interesting data set to many researchers and to challenge them to use statistical graphics to describe and visualize the data concisely on a single poster. The session helps highlight the importance of statistical graphics in data analysis.

This year's Data Expo was organized by Hadley Wickham, who assembled a truly massive set of data from the Research and Innovative Technology Administration (RITA) which coordinates the U.S. Department of Transportation (DOT) research programs. The data (available from `http://stat-computing.org/dataexpo/2009/`) consist of 123 million records of U.S. domestic commercial flights between 1987 and 2008. Each flight contains information about 29 variables, including the following:

- Dates: day of week, date, month, and year
- Arrival and departure times: actual and scheduled
- Origin and destination: airport code, latitude, and longitude
- Carrier: American, Aloha Air, ..., US Air

SAS software excels at handling massive data. This proved to be an advantage: the poster by Wicklin and Allison (2009) was awarded first place. This article describes several graphs in the poster. You can browse the electronic version of all the entries at `http://stat-computing.org/dataexpo/2009/posters/`.

The poster graphically presents ways in which flight delays and cancellations vary in time, among airports, and among airline carriers. The article also describes graphical methods and features of the data that elicited the most comments from visitors to the Data Expo.

All but one of the graphs in this paper were created by SAS/IML® Studio (formerly named SAS® Stat Studio) which is new software in SAS 9.2. SAS/IML Studio is intended for data analysts who are familiar with SAS/STAT® software, but who need a versatile programming environment to develop new computational algorithms or statistical graphics. See Wicklin (2008) for more information on SAS/IML Studio.

2 Data Reduction

When presenting data graphically, it is important to be able to quickly convey the main features in the data. But how can you summarize 21 years worth of data on the delays or cancellations of 123 million flights?

The DOT defines a departing flight as "delayed" if it departs more than 15 minutes after its scheduled departure time. The 15 minute window is also used for defining when an arriving flight is delayed.

For these data, it is useful to reduce the size of the data by computing a descriptive statistic such as the mean length of a delay or the percentage of flights delayed. These statistics are calculated over an appropriate aggregation unit such as a date, a year, an airport, or a carrier. This often proved to be an important first step in understanding and conveying gross trends in the data. All but two graphs in the poster displayed descriptive statistics, rather than the raw data.

For example, one quantity displayed in several graphs is the "percentage of flights delayed" (PFD) for a given day. (This is a standard quantity measured and reported by the DOT; see `http://www.bts.gov/`.) If a certain day has 30,000 flights and 6,000 of those are more than 15 minutes late, then the PFD for that day is 20%. Plotting the PFD for each day instead of the original data results in a significant reduction in the volume of data, but at the usual risk of losing details that might be important.

3 Temporal Effects: A 21-Year Summary

You can use a graphical layout called a *calendar plot* to present how the PFD varies over 21 years. A calendar plot for five years of the data is shown in Figure 1. For each year, there are seven rows that represent the days of the week and usually 53 columns that represent each week in the year. The color of each day represents the value of the variable for that day. The earliest reference I have seen for a calendar plot is Mintz, Fitz-Simons, and Wayland (1997). A SAS macro for computing the plot was presented in Zdeb and Allison (2005).

The calendar plot is a choropleth map: each year is a "country," each month is a "state," and each day is a "county." By thinking of the calendar plot as a choropleth map, you can use ideas from the cartographic literature to guide the design of the plot. For example, you can follow Pickle et al. (1996) and choose a color for each day as determined by the quantile of the PFD for each day.

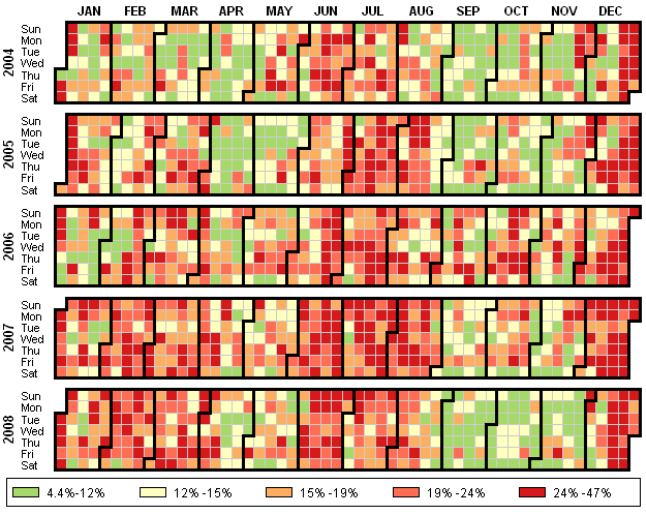

Figure 1: Calendar Plot for Percentage of Flights Delayed (2004–2008)

