

An Analysis of Days of the Week in the National Airspace System

Brendan Hogan
University of Virginia
STAT513/SYS613 Applied Multivariate Statistics
Final Project Report

May 5, 2006

Abstract

The National Airspace System in the United States consists of complex interactions between airlines, airports, airspace, and the air traffic controllers handling the flights. Because it is difficult to model in a meaningful way analytically, many studies of proposed changes to this system are done through simulation. This requires generating input data, including lists of the actual flights to be evaluated with their departure and arrival times, origin and destination airports, and other features of the flight depending on the goals of the simulation study. Typically this is done by starting with an actual day of air traffic and then modifying to fit the task at hand, for example by forecasting to a future volume of traffic or by adjusting the routes flown for an airspace study. A critical step in this process is the choice of the specific dates that are used as the basis for the simulated air traffic.

This present work attempts to aid this process by examining natural groupings that may exist between air traffic activity on different days of the week. If there are distinct groups of days of the week in the National Airspace System, then selecting a baseline traffic date from each of these groups will capture different aspects of flight activity. On the other hand, it would be possible to select a date close to the center of one of these groups and have it be representative of the air traffic patterns in the group at large. We show that this is indeed possible through a principal component analysis of 2005 Air Route Traffic Control Center data. Additionally, it seems that Wednesdays may be the day of the week most representative of weekday traffic activity as a whole.

1 Introduction

Simulation studies of the National Airspace System (NAS) are an effective and commonly used technique in the analysis of system performance questions. As an example of a question commonly evaluated through simulation, consider a 'what-if' scenario of how air traffic delays might react to the forecasted level of flights ten years from now both with and without proposed improvements in airport capacity. In all simulation studies such as this one it is critically important to provide input data that is representative of the aspects of the system being studied [6]. An analysis would be unreasonably biased if the input data fed into the simulation model included some atypical events that were not intended to be part of the study. Therefore it is important to establish the goals of the simulation study and develop the input data accordingly.

Often times the goal in a simulation study involves some analysis based on a 'typical day' of traffic in the NAS. In this case an analyst would choose a date for the basis of the input traffic that is representative of an average day. Previous work towards the generation of simulation input traffic has used Thursdays and Saturdays as representative weekdays and weekend days, respectively [2]. However no formal justification was done on whether this choice of days truly captures the desired variability in the system. This present work is an attempt to fill in that gap by examining any natural groupings that exist between air traffic activity on different days of the week, and to use that as the foundation of choosing days that are truly representative of weekday and weekend activity.

There has been some prior work done in clustering days in the NAS into functional groups, but this has largely involved the combination of traffic activity with weather and delays. Callaham, et. al. [3] did a clustering of weather day-types and traffic day-types based on the proportion of flights on each day that were cancelled, diverted, or otherwise severely impacted by the environment. Krozel, et. al. [5] also did a clustering of types of days in the NAS that was heavily weighted by the overall traffic volume and the total minutes of Ground Delay Program delay distributed to flights on each day. While it is useful to know that days in the NAS have successfully been grouped according to the interaction of volumes, weather, and delays, this current work has a slightly different focus.

At a level of abstraction higher than weather and delay effects, we aim to identify overall patterns in NAS activity by day of the week while considering regional impacts. As mentioned above, many studies of the NAS are interested in the big picture of traffic trends on a typical day. For these studies the details of delays and weather that flights encountered are not important, but rather it is the volume of traffic in particular areas of the country that are important. We describe a principal component analysis that explains most of the variability in the system through a lower-dimension representation, and allows for identification of day of the week patterns in NAS activity. The rest of the paper is outlined as follows. Section 2 describes the acquisition of the data for the study, Section 3 details the principal component analysis of this data, Section 4 identifies day of the week patterns based on the principal components, and Section 5 offers some conclusions.

2 Data Collection and Processing

This work consists of an analysis of traffic activity organized by Air Route Traffic Control Centers (ARTCCs, or Centers for short). There are 20 ARTCCs that handle en-route air traffic in the Continental United States as shown in Figure 1. The names and three-letter identifiers of these Centers are given in Appendix A for reference.

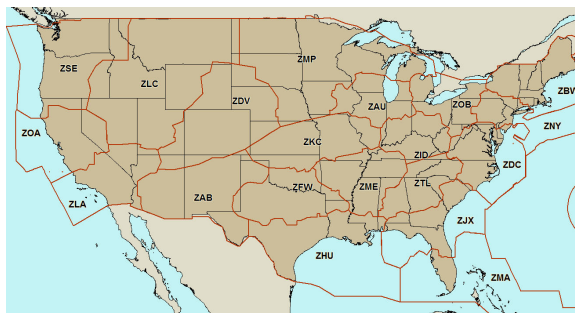


Figure 1: The 20 ARTCCs handling en-route traffic in the Continental United States

The data for this study comes from the Federal Aviation Administration’s Operations Network (OPSNET) database [1]. Specifically, the daily air traffic counts by Center were obtained for all of calendar year 2005. The data obtained in its raw form from the database contains counts of departures as well as overflights broken out by each of the four user classes: air carrier, air taxi, general aviation, and military. For the purpose of this study we are considering the total operations count across the four user classes and consisting of both departures and overflights. Some processing was done to extract just this total count from the raw file for each center and date and also to add a label for the day of week (see Appendix C.1 for details). This data was then organized as a matrix of 365 observations of 20 variables each as shown below.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,20} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ x_{365,1} & x_{365,2} & \cdots & x_{365,20} \end{bmatrix}$$

To give the reader a sense for the scale of this data, the sample mean vector containing the average daily number of aircraft handled by Center in 2005 is given below and the corresponding sample correlation matrix is given in Appendix B.

$$\bar{\mathbf{x}}' = \begin{bmatrix} \text{ZAB ZAU ZBW ZDC ZDV ZFW ZHU ZID ZJX ZKC ZLA ZLC ZMA ZME ZMP ZNY ZOA ZOB ZSE ZTL} \\ 4828 7940 5120 8437 5031 5846 5814 7902 7005 5707 6269 4251 6852 6322 5867 8422 4666 8274 3578 8865 \end{bmatrix}$$

3 Principal Component Analysis of Daily ARTCC Traffic

Principal component analysis is a common technique for reducing the dimension of a data set, therefore making it easier to work with and interpret in later steps of an analysis [4]. The goal in doing this is to represent the variance-covariance structure of a set of variables through a few linear combinations of these variables. Note that a complete set of as many linear combinations as there are variables, in this case 20, is required to capture 100% of the variance in a dataset. However, often the vast majority of the variance in the dataset can be explained by the first few principal components.

As shown in the vector $\bar{\mathbf{x}}$ in the previous section, there are large differences in the mean daily air traffic volumes between the ARTCCs. For that reason we perform the principal component analysis with respect to the standardized variables in order to eliminate any undesired effects due to the scaling. The standardized variables are defined as

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}}$$

where s_{jj} is the $j, j - th$ component of the sample standard deviation matrix, and the indices range from across the $i = 1, 2, \dots, n = 365$ observation days and the $j = 1, 2, \dots, p = 20$ variable ARTCC volumes. In terms of the principal component analysis, computing the components based on the sample correlation matrix of the natural variables is equivalent to computing the components based on the covariance matrix of the standardized variables. For that reason we will base our analysis off the sample correlation matrix of the natural variables (see Appendix B).

As the eigenvalues of this sample correlation matrix represent the amount of variability in the dataset that can be explained by each principal component, a screeplot of the eigenvalues will give some indication of how well the principal component analysis will perform. This plot is given in Figure 2 below and the sharp elbow in the graph suggests a large proportion of the variability in the data can be explained by the first two principal components.

Johnson and Wichern [4] state that the proportion of the standardized sample variance explained by the i th sample principal component is given by $\frac{\hat{\lambda}_i}{p}$, where $\hat{\lambda}_i$ is the estimate of the i th eigenvalue of the correlation matrix and p is the number of variables, in this case 20. Using this formula we can examine the cumulative percentage of the total variance that would be explained for each number of principal components considered. In effect this is the gain that you would get from including each principal component in a model of the overall system variability, and is shown in the table below.

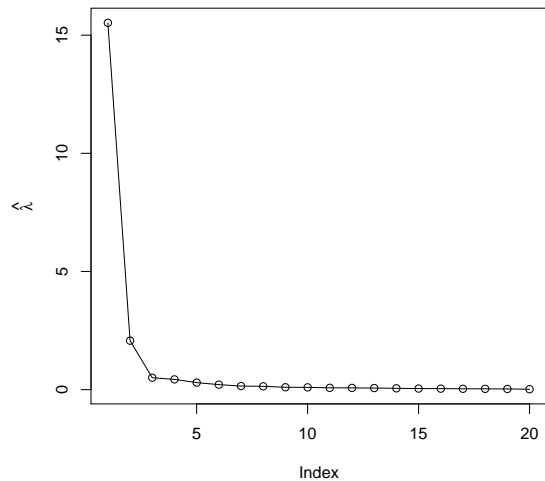


Figure 2: Screeplot of eigenvalues based on the sample correlation matrix for ARTCC counts