In Figure 1, the colors are determined by quintiles of the PFD for all days in the five-year time period. The colors are chosen to suggest a traffic light color scheme: the first quintile of the data is colored green for "go"; the second and third quintiles are yellow and pale orange for "caution"; a

darker orange and red are used for the upper quintiles to signify extreme delays. The actual colors are based on a diverging color scheme from `ColorBrewer.org` (Brewer, 2006). This web site is a valuable resource for creating color schemes because Brewer's schemes have the important property that each color is equally-perceived: no one color catches the eye and draws the viewer's attention. The colors are also designed to be distinguishable to the color blind.

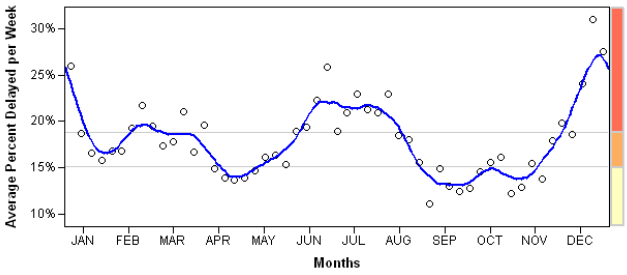

Figure 2: Average Weekly Percentage of Flights Delayed (2004–2008)

A few aspects of the data are apparent by looking at the color patterns of columns: the PFD is relatively low in the spring and fall, whereas summer and the last two weeks of December have many days with a high PFD. The PFD in January–March can be quite variable, presumably due to winter storms. If you average the values of the calendar plot for each week (that is, take the average of each column), you obtain the scatter plot in Figure 2 which plots the average weekly PFD versus the week of the year. A loess smoother (adjusted for periodicity in the data) is overlaid on the plot so that it is easier to see that, on average, the PFD is high during the summer and Christmas seasons but low during the spring and autumn. In a paneled layout, this figure can be placed underneath the calendar plot, similar to the arrangement used by Peng (2008).

A calendar plot for the percentage of flights canceled (PFC) can be similarly constructed and shows similar features. However, the most striking feature of the calendar plot for canceled flights is related to the terrorist attacks on September 11, 2001, as shown in Figure 3. In the two weeks prior to 9/11, an average of 2.3% of flights were canceled. By the middle of October, 2001, the average PFC was down to 1% of flights, due in part to a 17% reduction in the number of scheduled flights per day. The daily PFC remained low until 2004.

As suggested by Pickle et al. (1996), it is a good idea to plot the distribution of the variable that you are using to color the map. Figure 4 shows the distribution of the PFC variable for the five years shown in Figure 3. (The distribution over the full 21 years is similar, but was not included in the poster due to space considerations.) This distribution exhibits a long tail.

Following Pickle et al. (1996), there is a color bar beneath the density plot and the horizontal axis is truncated at the 99th percentile of the data. (The actual maximum is indicated in the graph.) The color bar shows the values that correspond to the colors used in the calendar plot; this gives the viewer a visual impression of the varying lengths of the quantiles. An investigation of the days with the most cancellations reveals that they are primarily related to the 9/11 attacks or to extreme weather events.

In summary, you can use the calendar plot to reveal features and seasonal trends for 21 years of airline delays and cancellations.

4 Carrier Effects: Multivariate Visualization of Time Series

The calendar plot is useful for showing how a single quantity varies during long periods of time. However, it is not so useful for comparing multiple quantities that vary in time. For example, suppose

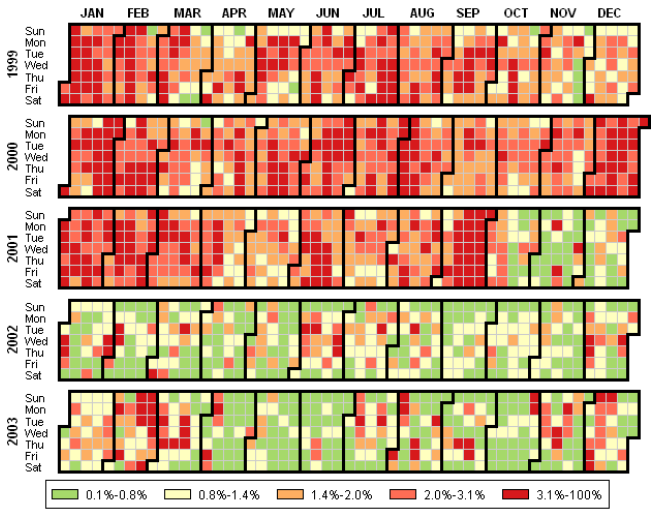

Figure 3: Calendar Plot for Percentage of Flights Canceled (1999–2003)