Index (i)	Variance ($\hat{\lambda}_i$)	Cumulative Percentage of Total Variance
1	15.522	77.6
2	2.073	88.0
3	0.505	90.5
4	0.433	92.7
5	0.297	94.1
6	0.211	95.2
7	0.151	96.0
8	0.140	96.7
9	0.100	97.2
10	0.095	97.6
11	0.077	98.0
12	0.073	98.4
13	0.066	98.7
14	0.055	99.0
15	0.045	99.2
16	0.042	99.4
17	0.035	99.6
18	0.033	99.8
19	0.029	99.9
20	0.017	100.0

Given the elbow in the screeplot of Figure 2, and the fact that the first two principal components together represent 88 percent of the variance in the sample, we are satisfied to use these two principal components as the basis of the analysis in the next section. The coefficients of these principal components are found by computing the associated eigenvectors ($\hat{\mathbf{e}}_i$) of the sample correlation matrix. For reference, these are given in the table below. The actual principal components would then be computed as $\hat{y}_i = \hat{\mathbf{e}}_i' \mathbf{x}$.

Variable	\hat{e}_1	\hat{e}_2
ZAB	-0.243	0.091
ZAU	-0.242	-0.108
ZBW	-0.213	-0.198
ZDC	-0.233	0.138
ZDV	-0.222	-0.074
ZFW	-0.243	-0.040
ZHU	-0.229	0.142
ZID	-0.243	-0.042
ZJX	-0.142	0.548
ZKC	-0.244	-0.071
ZLA	-0.219	0.126
ZLC	-0.230	-0.115
ZMA	-0.012	0.648
ZME	-0.240	-0.001
ZMP	-0.235	-0.196
ZNY	-0.222	0.157
ZOA	-0.233	-0.020
ZOB	-0.245	-0.108
ZSE	-0.230	-0.195
ZTL	-0.229	0.158

Note that the first principal component is in effect, a weighted sum of the traffic counts from each ARTCC. This gives a level of the overall traffic volume in the NAS on a given day. The negative coefficients of this principal component indicate that high volume days will have small, negative values in the principal component and low volume days will have large, positive values of the principal component. This is due to the standardization of the variables (subtract the mean and divide by the standard deviation) that was done in a previous step.

Note that the second principal component includes some positive coefficients and some negative coefficients. The effect of this is a regional indicator of the traffic levels on a given day. To aid in the interpretation of this second principal component, consider Figure 3 which shows the ARTCCs symbolized according to their coefficient value in the second principal component. From this graphic the spatial nature of the second principal component is clear, with the southern-most Centers having large coefficient values and northern-most Centers having small values. Therefore the combination of the first two principal components can capture both the overall traffic volume in the NAS, as well as geographic variations in that overall total.

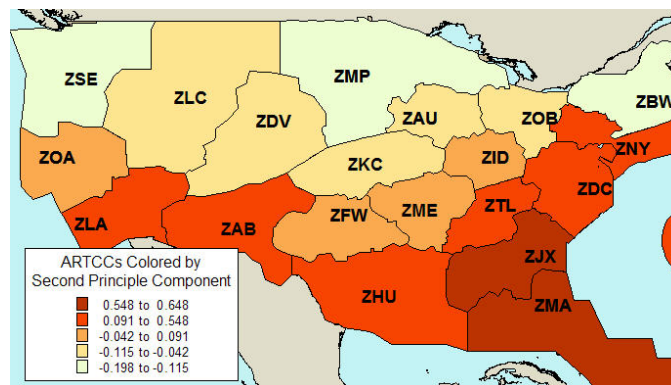


Figure 3: Regional effects implicit in the second principal component

4 Examination of Natural Groupings in the Principal Component Space

4.1 Initial Groupings

As we have shown in the previous section that the first two principal components together account for 88% of the variance in our data set, a scatter plot of the data in this principal component space should reveal any natural groupings of the data. As an initial check of the feasibility of this approach in distinguishing days of the week, we present this scatter plot labeled according to whether the points are weekdays or weekends in Figure 4. This does reveal a natural distinction between these two sets of days, and this is largely in the dimension of the first principal component representing overall volume in the NAS. This is an intuitive result and prepares us to move ahead with the analysis, but first we consider a couple of the outliers to gain insight into how abnormal data behaves in the principal component space.