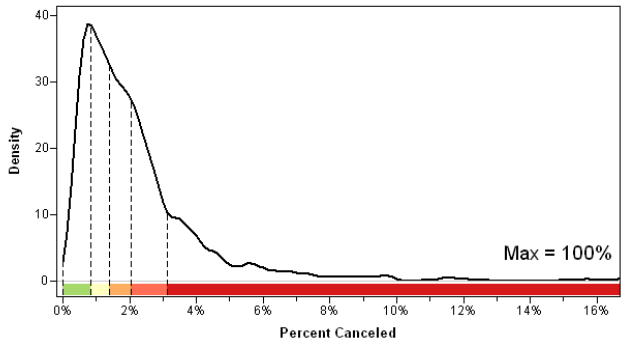

Figure 4: Distribution of the Percentage of Flights Canceled (1999–2003)

you are interested in comparing the mean length of delays for flights among the different carriers. Are there some carriers that are almost always on time, while others are habitually late? To investigate this question, you can construct a multivariate time series plot as described by Peng (2008).

The plot is displayed in Figure 5 for 20 major airline carriers during 2007. That year is chosen because there was a major winter storm (the "Valentine's Day Blizzard") that affected the entire eastern half of North America and paralyzed transportation in all forms (Wikipedia, 2009). The effect of the storm is visible as a dark brown vertical line in the middle of February.

Following Peng and a suggestion from John Sall (personal communication), the carriers are sorted according to the mean delay for all flights during the year. This makes it easy to see that Aloha Airlines (AQ) and Hawaiian Airlines (HA) have superior on-time performance, presumably because of short flight times and fair weather! The next best performing airlines in 2007 according to this metric were Southwest (WN), Frontier (F9), Delta (DL), AirTran (FL), Express (9E), and JetBlue (B6).

Again, it is useful to use ideas from choropleth maps to choose colors for these data. For this plot the colors are based on seven quantiles and the colors are based on a blue-brown color ramp developed Cindy Brewer and used successfully in Pickle et al. (1996). (This double-ended color scheme does a good job of simultaneously depicting both high and low values.) The blue shades correspond to days for which the mean daily delay for a particular carrier was less than five minutes. The brown colors

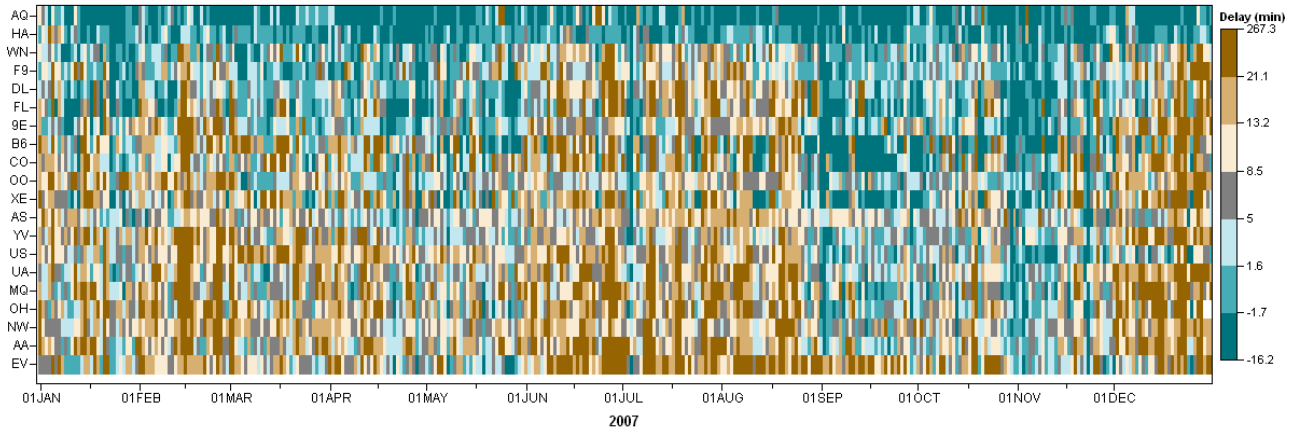

Figure 5: Multivariate Time Series of Mean Delay for Each Carrier (2007)

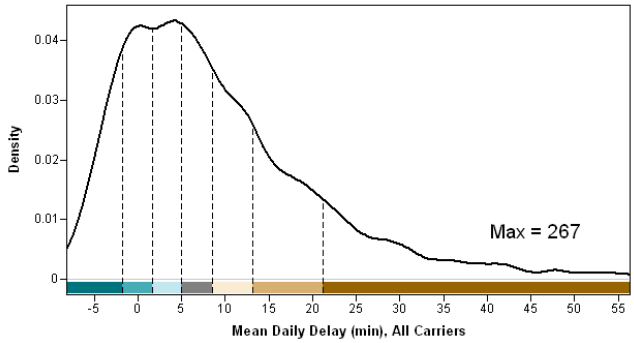

Figure 6: Distribution of the Mean Daily Delay for Each Carrier (2007)

correspond to days and carriers for which the mean delay was between 13 and 21 minutes (light brown) or more than 21 minutes (dark brown). The distribution of mean daily delays is shown in Figure 6. You can see that the median value of the distribution corresponds to a mean delay of seven minutes, but that the distribution has a long tail.

A box plot (not shown) of the mean delays for each day of the year could be placed on the right side of Figure 5 as suggested by Peng (2008) and used in a cartographic context by Carr, Wallin, and Carr (2000). The box plot enables you to see the within-carrier variation of the daily means throughout the year and conclude, for example, that even though the delays for Frontier airlines are usually low, there is a large amount of variation in the daily means.

Two features of this multivariate time series plot are striking. The first is the "outlying" behavior of Aloha and Hawaiian Airlines. The second is the well-defined vertical bands of brown (long delays) for most carriers in February, in the summer, and in the latter half of December. This is in marked contrast to the faint blue bands in April and May, and the heavier blue bands in September through early November.

These qualitative trends can be visualized further by smoothing the time series, as shown in Figure 7. Each point in this plot is the mean of the mean daily delays for the data shown in Figure 5, excluding the data for Aloha Airlines (AQ), Hawaiian Airlines (HA), and Alaska Airlines (AS). A loess smoother (adjusted for periodicity in the data) is overlaid on the plot. You can see that, on average, the mean delays in the spring and fall are shorter than the mean delays during the summer and during

the latter half of December. The Valentine's Day Blizzard (14FEB2007–16FEB2007) is apparent in the graph. The graph also highlights certain days near U.S. holidays: Independence Day (04JUL2007), the day before Labor Day (02SEP2007), and the Friday after Thanksgiving (23NOV2007) have short delays, whereas the day before Christmas Eve (23DEC2007) is associated with long delays.

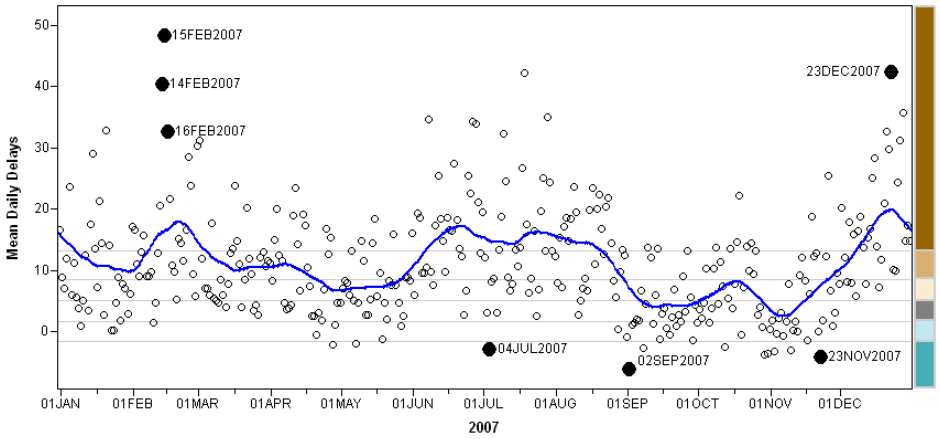

Figure 7: Average Mean Daily Delay (2007)

In summary, the multivariate time series plot enables you to see gross similarities and differences between the mean daily delays of the 20 carriers.