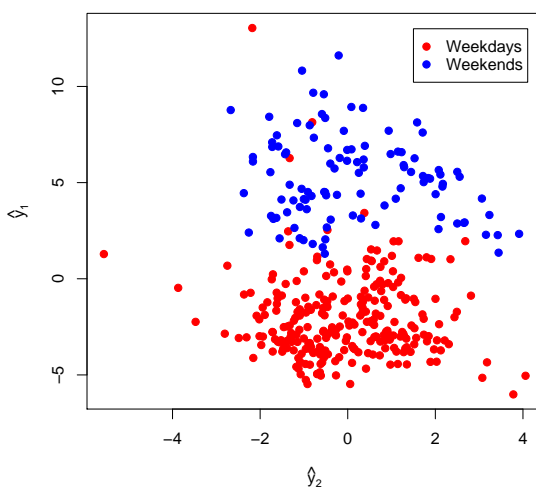


Figure 4: Scatter plot of the principle components \hat{y}_1 and \hat{y}_2 of the ARTCC count data, colored by weekday or weekend

4.2 Analysis of Outliers

We examine a couple of outliers (see Figure 5) to determine the cause of their abnormal response in either the first or second principal component. The first outlier we consider is at the top of Figure 5 and corresponds to Thursday November 24, 2005. This date was in fact the Thanksgiving holiday, which is historically a low operations day for airlines. Then it makes sense in light of the negative coefficients of our first principal component that this date would stick out by itself with a large value of \hat{y}_1 .

The second outlier we consider is on the left of Figure 5 and corresponds to Monday October 24, 2005. There was no holiday or other major event on this date so we must look a little deeper into the cause of the outlier. Consider the traffic counts by center on this date (labeled x_o for outlier) in comparison to the overall sample means shown below.

$$\bar{x}' = \begin{bmatrix} \text{ZAB ZAU ZBW ZDC ZDV ZFW ZHU ZID} & \text{ZJX} & \text{ZKC ZLA ZLC} & \text{ZMA} & \text{ZME ZMP ZNY ZOA ZOB ZSE ZTL} \\ 4828 & 7940 & 5120 & 8437 & 5031 & 5846 & 5814 & 7902 & 7005 & 5707 & 6269 & 4251 & 6852 & 6322 & 5867 & 8422 & 4666 & 8274 & 3578 & 8865 \end{bmatrix}$$

$$x_o' = \begin{bmatrix} \text{ZAB ZAU ZBW ZDC ZDV ZFW ZHU ZID} & \text{ZJX} & \text{ZKC ZLA ZLC} & \text{ZMA} & \text{ZME ZMP ZNY ZOA ZOB ZSE ZTL} \\ 4797 & 8102 & 5145 & 7403 & 4917 & 5891 & 5499 & 7289 & 4113 & 5599 & 6515 & 4101 & 1527 & 5926 & 5933 & 7873 & 4908 & 8441 & 3644 & 7850 \end{bmatrix}$$

It is clear from this comparison the regional nature of the abnormality on this date, specifically in Jacksonville (ZJX) and Miama (ZMA) centers, hence the outlier with respect to the second principal component \hat{y}_2 . This points towards a localized weather event on that day and in fact Hurricane Wilma struck southern Florida on the morning of October 24, 2005 as a Category 3 Hurricane [7]. It can be seen from the storm positions in Figure 6 why this

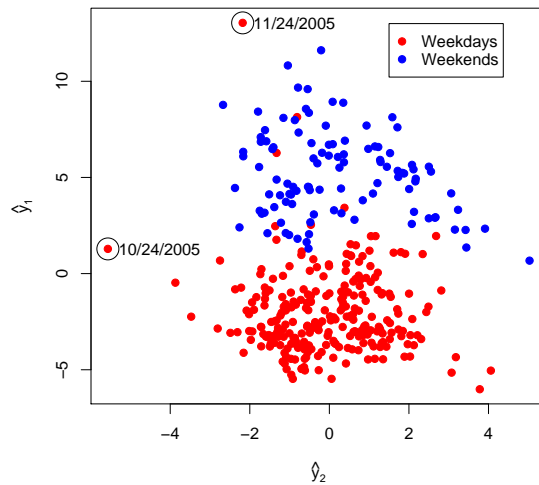


Figure 5: Scatter plot of the principle components \hat{y}_1 and \hat{y}_2 of the ARTCC count data, highlighting the outliers

was such an outlier on October 24, while there was no major regional drop in operations on the day before or the day after. Through this examination of the outliers we gain some understanding of the behaviour of the principal components as well as some confidence that they are capturing the variability of the data set in a reasonable way.

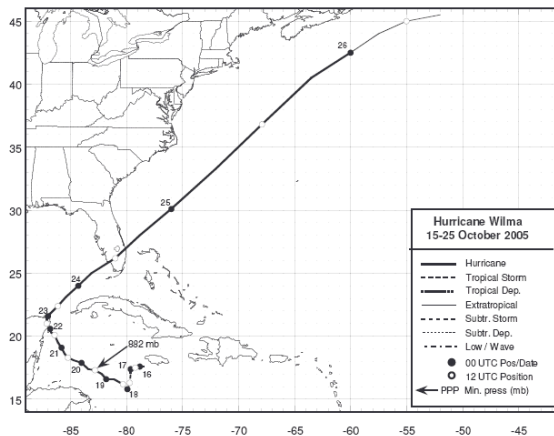


Figure 6: Storm positions for Hurricane Wilma [7]

4.3 Day of Week Groupings

We present in Figure 7 a scatter plot in the principal component space with the data points labeled according to the day of week. There are some noticeable patterns by specific day of the week in this plot, although it is busy with all 365 data points displayed. Therefore we present Figure 8 which shows the mean data point in the principal component space for each day of the week, along with the weekday and weekend means for reference. Trends in the day of week are much more clear from this view. For example the mean Wednesday appears slightly closer to the overall mean weekday, which is an interesting result for those involved with choosing a 'representative' day of air traffic for simulation studies or other purposes. Also interesting to note is the slight dispersion in both the first and second principal component dimensions of the mean of each weekday. This suggests some trends both in overall volume in the NAS as well as more geographically localized trends in day of week. This will be interesting to examine in further detail in future studies.

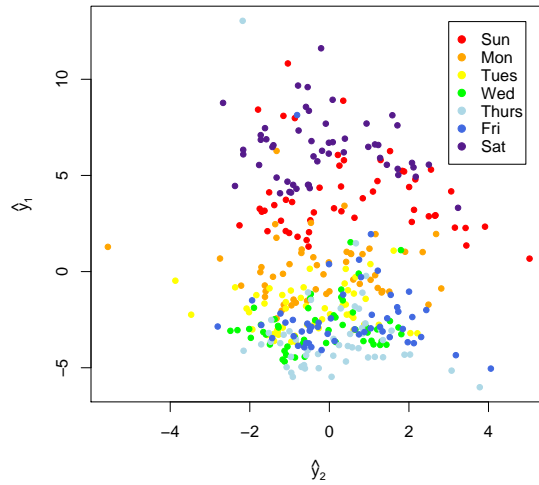


Figure 7: Scatter plot of the principle components \hat{y}_1 and \hat{y}_2 of the ARTCC count data, colored by day of week

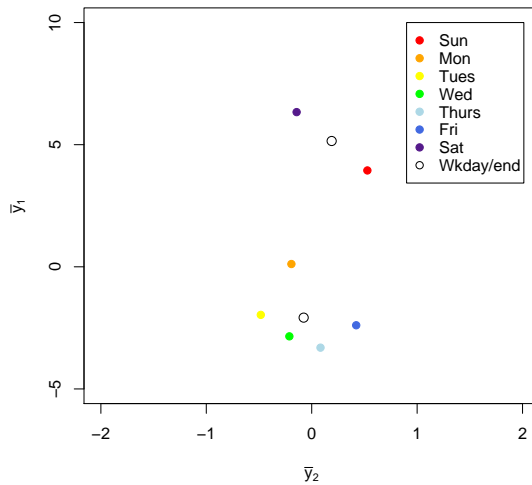


Figure 8: Scatter plot of the principle components means \bar{y}_1 and \bar{y}_2 for each day of the week of the ARTCC count data

5 Conclusions

We have shown principal component analysis to be an effective technique for the analysis of NAS traffic levels by ARTCC, where 88% of the total variance in the system can be explained by the first two principal components of the variables. In doing this we have provided a foundation for the further analysis of trends in air traffic activity by day of week, and scatter plots in the principal component space have been useful in illustrating those trends. There is much remaining work that could be done in this area and the successful results presented in this work provide the motivation for exploring these directions further. Among the future avenues for work would be an analysis of why there may be trends in the overall volume and spatial dispersion of ARTCC activity by day of week. Also, a more rich analysis could be done by including the user class variables (air carrier, air taxi, general aviation, and military) that the FAA OPSNET data contains in addition to the overall counts to provide insight into trends by specific users of the NAS. In addition, some other multivariate statistical techniques could be applied to this data such as

linear discriminant analysis to determine whether it is more or less successful in illustrating groupings of day of week activity than the principal component analysis described here.