5 Airport Effects: Interactive Charts

You can create a graph similar to Figure 5 for airports. In fact, you can create two such graphs: one for flights that originate at the airport, and another for flights for which the airport is the destination. However, you can also present the relationship between airports and delays by using interactive and dynamically linked graphics.

The interactive graphic takes the form of a map of the continental U.S. with a single airport selected, as shown in Figure 8. This graph is nicknamed the *splay plot* because it resembles splayed fingers. All flights that originate from the selected airport are grouped according to their destination and are colored according to the PFD for each destination. Note that some routes are consistently on time, whereas others are frequently delayed.

In the interactive version of Figure 8, you can click on the splay plot to select a different airport. The plot uses SAS/Intrnet® software to detect the click and to run a SAS program to create a splay plot for the new airport. (See the poster for an example.) So that the colors on the splay plot do not change from one airport to another, the colors are hard-coded into five categories: less than 10% of flights delayed (green), between 10% and 15% of flights delayed (purple), and so on, up to more than 25% of flights delayed (red).

The splay plot could be improved in several ways. It could include a density estimate similar to Figure 4 that shows the distribution of the coloring variable for each route. At the Data Expo, Simon Urbanek suggested using semitransparent lines in Figure 8.

A second approach to presenting the relationship between airports and delays is to use dynamically linked graphs. Figure 9 shows this approach in the SAS/IML Studio application. This figure shows the PFD for each major airport and for each year. The linked graphics enable you to select observations according to certain criteria. For example, you can select airports with a large PFD by selecting

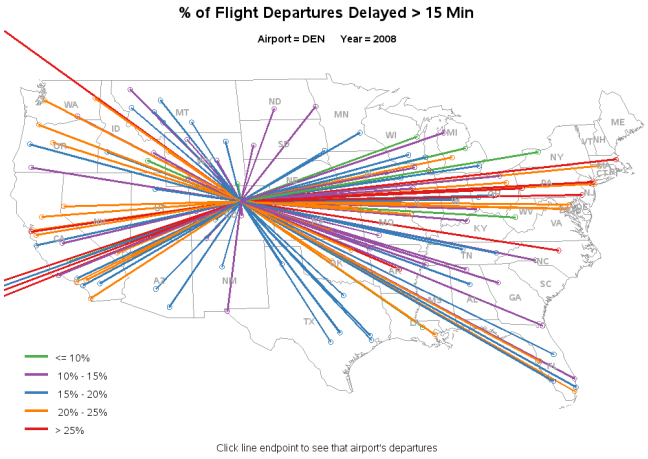

Figure 8: Percentage of Flights Delayed (Denver, 2008)

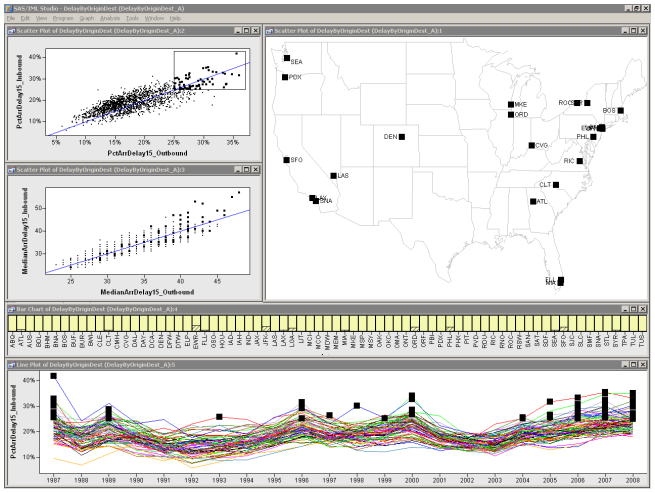

Figure 9: Relationships between Delays, Airports, and Years (1987–2008)

markers in the scatter plot in the upper-left plot, as shown in the figure. Or you can select certain airports or certain years or both and examine the PFD for those selected observations.

Interactively exploring Figure 9 revealed a difference between "hub" and "non-hub" airports. Hub airports appear to give preferential treatment to inbound flights. An example of this appears in the scatter plot shown in Figure 10 where the percentage of inbound flights that are delayed (PIFD) is plotted against the percentage of outbound flights that are delayed (POFD) for all major airports and all years. The large blue markers correspond to flights delayed at Chicago O'Hare during 1997–2008. Note that for each year the PIFD is less than the POFD. (The diagonal line is the identity line.)