A Air Route Traffic Control Center Identifiers

ID	ARTCC Name
ZAB	Albuquerque Center
ZAU	Chicago Center
ZBW	Boston Center
ZDC	Washington (D.C.) Center
ZDV	Denver Center
ZFW	Fort Worth Center
ZHU	Houston Center
ZID	Indianapolis Center
ZJX	Jacksonville Center
ZKC	Kansas City Center
ZLA	Los Angeles Center
ZLC	Salt Lake Center
ZMA	Miami Center
ZME	Memphis Center
ZMP	Minneapolis Center
ZNY	New York Center
ZOA	Oakland Center
ZOB	Cleveland Center
ZSE	Seattle Center
ZTL	Atlanta Center

B Sample Correlation Matrix for 2005 ARTCC Daily Counts

	ZAB	ZAU	ZBW	ZDC	ZDV	ZFW	ZHU	ZID	ZJX	ZKC	ZLA	ZLC	ZMA	ZME	ZMP	ZNY	ZOA	ZOB	ZSE	ZTL
ZAB	1.00	0.88	0.71	0.88	0.82	0.94	0.91	0.90	0.62	0.92	0.86	0.84	0.17	0.92	0.84	0.84	0.89	0.89	0.84	0.87
ZAU	0.88	1.00	0.84	0.84	0.85	0.91	0.81	0.93	0.42	0.93	0.80	0.88	-0.10	0.89	0.94	0.80	0.86	0.97	0.89	0.83
ZBW	0.71	0.84	1.00	0.78	0.78	0.76	0.67	0.81	0.25	0.79	0.64	0.79	-0.20	0.72	0.87	0.78	0.72	0.88	0.83	0.70
ZDC	0.88	0.84	0.78	1.00	0.73	0.85	0.83	0.87	0.67	0.84	0.82	0.76	0.18	0.84	0.78	0.94	0.80	0.88	0.75	0.91
ZDV	0.82	0.85	0.78	0.73	1.00	0.82	0.76	0.81	0.40	0.84	0.72	0.88	0.04	0.79	0.88	0.70	0.80	0.84	0.84	0.71
ZFW	0.94	0.91	0.76	0.85	0.82	1.00	0.89	0.92	0.48	0.94	0.82	0.86	-0.02	0.94	0.89	0.79	0.89	0.92	0.88	0.83
ZHU	0.91	0.81	0.67	0.83	0.76	0.89	1.00	0.84	0.64	0.84	0.79	0.77	0.23	0.88	0.77	0.80	0.81	0.82	0.77	0.84
ZID	0.90	0.93	0.81	0.87	0.81	0.92	0.84	1.00	0.49	0.94	0.77	0.86	-0.02	0.94	0.90	0.82	0.84	0.94	0.89	0.87
ZJX	0.62	0.42	0.25	0.67	0.40	0.48	0.64	0.49	1.00	0.46	0.62	0.39	0.73	0.53	0.30	0.64	0.48	0.42	0.29	0.72
ZKC	0.92	0.93	0.79	0.84	0.84	0.94	0.84	0.94	0.46	1.00	0.80	0.88	-0.05	0.93	0.92	0.78	0.88	0.94	0.90	0.84
ZLA	0.86	0.80	0.64	0.82	0.72	0.82	0.79	0.77	0.62	0.80	1.00	0.74	0.16	0.80	0.73	0.76	0.86	0.79	0.70	0.81
ZLC	0.84	0.88	0.79	0.76	0.88	0.86	0.77	0.86	0.39	0.88	0.74	1.00	-0.06	0.84	0.90	0.70	0.86	0.88	0.88	0.75
ZMA	0.17	-0.10	-0.20	0.18	0.04	-0.02	0.23	-0.02	0.73	-0.05	0.16	-0.06	1.00	0.02	-0.18	0.24	0.01	-0.10	-0.18	0.18
ZME	0.92	0.89	0.72	0.84	0.79	0.94	0.88	0.94	0.53	0.93	0.80	0.84	0.02	1.00	0.86	0.78	0.87	0.89	0.86	0.88
ZMP	0.84	0.94	0.87	0.78	0.88	0.89	0.77	0.90	0.30	0.92	0.73	0.90	-0.18	0.86	1.00	0.73	0.83	0.93	0.92	0.76
ZNY	0.84	0.80	0.78	0.94	0.70	0.79	0.80	0.82	0.64	0.78	0.76	0.70	0.24	0.78	0.73	1.00	0.75	0.85	0.70	0.83
ZOA	0.89	0.86	0.72	0.80	0.80	0.89	0.81	0.84	0.48	0.88	0.86	0.86	0.01	0.87	0.83	0.75	1.00	0.87	0.85	0.80
ZOB	0.89	0.97	0.88	0.88	0.84	0.92	0.82	0.94	0.42	0.94	0.79	0.88	-0.10	0.89	0.93	0.85	0.87	1.00	0.91	0.83
ZSE	0.84	0.89	0.83	0.75	0.84	0.88	0.77	0.89	0.29	0.90	0.70	0.88	-0.18	0.86	0.92	0.70	0.85	0.91	1.00	0.74
ZTL	0.87	0.83	0.70	0.91	0.71	0.83	0.84	0.87	0.72	0.84	0.81	0.75	0.18	0.88	0.76	0.83	0.80	0.83	0.74	1.00

C Computer Code

C.1 format_OPSNET_centers.pl

```
#!/usr/bin/perl
# Used to format a raw file of OPSNET ARTCC daily counts so it can be easily imported
# and worked with in R for multivariate analysis. For each day (rows) there should
# be 20 columns corresponding to the total daily aircraft handled by each ARTCC.
# This code also adds a column containing the day of week.

# Author: Brendan Hogan
# Date: 4/27/06

use Time::Local;

# set variables for input indices based on columns in OPSNET file
$dateIdx = 0; # Date in YYYYMMDD format
$artccIdx = 1; # ARTCC
$totalOpsIdx = 21; # total departure and overflights in ARTCC

# loop through input file, add data to hash structure
while (<>) {
    next if /^YYYYMMDD/; # ignore the header line
    @vars = split(/,/ , $_);

    $thisDate = $vars[$dateIdx];
    $thisArtcc = $vars[$artccIdx];
    $thisArtcc =~ s/\s+$/; # remove trailing whitespace in field
    $thisOpsCount = $vars[$totalOpsIdx];

    $ops{$thisDate}{$thisArtcc}[0] = $thisOpsCount;
} # end while more flight records

# loop through hash data structure, format and write to output
$printHdr = 1;
foreach $myDate (sort keys %ops) {
    # convert day of week (zero-based, starting with Sunday)
    $tmpDate = timegm(0, 0, 12, substr($myDate,6,2),
                    substr($myDate,4,2)-1, substr($myDate,0,4)-1900);
    $wday = (localtime $tmpDate)[6];

    # print header line the first time through this loop
    if ($printHdr == 1) {
        print "DATE DOW ";
        foreach $myArtcc (sort keys %{$ops{$myDate}}) {
            print $myArtcc, " ";
        }
        print "\n";
        $printHdr = 0;
    }

    print $myDate, " ", $wday, " ";

    # loop through ARTCCs
    foreach $myArtcc (sort keys %{$ops{$myDate}}) {
        print $ops{$myDate}{$myArtcc}[0], " ";
    }
}
```

```

    print "\n";
}

```

C.2 R language script

```

# R code for processing daily ARTCC count data and preparing graphics for report
# Author: Brendan Hogan
# Date: 5/02/06

# read data, header, place in matrix
tmp <- scan("C:\\Brendan\\UVA\\stat513\\project\\opsnet_2005_artccs_forR.txt", skip=1)
myHdr <- scan("C:\\Brendan\\UVA\\stat513\\project\\opsnet_2005_artccs_forR.txt", what="character", nlines=1)
artccData <- matrix(tmp, ncol=22, byrow=T)

# set vectors holding exact dates, and day of week to be used later for labeling plots
inputDates <- artccData[,1]
inputDOW <- artccData[,2]
artccs <- artccData[,3:22]
dimnames(artccs) <- list(inputDOW, myHdr[3:22])

# calculate sample mean vector, covariance matrix, and correlation matrix
artccs.xbar <- cbind(colMeans(artccs))
artccs.cov <- cov(artccs)
artccs.cor <- cor(artccs)

# scree plot of variability explained by each principal component
plot(eigen(artccs.cor)$values, ylab=expression(hat(lambda)), type="o")

# calculate cumulative percentage of total variance
cum.var.artccs.cor <- rep(0, 20)
for (i in 1:20) {
  cum.var.artccs.cor[i] <- sum(eigen(artccs.cor)$values[1:i])/20
}

# output coefficients of first two principal components
dummy <- matrix(cbind(eigen(artccs.cor)$vectors[,1],
                     eigen(artccs.cor)$vectors[,2]), nc=2, byrow=FALSE)
dimnames(dummy) <- list(myHdr[3:22], c("1", "2"))

# need standardized variables first, before computing principal components
artccs.submean <- artccs - matrix(cbind(rep(1,nrow(artccs)))) %*% t(artccs.xbar), nc=ncol(artccs))
artccs.stand <- artccs.submean %*% solve(diag(artccs.sd))

# define variables holding the principal components
pc1 <- artccs.stand %*% cbind(eigen(artccs.cor)$vectors[,1])
pc2 <- artccs.stand %*% cbind(eigen(artccs.cor)$vectors[,2])

# label weekdays different than weekends on scatter plot in principal component space scatter plot
plot(pc2, pc1, type="n", xlab=expression(hat(y)[2]), ylab=expression(hat(y)[1]))
points(pc2[inputDOW>=1 & inputDOW<=5], pc1[inputDOW>=1 & inputDOW<=5], pch=19, col="red")
points(pc2[inputDOW==6 | inputDOW==0], pc1[inputDOW==6 | inputDOW==0], pch=19, col="blue")
legend(1.5, 13, legend=c("Weekdays","Weekends"), col=c("red","blue"), pch=c(19,19))

# checking on the outliers
inputDates[pc1>12]      # 20051124 -> Thanksgiving Day, v. low volume of flights
inputDates[pc2 < (-5)] # 20051024 -> a Monday, not sure what's up with it