Philip Easterling, a former analyst with a major US airline, explained after the Data Expo that this is deliberate: hub carriers tend to hold outbound flights when an inbound connection is late, so that a single late inbound flight can result in dozens of outbound delays. For this same reason, air traffic controllers give preferential treatment to landing inbound flights at hubs.

In contrast, the non-hub airports tend to try to get outbound flights off the ground on time so that they can arrive at the hubs on time. For the specific example of Seattle, shown in Figure 10, Easterling explained that arriving flights are often delayed due to clouds and fog, but that the ground crews work

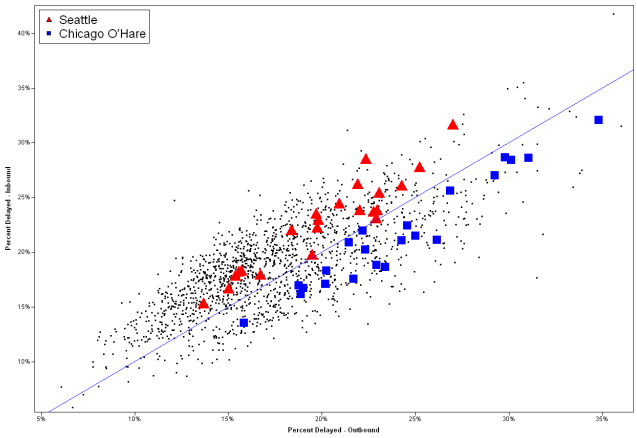

Figure 10: Differences between Hub and Non-Hub Airports (1987–2008)

hard to "turn around" these planes so that they depart on time.

6 Conclusions

The data set used for the 2009 Data Expo is exceedingly rich. It would be easy to fill up *two* posters with graphs of these data! This article briefly describes a few graphs that elicited the most comments from viewers of the poster. The two main approaches used in this article are (1) to reduce the dimensions of the data by plotting descriptive statistics aggregated over dates, years, airports, or carriers; and (2) to use design principles from cartographic research to aid the visualization of multivariate time series. These techniques enable you to create graphs that exhibit relationships among a dozen variables in a data set with 123 million observations.

7 Acknowledgements

Figure 8 was created by Robert Allison by using SAS/GRAPH® software. I thank Robert for help in preparing the original poster and for introducing me to calendar plots. Bob Rodriguez provided helpful comments on the graphics. Peter Borysov helped to write programs in SAS/IML Studio for graphs in the poster that were originally created with SAS/GRAPH software. Linda Pickle reviewed a draft of this article and made helpful suggestions. Finally, I thank the management at SAS, particularly Joe Hutchinson and Radhika Kulkarni, for supporting my participation in the Data Expo.

References

Brewer, Cynthia A. 2006. ColorBrewer. URL `http://www.ColorBrewer.org`.

Carr, Daniel B., Wallin, John F., and Carr, D. Andrew. 2000. "Two New Tools for Epidemiological Applications: Linked Micromap Plots and Conditioned Choropleth Maps." *Statistics in Medicine*, 19:2521–2538.

Mintz, D., Fitz-Simons, T., and Wayland, M. 1997. "Tracking Air Quality Trends with SAS/GRAPH." In *Proceedings of the Twenty-Second Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Peng, Roger. 2008. "A Method for Visualizing Multivariate Time Series Data." *Journal of Statistical Software*, 25(1):1–17. URL `http://www.jstatsoft.org/v25/c01`.

Pickle, L. W., Mungiole, M., Jones, G. K., and White, A. A. 1996. *Atlas of United States Mortality*. National Center for Health Statistics, Hyattsville, MD. URL `http://www.cdc.gov/nchs/products/other/atlas/atlas.htm`.

Wicklin, Rick. 2008. "SAS Stat Studio: A Programming Environment for High-End Data Analysts." In *Proceedings of the SAS Global Forum 2008 Conference*, Cary, NC: SAS Institute Inc. URL `http://www2.sas.com/proceedings/forum2008/362-2008.pdf`.

Wicklin, Rick, and Allison, Robert. 2009. "Congestion in the Sky: Visualizing Domestic Airline Traffic with SAS Software." Poster presentation, Joint Statistical Meetings. URL `http://stat-computing.org/dataexpo/2009/posters/`.

Wikipedia. 2009. "February 2007 North America winter storm." URL `http://en.wikipedia.org/wiki/February_2007_North_America_Winter_Storm`.

Zdeb, Mike, and Allison, Robert. 2005. "Stretching the Bounds of SAS/GRAPH Software." In *Proceedings of the Thirtieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.