```

```

# label outliers for discussion in paper
points(pc2[inputDates==20051124 | inputDates==20051024],
       pc1[inputDates==20051124 | inputDates==20051024], pch=21, cex=3)
text(pc2[inputDates==20051124]+1.3, pc1[inputDates==20051124], "11/24/2005")
text(pc2[inputDates==20051024]+1.3, pc1[inputDates==20051024], "10/24/2005")

# now get days of the week to be symbolized differently
myDays = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")
myDaysAbbrev = c("Sun", "Mon", "Tues", "Wed", "Thurs", "Fri", "Sat")
myColors = c("red", "orange", "yellow", "green", "lightblue", "royalblue", "purple4")

plot(pc2, pc1, type="n", xlab=expression(hat(y)[2]), ylab=expression(hat(y)[1]))
for (i in 0:6) {
  points(pc2[inputDOW==i], pc1[inputDOW==i], pch=19, cex=0.75, col=myColors[i+1])
}
legend(3, 13, legend=myDaysAbbrev, col=myColors, pch=19)

# now show the means of each day's pc's
plot(pc2, pc1, type="n", xlab=expression(bar(y)[2]), ylab=expression(bar(y)[1]),
     xlim=c(-2,2), ylim=c(-5,10))
for (i in 0:6) {
  points(mean(pc2[inputDOW==i]), mean(pc1[inputDOW==i]), pch=19, cex=1, col=myColors[i+1])
}
points(mean(pc2[inputDOW==6 | inputDOW==0]), mean(pc1[inputDOW==6 | inputDOW==0]), pch=21, cex=1.25)
points(mean(pc2[inputDOW>=1 & inputDOW<=5]), mean(pc1[inputDOW>=1 & inputDOW<=5]), pch=21, cex=1.25)
legend(1, 10, legend=myDaysAbbrev, col=myColors, pch=19)

```

References

- [1] FAA APO. About OPSNET. <http://www.apo.data.faa.gov/getInfo.asp?id=opsnet>, Accessed April 26, 2006.
- [2] Dipasis Bhadra, Jennifer Gentry, Brendan Hogan, and Michael Wells. Future air traffic timetable estimator. *Journal of Aircraft*, 42(2):320–328, March-April 2005.
- [3] Michael B. Callaham, James S. DeArmon, Arlene M. Cooper, Jason H. Goodfriend, Debra Moch-Mooney, and George H. Solomos. Assessing NAS performance: Normalizing for the effects of weather. In *4th USA/Europe Air Traffic Management R&D Symposium*, 2001.
- [4] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.
- [5] Jimmy Krozel, Bob Hoffman, Steve Penny, and Taryn Butler. Selection of datasets for NAS-wide simulation validations. Technical report, Metron Aviation, 2002.
- [6] Averill M. Law and David W. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, 3rd edition, 2000.
- [7] Richard J. Pasch, Eric S. Blake, Hugh D. Cobb III, and David P. Roberts. Tropical Cyclone Report Hurricane Wilma 15-25 October 2005. Technical Report TRC-AL252005_Wilma, National Hurricane Center, January 2006.